



Diagnostic accuracy of deep learning for detecting dental caries on bitewing and periapical radiographs: a systematic review

Precisión diagnóstica del aprendizaje profundo para detectar caries dental en radiografías de aleta de mordida y periapicales: revisión sistemática

Greta Cordovi Vallongo¹ 

Marta María Capote Fernández¹ 

¹ Universidad de Ciencias Médicas de La Habana. La Habana, Cuba.

Submitted: 03-02-2025 | Revised: 05-04-2025 | Accepted: 27-07-2025 | Published: 23-08-2025

Corresponding Author: Marta María Capote Fernández 

ABSTRACT

Introduction: Deep learning models have been proposed to automate caries detection on intraoral radiographs, but reported accuracy varies widely and the evidence has not been synthesized specifically for bitewing and periapical projections. **Methods:** Following PRISMA 2020, PubMed, OpenAlex, and Crossref were searched for studies published from 2018 to 2025 that evaluated deep learning models for detecting, classifying, or segmenting dental caries on bitewing or periapical radiographs against a reference standard. Risk of bias was assessed with QUADAS-2. **Results:** Thirty-nine studies were included, most published in 2024 and 2025 and most using bitewing radiographs and YOLO architectures. Reported sensitivity ranged from 13.7% to 99.4%, specificity from 70.6% to 99.9%, and the area under the curve reached 0.954. The single histologically validated study reported sensitivity of 13.7% and specificity of 95.9%. **Conclusions:** Deep learning attains high but highly heterogeneous accuracy for caries detection on intraoral radiographs. Estimates often exceed the accuracy expected of expert readers but weaken when validated against histology, and prospective external validation is still lacking.

Keywords: deep learning; dental caries; radiography dental; diagnosis computer-assisted; sensitivity and specificity; systematic review.

RESUMEN

Introducción: Se han propuesto modelos de aprendizaje profundo para automatizar la detección de caries en radiografías intraorales, pero la precisión reportada es muy variable y no se ha sintetizado de forma específica para proyecciones de aleta de mordida y periapicales. **Métodos:** Siguiendo PRISMA 2020, se buscó en PubMed, OpenAlex y Crossref estudios publicados entre 2018 y 2025 que evaluaran modelos de aprendizaje profundo para detectar, clasificar o segmentar caries dental en radiografías de aleta de mordida o periapicales frente a un estándar de referencia. El riesgo de sesgo se evaluó con QUADAS-2. **Resultados:** Se incluyeron treinta y nueve estudios, la mayoría publicados en 2024 y 2025 y la mayoría con radiografías de aleta de mordida y arquitecturas YOLO. La sensibilidad reportada osciló entre 13,7% y 99,4%, la especificidad entre 70,6% y 99,9%, y el área bajo la curva alcanzó 0,954. El único estudio validado histológicamente reportó sensibilidad de 13,7% y especificidad de 95,9%. **Conclusiones:** El aprendizaje profundo alcanza una precisión alta pero muy heterogénea para la detección de caries en radiografías intraorales. Las estimaciones suelen superar la precisión esperada de lectores expertos, pero se debilitan al validarse frente a histología, y aún falta validación externa prospectiva.

Palabras clave: aprendizaje profundo; caries dental; radiografía dental; diagnóstico por computador; sensibilidad y especificidad; revisión sistemática.

INTRODUCTION

Untreated caries of permanent teeth remains among the most prevalent conditions worldwide and contributes a substantial share of the global burden of oral disorders, a burden that has persisted across regions despite decades of preventive programs and water fluoridation.⁽¹⁾ The condition is largely preventable and treatable when detected early, which makes accurate detection a public health priority rather than a purely technical concern. Intraoral radiographs, particularly bitewing and periapical projections, are indispensable for detecting proximal and occlusal lesions that visual and tactile examination alone misses, and standardized frameworks such as the International Caries Detection and Assessment System organize this assessment by staging lesions from initial enamel changes to extensive cavitation.⁽²⁾

Visual-tactile examination and conventional radiographic interpretation are nonetheless subject to substantial observer variability. Detection of approximal lesions on bitewings is especially difficult because the lesions lie hidden between adjacent teeth, and agreement among examiners is only moderate for early enamel lesions, which are precisely the lesions for which minimally invasive treatment is most effective. Radiographic interpretation is also imperfect: a proportion of proximal lesions are missed on bitewings, and the false-positive rate varies with lesion depth and examiner experience. This variability can delay treatment, can promote overtreatment or undertreatment, and carries measurable clinical and economic costs, which is why automated and reproducible detection has attracted sustained interest.

Deep learning, and convolutional neural networks in particular, has been applied increasingly to dental imaging over the past decade. Reported performance has approached or exceeded that of experienced clinicians across detection, classification, and segmentation tasks, and the technology has moved from small proof-of-concept demonstrations on a few hundred images to models trained and tested on several thousand images.⁽³⁻⁵⁾ The clinical appeal is straightforward: a reliable automated second reader could reduce missed proximal lesions, standardize interpretation across examiners of different experience, and free the clinician to focus on treatment decisions rather than on the detection step. Object detectors, segmentation networks, and classification architectures have all been proposed for radiographic caries detection. Object detectors such as the YOLO family localize lesions with bounding boxes, segmentation networks such as U-Net delineate lesion boundaries, and classification networks assign an image or a region to a lesion category. Each formulation answers a different clinical question, from whether caries is present to where the lesion lies and how deep it extends, and the choice of formulation determines which accuracy metric is reported.

Several systematic reviews have summarized artificial intelligence for caries detection, but they differ in scope, imaging modality, and search period, and a number of them combine intraoral radiographs with panoramic radiographs or clinical photographs, which complicates the interpretation of pooled accuracy.⁽⁶⁻⁹⁾ Reviews that include panoramic radiographs dilute the evidence specific to bitewings, where the diagnostic challenge and the clinical workflow differ, and reviews published before 2024 cannot reflect the substantial number of detection studies that have since appeared. Many predate the rapid expansion of primary studies published in 2024 and 2025, and none focuses specifically on intraoral radiographs while distinguishing bitewing from periapical projections. A synthesis confined to intraoral projections and updated through 2025 is therefore needed to inform clinicians and to guide regulatory evaluation of these tools. The aim of this systematic review was to synthesize the diagnostic accuracy of deep learning models for detecting dental caries on bitewing and periapical radiographs relative to expert annotation, and to characterize the architectures, reference standards, and methodological quality of the included studies.

METHODS

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement.⁽¹⁰⁾

Studies were eligible when they were original investigations that developed or evaluated a deep learning or machine learning model for detecting, classifying, or segmenting dental caries on bitewing or periapical intraoral radiographs and reported at least one quantitative performance measure against a reference standard. The accepted performance measures were sensitivity, specificity, area under the receiver operating characteristic curve, Dice or intersection over union, accuracy, or F1-score. Studies whose primary input was a panoramic radiograph, cone-beam computed tomography, or a clinical photograph were excluded, as were reviews, editorials, protocols,

and studies reporting no quantitative metric. The publication window was 2018 to 2025. The lower bound of 2018 was chosen because deep learning studies of radiographic caries detection appeared mainly from that year onward, and the upper bound of 2025 matches the journal volume for which this review was prepared. The task definitions were deliberately broad: detection comprised studies that localized caries lesions, classification comprised studies that assigned a lesion or image to a severity or anatomic category, and segmentation comprised studies that delineated the lesion pixel by pixel, with studies combining tasks assigned to all relevant categories.

