DASINYA JOURNAL for engineering and informatics

https://doi.org/00.0000/xxxxx
Dasinya Journal for Engineering and Informatics. **2025**, 1, 7.

Review

A Review on Deep Learning Frameworks for Dental Anomaly and Disease Classification

Dawlat Abdulkarim Ali[1,*], **Haval Tariq Sadeeq**[2]

[1] *Department of Information Technology, Technical College of Informatics, Akre University for Applied Science, Kurdistan Region-Iraq*[1]; dawlat.ali@dpu.edu.krd
[2] *Artificial Intelligence Department, Technical College of Duhok, Duhok Polytechnic University, Duhok 42001, Kurdistan Region, Iraq* [2]; haval.tariq@dpu.edu.krd
* Correspondence: dawlat.ali@dpu.edu.krd

Abstract

Oral anomalies and dental diseases affect billions of people worldwide, yet diagnosis often relies on manual interpretation of radiographs and clinical images, which is time-consuming and prone to variability. Advances in deep learning (DL) have opened new opportunities for accurate, efficient, and scalable dental diagnostics. This review examines state-of-the-art DL frameworks applied to dental imaging modalities, including intraoral RGB photographs, bitewing and periapical radiographs, panoramic radiography, and cone-beam computed tomography (CBCT). The analysis covers preprocessing pipelines, backbone architectures (convolutional neural networks and vision transformers), task designs (classification, detection, segmentation, hybrid models), and strategies for addressing data imbalance, calibration, and uncertainty. Findings reveal that modality-specific preprocessing enhances reliability, hybrid CNN-Transformer models improve performance for wide-field or complex tasks, and segmentation-assisted classification increases sensitivity to subtle lesions. Moreover, calibrated probability outputs, robust evaluation metrics (ROC-AUC, PR-AUC), and external validation are essential for clinical readiness. The review identifies critical gaps—limited cross-site generalization, under-reported calibration, and scarce real-world validation—and outlines future directions such as label-efficient learning, federated training, and calibration-first pipelines. With these safeguards, DL-based systems can evolve from experimental tools to trustworthy clinical aids that strengthen diagnostic accuracy and decision support in dentistry.

Keywords: Dental imaging; deep learning; convolutional neural networks (CNNs); vision transformers (ViT); Class imbalance; probability calibration.

1. Introduction

Oral and dental health is a vital component of overall well-being, yet dental anomalies and oral diseases remain among the most prevalent chronic conditions worldwide. Approximately 3.5 billion people are affected, with the burden driven primarily by untreated dental caries (~2.5 billion) and severe periodontitis (~1 billion) [1,2]. When untreated, these conditions lead to pain, infection, tooth loss, and systemic complications that impair nutrition, speech, and quality of life; the burden is amplified in low- and middle-income countries where preventive services and advanced diagnostics are limited [1]. Traditional diagnosis relies on clinical assessment and manual interpretation of

Received: September 3, 2025; Revised: October 8, 2025; Accepted: October 10, 2025; Published: October 12, 2025; Available online: October 12, 2025.
Citation: To be added by editorial staff during production.

radiographs, which is time-consuming, subjective, and prone to inter-observer variability; overlapping anatomy, image noise, and early lesions further complicate detection[3].

Against this backdrop, artificial intelligence (AI)—particularly deep learning (DL)—has accelerated progress in medical image analysis, including transformer-based vision models such as ViT and Swin that capture long-range context [4,5]. In dentistry, preliminary work demonstrates AI on panoramic radiographs for multi-condition screening [6],while broader surveys of medical imaging emphasize scalable, data-efficient pipelines that transfer to clinical tasks [7,8].The evidence base spans panoramic staging/measurement on OPG [9,10], multi-label screening from intraoral RGB photographs [11,12], and three-dimensional analyses using CBCT and CBCT–IOS fusion to enrich anatomical context [13,14]. Persistent methodological gaps remain [7] highlights the need for transparent intended-use claims, external testing, and pre-specified operating points, alongside evaluation under class imbalance where PR-AUC complements ROC-AUC to reflect clinically meaningful decision thresholds[7].

Gap and novelty: Previous reviews have not consistently addressed probability calibration, explainability, and deployment readiness across all major dental imaging modalities. This review targets that gap by integrating: (i) modality-aware preprocessing and class-imbalance remedies; (ii) calibrated, threshold-ready probabilities (ECE, reliability diagrams) reported at clinically constrained operating points; and (iii) deployment artifacts (latency, memory footprint, structured outputs) together with external, site-stratified validation aligned with contemporary reporting guidance.

Objectives: This review aims to:

1. compare CNN and transformer frameworks across dental modalities and tasks.
2. consolidate imbalance-aware objectives and calibration metrics (PR-AUC, Cohen's κ, ECE, reliability diagrams) with clinically meaningful operating points.
3. Summarize deployment and reporting practices, including external validation, efficiency reporting, and integration into clinical systems.

Organization of the paper :Section 2 provides background and the theoretical framework; Section 3 reviews the literature by modality and task; Section 4 analyzes model choices and trade-offs; Section 5 outlines challenges; Section 6 describes future directions; Section 7 presents actionable recommendations; and Section 8 concludes.

2. Background and Theoretical Framework

2.1 Machine Learning (ML): a concise orientation

Machine learning (ML) studies algorithms that improve at a task through experience (data) rather than hand-written rules. Instead of prescribing decision logic, we provide examples and let the model infer patterns that map inputs to outputs.

Core idea: ML assumes useful regularities exist in the data and seeks to approximate the unknown function that generated them. The central challenge is generalization—performing well on new cases, not only on the training examples.

Data, features, and representations: Classic ML relied on human-designed features; modern approaches increasingly learn representations directly from raw inputs (via deep models or self-supervised objectives), reducing manual engineering[7,8].

Model families (high level):

Linear models (logistic/linear regression): simple, interpretable baselines; effective with near-linear relations or limited data;

Kernel methods (SVM, Gaussian processes): capture non-linear structure via similarity functions;

Tree ensembles (Random Forests, Gradient Boosting): robust to mixed types/outliers; strong tabular baselines;

Neural networks (feed-forward, CNNs, Transformers): flexible function approximators that scale with data and compute.

Why ML works well now: The confluence of three reinforcing trends—(1) larger datasets, (2) more compute, and (3) better algorithms (optimization, architectures, regularization)—has enabled rich, transferable representations across vision, language, and structured data.

2.2 Deep Learning (DL): encoders, tasks, and objectives

Deep learning (DL) is a branch of ML that uses multi-layer neural networks to learn complex functions directly from raw data, with hierarchical features learned end-to-end[7]. Training adjusts weights to minimize a loss via back-propagation and stochastic gradient methods (e.g., SGD, AdamW).

2.2.1 Key backbones

- CNNs (VGG, ResNet, DenseNet, Inception/Xception, EfficientNet, MobileNet, ConvNeXt/RegNet): exploit local patterns; strong for images; compute-efficient and data-friendly; Representative references appear in §2.5.1.
- Vision Transformers (ViT, Swin): self-attention captures long-range context; well-suited to wide-field or high-resolution inputs; typically benefit from stronger pretraining[4,5].

2.2.2 Task heads

- Classification (image/ROI label) for fast screening;
- Detection (bounding boxes) for focal findings;
- Segmentation (pixel/voxel masks) when geometry/staging matters;
- Seg→Cls (segmentation-assisted classification) to boost sensitivity for subtle, small targets.

2.2.3 Training essentials

- Optimization/regularization: SGD/AdamW, residual connections, normalization, dropout, weight decay, and data augmentation (e.g., flips/rotations, MixUp, CutMix)[15,16];
- Losses under imbalance: class-weighted cross-entropy, Focal Loss for classification/detection[17], Dice/IoU-aware losses (e.g., Focal-Tversky and Generalized Dice for segmentation[18,19];
- Calibration: temperature scaling with reliability diagrams/ECE for decision-useful probabilities[20].

Evaluation & robustness: Use ROC-AUC and PR-AUC under skew; report sensitivity/specificity at clinically fixed thresholds with confidence intervals. Prevent leakage with subject/site-level splits. Self-/semi-supervised pretraining and careful augmentation improve cross-site transfer [8, 40].

2.3 *Clinical Overview of Dental Anomalies and Oral Diseases*

Global burden: Oral diseases are among the most common non-communicable conditions worldwide (~3.5 billion affected), driven mainly by untreated dental caries (~2.5 billion) and severe periodontitis (~1 billion). Consequences include pain, infection, tooth loss, and impaired nutrition, speech, and quality of life—especially in underserved settings [1,2]

Routine diagnosis: Clinical examination plus radiographic interpretation remains standard, yet early or subtle lesions (e.g., proximal caries, incipient periapical radiolucencies) are frequently missed, and inter-observer variability reduces reliability [3]. Imaging adds modality-specific cues: bitewings (interproximal enamel–dentin changes), periapicals (apical radiolucency), OPG (jaw-wide screening), CBCT (3D tooth–bone anatomy), and intraoral RGB (color/texture) [11],[21–24] Representative appearances are shown in Figure 1, and clinical targets are mapped to modality and deep-learning (DL) task types in Table 1.

