Cureus
Part of SPRINGER NATURE

Open Access Review Article

Role of Artificial Intelligence in Improving Treatment Planning Success in Pediatric Dentistry: A Systematic Review

Review began 11/13/2026
Review ended 11/20/2026
Published 11/25/2026

© Copyright 2026
Alzahrani et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.120259

Lama Zayed A. Alzahrani [1], Ahmed Fahad Alzahrani [2], Walaa Ali Altarouti [3], Fatema Hussain Eisa AlSoltan [4], Waad M. Alahmadi [5], Jenan Ehsan Ali Alali [6], Hayfaa Al-Shayea [7], Areej O. Yousif Abduldaiem [8], Laila Alanaz [9]

1. General Dentistry, King Abdul Aziz University Hospital, Jeddah, SAU 2. General Dentistry, Mass Almaali Medical Center, Taif, SAU 3. General Dentistry, Ibn Sina National College, Jeddah, SAU 4. General Dentistry, Riyadh Elm University, Riyadh, SAU 5. General Dentistry, Taibah University, Medina, SAU 6. General Dentistry, Vision College, Jeddah, SAU 7. Pediatric Dentistry, Prince Sultan Military Medical City, Riyadh, SAU 8. Dentistry, Prince Abdulrahman Advanced Dental Institute, Riyadh, SAU 9. Dentistry, Private Clinic, Riyadh, SAU

Corresponding author: Laila Alanaz, llm20010-17-@hotmail.com

Abstract

Artificial intelligence (AI) is increasingly being explored as a decision support tool in dentistry. Its application in pediatric dentistry may improve diagnostic accuracy and support treatment planning; however, evidence regarding its clinical performance remains limited. This systematic review aimed to evaluate the applications, performance, and methodological quality of AI-based tools used to support treatment planning in pediatric dentistry.

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. We searched PubMed, ScienceDirect, the Cochrane Library, and Google Scholar, supplemented by Google searches and manual screening of reference lists. Studies evaluating AI applications relevant to pediatric dental diagnosis, treatment planning, or clinical decision-making were eligible. A total of 4,235 records were identified. After duplicate removal, screening, and full-text assessment, 16 studies were included in the qualitative synthesis. Risk of bias was assessed using Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2), the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Analytical Cross-Sectional Studies, and Prediction Model Risk of Bias Assessment Tool for AI (PROBAST+AI), as appropriate to study design. Owing to substantial heterogeneity in study designs, AI applications, and outcome measures, findings were synthesized narratively.

The 16 included studies investigated diverse AI applications, including dental caries detection, tooth identification and segmentation, detection of developmental and structural abnormalities, orthodontic assessment, and AI-assisted clinical decision support. Reported performance varied considerably across applications and evaluation methods. For example, a lightweight deep-learning model achieved a balanced accuracy of 0.929 for pediatric caries screening, while another study reported 91% accuracy and 98% precision for caries detection in young permanent molars. Automated tooth detection and classification models also demonstrated promising performance, including an F1-score of 0.978 for second-molar detection and an overall mean average precision of 0.983 in one study. Nevertheless, performance was less consistent for certain diagnostic tasks, with some models demonstrating limited sensitivity or modest detection accuracy. Risk-of-bias assessment using QUADAS-2 identified five studies with a high overall risk of bias and nine with an unclear overall risk of bias. The separate assessments using JBI and PROBAST+AI also identified substantial methodological concerns, particularly regarding sample representativeness, reference-standard validity, and external validation.

AI demonstrates considerable potential to support diagnostic assessment and clinical decision-making in pediatric dentistry. However, the available evidence primarily establishes technical performance in specific diagnostic and image-analysis tasks rather than improvements in treatment success or patient outcomes. Methodological limitations, heterogeneous validation approaches, and insufficient external clinical evaluation currently restrict its routine clinical applicability. Larger, multicenter prospective studies using uniform reference standards and clinically meaningful outcomes are required to establish the effectiveness of AI-assisted pediatric dental treatment planning.