Information sources and search

PubMed, OpenAlex, and Crossref were searched for records published between 2018 and 2025. Queries combined terms for deep learning, dental caries, and intraoral radiography, and several synonymous strings were run in each source to maximize recall. The PubMed query set comprised three strings: ("deep learning" OR "convolutional neural network" OR CNN) AND "dental caries" AND (bitewing OR periapical OR "intraoral radiograph") AND (detection OR diagnosis OR classification); "artificial intelligence" AND (caries OR "cariou lesions") AND (radiograph OR radiography) AND (detection OR diagnostic); and "deep learning" AND (caries OR "cariou lesion") AND ("sensitivity" OR "specificity" OR "diagnostic accuracy" OR "area under"). Equivalent free-text queries were adapted to the OpenAlex and Crossref syntax, and each source returned up to 50 records per query. A Semantic Scholar query was attempted but returned a rate-limit error and contributed no records; the remaining three sources provided all retrieved records, which were deduplicated by digital object identifier and by title before screening.

Study selection

After duplicate removal, screening proceeded in two stages. Title and abstract screening excluded clearly irrelevant records, and uncertain records were carried forward rather than excluded for lack of information. Full text was available for 14 reports; for the remaining reports, eligibility and data extraction relied on the abstract, which reports the key performance metrics in most diagnostic accuracy studies.

Data extraction

For each included study the following items were extracted: first author and year, country, model architecture, imaging projection, sample size, reference standard, task, and reported diagnostic performance. Values were recorded as reported, without recalculation, and when a study reported several metrics for different lesion severities or dentitions, all were retained so that the range of performance could be described. Where a study reported accuracy separately for enamel and dentin lesions or for primary and permanent teeth, the stratum-specific values were recorded separately, so that the analysis could distinguish severity and dentition effects rather than collapsing them into a single average.

Risk of bias assessment

Risk of bias and applicability were assessed with QUADAS-2 across patient selection, the index test, the reference standard, and flow and timing.⁽¹¹⁾ Because most studies reported no prospective registration and no independent external validation, risk of bias was summarized narratively by domain rather than quantified per study with a traffic-light table.

Synthesis

A meta-analysis was not performed because the studies reported heterogeneous tasks, metrics, and reference standards, and pooling sensitivity or specificity across detection, classification, and segmentation studies would not be meaningful. A pooled analysis would additionally require each study to report the same accuracy metric with an estimable variance, and only a minority of the included studies reported the confidence intervals or the contingency tables needed to derive it. Findings were therefore synthesized narratively, grouped by imaging projection and by performance metric, and heterogeneity was described rather than statistically tested.

RESULTS

Study selection

The searches retrieved 350 records, of which 84 duplicates were removed and 266 unique records were screened (Figure 1). PubMed contributed 150 records, OpenAlex 100, and Crossref 100, with duplicates arising from records indexed in more than one source. Title and abstract screening excluded 194 records, leaving 72 reports assessed for eligibility. Of these, 33 were excluded: 9 because the imaging input was not intraoral radiography, 13 because no extractable accuracy metric was reported, 5 because the outcome was not caries, 4 because the report was not an original study, 1 because the publication year fell outside the window, and 1 because the reference could not be verified. The 13 exclusions for a missing accuracy metric reflect records whose abstract or full text did not report a quantitative measure, and they were excluded rather than carried forward with imputed values. Thirty-nine studies were included.

Characteristics of included studies

The included studies spanned 2020 to 2025, with the majority published in 2025 (21 studies) and 2024 (9 studies), and only one study published in 2020 (Table 1). This distribution documents the recent acceleration of the field: the number of studies more than doubled between 2023 and 2025.

Most studies evaluated models on bitewing radiographs (28 of 39); the remainder used periapical radiographs(8), both projections(2), or unspecified intraoral radiographs(1). The bitewing studies concentrated on approximal and interproximal lesions, whereas the periapical studies additionally addressed occlusal and secondary lesions. One study focused on primary molars,(12) one on pediatric patients aged 5 to 12 years,(13) and one on caries beneath restorations and orthodontic brackets.(14)

The YOLO family of object detectors was the most common architecture, appearing across successive versions from YOLOv5 to YOLOv11.(14-25)The included studies track the release cycle of these detectors: YOLOv5 and YOLOv7 appear in the 2021 to 2024 studies,(15,19) YOLOv8 in the 2023 to 2025 studies,(16,26) and YOLOv9 and YOLOv11 in the most recent 2025 reports.(17,18,21)Studies from different years are therefore not directly comparable, because the base detector improved between releases. Segmentation networks based on U-Net or on Mask R-CNN with a transformer backbone were also frequent.(27-31)A smaller group applied classification networks, including VGG19, Inception, and a ConvNeXt variant.(32-37)Several studies compared multiple detectors or ensembles and reported the best-performing variant.(25,38,39)In terms of task, 31 studies performed detection, 6 performed classification, and 3 performed segmentation.

The reference standard was expert reader annotation in nearly all studies, whether a single experienced dentist, a consensus panel, or calibrated specialists. Only one study validated against histology, the least biased criterion.(40) Geographically, the countries reported in the included studies were predominantly in Europe and Asia, with a small number from the Americas, although many studies did not report the country of origin, which constrains any assessment of whether the models generalize to populations and radiographic conventions beyond those studied. Sample sizes ranged from fewer than 150 images(23,33) to several thousand, and the largest population-based dataset comprised almost 14,000 bitewings drawn from a regional health study.(39)A few studies evaluated commercially available software rather than an in-house model, including Diagnocat, Second Opinion, and Denti.AI.(40-42)

The datasets used to train and evaluate these models differed in origin and in size. The smallest studies used between 100 and 150 radiographs,(23,33,43)whereas the largest used several thousand images, including a population-based cohort of almost 14,000 bitewings(39)and a multi-center collection of 3,989 bitewings.(25)Dataset composition also varied: some studies restricted the analysis to a single dentition or a single lesion location, such as approximal surfaces,(12,44) whereas others mixed primary and permanent teeth and multiple lesion types.(15,28) The training data therefore ranged from highly curated single-center collections to heterogeneous real-world collections, which is a plausible contributor to the spread of reported performance.

Risk of bias

Across QUADAS-2 domains, the reference standard was the principal source of concern because it was expert annotation rather than an independent criterion in nearly all studies, and only one study used histology.(40)Expert annotation, especially when produced by a single reader, risks misclassifying lesions and can inflate model estimates because the model is trained to reproduce

the same reader it is then compared against.