Brief disease primers (diagnostic signatures):

- Dental caries: enamel–dentin radiolucency; bitewings preferred for proximal lesions; conservative contrast handling preserves faint signals [21,22];
- Gingivitis & calculus: erythema/edema and mineralized plaque; in RGB, white balance and ROI-centric framing stabilize color cues [11];
- Periodontitis: crestal bone-level reduction and angular defects; measurement/segmentation on BW/OPG with stage-aware reporting [9,10];
- Periapical lesions: apical radiolucency ± cortical disruption; PA first, CBCT for 3D extent; Seg→Cls can improve sensitivity for small lesions[13,24] ;
- Tooth wear/erosion: glossy facets, enamel loss, cupping; standardized RGB capture mitigates illumination bias [11];
- Oral mucosal ulcers: shallow ulcer base with erythematous halo and fibrin slough; careful annotation required due to visual variability[11].

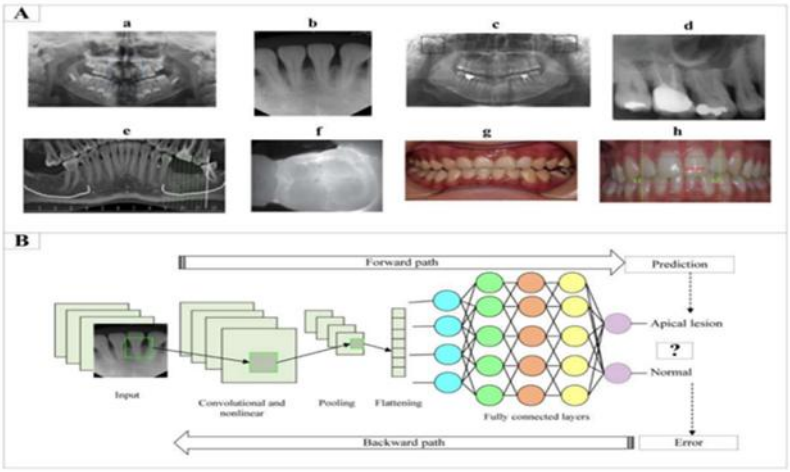

Figure 1. Representative findings across modalities (clinical montage): (A) intraoral RGB; (B) PA/BW radiographs; (C) OPG; (D) CBCT/IOS.

Table 1. Clinical targets mapped to primary imaging modality, radiologic signature, DL head(s), and Representative studies

Condition	Primary modality	Typical signature	DL task focus	Representative studies
Proximal /occlusal caries	Bitewing /Periapical	Enamel–dentin radiolucency	Classification/ Detection	[22], [3]
Periodontitis (bone loss)	Bitewing / OPG	Crestal bone-level reduction; angular defects	Measurement / Segmentation	[9], [10]
Periapical lesion	Periapical / CBCT	Apical radiolucency; cortical disruption	Detection + Segmentation	[13],[26]
Tooth wear / erosion	Intraoral RGB	Glossy wear facets; enamel loss; cupping	Grading / Classification	[11]
Developmental anomalies	OPG / CBCT	Missing/extra teeth; impactions	Multi-label classification	[6], [3]
Mucosal inflammation / ulcers	Intraoral RGB	Redness; ulcer base; fibrin slough	Lesion localization / Classification	[11]

[1] Note: OPG = orthopantomogram (panoramic radiograph); CBCT = cone-beam computed tomography; RGB = intraoral color imaging; DL = deep learning. "Typical signature" items are illustrative and may vary by exposure/positioning.

2.4 Imaging Modalities & AI Relevance

Modality-aware design: Tailor preprocessing and model choices to each modality's physics/geometry to preserve faint cues and avoid anatomical distortion [3].

The main practical points are:

- Intraoral RGB: white-balance/color-constancy → mild photometric jitter; ROI cropping; bounded augmentation[25]. see the preprocessing block in Figure 2.
- Periapical/OPG: conservative contrast (mild CLAHE/gamma), small affine transforms; avoid heavy blur that suppresses subtle radiolucencies [3]; key cautions are listed in Table 2;
- CBCT/IOS: isotropic resampling and intensity harmonization; MAR when appropriate (document potential intensity shifts); strict registration QA for CBCT ↔IOS fusion[23,24,26]; the Seg→Cls variant is sketched in Figure 3.

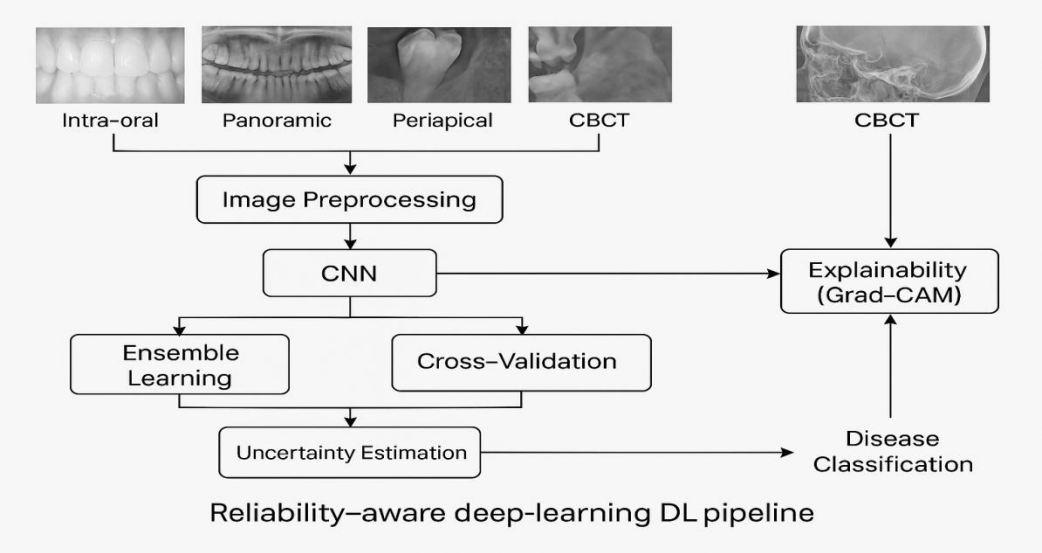

Figure 2. Reliability-aware workflow: preprocessing → encoder (CNN/Transformer) → uncertainty → calibration (ECE) → fixed operating points

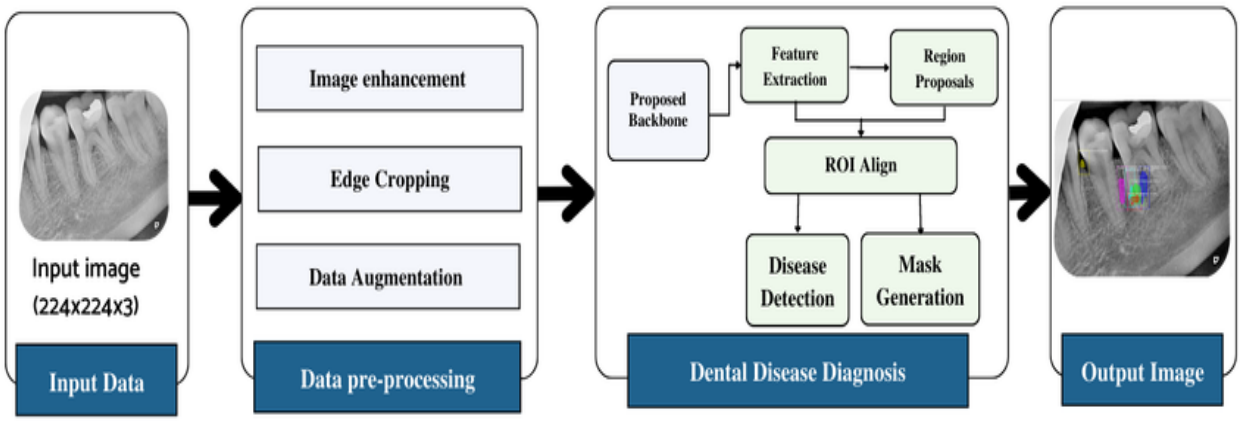

Figure 3. Segmentation-assisted pipeline (Seg→Cls): enhancement/cropping/augmentation → feature backbone → proposals/masks → calibrated decision with Grad-CAM overlays.

Table 2. Modalities, typical DL tasks, advantages/limitations, and practical notes.