Categories: Dentistry
Keywords: artificial intelligence, clinical decision-making, clinical decision support, deep learning, diagnostic accuracy, machine learning, pediatric dentistry, treatment planning

Introduction And Background

Artificial intelligence (AI) is rapidly reshaping clinical decision-making across healthcare, and pediatric dentistry is no exception. From deep learning algorithms that detect caries on radiographs to large language models (LLMs) that generate treatment recommendations, AI tools are increasingly positioned as adjuncts to clinician judgment in domains once considered exclusively human. The evidence base for AI in pediatric dentistry has expanded considerably in recent years, though it remains heterogeneous in quality and scope. A systematic review and meta-analysis by Shahbazi et al. [1] examining AI for radiographic dental age

estimation, which is a critical input for orthodontic and pediatric treatment planning, reported a pooled accuracy of 0.85 (95% CI 0.79-0.91) across seven studies, supporting the effectiveness of AI models in individuals under 25 years. At the diagnostic level, deep learning models have demonstrated strong performance for caries detection in children. A study using compact convolutional neural networks on pediatric oral photographs reported ResNet-18 achieving a balanced accuracy of 0.929 and EfficientNet-B0 achieving an area under the receiver operating characteristic curve (ROC-AUC) of 0.978 for early childhood caries screening [2]. Another 2025 study validating YOLOv8 on interproximal radiographs of children aged six to nine years reported 91% accuracy and 98% precision for caries detection in young permanent molars [3]. A school-based application of YOLOv8x for caries screening in children found diagnostic performance comparable to that of dentists while reducing examination time to approximately one-quarter of the original duration [4].

Beyond diagnosis, AI has been applied to tasks that directly inform treatment planning. Deep learning has been used for automatic tooth detection and numbering in cone-beam computed tomography (CBCT) images of patients with mixed dentition, a task the authors described as "the initial step in treatment planning" [5]. For developmental anomalies such as dens evaginatus, a ResNet-Transformer model trained on panoramic radiographs from patients aged 3-16 years achieved an F1-score of 0.887, outperforming dentists' average F1-score of 0.878 and improving diagnostic accuracy when used as a support tool [6]. In orthodontics, machine learning models have been developed to predict treatment pathways for Class II malocclusion (camouflage, growth modulation, or surgery), with support vector machine and random forest models achieving area under the precision-recall curve (PR-AUC) values of 1.00 for growth modulation and 0.92 for camouflage [7]. A separate multimodal deep learning framework for midpalatal suture assessment, a decision point between conventional expansion and surgical intervention, achieved 93.75% accuracy and was explicitly framed as a clinical decision support tool for less experienced clinicians [8].

The arrival of LLMs has further expanded the plausible scope of AI in pediatric dental decision-making. A 2025 pilot study comparing ChatGPT-generated diagnoses and treatment plans with those of four pediatric dentists found 80% accuracy for the AI, with no statistically significant difference from clinician performance ($p = 0.926$) and Cohen's kappa values ranging from 0.615 to 1.00 [9]. Comparative evaluations of different LLMs for pediatric dentistry queries have also emerged, with a 2025 study reporting that ScholarGPT provided more accurate and complete responses than ChatGPT-4o across clinical topics, and a 2026 comparative study assessing the feasibility of LLMs for pediatric dentistry queries [10]. Chatbot performance in pediatric fluoride education has likewise been evaluated, with ChatGPT-4o demonstrating superior reliability and informational quality compared with Gemini Pro and DeepSeek V3 [11]. These studies suggest that AI tools are moving beyond image analysis into the linguistic and reasoning domains that underpin treatment planning conversations and referral decisions.

Despite the growing interest in AI applications in pediatric dentistry, important uncertainties remain regarding their effectiveness as clinical decision support tools. Existing reviews have predominantly focused on diagnostic performance, including caries detection, dental anomaly identification, and disease-risk prediction, whereas evidence directly evaluating treatment recommendations and clinical decision-making remains comparatively limited [12,13]. Furthermore, recent literature on generative AI in pediatric dentistry has highlighted variable performance and the possibility of inaccurate or misleading recommendations, supporting the need for appropriate clinical oversight [14]. Other important concerns include limited external validation, differences in study populations and reference standards, small datasets, and uncertainty regarding the generalizability of AI-generated recommendations to routine clinical practice.

Therefore, a focused evaluation of AI applications in pediatric dental treatment planning is warranted to determine how these technologies are currently being used, assess their reported performance, and identify the limitations affecting their clinical applicability. The present systematic review aims to evaluate the role of AI in supporting treatment planning and clinical decision-making in pediatric dentistry, with particular emphasis on the applications, reported performance, and methodological quality of the available evidence.

Review

Methodology

Study Design and Reporting Guidelines

This systematic review evaluated the role of AI in improving treatment planning accuracy, clinical decision-making, and treatment outcomes in pediatric dentistry. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines to ensure transparent and comprehensive reporting of the review process. The methodology was developed to identify, evaluate, and synthesize available evidence regarding the application of AI technologies in pediatric dental treatment planning, including their diagnostic capabilities, predictive performance, treatment recommendations, and potential contributions to clinical success.