Patient and image selection raised further concerns, because cohorts were assembled retrospectively from single institutions in most studies, and the authors of many reports did not describe how images were selected or excluded. Flow and timing were generally acceptable because the index test and the reference standard were applied to the same images, but blinding was reported inconsistently, and few studies reported external validation on independent cohorts. Exceptions included population-based and multi-center datasets that offered a more representative test set. (25,30,39)

In the patient selection domain, most studies did not report consecutive or random sampling, and several used retrospectively collected images from a single department, which limits applicability to other populations and acquisition protocols. In the index test domain, the models were generally described in sufficient technical detail to be reproduced, but preprocessing, augmentation, and hyperparameter choices were reported inconsistently. The reference standard domain carried the greatest weight of concern: when the reference labels derive from expert readers, the model is evaluated against human judgement rather than against the true presence of a lesion, and agreement between readers is imperfect, so the target itself contains error. Only the histologically validated study addressed this directly.(40) In the flow and timing domain, almost all studies applied the model and the reference standard to the same set of images, which is appropriate, but very few held out a random subset from model development to serve as a truly independent test set, and several did not state whether the test images overlapped with the training distribution. The overall judgement was therefore that the risk of bias was moderate to high in most studies, driven principally by the reference standard and selection domains, and that the applicability of the findings to routine clinical populations was uncertain in the majority of reports.

Diagnostic accuracy

Twenty studies reported sensitivity, fifteen reported specificity, and six reported an area under the receiver operating characteristic curve (Table 2). Reported sensitivity ranged from 13.7% in the histologically validated study(40) to 99.4% for a model trained to detect caries under restorations and brackets.(14) Sensitivity at or above 90% was reported by a substantial group,(13,26,42,43,45-47) whereas another group reported sensitivity in the range of 51% to 73%.(18,24,30,41,44) Intermediate estimates were reported for primary molars, with sensitivity of 50.9% for dentin caries requiring treatment,(12) and for a multiclass periapical task with sensitivity of 17.4%.(20) A self-supervised approach attained a sensitivity of at least 45% with very few annotations.(35)

Sensitivity differed by imaging projection and by dentition. On bitewing radiographs, sensitivity ranged from 50.9% for dentin caries in primary molars(12) to 99.4% for caries under restorations and brackets,(14) and most detection studies reported values between 72% and 93%.(16,19,24,26,30,44,45) On periapical radiographs, sensitivity ranged from 17.4% in a multiclass task(20) to 94.0%.(43) For primary dentitions, recall was 0.78 for primary and 0.75 for permanent teeth in one segmentation study,(15) and sensitivity for primary molars was lower than the values reported for permanent teeth.(12) The studies that focused on children reported sensitivity between 88% and 89%.(13,46)

Stratification by lesion depth showed that enamel lesions were detected with different accuracy than dentin lesions. One detector reported recall of 96.0% for enamel and 80.1% for dentin lesions,(16) whereas another reported sensitivity of 0.74 for all secondary lesions and 0.81 for dentin lesions.(30) The direction of this difference was not uniform across studies, because the models differed in how they were trained to handle shallow relative to deep lesions.(18,41)

Specificity ranged from 70.6%(43) to 99.9%,(14) and most studies that reported both sensitivity and specificity obtained specificity above 85%.(19,30,42,44-46) The histologically validated study reported specificity of 95.9% alongside its low sensitivity,(40) which indicates that the model was conservative and missed many lesions rather than over-calling them.

The area under the curve ranged from 0.548 for the histologically validated study(40) to 0.954 for a system combining ResNet with a segmentation module.(48) Intermediate values were reported for secondary caries detection, with areas under the curve of 0.94 for all lesions and 0.95 for dentin lesions in one study,(30) and near 0.80 for primary and secondary caries in another.(31) A study that validated software against expert consensus reported an area under the curve of 0.94 for caries.(47)

A subset of the detection studies reported the full set of sensitivity, specificity, and accuracy. On bitewings, a software-assisted interpretation reported sensitivity of 91.2%, specificity of 87.1%, and accuracy of 89.4%;(45) a YOLOv7 model reported sensitivity of 83.3%, specificity of 96.1%, and accuracy of 93.4%;(19) a Faster R-CNN model reported sensitivity of 72%, specificity of 93%, and ac-

curacy of 87%;(44)and a modified YOLO detector reported sensitivity of 72.3%, specificity of 98.2%, and accuracy of 94.6%.(24)On periapical radiographs, a software evaluated across several conditions reported sensitivity of 91.0% and specificity of 87% for caries,(42)and a deep learning system reported sensitivity of 89.3% and accuracy of 90.8% in children.(13)Other studies reported a single or partial set of metrics, such as recall without specificity(16,17,49)or mean average precision without sensitivity,(21,39) which limited direct comparison.

Other detection studies reported object-detection metrics rather than sensitivity and specificity. A YOLOv11 model for proximal caries reported recall of 0.86, precision of 0.84, an F1-score of 0.85, and a mean average precision of 0.89.(17) An ensemble of four detectors reported an F1-score between 0.79 and 0.80 and average precision of 0.86.(25) A lightweight detector deployed as a smartphone application reported sensitivity of 0.75, precision of 0.85, and an F1-score of 0.80.(23)A U-Net evaluated on 50 bitewings reported precision of 63.3%, recall of 65.0%, and an F1-score of 64.1%.(29) A transfer learning model applied to bitewing films reported accuracy of 90.3% for caries,(37)and a model combining a super-resolution network with YOLOv8 reported a mean average precision of 0.57 across six conditions, with sensitivity of 17.4% and specificity of 88.7% for caries specifically.(20)

Segmentation and classification approaches reported related metrics. Segmentation networks obtained Dice or intersection over union coefficients between roughly 0.50 and 0.90,(15,27,28,36) with the highest values for advanced or cavitated lesions and the lowest for initial enamel lesions. Classification networks reported accuracy from 73.3% for a small Inception model(33)to 93.9% for a VGG19 model,(32)and 93.3% for a ConvNeXt model applied to the G.V. Black classification.(34)A comparison of four detectors on periapical radiographs reported F1-scores of 0.87 for YOLOv5, 0.86 for Trans-UNet, 0.82 for DETR, and 0.80 for U-Net.(38)A commercial software evaluated on periapical radiographs reached an accuracy of only 61% for caries,(50)which illustrates the wide spread of reported performance across tools and datasets.

Studies that compared models against clinicians generally found the model at least comparable to experienced readers,(13,18,49)and one detection model outperformed six dentists on recall and F1-score.(18)A large population-based evaluation of several detectors reported a best mean average precision of 0.65 and a false-negative rate of 0.15, lower than the best human annotator, which points to the gap between controlled and population-based performance.(39)A model evaluated for secondary caries staging reported sensitivity of 0.74 and an area under the curve of 0.94, and its severity staging correlated strongly with the reference standard.(30)

Commercially available software produced a wide range of results, which is relevant because these tools are the ones most likely to enter clinical practice. Diagnocat reported sensitivity of 0.51 to 0.76 and specificity of 0.88 to 0.97 across observers,(41)Second Opinion reported sensitivity of 91.0% and specificity of 87% for caries among several conditions detected on the same radiographs,(42) and Denti.AI, the only tool validated against histology, reported sensitivity of 13.7% and specificity of 95.9%.(40)An unnamed software evaluated on periapical radiographs reached an accuracy of 61% for caries.(50)These differences partly reflect the training data and operating thresholds chosen by each vendor rather than a single underlying capability of deep learning.