Modality	Typical tasks	Advantages	Limitations	Practical notes
Intraoral RGB	Caries, calculus, mucosal lesions, discoloration	Rich color /texture; low cost	Illumination/ specular glare; pose variability	Apply white-balance and color-constancy; crop to ROI; use bounded color jitter.
Periapical radiograph	Apical lesions; endodontic status; per-tooth assessment	High root/detail resolution	Sensitivity to projection geometry	Prefer conservative contrast operations (e.g., mild local contrast); avoid heavy blur; small affine only.
Bitewing radiograph	Proximal caries; crestal bone levels	Good interproximal visibility	Overlap; horizontal angulation errors	Minor rotations/affine only; document alignment protocol.
Occlusal radiograph	Impactions; supernumeraries	Wide occlusal field	Lower in-plane resolution	Use multi-scale encoders; moderate input size.
Panoramic (OPG/DPR)	Multi-finding screening; anomaly mining; staging	Global jaw context	Magnification; overlap/distortion	Multi-scale/long-range encoders; careful gray-scale normalization.

Cephalometric (lateral/PA)	**Skeletal relations; landmarking**	**Standardized projections**	**Landmark variability**	**Keypoint/segmentation pipelines; inter-rater consistency checks.**
CBCT (3D)	**Implants, pathology, root morphology**	**True volumetric anatomy**	**Dose; metal artifacts; voxel size variance**	**Isotropic resampling; MAR when needed; intensity harmonization across scans.**
IOS (3D surface)	**Occlusion; aligners; surface fusion**	**Accurate dental surfaces**	**No internal anatomy**	**Smooth meshes; rigid/non-rigid registration QA; fuse with CBCT/OPG if available.**
NILT / QLF / OCT / HSI	**Early caries/plaque/cracks; tissue typing**	**Non-ionizing; quantitative/spectral**	**Limited FoV; device availability; high dimensionality**	**Device calibration; dimensionality reduction; patch-based local 3D nets**

[2]Notes: ROI = region of interest; BW = bitewing; PA = periapical; OPG/DPR = panoramic radiography; CBCT = cone-beam computed tomography; IOS = intraoral surface scan; NILT = near-infrared light transillumination; QLF = quantitative light-induced fluorescence; OCT = optical coherence tomography; HSI = hyperspectral imaging; MAR = metal-artifact reduction; FoV = field of view. "Conservative contrast operations" = mild local contrast adjustments (e.g., CLAHE with gentle gamma) that preserve faint radiolucencies; "intensity harmonization" = matching intensity ranges across scanners/exams.

2.5 *CNN/Transformer Families & Heads*

Scope: This section summarizes widely used image encoders (classic CNNs, modern convnets, vision transformers) and links them to task heads (classification, detection, segmentation, Seg→Cls) that recur across dental imaging. The goal is a practical "when to use what" map tied to data scale, lesion size/contrast, field-of-view, and deployment constraints. Figures 2–3 visualize the surrounding workflow choices; Tables 3–4 give side-by-side comparisons.

2.5.1 Convolutional encoders (representative families)

Convolutional encoders (CNNs) are a practical default for dental imaging because they capture local textures and edges, run efficiently on common hardware, and transfer well from ImageNet. With limited or imbalanced datasets—typical in dentistry—CNN backbones often deliver strong, stable baselines for periapical, bitewing, and panoramic tasks. Use them when latency/memory matter or when global long-range context is not the primary bottleneck.

- ResNet-50. Residual skips stabilize deep training and transfer well; a dependable default for periapical/OPG classifiers and detectors. On small single-center sets, tighten regularization and calibrate probabilities to curb overfitting [27,20];
- VGG-16. Deep stacks of 3×3 convs with a large FC head; stable transfer features but heavy (~138 M params). Mostly a baseline now when memory permits [28];
- DenseNet-121. Dense connections encourage feature reuse and strong gradients with good parameter efficiency; watch activation memory during training [29]
- InceptionV3 / Xception. Multi-scale (factorized) convs and auxiliary heads capture wide-field context useful for OPG; prefer ≥299[2] inputs; Xception's depthwise separables are parameter-efficient [30,31];
- MobileNetV2/V3. Inverted residuals and NAS/SE refinements suit edge-class latency/power budgets (chairside/handheld). Report ECE and apply temperature scaling before fixing clinical thresholds;
- EfficientNet / EfficientNetV2. Compound scaling offers strong accuracy-efficiency; B0–B3 reliable on RGB/periapicals; larger variants need careful input sizing and memory planning [32,33]
- ConvNeXt / RegNet. "Modern conv" designs that match transformer-level accuracy with predictable compute; check batch-1 latency for high-res OPG multi-finding and pick RegNetX/Y to meet millisecond budgets [34–36].

2.5.2 Vision transformers (global-context encoders)

Unlike CNNs, vision transformers use self-attention to capture long-range context across patches

- ViT: Global self-attention over patch tokens; excellent long-range context but benefits from large pretraining or strong regularization on smaller dental datasets [4];

- Swin Transformer: Shifted-window attention yields hierarchical, high-resolution features well suited to detection/segmentation on OPG and 3D; typically more data-efficient than vanilla ViT in medical imaging [5].

Rule of thumb: Prefer convnets (ResNet/DenseNet/EfficientNet) for limited data, high-SNR radiographs, and tight latency; consider Swin/hybrids when long-range context is essential (panoramic, large-FoV, multi-finding) and compute allows.

2.5.3 Task heads and their clinical fit

- Classification (image/ROI label). Best for screening/global status; simple and fast but no localization. Pair with calibrated thresholds for triage [20,41];
- Object detection (boxes + scores). Targets focal findings (e.g., proximal caries, periapical cues). Focal Loss helps with class/anchor imbalance [17].
- Segmentation (pixel/voxel masks). Needed when geometry matters (bone-loss measurement, lesion extent). Combine CE with Dice/IoU-aware losses; Focal-Tversky can boost small-structure sensitivity [18,19];
- Seg→Cls (segmentation-assisted classification). Two-stage (masks → region features → final class) improves sensitivity to small/low-contrast lesions and supports Grad-CAM overlays; useful for subtle periapical pathology [9,12];
- DETR-style detectors. End-to-end set prediction with fewer hand-tuned priors; clean design but comparatively data-hungry and slower to converge [37].

2.5.4 Data augmentation — task/modality-aware recipes

Good augmentation in dental imaging should stay anatomically plausible and respect imaging physics. The goal is to boost generalization without washing out subtle diagnostic cues (e.g., faint interproximal radiolucencies or mild mucosal erythema). Below are conservative, low-risk defaults by modality and task. Parameters are deliberately modest; push them further only if you can justify with ablations and visual spot-checks[38].

General principles (apply everywhere)

1. Keep geometry believable. Use small rotations/translations/scale to avoid unrealistic tooth/bone deformation .
2. Protect diagnostic signal. Avoid heavy blur/sharpen and extreme photometric shifts that could hide early caries or apical changes.
3. Match real-world variability. Use site/device-aware photometrics (e.g., color constancy for RGB; gentle local contrast for X-ray) to mimic clinical capture differences].
4. Prevent leakage & document settings. Augment after patient-level splitting with site/scanner stratification; report exact operators/ranges; calibrate probabilities (reliability diagrams/ECE) before fixing thresholds.

(A) Radiographs (BW/PA/OPG)

- Geometry: rotations ≈±3–5°, tiny translations (≤3%), scale ≈0.97–1.03; horizontal flip only when left–right symmetry is clinically acceptable.
- Photometrics: mild local contrast (e.g., CLAHE clip 1.0–2.0; 8×8 grid) or gentle gamma ≈0.9–1.1 to counter exposure variability without over-enhancing edges.
- Notes: stay conservative to preserve faint proximal radiolucencies and apical signs; for OPG, pair with careful grayscale normalization.

(B) Intraoral RGB photographs

- Color pre-normalization: white balance or color-constancy (e.g., Gray-World/Shades-of-Gray) to reduce device/lighting drift.
- Framing: ROI-centric random crops (scale ≈0.85–1.00; aspect ≈0.9–1.1) to maintain tooth/gingival context.
- Conservative jitter: brightness ±0.10–0.18, contrast ±0.08–0.15, saturation ±0.10–0.20, hue ±5–10°, gamma 0.9–1.1.
- Avoid: strong blur/sharpen or aggressive color shifts that might mask enamel discoloration or mucosal erythema [3,24,25].

(C) CBCT volumes and CBCT↔IOS fusion

- Resampling & geometry: isotropic resampling to a clinically appropriate voxel size before augmentation; small 3D rotations ≈±5–10° and scale ≈±5% only.
- Intensity handling: site/scanner harmonization; if metal-artifact reduction (MAR) is used, document parameters and audit downstream impact because MAR changes intensity statistics.
- Fusion: enforce QA for CBCT↔IOS registration and report alignment metrics/failure modes [23,24].