Research Question and Eligibility Criteria

The research question was formulated using the population, intervention, comparison, and outcome (PICO) framework. The population comprised pediatric patients, generally defined as individuals aged 18 years or younger, requiring dental assessment, diagnosis, or treatment planning. The intervention included AI-based

technologies, such as machine learning, deep learning, artificial neural networks, convolutional neural networks, and other computational intelligence systems, utilized to support or improve treatment planning in pediatric dentistry. The comparison included conventional diagnostic approaches, treatment planning performed by dental professionals, and alternative AI-based methods, where applicable. The primary outcomes included treatment planning accuracy, diagnostic precision, predictive performance, appropriateness of treatment recommendations, and clinical treatment success. Secondary outcomes included reductions in diagnostic errors, improvements in clinical decision-making, treatment efficiency, and the potential for personalized treatment planning.

Studies were considered eligible if they investigated the application of AI in pediatric dental diagnosis or treatment planning and reported outcomes relevant to clinical decision-making or treatment planning performance. Studies involving pediatric dental conditions such as dental caries, developmental dental abnormalities, malocclusion, orthodontic conditions, dental trauma, and other relevant oral health conditions were considered. Both clinical and experimental studies were eligible if their findings were directly relevant to pediatric dental treatment planning.

Studies were excluded if they failed to investigate AI applications, focused exclusively on adult dental populations without pediatric-specific results, or evaluated AI applications unrelated to diagnosis or treatment planning. Editorials, letters to the editor, narrative reviews, systematic reviews, conference abstracts without sufficient methodological information, case reports, and studies lacking relevant outcome data were also excluded. Studies published in languages other than English were excluded.

Information Sources and Search Strategy

A comprehensive literature search was performed across the electronic databases PubMed/MEDLINE, ScienceDirect, and the Cochrane Central Register of Controlled Trials (CENTRAL) to identify relevant studies investigating AI applications in pediatric dentistry. Google Scholar was additionally searched to identify potentially relevant publications that might not have been indexed in the primary databases. The reference lists of eligible studies and relevant review articles were also examined manually to identify additional studies.

The search strategy incorporated Medical Subject Headings (MeSH) and relevant free-text keywords related to AI, machine learning, deep learning, pediatric dentistry, and treatment planning. Boolean operators, including AND and OR, were used to combine search terms and improve search sensitivity. The search terms included "artificial intelligence", "machine learning", "deep learning", "neural networks", "computer-assisted diagnosis", "pediatric dentistry", "paediatric dentistry", "child dental health", "dental treatment planning", "clinical decision-making", "treatment outcomes", and "predictive modeling". A detailed search strategy for each database is provided in the Appendices. No restrictions were initially imposed on publication year to ensure comprehensive identification of the available literature. The final search was conducted on October 30, 2026.

Study Selection and Screening Process

All records identified through the database searches were imported into reference management software, and duplicate publications were removed before screening. The study selection process was conducted in two stages. Two reviewers independently screened titles and abstracts against the predefined eligibility criteria to identify potentially relevant publications. Subsequently, the full texts of potentially eligible studies were retrieved and independently assessed by both reviewers to determine their suitability for inclusion in the systematic review. Disagreements were resolved through discussion and consensus, with consultation of a third reviewer where necessary. The same independent-review procedure was followed for methodological quality and risk-of-bias assessment. Reasons for excluding studies after full-text assessment were documented. The overall study identification, screening, eligibility assessment, and inclusion process was presented using a PRISMA 2020 flow diagram (Figure *1*).