Two studies anchor the ends of the performance distribution and show how much the reference standard matters. At the upper end, a model trained to detect caries beneath restorations and brackets reported sensitivity of 99.4% and specificity of 99.9% against an unspecified reference.(14)At the lower end, a commercially available software validated against histology reported sensitivity of 13.7% with specificity of 95.9%.(40)The gap between these estimates is unlikely to reflect a comparable gap in underlying model capability; it reflects differences in task difficulty and, critically, in how the ground truth was defined. Taken together, the accuracy findings support three generalizations: deep learning detects caries on intraoral radiographs with high accuracy relative to expert readers in most studies, with sensitivity above 80% and specificity above 85% in most detection studies that reported both; the estimates are heterogeneous and depend on the reference standard, the unit of analysis, the dentition, and the lesion depth, so a single pooled figure is not a reliable guide to the performance of any particular model; and evidence for performance against an independent histologic criterion is limited to one study and indicates substantially lower sensitivity, so the high figures reported in the majority of studies should be read as performance against human judgement rather than against the true presence of disease.

The field also evolved over the review period. The earliest studies used comparatively small datasets and reported lower or more variable performance,(24,27,29,33,37)whereas studies published in 2024 and 2025 applied more recent architectures to larger, and sometimes population-based, datasets and reported generally higher accuracy.(17,25,39,48)The reporting of threshold-independent measures such as the area under the curve also increased in the more recent reports,(30,47,48) which allows a fairer comparison across models than sensitivity and specificity alone.

DISCUSSION

This review found that deep learning models for caries detection on intraoral radiographs achieve high but markedly heterogeneous accuracy. Sensitivity and specificity reach the levels reported for experienced examiners or above in many studies, yet the estimates weaken substantially when the reference standard is histology rather than expert annotation, and prospective external validation remains scarce.

The high estimates align with earlier reviews. A systematic review and meta-analysis of artificial intelligence platforms for caries detection reported pooled sensitivity and specificity near 88% across clinical images and radiographs,(7) and a review restricted to approximal caries on bitewings reported similarly high overall accuracy.(6)The umbrella review by Negi and colleagues reached the same conclusion, that model performance is generally high but that the certainty of the evidence is limited by risk of bias and heterogeneity.(8)Earlier syntheses likewise found sensitivity and specificity predominantly above 80%.(4,5) The present review updates those estimates with the large cohort of studies published in 2024 and 2025 and, by restricting the analysis to intraoral projections, isolates the task in which deep learning is most directly comparable to routine radiographic diagnosis. The contrast with a review of caries detection on oral photographs, which found lower and more variable performance, reinforces that the imaging modality, and not only the algorithm, drives reported accuracy.(51)

The earlier meta-analyses provide useful anchors. Carvalho and colleagues pooled studies of approximal caries on bitewings and reported high diagnostic accuracy, with sensitivity and specificity near 0.9, but noted that most studies used expert readers as the reference standard.(6)Abbott and colleagues pooled studies across clinical images and radiographs and reported sensitivity and specificity near 0.88, with substantial heterogeneity that they attributed to the diversity of inputs and tasks.(7)Mohammad-Rahimi and colleagues reported high accuracy in an earlier review of deep learning for caries detection and flagged the same absence of external validation.(4)Reyes and colleagues, reviewing machine learning for caries diagnosis and prognosis, likewise reported high performance with limited evidence of clinical translation.(5)The present review confirms those conclusions with a larger and more recent body of evidence and adds the observation that performance against histology is markedly lower than performance against expert annotation.

The present estimates are also consistent with the wider diagnostic accuracy literature for artificial intelligence in dental imaging. Reviews of deep learning for tooth identification, for periodontal bone loss, and for periapical lesion detection have each reported high pooled accuracy alongside the same recurring limitation, namely that reference standards rely on human readers and that external validation is uncommon.(3,9) This convergence across tasks suggests that the main uncertainty in dental artificial intelligence is not whether networks can learn radiographic patterns, but whether performance measured in development translates to unseen patients, devices, and clinical settings.

The heterogeneity has several sources. First, the task differed across studies, because detection, classification, and segmentation of caries are not interchangeable, and metrics such as the F1-score, mean average precision, and the Dice coefficient are not directly comparable to sensitivity and specificity. Second, the reference standard varied from a single reader to consensus panels to histology, and the one histologically validated study reported sensitivity of only 13.7%, a reminder that expert annotation may overestimate the presence of lesions that deep learning is then trained to reproduce.(40)Third, lesion severity and dentition differed, with primary teeth and early enamel lesions being harder to detect than cavitated lesions in permanent teeth, a pattern visible in the low sensitivity for dentin lesions in primary molars.(12)Fourth, sample sizes ranged from fewer than 150 to almost 14,000 images, and models trained and tested within a single institution tend to overstate performance relative to external cohorts, as the population-based evaluation suggests.(39)Fifth, some studies reported the performance of commercial software, whose fixed training data and thresholds differ from in-house models and can yield lower observed accuracy, as in the 61% accuracy reported for one tool.(50)

The mechanisms by which deep learning detects caries also bear on the observed heterogeneity. Convolutional networks learn to associate local radiographic patterns, such as radiolucencies at the proximal surface or beneath the enamel-dentin junction, with lesion labels, but they are sensitive to image acquisition, contrast, and the presence of restorations, brackets, and overlapping anatomy.(3)Most studies used transfer learning, initializing their networks from weights pretrained on natural images, and applied data augmentation to expand small training sets; these practices improve performance on small dental datasets but do not guarantee generalization to a different scanner, population, or lesion distribution, which is why internal estimates tend to exceed exter-

nal ones.(3)This dependence on image characteristics explains why models that perform well on curated single-center datasets degrade on population-based cohorts with greater variability,(39) and why caries beneath restorations or in primary teeth, where anatomy and appearance differ, yields different accuracy than cavitated lesions in permanent teeth.(12,14)

The strengths of this review are the systematic search across three bibliographic databases, the explicit 2018 to 2025 window, the focus on intraoral projections, and the recording of performance values as reported without recalculation. A further strength is the separation of bitewing from periapical studies and the description of performance by projection, which reveals differences that pooled analyses obscure, together with the distinction among detection, classification, and segmentation, which is necessary because these tasks answer different clinical questions. The review also has limitations. Full text was available for only 14 of the 72 reports assessed, so eligibility and data extraction for the remainder relied on abstracts, which may omit details needed for a complete risk-of-bias appraisal; future reviews should retrieve full text for all reports. Screening and extraction were performed by a single author with artificial intelligence assistance rather than two independent reviewers, which raises the risk of selection error; a dual-reviewer design with adjudication is recommended. A meta-analysis was not possible given the heterogeneous metrics and reference standards; a future meta-analysis restricted to studies reporting sensitivity and specificity against a common reference standard would quantify the pooled accuracy. Publication bias may also inflate the observed performance, because studies with unfavorable results are less likely to be published, and registration of diagnostic accuracy studies would mitigate this. A further limitation concerns the reliance on the reported accuracy figures themselves. The included studies used different definitions of a true positive, whether at the image, tooth, surface, or lesion level, and the choice of unit of analysis changes the resulting sensitivity and specificity. Some studies counted a detection as correct when the model identified the tooth but not the exact surface, which inflates apparent accuracy relative to a surface-level or lesion-level criterion. The present review could not harmonize these definitions, and the reported ranges therefore reflect the unit of analysis chosen by each study as much as the underlying performance of the models.