(D) Detection and segmentation heads

- Sampling: class- and ROI-balanced crops/patches to counter foreground sparsity (anchors/proposals).
- Deformations: small affine/elastic only; avoid shape warps that would invalidate measurement tasks (e.g., bone-loss staging).
- Loss coupling: use Focal Loss for detection, and Generalized Dice or Focal-Tversky for imbalanced/small masks in segmentation [17–19].

(E) Mix-based regularizers (use sparingly)

- MixUp/CutMix: helpful on small, heterogeneous cohorts to stabilize decision boundaries; keep strengths modest (typical α≈0.2–0.4; CutMix probability ≤0.2) so you don't wash out faint signals [15,16,21]
- CoarseDropout: a single small hole (≤24–32 px in 2D) at low probability to encourage robustness without erasing key anatomy.

Reporting, calibration, and safeguards

- Qualitative verification: include a montage of augmented samples per modality (in the supplement) to visually confirm plausibility;
- Ablations: report no-aug vs proposed-aug; under class imbalance, include PR-AUC alongside ROC-AUC, and report sensitivity/PPV at pre-specified specificity (e.g., ≥0.90) with 95% CIs [7,27];
- Calibration: augmentation may cut variance but does not guarantee calibrated probabilities; apply temperature scaling and report reliability diagrams/ECE before fixing clinical thresholds[20].

Take-home: Prefer small, physics-respecting transforms tuned to each modality and task. For radiographs, emphasize conservative contrast and minimal geometry; for RGB, stabilize color; for CBCT, prioritize resampling/harmonization and registration QA. Couple these recipes with imbalance-aware losses, calibration, and transparent reporting to achieve clinically meaningful, reproducible gains[37].

2.5.5 Choosing encoders and heads — practical guidance

Selecting a backbone and prediction head should reflect the task (classification, detection, segmentation), dataset scale/imbalance, and deployment constraints. The checklist below summarizes pragmatic defaults and reporting practices.

Pick the backbone/head to match task, data scale/imbalance, and deployment limits.

- Small, imbalanced datasets: ResNet-50, DenseNet-121, or EfficientNet-B0/B3; class-aware training (class weights or Focal) with mild label smoothing; avoid double-weighting; calibrate with temperature scaling; report PR-AUC and sensitivity/PPV at fixed specificity (≥0.90) with 95% CIs; include a small external test when available [27,32, 20].
- Wide-field OPG: InceptionV3, ConvNeXt, or Swin with detection/segmentation heads; many tasks are multi-label—use BCE/Focal-BCE and report mAP/AP and macro-F1; measure throughput and batch-1 latency at clinical resolution [31,34,5].
- Subtle, small, low-contrast lesions: segmentation or Seg→Cls; consider Focal-Tversky/unified-focal; gentle CLAHE can help—quantify on validation [18,25].
- Edge (chairside): efficient backbones (EfficientNet-B0/B3 or compact variants); plan pruning and INT8 quantization; profile batch-1 latency/memory/power; assess calibration (ECE) before locking thresholds [32, 20].
- 3D CBCT & fusion: Swin or hybrid pyramids feeding 3D/2.5D segmenters; document MAR, intensity harmonization, and resampling; validate registration (TRE, HD95/Chamfer) and report timing/memory [23, 26].

Reporting & calibration (all): Compute ECE with reliability diagrams; fix thresholds on validation, then report sensitivity/PPV (or mAP for detection, Dice/IoU for segmentation) with 95% CIs on internal/external tests. Include runtime, memory, and—when relevant—energy at clinical resolution [20,41].

Table 3. Encoder families— core idea, strengths/limits, typical dental fit, key refs.

Family	Core idea	Strengths	Limitations	Typical dental fit	Key refs
ResNet-50	Residual skips	Stable transfer, robust	Can overfit small cohorts	Periapical/OPG classifiers & detectors	[27]
DenseNet-121	Dense reuse	Param-efficient	Activation memory	Radiographs with limited data	[30]
InceptionV3/Xception	Factorized convs	Multi-scale context	Prefers ≥299² inputs	Wide-field OPG	[31,32]
EfficientNet-B0/B3	Compound scaling	Strong acc-efficiency	Larger variants need care	RGB/ periapical, screening	[33,34]
Mobile-friendly CNNs	Inverted residuals	Edge latency/power	Capacity limits	Chairside/handheld	[32]
ConvNeXt/RegNet	Modern conv design	Transformer-level acc.	Check batch-1 latency	OPG multi-finding	[34-36]
ViT/Swin	Self-attention	Global context	Data/pretrain hungry	OPG/3D; detection/seg	[4,5]

Table 4. Task heads and when to use them.

Head	Output	Prefer when…	Pros	Cons	Dental use	Refs
Classifier	Image/ROI label	Screening /	Simple, fast	No localization	RGB multi-label; OPG screening.	[7,28]
Detector	Boxes + scores	Focal lesions / triage	Localizes	Misses shape	Caries/periapical on BW/PA.	[37]
Segmenter	Pixel/voxel mask	Geometry / staging	Precise extent	Annotation cost	Bone loss; lesion masks.	[19,18]
Seg→Cls	Mask→features→class	Small / low-contrast	Boosts sensitivity	Two-stage	Periapical radiolucencies.	[9,12]
DETR-style	Set of objects	Fewer priors	Clean design	Data-hungry	OPG multi-finding	[39]

2.6 Class Imbalance and Probability Calibration

Why imbalance matters: Dental datasets are typically skewed (many healthy/mild cases, fewer severe or rare conditions). Under skew, models can look "good" on accuracy while failing to detect minority classes. Two levers are used together: (A) data-level rebalancing and (B) loss-/threshold-level reweighting.

(A) Data-level rebalancing:

- Stratified k-fold and patient-level splits keep prevalence consistent and prevent leakage across views of the same subject;
- Class-aware sampling / minority oversampling paired with stronger augmentation for rare classes (e.g., MixUp, CutMix, careful photometric/affine; Albumentations) helps reduce variance without memorizing artifacts;
- For detection/segmentation, hard-example mining and patch/ROI balancing reduce anchor/foreground sparsity.

(B) Losses, smoothing, and thresholds:

- Class-weighted CE or Focal for skewed classification; Focal-Tversky/Generalized Dice for tiny/sparse masks; mild label smoothing can temper over-confidence on small, heterogeneous sets [17–19].
- Fix operating thresholds on the validation set, then report sensitivity and PPV at a pre-specified specificity (e.g., $\geq$ 0.90) with 95% confidence intervals [41].

Calibration for decision-useful probabilities: Neural networks are often miscalibrated (over-confident). Temperature scaling is a simple, effective post-hoc method to reduce ECE, and reliability diagrams plus the Brier score communicate probability trustworthiness. For safety-critical use, adopt selective prediction (abstain under low confidence) to trade coverage for risk [20,40].

Table 5. Class-imbalance remedies at a glance.

Lever	What it does	Prefer when…	Caveats	Refs
Class-aware sampling / minority oversampling	**Increases rare-class exposure per epoch**	**Severe skew; small datasets**	**Risk of overfitting without strong augmentation**	**[21,38]**
MixUp / CutMix (with standard aug)	**Regularizes decision boundary; combats label noise**	**Limited labels; heterogeneous capture**	**Tune mix ratios; preserve faint radiolucencies on X-ray**	**[15,16,21]**
Class-weighted CE	**Penalizes minority errors more**	**Any skew; simple baseline**	**Can still be over-confident**	**[17]**
Focal Loss (cls.)	**Down-weights easy negatives; focuses on hard positives**	**Detection/cls. with many negatives**	**Tune γ and α; watch convergence**	**[17]**
Unified-Focal / Focal-Tversky (seg.)	**Emphasizes small/sparse masks**	**Tiny lesions; bone-loss edges**	**Balance with Dice/CE for stability**	**[18,19]**
Label smoothing	**Reduces over-confidence, noise sensitivity**	**Small/heterogeneous labels**	**Too much can blur minority signals**	**[30]**
Thresholds at fixed specificity	**Clinically aligned operation**	**Screening/triage workflows**	**Must be set on validation, then locked**	**[41]**
Temperature scaling + ECE	**Calibrates probabilities**	**Before deployment**	**Re-tune if distribution shifts**	**[20]**

2.7 Explainability and Multimodal Fusion

Explainability (XAI): what it is—and is not. Post-hoc methods help clinicians judge plausibility (did the model look at the right place?) and curate error galleries; they do not guarantee correctness. Use multiple views and sanity checks, and interpret XAI alongside metrics and external validation [41–43]

Figure 4 illustrates the Grad-CAM pipeline used in this review—covering target-layer selection, heat-map computation, and upsampling/overlay—and serves as a reference for the plausibility panels reported later.