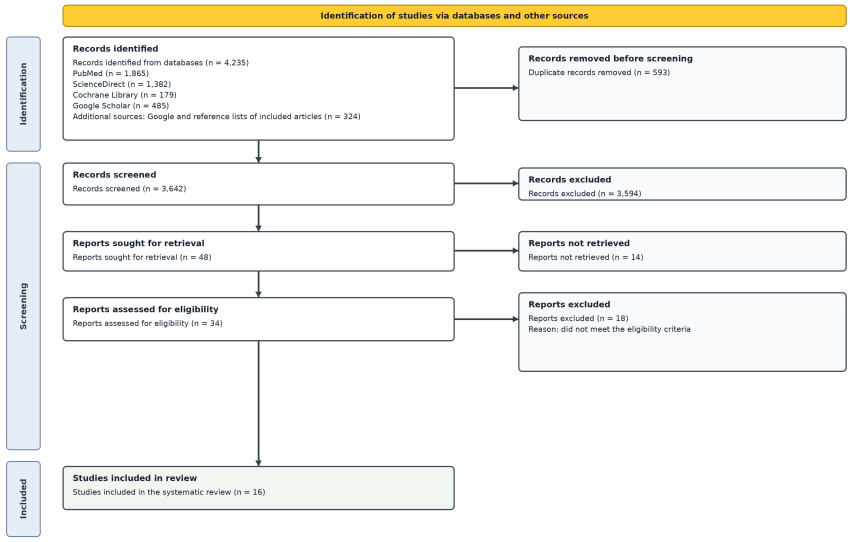


FIGURE 1: PRISMA flowchart showing the study selection process

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Data Extraction

Data extraction was performed using a standardized data extraction form developed specifically for this review. Relevant information was collected from each included study, including the first author's name, year of publication, country of study, study design, sample size, and pediatric population characteristics. Details regarding the AI technology, algorithm, computational model, and specific dental application were also extracted.

Additional information included the dental condition investigated, type of diagnostic or treatment planning task, reference standard or comparator, and the outcomes evaluated. Particular attention was given to measures of AI performance, including accuracy, sensitivity, specificity, precision, recall, AUC, and other reported performance indicators. Where available, information regarding the influence of AI-assisted planning on treatment recommendations, clinical decision-making, treatment success, and patient-related outcomes was also recorded.

Outcomes of Interest

The primary outcome of this systematic review was the effectiveness of AI-based technologies in improving treatment planning performance and clinical decision-making in pediatric dentistry. This included the accuracy and reliability of AI-generated treatment recommendations, predictive capabilities, and agreement between AI-assisted assessments and decisions made by experienced dental professionals. Secondary outcomes included improvements in diagnostic accuracy, early identification of dental conditions, reduced treatment planning errors, optimized treatment approaches, and potential improvements in clinical efficiency. Studies reporting direct clinical treatment outcomes were also evaluated to determine whether AI-assisted planning translated into measurable improvements in treatment success. The analysis maintained a clear distinction between algorithmic diagnostic performance and actual clinical treatment success.

Quality Assessment and Risk of Bias

The methodological quality and risk of bias of the included studies were assessed using appraisal instruments selected according to their study designs and primary objectives. The Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool was used for 14 studies evaluating diagnostic or image-classification performance. QUADAS-2 evaluates four domains: patient selection, index test, reference standard, and flow and timing. Each domain was assessed for potential risk of bias, while concerns regarding applicability were considered in relation to patient selection, the index test, and the reference standard. Domain-level judgments were classified as low, high, or unclear risk of bias, based on the information reported in each study.

The Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Analytical Cross-Sectional Studies was applied to the study by Naik et al. [9], which evaluated AI-generated pediatric dental diagnoses and treatment plans in a cross-sectional pilot study. The checklist assessed the clarity of inclusion criteria, the description of the study population and setting, the validity and reliability of exposure and outcome measurements, the identification and management of potential confounding factors, and the

appropriateness of the statistical analysis. Each item was evaluated according to the JBI response categories, including yes, no, unclear, and not applicable, where appropriate.

The Prediction Model Risk of Bias Assessment Tool for AI (PROBAST+AI) was applied to the prognostic component of Yelkenci et al. [15], which investigated the prediction of spontaneous space closure following first permanent molar extraction. The assessment considered four domains: participants, predictors, outcome, and analysis. Particular attention was given to participant selection and representativeness, the validity and measurement of predictors, the definition and assessment of the predicted outcome, and the appropriateness of the model-development and validation procedures.

Risk-of-bias assessments were conducted independently by two reviewers. Disagreements were resolved through discussion and consensus, with consultation of a third reviewer where necessary. Judgments were based on the methodological information reported in the full-text publications and any available supplementary material. When insufficient information was available to establish whether a methodological criterion had been adequately addressed, the relevant domain or item was classified as unclear rather than being classified as having a low risk of bias. Because the included studies differed in design and the appraisal instruments assessed different methodological domains, the results were presented separately for each instrument. No numerical scores were calculated or pooled across the different tools. Overall judgments were interpreted in accordance with the relevant instrument's guidance, and identified methodological limitations were considered when interpreting reported AI performance and its potential clinical applicability.

Data Synthesis and Analysis

A structured narrative synthesis was undertaken to summarize the findings of the included studies. The studies were grouped according to their principal clinical applications and AI tasks, including caries detection and classification, tooth detection and segmentation, orthodontic treatment planning, prediction of dental development, assessment of dental trauma, and other pediatric dental applications. Within these groups, study characteristics and findings were compared descriptively, considering the study population, sample size, AI technology, input modality, clinical task, reference standard, validation approach, and reported performance outcomes.