For clinical practice, the findings support the use of deep learning as a decision-support adjunct rather than a replacement for radiographic interpretation. The models are most consistent for cavitated and dentinal lesions, whereas performance for early enamel lesions and for histologically confirmed lesions remains uncertain, so clinicians should interpret negative model outputs with caution in high-risk patients. For research and regulation, the field would benefit from prospective studies with external validation, histologic or independent reference standards, and reporting standards that separate detection from classification and segmentation. The implications also differ by stakeholder. For clinicians, the results support cautious use of these tools as a second reader, particularly for dentinal and approximal lesions where the models are most consistent, while reserving human judgement for early enamel lesions and for cases with restorations or atypical anatomy. For educators, the finding that a detection model outperformed a group of dentists on recall, and that software support improved the diagnostic capacity of learners, suggests a role for these tools in training and calibration. For regulators, the scarcity of histologically validated and externally tested studies means that claims of diagnostic accuracy should be treated as device-specific rather than general, and that post-market surveillance will be needed to confirm performance in routine use. The adoption of artificial intelligence in oral radiology is advancing, yet clinicians remain cautious about interpretability and trust, which emphasizes the need for evidence generated under conditions that resemble routine practice.(9,52,53)From a translational perspective, the commercial availability of several of the evaluated tools means that the question is no longer whether deep learning can detect caries, but whether the tools cleared or marketed for clinical use perform as claimed on the populations and image types encountered in practice. The wide spread observed even among commercial tools, from an accuracy of 61% to sensitivity and specificity above 90%, indicates that the evidence base for individual products is thin, and that clinicians and purchasers cannot assume that any deep learning tool performs at the level suggested by the published literature as a whole.

For future research, the most informative studies would be prospective, register caries detection models across multiple centers and acquisition devices, report sensitivity and specificity at the lesion and the surface level against an independent reference such as histology or a validated consensus, and publish the operating threshold. Standardized reporting would permit a quantitative meta-analysis and would clarify whether the differences observed here reflect true differences in model quality or only differences in measurement. Until such studies are available, systematic reviews of this literature should interpret pooled estimates cautiously and should separate intraoral from panoramic radiographs and detection from segmentation. For readers and authors of future

reviews, two methodological lessons follow. First, using abstracts as a data source, while necessary when full text is paywalled, can bias a review toward studies whose abstracts report favorable numbers, so the proportion of abstract-only studies should be reported. Second, the unit of analysis and the reference standard should be extracted and used as stratification variables, because these choices explain more of the observed variation than the choice of architecture.

Taken together, the evidence supports a measured conclusion. Deep learning is a capable but not yet self-sufficient tool for radiographic caries detection, and its value in practice will depend less on further gains in headline accuracy than on closing the gap between development and deployment, through external validation, transparent reference standards, and studies that measure the effect of model assistance on actual clinical decisions and patient outcomes. The present review provides a current, projection-specific synthesis against which those future studies can be compared.

CONCLUSIONS

This systematic review synthesized the diagnostic accuracy of deep learning for caries detection on intraoral radiographs and found that models achieve high but highly heterogeneous sensitivity and specificity. The most frequent architectures were object detectors and segmentation networks applied to bitewing radiographs, and the reference standard was expert annotation in nearly all studies. Risk of bias was moderate to high across studies, chiefly because of unblinded expert reference standards, convenience sampling, and a scarcity of external validation. The single histologically validated study indicated that performance against an independent criterion is substantially lower than performance against expert annotation. Prospective, externally validated studies with histologic reference standards are required before these models can be recommended for unassisted clinical use.

First author, year	Country	Architecture	Imaging	Sample	Reference standard	Task
Ahmed 2025	NR	NR	bitewing	554 samples	modified KAU classification	classification
Ahmed 2025	Pakistan	NR	bitewing and periapical	240 adults	radiologist + endodontist consensus	detection
Albadr 2025	NR	CNN	bitewing	20 bitewings (test)	experienced dentists	detection
Alsolamy 2025	NR	YOLOv11	bitewing	1354 bitewings	restorative dentistry consultant	detection
Ayhan 2025	NR	YOLOv9c	bitewing	NR	NR	detection
Bayati 2025	Iran	YOLOv8	bitewing	552 bitewings (1506 images)	university annotation	detection
Chaudhari 2025	NR	ESRGAN + YOLOv8	periapical	100 images (test)	NR	detection
de Brito 2025	Brazil	YOLOv8 + EfficientNet-B0	bitewing	NR	radiology professor (gold standard)	detection and classification
Gucclu 2025	NR	YOLOv5x	bitewing	2000 bitewings (5704 lesions)	polygon annotation (Craniocatch)	segmentation
Haidar 2025	NR	AI software (unnamed)	periapical	283 periapical radiographs	two dentists >10 years	detection
Ibraheem 2025	NR	Second Opinion v1.1 (CADE)	periapical	300 periapical radiographs (1030 teeth)	two oral radiologists	detection
Malik 2025	NR	NR	bitewing	120 children (6-14 years)	calibrated pediatric dentists	detection
Mao 2025	Taiwan	YOLOv8 + AlexNet/ Inception-v3	bitewing	505 bitewings	NR	detection
Noor 2025	Pakistan	NR	intraoral	100 radiographs	dental radiologists	detection
Parathanath 2025	India	CBMNet (ConvNeXt-Tiny + attention)	periapical	1103 periapical radiographs	G.V. Black classes (experts)	classification
Patel 2025	NR	CNN (assisted software)	bitewing	400 bitewings	consensus with clinical validation	detection