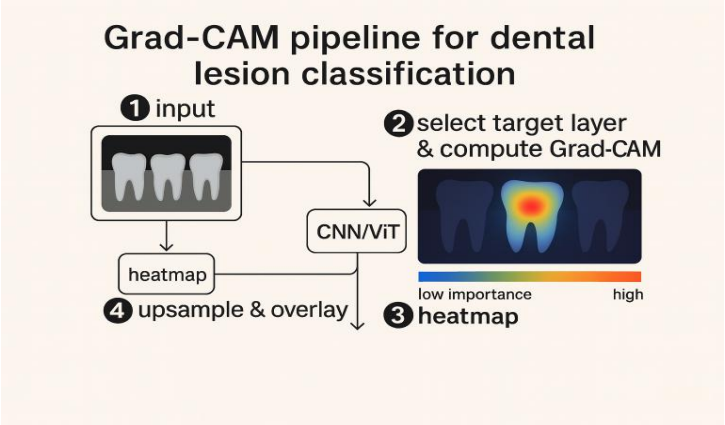

Figure 4. Schematic of the Grad-CAM workflow used in this review: (1) select the target convolutional layer; (2) compute class-discriminative gradients and weight the feature maps; (3) upsample and overlay the heat map on the input image for clinician-readable plausibility checks [43].

1. Common tools:

- Grad-CAM: fast class-discriminative heatmaps—good for plausibility overlays and quick QA; layer-dependent and resolution-limited [41];
- Integrated Gradients: axiomatic attributions; useful for aggregated trends; sensitive to baseline choice[42];
- SHAP: consistent feature contributions; informative cohort-level analysis; computationally heavier[43]

Table 6. Popular XAI methods.

Method	Strengths	Limitations	Dental use cases	Refs
Grad-CAM	Fast, intuitive overlays	Layer/resolution dependent	Lesion plausibility; failure analysis panels	**[41]**
Integrated Gradients	Axiomatic, path integrated	Baseline choice sensitivity	Aggregate attribution trends	[42]
SHAP	Consistent local→global contributions	Compute cost	Cohort-level factor analysis, reader studies	[43]

2. Multimodal fusion (OPG/PA/CBCT-IOS): Early fusion (with reliable registration) exploits complementary cues; late/attention fusion is safer when modalities are heterogeneous or missing. For CBCT-IOS, enforce registration QA, isotropic resampling, and intensity harmonization; if MAR is applied, document potential intensity shifts and audit downstream bias [23,26]. Include ablations vs. single-modality baselines and report site/scanner-wise results [27].

2.8 *Evaluation and Reporting*

2.8.1 Data splits and leakage prevention

Use patient-level splits with site/scanner stratification; never allow multiple images from the same patient to cross train/validation/test. Include at least one external test cohort to quantify distribution shift [27].

2.8.2 Metrics under class imbalance

- Classifiers: ROC-AUC + PR-AUC; per-class precision/recall/F1; report sensitivity/PPV at a fixed specificity (e.g., ≥ 0.90) with 95% CIs[40];
- Detectors: AP/mAP at relevant IoU thresholds; baselines include Faster R-CNN and DETR; see dental exemplars in §3 and Table 7 [37,39];
- Segmenters: Dice and IoU, with small-structure analyses (e.g., per-tooth bone-loss edges) [18,19];
- Calibration: reliability diagrams, ECE, optionally Brier; apply temperature scaling before fixing thresholds [20].

2.8.3 Statistical testing and uncertainty

Use patient-level bootstrapping to derive confidence intervals (CIs) and paired tests for matched designs. Quantify predictive uncertainty—e.g., Monte-Carlo Dropout—to enable selective prediction and risk–coverage analyses, so low-confidence cases can be flagged or deferred [44].

Generalization and shift robustness:

Report site-wise performance and cross-site deltas. When target labels are unavailable, benchmark unsupervised domain adaptation (DANN; Deep CORAL)[45] and test-time adaptation (TENT) as robustness baselines [46],[[47]. Document any preprocessing that alters intensity statistics—such as metal-artifact reduction (MAR) in CBCT—and analyze downstream impact [26].

2.8.4 Computational efficiency and deployment artifacts.

Provide batch-1 latency and memory footprint on intended hardware, plus throughput at clinical resolution. Release structured outputs (DICOM-SR/JSON) and frozen thresholds for audits and PACS/RIS integration.

2.8.5 Ethics, privacy, and fairness (brief).

Favor multi-site collaboration—including federated training—to expand diversity without centralizing PHI, and preserve site-level audit trails to enable accountability. When privacy-enhancing technologies are used, report formal parameters (e.g., (ε,δ)(\varepsilon,\delta)(ε,δ) for differential privacy) alongside the measurable utility impact at clinically relevant operating points. Finally, publish subgroup audits (age/sex/device/site) and probe for shortcut learning (e.g., acquisition markers, metal artifacts) to ensure equitable and robust performance across populations.

3. Literature Review and Critical Analysis

This review synthesizes evidence by modality (intraoral RGB; panoramic radiography OPG/DPR; periapical/bitewing; 3D CBCT and CBCT–IOS fusion) and by task design (classification, detection, segmentation, Seg→Cls). We prioritize studies reporting explicit metrics and clinically interpretable operating points, with thresholds pre-specified on validation and results summarized at those operating points (e.g., specificity ≥ 0.90 with sensitivity/PPV and 95% CIs) per contemporary reporting guidance [40].

3.1 Intraoral RGB (screening, grading, tele-dentistry)

Across standardized photo capture, multi-label screening and targeted grading are consistently feasible. Preprocessing typically combines white-balance or color constancy with bounded photometric jitter and ROI-centric crops. Mix-style regularizers appear when labels are limited. Lightweight backbones dominate for accuracy–efficiency, and several works add saliency to aid plausibility review.

Representative evidence includes early gingivitis detection from intraoral photos using CNN/detector pipelines [12], broad multi-label screening at the image level with explicit macro-F1/PR-AUC reporting [11], and condition-specific grading of tooth wear with high agreement [73].

Further, lightweight ensembles (VGG/MobileNet/Inception) achieve strong internal accuracy at low latency [78], while fuzzy rank-based ensembles with uncertainty targeting heterogeneous capture report robust performance on public sets [79]. Chairside-oriented MobileNetV2 models augmented with Grad-CAM illustrate edge-efficient inference and clinician-readable overlays [83]. A comparative transfer-learning benchmark focused on dental disease classification helps position backbone trade-offs specifically for RGB tasks [48].Representative RGB studies and their key outcomes are summarized in Table 7.

3.2 Panoramic radiographs (OPG/DPR): multi-finding screening and staging

OPG offers jaw-wide context but is sensitive to magnification and overlap. Baseline screeners underscore the value of careful grayscale normalization and multi-scale context [6]. Staging studies align predictions to clinical definitions and increasingly emphasize calibrated operating points and external cohorts as the logical next step [10],[18]. Architecturally, modern convnets (e.g., ConvNeXt) and attention models (e.g., Swin) capture long-range structure at higher input resolutions; hybrid CNN+ViT approaches also appear with confidence estimation to support thresholding and triage workflows [54]. Pediatric OPG work explores age-aware modeling and highlights cross-site shift as a key limitation [82]. Methods that pair deep CNN features with classical classifiers (e.g., SVM) can deliver high agreement when labels are limited, albeit with two-stage complexity [53]. Broadly, reported results suggest competitive AUC/F1 on internal cohorts, with backbone and resolution choices materially influencing performance [13].see Table 7.

3.3 Periapical/bitewing radiographs: detection and Seg→Cls

Foundational pipelines established feasibility for tooth detection/numbering on periapical radiographs, providing reference baselines and mAP by tooth index [56][49]. For lesions, instance/semantic segmentation—alone or as a front-end to a final classifier (Seg→Cls)—consistently improves sensitivity to small, low-contrast targets relative to detection-only approaches, at increased annotation cost [57],[58]. Modern convnets trained with gentle radiograph augmentation preserve faint radiolucencies and yield competitive AUC/PR-AUC, while emphasizing the need for cross-site testing [50].

3.4 3D CBCT and CBCT-IOS fusion

CBCT contributes volumetric tooth–bone detail; IOS adds accurate surface geometry. Fusion improves anatomical completeness when registration QA and intensity harmonization are enforced, with studies reporting higher planning accuracy versus single-modality inputs [13]. Metal-artifact reduction (MAR) enhances visibility yet can alter intensity statistics; both parameters and downstream effects should be documented and audited [26]. Cross-cutting methods: imbalance, calibration, and explainability.