AI performance was summarized using the measures reported in the original publications, including accuracy, sensitivity, specificity, AUC, precision, recall, F1-score, mean average precision (mAP), Dice similarity coefficient, and intersection over union (IoU), as applicable. These measures were interpreted according to the specific task they evaluated and were not treated as interchangeable indicators of performance. Diagnostic classification measures, including sensitivity and specificity, were considered separately from technical image-analysis measures, such as object detection and segmentation performance. Similarly, anatomical detection or image-classification performance was distinguished from prognostic accuracy and the clinical effectiveness of AI-generated treatment recommendations. Where studies compared AI outputs with assessments by dental professionals or conventional diagnostic approaches, the findings were considered descriptively. Particular attention was given to whether reported improvements in technical performance translated into more appropriate clinical decisions or demonstrable patient-related outcomes.

Quantitative synthesis, including meta-analysis, was considered potentially appropriate when studies demonstrated sufficient clinical and methodological comparability in their populations, AI applications, tasks, reference standards, outcome definitions, and performance measures. However, these conditions were not sufficiently met across the included studies. The investigations evaluated fundamentally different tasks, including image classification, object detection, semantic segmentation, prognostic modeling, and AI-generated clinical recommendations. Further differences in imaging modalities, dataset composition, class distribution, validation procedures, units of analysis, and reference-standard definitions limited comparability across findings. In particular, measures such as accuracy, AUC, F1-score, Dice similarity coefficient, IoU, and mAP quantify different dimensions of model performance and cannot be meaningfully combined into a single pooled estimate. Therefore, meta-analysis and meta-regression were not performed.

Heterogeneity was examined descriptively by comparing study designs, participant and dataset characteristics, AI architectures, clinical applications, imaging modalities, reference standards, validation strategies, and outcome measures. For diagnostic studies, sensitivity, specificity, and false-negative findings were considered where reported, particularly for caries detection and other clinically consequential diagnostic tasks. The reported class distribution and potential class imbalance were also considered when available. Where these parameters were not reported, their absence was considered when interpreting the reliability and clinical applicability of the findings. The methodological quality and risk-of-bias assessments were also considered when interpreting the reported model performance and its potential generalizability to routine pediatric dental practice.

Reported point estimates and corresponding 95% confidence intervals were extracted from the original publications where available. Studies that did not report confidence intervals were identified accordingly. Confidence intervals were not independently calculated or imputed when they were not reported or when the information required for their valid calculation was unavailable. No additional statistical analyses, recalculations of model-performance estimates, or inferential statistical comparisons were performed by the review team; numerical performance estimates were extracted from the original publications and

synthesized descriptively.

Formal statistical assessment of reporting bias was not performed because no quantitative synthesis was undertaken, and the substantial heterogeneity of the included studies did not support meaningful funnel plot analysis or related statistical tests. Nevertheless, the potential for selective outcome reporting and publication bias was considered when interpreting the findings. A formal assessment of evidence certainty was also not undertaken, given the diversity of study designs, clinical applications, outcome measures, and validation approaches. These limitations were considered when drawing conclusions about the reliability, generalizability, and clinical applicability of the available evidence.

Results

Study Selection

The literature search identified 4,235 records from electronic databases and additional sources, including Google searches and manual reference list screening. Following the removal of 593 duplicate records, 3,642 records remained for title and abstract screening. Of these, 3,594 records were excluded because they did not meet the predefined eligibility criteria. Forty-eight reports were subsequently sought for retrieval, of which 14 could not be obtained. The remaining 34 full-text reports were assessed for eligibility. Following full-text evaluation, 18 reports were excluded because they did not satisfy the inclusion criteria. Ultimately, 16 studies were included in the qualitative synthesis of this systematic review.

Characteristics of Included Studies

A total of 16 studies published between 2023 and 2026 were included in this systematic review. The studies were conducted across several geographical settings, including Türkiye, India, Indonesia, Saudi Arabia, Brazil, the United Kingdom, and the Netherlands. Study designs and sample characteristics varied considerably, including retrospective image-based investigations, cross-sectional studies, comparative model-development studies, and clinical pilot research. The datasets ranged from small clinical samples to several thousand dental photographs, radiographs, and digital dental impressions.

The investigated AI applications covered several areas of pediatric dentistry. Five studies primarily evaluated AI-assisted caries detection or classification using intraoral photographs and dental radiographs [2-4,16,17]. Other investigations focused on automated tooth identification, classification, segmentation, and the detection of developmental or structural abnormalities using panoramic radiographs, CBCT, and digital dental impressions [5,18-21]. A further group explored applications with more direct implications for clinical decision-making, including extraction-related assessment, furcation lesion detection, prediction of spontaneous space closure, orthodontic assessment, and AI-generated diagnoses and treatment recommendations [9,12,15,22-24].