First author, year	Country	Architecture	Imaging	Sample	Reference standard	Task
Rodrigues 2025	NR	Denti.AI	bitewing	171 proximal surfaces	histology	detection
Salehizeinabadi 2025	NR	YOLOv11	bitewing	730 bitewings (1115 lesions)	experienced dentists (Roboflow)	detection
Tan 2025	NR	DINO and YOLOv7	bitewing	1023 bitewings (primary molars)	two examiners	detection
van Nistelrooij 2025	Netherlands	Mask R-CNN (Swin Transformer)	bitewing	413 bitewings (2612 restored teeth)	two assessors + two experts	detection
Yavsan 2025	Turkey	CNN	periapical	415 periapical radiographs (830 labels)	pediatric dentist	detection
Ayhan 2024	Turkey	YOLOv7	bitewing	1170 bitewings	NR	detection
Chaves 2024	Netherlands	Mask R-CNN (Swin Transformer)	bitewing	425 bitewings	NR	detection
Dhanak 2024	India	EfficientDet-Lite1	bitewing	100 bitewings (80 lesions)	known lesions (ground truth)	detection
ForouzesheFar 2024	Iran	VGG19 (and AlexNet, ResNet50, VGG16)	bitewing	713 images	NR	classification
Karakus 2024	NR	YOLOv8	bitewing	860 bitewings	two oral radiologists	detection
Liu 2024	NR	ResNet + SAM	periapical	4278 periapical radiographs	medical experts	detection
Perez de Frutos 2024	Norway	RetinaNet, YOLOv5, EfficientDet	bitewing	13,887 bitewings (197 consensus)	six experts / consensus	detection
Szabo 2024	NR	CNN (Diagnocat)	bitewing and periapical	323 teeth	observer consensus	detection
Ying 2024	NR	YOLOv5, DETR, U-Net, Trans-UNet	periapical	NR	five dentists	detection
Azhari 2023	NR	U-Net ensemble (ResNet50/ ResNeXt101/VGG19)	bitewing	771 bitewings	general dentists	segmentation
Kunt 2023	NR	ensemble (YOLOv5, Faster R-CNN, RetinaNet, EfficientDet)	bitewing	3989 bitewings (7257 lesions)	NR	detection
Chen 2022	NR	Faster R-CNN	bitewing	978 bitewings (10,899 surfaces)	two endodontists + radiologist	detection
Taleb 2022	NR	self-supervised	bitewing	38,000 bitewings (pretraining); 343 test	3 professionals + senior dentist	classification
Bayraktar 2021	NR	modified YOLO	bitewing	1000 bitewings	two experienced dentists	detection
Lee 2021	South Korea	U-Net	bitewing	304 (training) / 50 (evaluation)	NR	detection
Mao 2021	NR	AlexNet (transfer learning)	bitewing	NR	NR	detection
Moran 2021	NR	Inception (vs ResNet)	bitewing	112 bitewings	experts	classification
Khan 2020	NR	U-Net + DenseNet121	periapical	206 periapical radiographs	3 specialist clinicians	segmentation

First author, year	Reported performance
Ahmed 2025	IoU 0.55; F1 0.535
Ahmed 2025	sensitivity 91.7%; specificity 89.2%; AUC 0.94
Albadr 2025	recall 0.77; precision 0.87; F1 0.82 (in vitro)
Alsolamy 2025	recall 0.864; precision 0.844; F1 0.851; mAP 0.888
Ayhan 2025	sensitivity 0.727; specificity 0.726; F1 0.687 (outperformed six dentists)
Bayati 2025	recall 79.8%; precision 84.8%; F1 82.2%
Chaudhari 2025	sensitivity 0.174; specificity 0.887 (multiclass)
de Brito 2025	accuracy 91% (YOLOv8) / 89% (EfficientNet-B0)
Guclu 2025	recall 0.78 primary / 0.75 permanent; AUC 0.87 / 0.76; F1 0.79 / 0.76
Haidar 2025	accuracy 61% (caries)
Ibraheem 2025	sensitivity 91.0%; specificity 87%
Malik 2025	sensitivity 88.3%; specificity 90.8%; accuracy 89.6%
Mao 2025	sensitivity 99.4%; specificity 99.94%; accuracy 95.8-99.2%
Noor 2025	sensitivity 94.0%; specificity 70.6%; accuracy 90.0%
Parathanath 2025	accuracy 93.3%; per-class sensitivity 0.90-0.94
Patel 2025	sensitivity 91.2 +/- 3.5%; specificity 87.1 +/- 2.9%; accuracy 89.4 +/- 3.2%
Rodrigues 2025	sensitivity 13.7%; specificity 95.9%; AUC 0.548 (histology)
Salehizeinabadi 2025	mAP50 0.52-0.80 by severity; inter-rater DSC 0.85
Tan 2025	sensitivity 0.509; specificity >0.99; accuracy 0.98
van Nistelrooij 2025	sensitivity 0.737; specificity 0.966; AUC 0.940 (secondary caries)
Yavsan 2025	sensitivity 89.3%; accuracy 90.8%; F1 90.2% (children 5-12 years)
Ayhan 2024	sensitivity 0.833; specificity 0.961; accuracy 0.934
Chaves 2024	AUC 0.806 primary / 0.804 secondary; F1 0.689 / 0.719
Dhanak 2024	sensitivity 0.75; precision 0.846; F1 0.795
ForouzeshFar 2024	accuracy 93.9% (VGG19)
Karakus 2024	sensitivity 0.932; precision 0.977; F1 0.954
Liu 2024	AUC 0.954 (95% CI 0.924-0.980); accuracy 0.885
Perez de Frutos 2024	YOLOv5 mAP 0.647; F1 0.548; false-negative rate 0.149
Szabo 2024	sensitivity 0.51-0.76; specificity 0.88-0.97; accuracy 0.76-0.86
Ying 2024	F1 0.87 (YOLOv5) / 0.86 (Trans-UNet) / 0.82 (DETR) / 0.80 (U-Net)
Azhari 2023	per-class IoU 0.08-0.98; advanced caries most precise
Kunt 2023	F1 0.79-0.80; precision 0.83; recall 0.77; AP 0.86
Chen 2022	sensitivity 0.72; specificity 0.93; accuracy 0.87
Taleb 2022	sensitivity >=45%; +6 p.p. with pretraining
Bayraktar 2021	sensitivity 72.3%; specificity 98.2%; accuracy 94.6%
Lee 2021	precision 63.3%; recall 65.0%; F1 64.1%
Mao 2021	accuracy 90.3% (caries)
Moran 2021	accuracy 73.3%
Khan 2020	Dice 0.569; mIoU 0.501 (validation)