3.5 Cross-cutting observations

Across modalities, three patterns recur. First, conservative, modality-aware preprocessing (normalization for OPG; color stabilization for RGB) supports stable training and plausible overlays [6],[11],[83]. Second, segmentation or Seg→Cls tends to boost sensitivity for subtle, small targets in periapical tasks, with dataset resources emerging to standardize comparison and reporting [57], [51]. Third, hybrid or attention-augmented encoders help capture long-range context in OPG; several groups pair these with confidence measures to aid threshold selection and reader workflows [54],[82]. External, cross-site validation remains the main limitation cited across studies.

Table7. Representative recent studies (abridged, organized by modality).

Modality	Study	Problem & design	Key contribution	Limitations/notes	Results (brief)
RGB	**Alalharith et al., 2020 [12]**	**Gingivitis detection from intraoral photos (CNN/detector)**	**Standardized capture + ROI improves grading/detection**	**Needs imbalance handling & calibration**	**AUC/Acc. improved vs. naïve preprocessing**
RGB	**Park et al., 2022 [11]**	**Multi-label intraoral photo screening**	**Feasible broad screening on RGB**	**Heterogeneous capture; threshold effects**	**Macro-F1/PR-AUC reported; PPV depends on threshold.**
RGB	**Pang et al., 2025 [73]**	**Tooth-wear grading (CNN)**	**Sensitive to subtle enamel wear; clinically aligned grading**	**Single condition; not multi-condition**	**High κ and grading agreement.**
RGB	**Hussain et al., 2023 [78]**	**Lightweight ensemble (VGG/MobileNet/Inception)**	**High accuracy with low-latency models**	**Depends on ensemble policy; shift risk**	**Acc. >90% (internal); low latency.**
RGB	**Razmjouei et al., 2025 [79]**	**Fuzzy rank-based ensemble + uncertainty**	**Robust fusion under heterogeneity**	**Ensemble inference overhead**	**Acc. ~91–97% on public sets.**
RGB	**Taşkın, 2024 [83]**	**MobileNetV2 + Grad-CAM (edge)**	**Chairside-efficient with saliency maps**	**Backbone capacity limits**	**Acc. >85% (task-specific).**
RGB	**Ikhwani et al., 2024 [48]**	**Comparative transfer learning for dental disease classification**	**Side-by-side backbone benchmarking for RGB tasks**	**Dataset diversity/standardization**	**Competitive accuracy across TL backbones.**
OPG	**Zhu et al., 2023 [6]**	**OPG multi-disease CNN**	**Normalization + multi-scale context**	**Limited interpretability reporting**	**Competitive AUCs (internal).**

OPG	**Almalki et al. (2022) [13]**	**OPG model benchmark**	**Baselines; effect of preprocessing**	**Dataset variability**	**Back-bone/resolution materially affect AUC/F1.**
OPG	**Hsieh & Cheng, 2024 [53]**	**CNN features + SVM on OPG**	**High κ with limited labels**	**Two-stage complex-ity**	**κ >0.8 (internal split).**
OPG	**Li (2025); Shon (2022) [10], [18]**	**Periodontitis staging on OPG**	**Clinically aligned outcomes**	**Geometry sensitivi-ty; calibration need-ed**	**Stage-wise F1/κ; sensitivity at set specificity.**
OPG	**Parkhi et al., 2025 [54]**	**CNN+ViT with confi-dence on OPG**	**Calibrated lo-cal+global features**	**External cohorts pending**	**AUC/PR-AUC improved vs. CNN-only**
OPG	**Pham, 2025 [82]**	**Pediatric OPG trans-formers**	**Age-aware modeling**	**Cross-site shift**	**AUC stable within site; drops cross-site.**
Periap-ical	**Chen et al., 2019 56]**	**Tooth detec-tion/numbering**	**Foundational detec-tion pipeline**	**Older backbones**	**mAP per-tooth numbering re-ported.**
Periap-ical	**Fatima et al., 2023 [57]**	**Instance segmentation of periapical lesions**	**Masks improve sen-sitivity for small/low-contrast lesions**	**Annotation cost**	**Higher small-lesion re-call vs. detec-tion-only.**
Periap-ical	**Thalji et al., 2024 [51]**	**Segmented periapical dataset**	**Resource for fair compari-sons/standardized reporting**	**Label distribution skew**	**Enables stand-ardized IoU/Dice.**
Periap-ical	**Liu et al., 2024 [50]**	**Periapical lesion detec-tion with ConvNeXt**	**Modern convs + gen-tle augmentation**	**Needs cross-site testing**	**Competitive AUC/PR-AUC; preserves faint lesions.**
3D/Fusi on	**Hegazy et al., 2023 [26]**	**CBCT MAR**	**Artifact reduction improves inputs**	**MAR side effects to track**	**Improved SNR and downstream accuracy**
3D/Fusi on	**Liu et al., 2023 [13]**	**Deep CBCT↔IOS fu-sion**	**Better anatomical completeness**	**Registration QA mandatory**	**Higher planning accuracy vs. sin-gle modality**

Abbreviations: AUC = area under the ROC curve; PR-AUC = area under the precision–recall curve; PPV = positive predictive value; κ = Cohen's kappa; mAP = mean average precision; IoU = intersection over union; ECE = expected calibration error; SNR = sig-nal-to-noise ratio; TL = transfer learning.

4. Discussion and Model Analysis

In this section, we translate the literature synthesis into practical design guidance. We begin with backbone fami-lies and their trade-offs (§4.1), then map clinical questions to task heads (§4.2), formalize operation under class imbal-ance and probability calibration (§4.3–§4.4), and close with engineering considerations for deployment (§4.5).

4.1 Backbone families: practical trade-offs

- ResNet-50 / DenseNet-121. Reliable defaults for radiographs and RGB under constrained data; calibrate predic-tions to mitigate over-confidence on single-center cohorts [27,28].

- EfficientNet-B0/B3, MobileNetV2/V3. Strong accuracy–efficiency for chairside/edge use; report ECE and apply temperature scaling before fixing thresholds [32,20]
- ConvNeXt / Inception / Swin. Favor when long-range, multi-scale context is essential (OPG, multi-finding). Check batch-1 latency and memory footprint at clinical resolution [34,30]
- ViT. Powerful global context with large-scale pretraining, hybrids or Swin often prove more data-efficient in medical imaging [4].

Takeaway: Start conv-first for limited data or edge constraints; escalate to Swin/hybrids for large-FOV OPG or 3D contexts when compute and data allow. A side-by-side of backbone families, data needs, and dental fit is provided in Table 8.

4.2 Task heads vs. clinical questions

- Classification: best for screening; requires calibrated thresholds and a clear intended use (triage vs. confirmatory) [20].
- Detection: localizes focal findings (e.g., proximal caries). Use Focal Loss for class/anchor imbalance [17].
- Segmentation: needed when geometry/staging matters (bone loss, lesion extent); report Dice/IoU and small-structure analyses [18,19].
- Seg→Cls (segmentation-assisted classification): two-stage (masks → region features → class) that improves sensitivity to small, low-contrast lesions and enables plausibility overlays; the workflow is illustrated in Figure 3. Trade-offs are extra annotation and two-stage complexity [9,12].
- DETR-style: cleaner priors with end-to-end set prediction, but typically more data-hungry and slower to converge [39].

To turn these principles into quick, actionable choices, Table 8 provides a concise map that links each clinical question to the most suitable head (Classifier / Detector / Segmenter / Seg→Cls / DETR-style), detailing outputs, preferred use cases, key pros/cons, typical dental applications, and practical notes (e.g., calibration, class-imbalance handling). Read Table 8 alongside Figure 3 and report under imbalance with PR-AUC in addition to ROC-AUC, evaluated at a pre-specified specificity (e.g., ≥0.90) with sensitivity, PPV, and 95% CIs.

Table 8. Task heads: output, when to prefer, pros/cons, and typical dental use
(Families are representative, not exhaustive. Select according to data scale, resolution, and deployment constraints.)

Head	Output	Prefer when…	Pros	Cons	Typical dental use	Notes
Classifier	**Image/region label**	**Screening; global status**	**Simple; fast**	**No location**	**RGB multi-label; OPG screening**	**Calibrate thresholds**
Detector	**Boxes + scores**	**Focal lesions; triage**	**Localizes findings**	**Misses shape**	**Caries/periapical cues on BW/PA**	**Focal loss helps [17]**
Segmenter	**Pixel/voxel mask**	**Geometry/staging needed**	**Precise extent**	**Annotation cost**	**Bone loss; periapical masks**	**Report Dice/IoU[18,19]**
Seg→Cls	**Mask features → class**	**Small/low-contrast lesions**	**Boosts sensitivity**	**Two-stage complexity**	**Periapical radiolucencies**	**Good for subtle cues [9,12]**
DETR-style	**Set of objects**	**End-to-end, fewer priors**	**Clean design**	**Data-hungry**	**Panoramic multi-finding**	**Longer training [39]**

Metrics note. For detection, add AP/mAP alongside PR-AUC; for classifiers, report PR-AUC and sensitivity/PPV at pre-specified specificity.