The AI technologies employed included convolutional neural networks, You Only Look Once (YOLO)-based object-detection models, transformer architectures, U-Net variants, conventional machine-learning classifiers, and LLMs. Reference standards differed across investigations and included expert annotations, specialist consensus, orthodontist assessments, dentist-generated treatment plans, and predefined radiographic or developmental criteria. Similarly, validation strategies included internal training, validation, and testing partitions, cross-validation, and, in a limited number of studies, independent validation datasets.

The reported outcomes encompassed accuracy, sensitivity, specificity, precision, recall, F1-score, AUC, the Dice similarity coefficient, IoU, and mAP. These measures reflected different aspects of model performance and were therefore selected based on the specific task evaluated in each study.

Performance of AI applications in Pediatric Dentistry

AI-assisted dental caries detection: Caries detection was one of the most frequently investigated AI applications in the included studies. Different imaging modalities and deep-learning architectures were used to identify and classify carious lesions in primary and permanent dentition. Alangari and AlShenaifi [2] evaluated ResNet-18, MobileNetV3-Small, and EfficientNet-B0 using pediatric oral photographs. ResNet-18 achieved a balanced accuracy of 0.929 and a weighted F1-score of 0.954, while EfficientNet-B0 demonstrated a ROC-AUC of 0.978 and a PR-AUC of 0.990.

Similarly, de Brito et al. [3] evaluated YOLOv8 and EfficientNet-B0 for caries detection in young permanent molars using interproximal radiographs. The YOLOv8 model achieved 91% accuracy and 98% precision, whereas the classification model achieved 89% accuracy. Fadilah et al. [4] investigated YOLOv8x for detecting caries from intraoral photographs of schoolchildren. The model achieved an mAP of 45.8%, with an overall precision of 72.6% and a recall of 41.1%. Although class-specific sensitivity and specificity values were reported, the overall recall indicated that some lesions could remain undetected.

Azhari et al. [17] evaluated semantic segmentation models for identifying carious lesions in bitewing radiographs. Model performance differed substantially according to lesion category, with lower IoU values for primary-tooth caries categories than for sound tooth structures. Güçlü et al. [16] investigated YOLOv5x

for caries detection in bitewing radiographs of primary and permanent teeth and reported F1-scores of 0.79 and 0.76, respectively. The corresponding AUC values were 0.87 for primary teeth and 0.76 for permanent teeth. Collectively, these studies demonstrated the potential of AI-based image analysis for caries detection, although performance varied by imaging modality, lesion category, and model architecture.

Tooth detection, segmentation, and developmental assessment: Several studies investigated automated identification, classification, and segmentation of teeth in pediatric dental images. Jaiswal et al. [19] evaluated a convolutional neural network for tooth classification using panoramic radiographs from children and adolescents. Özüdoğru et al. [5] investigated nnU-Net v2 for automatic tooth detection and numbering in CBCT images of patients with mixed dentition. The latter study reported a numbering accuracy of 0.99, with precision of 0.86, recall of 0.84, and a Dice similarity coefficient of approximately 0.81.

van Nistelrooij et al. [20] evaluated an automated framework for tooth detection, segmentation, and labeling using digital dental impressions. The reported performance included a detection F1-score of 0.996, a segmentation Dice similarity coefficient of 0.969, and a macro-F1-score of 0.989 for tooth labeling. These findings indicated high technical performance in automated three-dimensional dental image processing. Gülşen et al. [21] also investigated YOLOv8-based detection of ectopic teeth in pediatric panoramic radiographs, although the extracted data did not provide sufficiently verified numerical performance estimates for detailed comparison.

Uzel et al. [18] evaluated AI-based segmentation and classification of pediatric panoramic radiographs. Although the study reported high classification accuracy across the evaluated datasets, segmentation performance was less consistent, with a reported recall of 38% and an F1-score of 55%. This difference highlighted the importance of assessing individual model tasks rather than relying on a single overall accuracy measure.

AI-assisted clinical decision-making and treatment planning: A smaller group of studies evaluated AI applications more directly related to treatment planning and clinical decision-making. Çetin et al. [23] investigated machine-learning classification of indications for first permanent molar extraction using panoramic radiographs. The model achieved 77.78% accuracy, with an AUC of 0.77 and a sensitivity of 0.81. However, the classification outcome was derived from developmental staging criteria rather than verified against actual clinical extraction decisions or subsequent treatment outcomes.