REFERENCES

1. GBD 2017 Oral Disorders Collaborators, Bernabe E, Marcenes W, Hernandez CR, Bailey J, Abreu LG, et al. Global, Regional, and National Levels and Trends in Burden of Oral Conditions from 1990 to 2017: A Systematic Analysis for the Global Burden of Disease 2017 Study. *J Dent Res*. 2020;99(4):362-373. doi: 10.1177/0022034520908533
2. Ismail AI, Sohn W, Tellez M, Amaya A, Sen A, Hasson H, et al. The International Caries Detection and Assessment System (ICDAS): an integrated system for measuring dental caries. *Comm Dent Oral Epid*. 2007;35(3):170-178. doi: 10.1111/j.1600-0528.2007.00347.x
3. Khanagar SB, Al-ehaideb A, Maganur PC, Vishwanathaiah S, Patil S, Baeshen HA, et al. Developments, application, and performance of artificial intelligence in dentistry – A systematic review. *Journal of Dental Sciences*. 2021;16(1):508-522. doi: 10.1016/j.jds.2020.06.019
4. Mohammad-Rahimi H, Motamedian SR, Rohban MH, Krois J, Uribe SE, Mahmoudinia E, et al. Deep learning for caries detection: A systematic review. *Journal of Dentistry*. 2022;122:104115. doi: 10.1016/j.jdent.2022.104115
5. Reyes LT, Knorst JK, Ortiz FR, Ardenghi TM. Machine Learning in the Diagnosis and Prognostic Prediction of Dental Caries: A Systematic Review. *Caries Res*. 2022;56(3):161-170. doi: 10.1159/000524167
6. Carvalho BKG, Nolden EL, 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. *Journal of Dentistry*. 2024;151:105388. doi: 10.1016/j.jdent.2024.105388
7. ABBOTT LP, SAIKIA A, ANTHONAPPA RP. ARTIFICIAL INTELLIGENCE PLATFORMS IN DENTAL CARIES DETECTION: A SYSTEMATIC REVIEW AND META-ANALYSIS. *Journal of Evidence-Based Dental Practice*. 2025;25(1):102077. doi: 10.1016/j.jebdp.2024.102077
8. Negi S, Mathur A, Tripathy S, Mehta V, Snigdha NT, Adil AH, et al. Artificial Intelligence in Dental Caries Diagnosis and Detection: An Umbrella Review. *Clinical & Exp Dental Res*. 2024;10(4):e70004. doi: 10.1002/cre2.70004
9. Demir K, Sokmen O, Karabey Aksakalli I, Torenek-Agirman K. Comprehensive Insights into Artificial Intelligence for Dental Lesion Detection: A Systematic Review. *Diagnostics*. 2024;14(23):2768. doi: 10.3390/diagnostics14232768
10. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;n71. doi: 10.1136/bmj.n71
11. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(8):529-536. doi: 10.7326/0003-4819-155-8-201110180-00009
12. Tan ZQ, Pinho IF, Banks R, Chen X, Pontes LRA, Braga MM, et al. Caries Detection in Primary Molars with Bitewing Radiographs through Deep Learning Based-Object Detectors. *Caries Res*. 2025;1-9. doi: 10.1159/000550079
13. Yavsan ZS, Orhan H, Efe E, Yavsan E. Diagnosis of approximal caries in children with convolutional neural networks based detection algorithms on radiographs: A pilot study. *AOS*. 2025;84:18-25. doi: 10.2340/aos.v84.42599
14. Mao YC, Lin YJ, Hu JP, Liu ZY, Chen SL, Chen CA, et al. Automated Caries Detection Under Dental Restorations and Braces Using Deep Learning. *Bioengineering*. 2025;12(5):533. doi: 10.3390/bioengineering12050533
15. Güçlü HS, Demirel A, Orhan K. Evaluation of a YOLOv5x-based deep learning model for interproximal caries segmentation on bitewing radiographs across primary and permanent teeth. *BMC Oral Health*. 2025;26(1):202. doi: 10.1186/s12903-025-07570-2
16. 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
17. Alsolamy M, Nadeem F, Azhari AA, Ahmed WM. Automated Detection, Localization, and Severity Assessment of Proximal Dental Caries from Bitewing Radiographs Using Deep Learning. *Diagnostics*. 2025;15(7):899. doi: 10.3390/diagnostics15070899
18. 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):1949-1961. doi: 10.1111/jerd.13470
19. Ayhan B, Ayan E, Bayraktar Y. A novel deep learning-based perspective for tooth numbering and caries detection. *Clin Oral Invest*. 2024;28(3):178. doi: 10.1007/s00784-024-05566-w
20. Chaudhari AY, Birwadkar P, Joshi S, Verma Y, Sindgi R. Classification of periapical dental X-ray using the YOLOv8 deep learning model. *MethodsX*. 2025;15:103721. doi: 10.1016/j.mex.2025.103721
21. Salehizeinabadi M, Neghab S, Ameli N, Baghi KK, Pacheco-Pereira C. Automated Classification of Dental Caries in Bitewing Radiographs Using Machine Learning and the ICCMS Framework. *International Journal of Dentistry*. 2025;2025(1):6644310. doi: 10.1155/ijod/6644310
22. de Brito D, dos Santos T, Costa S, dos Santos A, da Silva I, da Silva N, et al. Accuracy in diagnosing caries in young permanent molars using interproximal radiographic imaging and validation by artificial intelligence. *J Clin Exp Dent*. 2025;17(6):e648-e655. doi: 10.4317/jced.62396
23. Department of Conservative Dentistry and Endodontics, Government Dental College and Hospital, Ahmabad, Gujarat, India, Dhanak N, Chougule VT, Department of Paediatric and Preventive Dentistry, Bharati Vidyapeeth (Deemed to be University) Dental College and Hospital, Sangli, Maharashtra, India, Nalluri K, Apex, North Carolina, USA, et al. Artificial intelligence enabled smart phone app for real-time caries detection on bitewing radiographs. *Bioinformation*. 2024;20(3):243-247. doi: 10.6026/973206300200243
24. Bayraktar Y, Ayan E. Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs. *Clin Oral Invest*. 2022;26(1):623-632. doi: 10.1007/s00784-021-04040-1