4.3 Imbalance, thresholds, and calibration

Under skew, we recommend treating PR-AUC as mandatory alongside ROC-AUC and pre-specifying a fixed specificity (e.g., ≥ 0.90) with sensitivity/PPV and 95% CIs, following TRIPOD+AI and standard imbalanced-data practice [27,35]. Temperature scaling and reliability diagrams (ECE) turn scores into decision-useful probabilities [20]. For safety, adopt selective prediction to abstain on low-confidence cases [41].

4.4 *Interpretability and reader workflow*

Grad-CAM overlays guide plausibility checks and error curation; IG/SHAP add cohort-level insigh . Prospective reader studies are still sparse, but XAI panels are routinely requested by clinicians and can shorten adjudication in discordant cases when presented with structured summaries (per-tooth/per-region outputs) [42-44].

4.5 Engineering for deployment

Clinical viability requires disciplined engineering: pinned seeds/packages; saved configs; export to ONNX/TensorRT with FP16/INT8 as appropriate; batch-1 latency and memory footprint on target hardware; structured outputs (DICOM-SR/JSON); model cards and a fail-closed/abstention policy [27][52]

5. Challenges

This section consolidates the principal barriers to reliable dental AI—data/label quality, class imbalance, domain shift, multimodal fusion, calibration/uncertainty, and clinician-usable explainability—and frames concrete safeguards for each.

5.1 Data scarcity, label quality, and governance

Multi-site, diverse datasets remain rare; label noise (e.g., subtle proximal caries) is common. Best practice: patient-level splits, site/scanner stratification, ≥2 expert readers with blinded re-reads, and a clear hierarchical taxonomy; report κ for agreement [27].

5.2 Class imbalance and clinically aligned operation

Imbalance is the norm in dental imaging: Combine data-level remedies (minority oversampling + stronger augmentation) with loss-level choices (class-weighted CE, Focal Loss, and Unified Focal Loss for segmentation) [39]. Because clinical adoption hinges on specificity-constrained operation, thresholds must be pre-specified on validation data (e.g., specificity ≥ 0.90) and results reported at those thresholds (sensitivity and PPV with 95% CIs), not only overall AUCs [17-19,35].

5.3 Domain shift and cross-site generalization

Cross-site performance often drops due to device/protocol differences. Include external cohorts and site-wise reporting; benchmark domain adaptation and test-time adaptation baselines (DANN, Deep CORAL, TENT) [46,43,47]. For CBCT, disclose MAR and its effect on intensity statistics [26].

5.4 Multimodal fusion pitfalls

Fusion helps only with accurate registration and harmonized inputs. Enforce registration QA; analyze missing-modality scenarios; ablate against single-modality baselines [30,42].

5.5 Calibration, uncertainty, and selective prediction

Modern networks are over-confident; apply temperature scaling; quantify ECE; explore ensembles or MC-dropout to enable risk-coverage curves and abstention [34,44].

5.6 Explainability that clinicians can use

Grad-CAM, Integrated Gradients, and SHAP assist plausibility checks and error triage but are not proof of correctness [36–38,71,44]. Reader-friendly panels should align saliency with known radiologic signs and surface shortcut cues (acquisition markers, metal artifacts).

5.7 Reproducibility and reporting

Follow TRIPOD+AI: transparent splits; internal/external results; ROC-/PR-AUC; per-class precision/recall/F1; κ; Dice/IoU; calibration plots; 95% CIs; DeLong for correlated AUCs; decision thresholds; latency/memory [27,69].

5.8 Privacy, fairness, and auditing

Federated learning requires audit trails (data curation, update logs, fairness checks); if using differential privacy, report (ε, δ) and the utility trade-off (e.g., ΔPR-AUC at fixed specificity). Publish subgroup metrics (age/sex/device/site) and discuss shortcut risks [67,68,70,44].

6. Future Directions

- Label-efficient learning at scale. Combine self-supervised pretraining with semi/weak supervision to reduce annotation burden while improving recall and calibration across modalities [63].
- Routine shift-robustness. Make domain adaptation (DANN/Deep CORAL) and test-time adaptation (TENT) standard baselines; always include site-wise deltas on external cohorts [64,72,74].
- Calibration-first pipelines. Treat calibration and uncertainty as first-class outcomes — publish reliability diagrams/ECE, Brier, and define abstention policies tuned on validation [34].
- Clinically usable fusion. Standardize CBCT↔IOS protocols (registration QA metrics, harmonization) and document MAR side-effects; evaluate against robust single-modality baselines [30,31,42].
- Reader/workflow studies. Move beyond retrospective metrics to prospective, multi-center reader studies tracking time-to-decision, discordant-case triage, and the utility of XAI overlays [44,70].
- Privacy-preserving collaboration with auditing. Develop federated frameworks with verifiable site-level audits (quality, fairness, drift) and quantify DP trade-offs on sensitivity at fixed specificity [67,68].
- Decision rules under constraints. Provide practical "when-to-use-what" guidance (conv vs. transformer; classifier vs. detector vs. segmentation vs. Seg→Cls) keyed to data scale, lesion size/contrast, FOV, and latency/memory budgets (§§2.5, 4).

7. Recommendations (Actionable Checklist)

In summary, we recommend:

7.1 Data & labeling

- Split at the patient level, stratify by site/scanner; include ≥ 1 external cohort [27,44];
- Publish a labeling protocol (taxonomy + decision rules) and an adjudication flow (≥2 experts; ≈10% blinded re-reads); report inter-rater κ;
- When dense masks are costly, combine self-supervised pretraining with semi/weak supervision rather than shrinking scope.

7.2 Objectives & metrics

- Treat PR-AUC as mandatory alongside ROC-AUC for imbalanced problems.
- Pre-specify thresholds on validation (e.g., specificity ≥ 0.90), then report sensitivity and PPV with 95% CIs at those fixed thresholds on internal and external tests.
- Use task-appropriate metrics: mAP/AP (detection), Dice/IoU (segmentation), and per-class precision/recall/F1 (classification).

7.3 Calibration & uncertainty

- Apply temperature scaling (or isotonic) and publish reliability diagrams with ECE (optionally Brier).
- Define a selective-prediction policy (when to abstain) and quantify the risk–coverage trade-off.

7.4 Shift robustness

- Provide site-wise performance and cross-site deltas; where relevant, add domain adaptation or test-time adaptation baselines.

- For CBCT, document MAR usage and intensity harmonization and analyze their impact.

7.5 Backbones & heads (fit to constraints)

- Choose encoders/heads by data scale, lesion size/contrast, resolution, and latency/memory budget (see §4 tables).
- Prefer Seg→Cls when small, low-contrast lesions are clinically critical.

7.6 Explainability (XAI)

- Provide clinician-readable panels (Grad-CAM / Integrated Gradients / SHAP), sanity checks, and failure galleries linking saliency to recognized radiologic signs.

7.7 Engineering & deployment

- Report batch-1 latency, memory footprint, and throughput at clinical resolution on target hardware.
- Emit structured outputs (DICOM-SR/JSON) with per-tooth/per-region fields to ease PACS/RIS integration.
- Release model cards (intended use, cohorts, thresholds, limitations) and document fail-closed/abstention behavior.
- Pin seeds/packages, track configs; export for inference (ONNX/TensorRT; consider FP16/INT8).

7.8 Privacy & fairness

- In federated settings, define who audits site contributions, fairness, and drift; publish audit summaries.
- If using differential privacy, report ε/δ and the performance impact at fixed specificity.

7.9 Reproducibility & transparency (add)

- Provide code and exact configs, dataset split manifests, and versioned model artifacts sufficient for third-party replication.
- Align reporting with TRIPOD+AI items (checklist in supplement).

7.10 Post-deployment monitoring (add)

- Establish a plan for monitoring calibration and performance drift, periodic re-calibration, and subgroup audits; log abstentions and clinician overrides.

7.11 Practical deployment rule

If, after calibration, specificity-constrained targets on an external cohort are not achieved, the model must be deployed only as a second reader with selective abstention—never as an autonomous gatekeeper. Promotion to autonomous use should occur only after the model meets and sustains those externally validated, specificity-constrained targets under post-deployment monitoring.

8. Conclusion

This review synthesizes DL methods for assessing dental anomalies and diseases across intraoral RGB, BW/PA radiographs, panoramic OPG, and 3D CBCT/IOS, emphasizing modality-aware preprocessing, imbalance-aware objectives, and calibration as prerequisites for decision-useful AI. Compared with prior surveys, we center clinically constrained operation—pre-specified, high-specificity thresholds with calibrated probabilities and selective abstention—alongside engineering artifacts (latency, memory) and structured outputs for integration. Persistent gaps include small single-center datasets, limited cross-site robustness, under-reported calibration/uncertainty, and scarce prospective reader/workflow studies. Moving forward, label-efficient learning, routine shift-robustness baselines, and federated, auditable collaboration are essential. Critically, site-stratified external validation should be a gating criterion before deployment. With these guardrails, dental AI can progress from retrospective promise to dependable, safety-preserving and workload-reducing support that augments—rather than replaces—expert judgment.