Karamüftüoğlu et al. [12] evaluated YOLO-based and transformer-based models for detecting furcation lesions in primary molars. The best-performing model, RT-DETR-X, achieved an mAP@0.5 of 0.434. Although the study linked radiographic findings to potential treatment recommendations, the recommendations were based on radiographic classification categories and were not validated against subsequent patient outcomes. Yelkenci et al. [15] investigated automated assessment of second- and third-molar characteristics to classify the likelihood of spontaneous space closure following first permanent molar extraction. The YOLOv8n model achieved an overall mAP@0.5 of 0.983, with F1-scores of 0.978 for second-molar detection and 0.904 for third-molar detection. However, these values primarily reflected anatomical detection performance. The prognostic categories were assigned using a predefined decision matrix rather than validated against observed post-extraction outcomes, limiting conclusions regarding the model's actual predictive ability. Naik et al. [9] compared diagnoses and treatment plans generated by ChatGPT 3.5 with those developed by pediatric dentists in a pilot study involving 10 children. The study reported no statistically significant difference between AI-generated and clinician-generated assessments ($p = 0.926$), with agreement varying across individual dentists. Nevertheless, the small sample and exploratory design did not establish equivalence or non-inferiority between ChatGPT and pediatric dentists.

Herawati et al. [24] compared four multimodal LLMs for orthodontic assessment using standardized intraoral photographs from 50 children. Agreement with the orthodontist varied across models, with kappa values ranging from −0.15 to 0.63 and AUC values from 0.56 to 0.85. The absence of clinical and radiographic information in the assessment limited the extent to which these results could be interpreted as evidence of comprehensive orthodontic treatment-planning capability. Pham and Al-Hebshi [22] also explored a text-driven deep-learning framework for classifying pediatric dental conditions from panoramic radiographs, reporting different classification accuracies across the evaluated architectures, with the one-dimensional convolutional neural network (1D-CNN), bidirectional encoder representations from transformers (BERT), and long short-term memory (LSTM) models achieving accuracies of 84%, 77%, and 57%, respectively.

Overall, the findings suggested that AI performance was generally more promising for defined image-based detection, segmentation, and classification tasks than for complex clinical decisions requiring integration of multiple patient-specific factors. Importantly, the reported technical performance measures did not consistently demonstrate improvements in treatment selection, treatment success, or patient-related outcomes. Table *1* summarizes the principal characteristics and performance estimates of these studies. Detailed performance estimates are provided in the Appendices.

Study	Country	Design and sample	AI application	Comparator/reference standard	Validation and principal findings	Risk of bias
Azhari et al.		Diagnostic model study; 771 bitewing radiographs	Semantic-segmentation models	Manual annotations by	Pediatric IoU was high for sound surfaces (97%) but low for	Unclear