25. Kunt L, Kybic J, Nagyová V, Tichý A. Automatic caries detection in bitewing radiographs: part I—deep learning. *Clin Oral Invest*. 2023;27(12):7463-7471. doi: 10.1007/s00784-023-05335-1
26. Karakuş R, Öziç MÜ, Tassoker M. AI-Assisted Detection of Interproximal, Occlusal, and Secondary Caries on Bite-Wing Radiographs: A Single-Shot Deep Learning Approach. *J Digit Imaging Inform med*. 2024;37(6):3146-3159. doi: 10.1007/s10278-024-01113-x
27. Khan HA, Haider MA, Ansari HA, Ishaq H, Kiyani A, Sohail K, et al. Automated feature detection in dental periapical radiographs by using deep learning. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*. 2021;131(6):711-720. doi: 10.1016/j.oooo.2020.08.024
28. Azhari AA, Helal N, Sabri LM, Abduljawad A. Artificial intelligence (AI) in restorative dentistry: Performance of AI models designed for detection of interproximal carious lesions on primary and permanent dentition. *DIGITAL HEALTH*. 2023;9:20552076231216681. doi: 10.1177/20552076231216681
29. Lee S, Oh SI, Jo J, Kang S, Shin Y, Park JW. Deep learning for early dental caries detection in bitewing radiographs. *Sci Rep*. 2021;11(1):16807. doi: 10.1038/s41598-021-96368-7
30. van Nistelrooij N, Chaves ET, Cenci MS, Cao L, Loomans BAC, Xi T, et al. Deep Learning-Based Algorithm for Staging Secondary Caries in Bitewings. *Caries Res*. 2024;59(3):163-173. doi: 10.1159/000542289
31. Chaves ET, Vinayahalingam S, van Nistelrooij N, Xi T, Romero VHD, Flügge T, et al. Detection of caries around restorations on bitewings using deep learning. *Journal of Dentistry*. 2024;143:104886. doi: 10.1016/j.jdent.2024.104886
32. ForouzeshtFar P, Safaei AA, Ghaderi F, Hashemikamangar SS. Dental Caries diagnosis from bitewing images using convolutional neural networks. *BMC Oral Health*. 2024;24(1):211. doi: 10.1186/s12903-024-03973-9
33. Moran M, Faria M, Giraldo G, Bastos L, Oliveira L, Conci A. Classification of Approximal Caries in Bitewing Radiographs Using Convolutional Neural Networks. *Sensors*. 2021;21(15):5192. doi: 10.3390/s21155192
34. Parathanath AJ, A M. CBMNet: a dual-attention enhanced ConvNeXt model for accurate G.V. Black type I–III classification in intraoral periapical radiographs. *Sci Rep*. 2025;15(1):39287. doi: 10.1038/s41598-025-22975-3
35. Taleb A, Rohrer C, Bergner B, De Leon G, Rodrigues JA, Schwendicke F, et al. Self-Supervised Learning Methods for Label-Efficient Dental Caries Classification. *Diagnostics*. 2022;12(5):1237. doi: 10.3390/diagnostics12051237
36. Ahmed WM, Azhari AA, Fawaz KA, Ahmed HM, Alsadah ZM, Majumdar A, et al. Artificial intelligence in the detection and classification of dental caries. *The Journal of Prosthetic Dentistry*. 2025;133(5):1326-1332. doi: 10.1016/j.prosdent.2023.07.013
37. Mao YC, Chen TY, Chou HS, Lin SY, Liu SY, Chen YA, et al. Caries and Restoration Detection Using Bitewing Film Based on Transfer Learning with CNNs. *Sensors*. 2021;21(13):4613. doi: 10.3390/s21134613
38. Ying S, Huang F, Shen X, Liu W, He F. Performance comparison of multifarious deep networks on caries detection with tooth X-ray images. *Journal of Dentistry*. 2024;144:104970. doi: 10.1016/j.jdent.2024.104970
39. Pérez de Frutos J, Holden Helland R, Desai S, Nymoen LC, Langø T, Remman T, et al. AI-Dentify: deep learning for proximal caries detection on bitewing x-ray - HUNT4 Oral Health Study. *BMC Oral Health*. 2024;24(1):344. doi: 10.1186/s12903-024-04120-0
40. Rodrigues N, Martinez-Rus F, Miguel-Calvo A, Pradies G, Salido MP. Accuracy Assessment of Human and Artificial Intelligence-Assisted Bitewing Radiography and Near-Infrared Reflectance Imaging-Based Methods for Interproximal Caries Detection: A Histological Validation. *Caries Res*. 2025;60(2):173-184. doi: 10.1159/000546644
41. Szabó V, Szabó BT, Orhan K, Veres DS, Manulis D, Ezhov M, et al. Validation of artificial intelligence application for dental caries diagnosis on intraoral bitewing and periapical radiographs. *Journal of Dentistry*. 2024;147:105105. doi: 10.1016/j.jdent.2024.105105
42. Ibraheem WI, Jain S, Ayoub MN, Namazi MA, Alfaqih AI, Aggarwal A, et al. Assessment of the Diagnostic Accuracy of Artificial Intelligence Software in Identifying Common Periodontal and Restorative Dental Conditions (Marginal Bone Loss, Periapical Lesion, Crown, Restoration, Dental Caries) in Intraoral Periapical Radiographs. *Diagnostics*. 2025;15(11):1432. doi: 10.3390/diagnostics15111432
43. Noor N, Mirza A, Faraid V, Ullah I. Role of Artificial Intelligence in Early Detection of Dental Caries using Radiographic Images. *JBKCD*. 2025;3(1):35-41. doi: 10.65293/jbkcd.v3i01.46
44. Chen X, Guo J, Ye J, Zhang M, Liang Y. Detection of Proximal Caries Lesions on Bitewing Radiographs Using Deep Learning Method. *Caries Res*. 2022;56(5-6):455-463. doi: 10.1159/000527418
45. Patel HB, Marwaha J, Hirani N, Suthar P, Arya A, Pandey N. Diagnostic Accuracy and Efficiency of AI-Assisted Radiographic Interpretation Compared to Conventional Methods in Early Detection of Dental Caries. *J Pharm Bioallied Sci*. 2025;17(Suppl 4):S3192-S3194. doi: 10.4103/jpbs.jpbs_1278_25
46. Malik AI, Adeel-ur-Rehman, Saad Umar, Muhammad Zubair, Zainab Sajjad, Ali Ghaffar, et al. DIAGNOSTIC ACCURACY OF AI-BASED VERSUS CONVENTIONAL RADIOGRAPHIC CARIES DETECTION IN PEDIATRIC PATIENTS: A CROSS-SECTIONAL STUDY. *IJHR*. 2025;3(3 (Health & Rehabilitation)):29-34. doi: 10.71000/fdmjfp07
47. Saher Ahmed, Rabia Masood, Ifrah Rashid, Amna Jamil Haq, Rehmeen Asif, Syeda Tooba Sajjad, et al. Diagnostic Accuracy of AI in Radiographic Detection of Dental Caries and Periapical Lesions. *LMJHCR*. 2025:e55. doi: 10.61919/4w0nfp45
48. Liu Y, Xia K, Cen Y, Ying S, Zhao Z. Artificial intelligence for caries detection: a novel diagnostic tool using deep learning algorithms. *Oral Radiol*. 2024;40(3):375-384. doi: 10.1007/s11282-024-00741-x
49. Albadr AK. Development and Comparative Analysis of an AI Based Convolutional Neural Network Algorithm versus Dentists of Varying Experience in Detecting Dental Caries using Bitewing X-ray Images: An In vitro Study. *J Pharm Bioallied Sci*. 2025;17(Suppl 4):S3364-S3366. doi: 10.4103/jpbs.jpbs_1423_25

50. Haidar A, Alamgir W, Haider I, Siddiqui SN, Anwar MA, Khan B, et al. Evaluating the Diagnostic Accuracy of Artificial Intelligence in Periapical Radiographs. *J Rehman Coll Dent*. 2025;6(03):98-104. doi: 10.52442/jrcd.v6i03.151
51. Moharrami M, Farmer J, Singhal S, Watson E, Glogauer M, Johnson AEW, et al. Detecting dental caries on oral photographs using artificial intelligence: A systematic review. *Oral Diseases*. 2024;30(4):1765-1783. doi: 10.1111/odi.14659
52. Saif N, Elkoumi A, Shaalan AK. Use of artificial intelligence in oral radiology: a multicenter cross-sectional study in Egypt. *BMC Oral Health*. 2025;25(1):1888. doi: 10.1186/s12903-025-07269-4
53. Haider M. Developing an AI-Powered Algorithm for Automated Detection and Classification of Dental Caries from Intraoral Radiographs: A Machine Learning Approach. *J Pharm Bioallied Sci*. 2024;16(Suppl 4):S3089-S3091. doi: 10.4103/jpbs.jpbs_1097_24
54. Material Complementario dejar para maquetar luego manual en word
55. Table 1. Characteristics of the 39 included studies evaluating deep learning for caries detection on intraoral radiographs.
56. Note: NR = not reported; CNN = convolutional neural network; YOLO = you only look once; DSC = Dice similarity coefficient.
57. Source: Compiled from data extracted from the included studies.
58. Table 2. Reported diagnostic performance of deep learning models for caries detection on intraoral radiographs.
59. Note: AUC = area under the receiver operating characteristic curve; F1 = F1-score; mAP = mean average precision; IoU = intersection over union; CI = confidence interval.
60. Source: Compiled from data extracted from the included studies.

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Greta Cordovi Vallongo, Marta María Capote Fernández.

Writing–review and editing: Greta Cordovi Vallongo, Marta María Capote Fernández.