References

1. Organization, W.H. Global Oral Health Status Report: Towards Universal Health Coverage for Oral Health by 2030; World Health Organization, 2022;
2. Bernabé, E.; Marcenes, W.; Hernandez, C.R.; et al. Global, Regional, and National Levels and Trends in Burden of Oral Conditions from 1990 to 2017. Journal of Dental Research 2020, 99, 362–373, doi:10.1177/0022034520908533.
3. Whaites, E.; Drage, N. Essentials of Dental Radiography and Radiology; 6, Ed.; Elsevier, 2023; ISBN 9780702074610.
4. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.; 2021.
5. Liu, Z.; et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows.; 2021; pp. 9992–10002.
6. Zhu, J.; Chen, Z.; Zhao, J.; et al. Artificial Intelligence in the Diagnosis of Dental Diseases on Panoramic Radiographs: A Preliminary Study. BMC Oral Health 2023, 23, 358, doi:10.1186/s12903-023-03158-7.
7. Galić, I.; Habijan, M.; Leventić, H.; Romić, K. Machine Learning Empowering Personalized Medicine: A Comprehensive Review of Medical Image Analysis Methods. Electronics 2023, 12, 4411, doi:10.3390/electronics12194411.
8. Azizi, S.; Mustafa, B.; Ryan, F.; et al. Big Self-Supervised Models Advance Medical Image Classification.; 2022; pp. 3478–3488.
9. Shon, H.S.; Kong, V.; Park, J.S.; et al. Deep Learning Model for Classifying Periodontitis Stages on Dental Panoramic Radiography. Applied Sciences 2022, 12, 8500, doi:10.3390/app12178500.
10. Razi, T.; et al. Accuracy of Bone-Loss Measurement on Digital Panoramics vs Bitewings. Imaging Science in Dentistry 2022, 52, 167–175, doi:10.5624/isd.20210190.
11. Park, W.; Kim, H.; Lee, J.; et al. Deep Learning for Screening Multiple Oral Conditions from Intraoral Photographs. Scientific Reports 2022, 12, 16413, doi:10.1038/s41598-022-21055-3.
12. Alalharith, D.M.; et al. A Deep Learning-Based Approach for the Detection of Early Signs of Gingivitis in Orthodontic Patients Using Faster R-CNN. International Journal of Environmental Research and Public Health 2020, 17, 8447, doi:10.3390/ijerph17228447.
13. Liu, J.; et al. Deep Learning-Enabled 3D Multimodal Fusion of Cone-Beam CT and Intraoral Mesh Scans for Clinically Applicable Tooth-Bone Reconstruction. Patterns 2023, 4, doi:10.1016/j.patter.2023.100825.
14. Han, X.; et al. Digital Registration versus Cone-Beam Computed Tomography for Evaluating Implant Position: A Prospective Cohort Study. BMC Oral Health 2024, 24, 304, doi:10.1186/s12903-024-03938-3.
15. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the ICCV; 2019; pp. 6023–6032.
16. Zhang, H.; Cissé, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. In Proceedings of the ICLR; 2018.
17. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the ICCV; 2017; pp. 2980–2988.
18. Abraham, N.; Khan, N.M. A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation. In Proceedings of the ISBI; 2019; pp. 683–687.
19. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Cardoso, M.J. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In DLMIA/ML-CDS (LNCS 10553); 2017; pp. 240–248.
20. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the ICML (PMLR); 2017; pp. 1321–1330.
21. Pitts, N.B.; et al. Bitewing Radiography for Proximal Caries—Best Practice. Community Dental Health 2019, 36, 206–214.
22. Farman, A.G.; Farman, T.T. Extraoral and Panoramic Imaging: Principles and Applications. Dental Clinics of North America 2008, 52, 715–728, doi:10.1016/j.cden.2008.03.003.
23. Baan, F.; Bruggink, R.; Nijsink, J.; Maal, T.; Ongkosuwito, E. Fusion of Intra-Oral Scans in Cone-Beam Computed Tomography Scans. Clinical oral investigations 2021, 25, 77–85.
24. Kim, Y.-J.; Ahn, J.-H.; Lim, H.-K.; Nguyen, T.P.; Jha, N.; Kim, A.; Yoon, J. Novel Procedure for Automatic Registration between Cone-Beam Computed Tomography and Intraoral Scan Data Supported with 3D Segmentation. Bioengineering 2023, 10, 1326.
25. Buslaev, A.; Iglovikov, V.; Khvedchenya, E.; et al. Albumentations: Fast and Flexible Image Augmentations. Information 2020, 11, 125, doi:10.3390/info11020125.
26. Hegazy, M.A.; Cho, M.H.; Cho, M.H.; Lee, S.Y. Metal Artifact Reduction in Dental CBCT Images Using Direct Sinogram Correction Combined with Metal Path-Length Weighting. Sensors 2023, 23, 1288.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition.; 2016; pp. 770–778.

28. Krichen, M.; Ketata, R.; Alshammari, R.; Alshammari, M. Convolutional Neural Networks: A Survey. Computers 2023, 12, 151.
29. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks.; 2017; pp. 4700–4708.
30. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision.; 2016; pp. 2818–2826.
31. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions.; 2017; pp. 1251–1258.
32. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.; 2019; pp. 6105–6114.
33. Tan, M.; Le, Q.V.; Chen, B.; Pang, R.; Others EfficientNetV2: Smaller Models and Faster Training.; 2021; pp. 10096–10106.
34. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A Convnet for the 2020s. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022; pp. 11976–11986.
35. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. ConvNeXt V2: Co-Designing and Scaling ConvNets with Masked Autoencoders.; 2023.
36. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.; Dollár, P. Designing Network Design Spaces (RegNet).; 2020; pp. 10428–10436.
37. Carion, N.; Massa, F.; Synnaeve, G.; et al. End-to-End Object Detection with Transformers. In Proceedings of the ECCV; 2020.
38. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. Journal of Big Data 2019, 6, 60, doi:10.1186/s40537-019-0197-0.
39. Hegazy, A.; et al. Metal-Artifact Reduction in Dental CBCT and Its Downstream Effects on Analysis. Physics in Medicine & Biology 2023, 68, 145007, doi:10.1088/1361-6560/ace8d8.
40. US Food and Drug Administration Good Machine Learning Practice for Medical Device Development: Guiding Principles. FDA webpage 2021.
42. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks (Integrated Gradients).; 2017; pp. 3319–3328.
43. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the NeurIPS; 2017.
44. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the ICML; 2016.
45. Ganin, Y.; Ustinova, E.; Ajakan, H.; et al. Domain-Adversarial Training of Neural Networks. Journal of Machine Learning Research 2016, 17, 1–35.
46. Sun, B.; Saenko, K. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In Proceedings of the ECCV Workshops; 2016.
47. Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; Darrell, T. TENT: Fully Test-Time Entropy Minimization. In Proceedings of the ICLR; 2021.
48. Ikhwani, Y.; Noersasongko, E.; Soeleman, M.A.; others Comparative Performances of the Convolutional Neural Network Based Transfer Learning Models for Classification of Dental Disease. In Proceedings of the 2024 International Seminar on Application for Technology of Information and Communication (iSemantic); IEEE, 2024; pp. 445–450.
49. IHE Radiology Technical Committee IHE Radiology Technical Framework Supplement: AI Results (AIR). 2025.
50. Liu, J.; Liu, X.; Shao, Y.; Gao, Y.; Pan, K.; Jin, C.; Ji, H.; Du, Y.; Yu, X. Periapical Lesion Detection in Periapical Radiographs Using the Latest Convolutional Neural Network ConvNeXt and Its Integrated Models. Scientific Reports 2024, 14, 25429, doi:10.1038/s41598-024-75748-9.
51. Thalji, N.; Aljarrah, E.; Almomani, M.H.; Raza, A.; Migdady, H.; Abualigah, L. Segmented X-Ray Image Data for Diagnosing Dental Periapical Diseases Using Deep Learning. Data in Brief 2024, 54, 110539, doi:10.1016/j.dib.2024.110539.
52. Minderer, M.; Djolonga, J.; Romijnders, R.; Hubis, F.; Zhai, X.; Houlsby, N.; Tran, D.; Lucic, M. Revisiting the Calibration of Modern Neural Networks.; Curran Associates, Inc., 2021; pp. 15682–15694.