(2023) [17]	Saudi Arabia	(554 adult, 217 pediatric).	for caries detection in bitewing images.	general dentists.	primary-caries categories (8-25%), indicating limited segmentation performance for lesions.	(QUADAS-2)
de Brito et al. (2025) [3]	Brazil	Retrospective image-based study; 500 selected radiographs from approximately 5,000 screened; children aged 6-9 years.	YOLOv8 for detection and EfficientNet-B0 for classification of first permanent molar findings.	Radiology professor calibration/annotation; inter-examiner agreement reported as 97.4% (κ = 0.88).	YOLOv8 accuracy was 91%, and precision was 98%; classifier accuracy was 89%. Selection of a subset from a larger image pool may affect representativeness.	High (QUADAS-2)
Alangari and AlShenaifi (2026) [2]	Saudi Arabia	Image-based diagnostic model study; 435 intraoral photographs from children aged 3-14 years.	ResNet-18, MobileNetV3-Small, and EfficientNet-B0 for image classification.	Dentist-annotated reference labels; patient-level stratified split.	ResNet-18 balanced accuracy 0.929 and weighted F1 0.954; sensitivity 1.00. EfficientNet-B0 ROC-AUC 0.978 and PR-AUC 0.990.	Unclear (QUADAS-2)
Fadilah et al. (2025) [4]	Indonesia	Prospective cross-sectional analytical study; 3,221 intraoral photographs from schoolchildren.	YOLOv8x/HI Bogi for ICDAS-based caries detection and grading.	Eight dentists annotated images; two ICDAS experts verified labels.	mAP 45.8%, precision 72.6%, recall 41.1%; class-specific sensitivity 82.31-96.45% and specificity 77.10-99.33%. AI assessment was reported to be approximately four times faster.	High (QUADAS-2)
Çetin et al. (2026) [23]	Türkiye	Retrospective observational study; 1,000 panoramic images from children aged 8-10; 176 selected for analysis.	Gabor/HOG feature extraction with support vector machine classification to estimate extraction indication.	Label based on Demirjian developmental stages E-F; not actual post-extraction clinical outcomes.	Accuracy 77.78%, AUC 0.77, and sensitivity 0.81 for the indicated group. The reference label is a proxy for clinical indication.	High (QUADAS-2)
Gülşen et al. (2026) [21]	Türkiye	Image-based study of pediatric mixed-dentition panoramic radiographs; 1,059 panoramic radiographs that include 1,632 ectopic canines and 245 ectopic molars	YOLOv8-based detection of ectopic canines/molars.	Specialist-calibrated image annotations.	Patient-level split into training (80%), validation (10%), and testing (10%) sets. YOLOv8x achieved an overall mAP@0.5 of 0.756. For ectopic canines, precision, recall, F1-score, and Dice similarity coefficient were 0.786, 0.771, 0.778, and 0.768, respectively; for ectopic molars, the corresponding values were 0.812, 0.650, 0.722, and 0.757, respectively. Interobserver agreement was excellent (κ = 0.83 for canines; κ = 0.93 for molars).	Unclear (QUADAS-2)
Herawati et al. (2026) [24]	Indonesia	Comparative cross-sectional study; 50 children aged 9-12 years; five standardized intraoral photographs per child.	Four multimodal LLMs for orthodontic assessment/decision support.	Orthodontist assessment.	Agreement κ ranged from −0.15 to 0.63; AUC ranged from 0.56 to 0.85. Assessment was based on photographs without clinical or radiographic information.	Unclear (QUADAS-2)
Jaiswal et al. (2023) [19]	India	Retrospective image-based study; 1,293 initial orthopantomograms, of which 620 were selected; age 4-16 years, including 314 mixed-dentition images.	CNN-based tooth classification on panoramic radiographs.	Classification into deciduous and permanent dentition categories using the selected and labeled radiographic dataset.	70:15:15 training, validation, and testing split. EfficientNet-B0 and EfficientNet-B3 achieved 98% accuracy, precision, recall, and F1-score.	High (QUADAS-2)
Karamüftüoğlu et al. (2025) [12]	Türkiye	Retrospective study; 387 panoramic radiographs from children aged 3-13 years; 70/20/10 train/validation/test split.	YOLOv12x, RT-DETR-L and RT-DETR-X for detection/classification of primary-molar furcation lesions, linked to treatment recommendations.	Annotated lesion categories; expert-annotated samples used for qualitative assessment.	RT-DETR-X achieved the highest mAP@0.5 (0.434). The treatment recommendation component was linked to radiographic categories rather than validated against subsequent patient outcomes.	Unclear (QUADAS-2)
Naik et al. (2025) [9]	India	Cross-sectional pilot; 10 cooperative children aged 6-12 years; four pediatric dentists independently assessed standardized scenarios.	ChatGPT 3.5 was used to generate pediatric dental diagnoses and treatment plans.	Dentist-generated plans; reviewed by a fifth dentist and validated by a senior pediatric dentist (blinded).	No statistically significant difference was reported ($p = 0.926$); the very small sample does not establish equivalence or non-inferiority.	High (JBI cross-sectional)
Özüdoğru et al. (2025) [5]	Türkiye	Retrospective study; 49 CBCT volumes from children with mixed dentition; 90/10 train/test	nnU-Net v2 for tooth segmentation and numbering.	Consensus ground truth from three senior pediatric dentists and one oral/maxillofacial radiologist, checked by a	Numbering accuracy 0.99; precision 0.86; recall 0.84; Dice 0.81; Jaccard 0.81; 95% Hausdorff distance 1.93.	High (QUADAS-2)

		split.		senior radiologist.		
Pham and Al-Hebshi (2026) [22]	United Kingdom	Comparative computational model-development study; 58 selected panoramic radiographs (29 caries and 29 periapical infection images) from an original 100-image dataset.	Radiograph-to-text transformer followed by 1D-CNN, LSTM, or BERT classification; compared with image-based pretrained CNNs.	Original dataset annotations were established through six dental experts, including independent annotation and consensus; a dentist qualitatively reviewed the generated text for anatomical plausibility.	Repeated 90:10 partitioning and five-fold cross-validation. 1D-CNN achieved 84% accuracy (AUC 0.93), BERT 76.67%, and LSTM 56.67%; image-based models achieved 50.00-66.67% accuracy in the reported holdout evaluation.	Unclear (QUADAS-2)
Güçlü et al. (2025) [16]	Türkiye	Retrospective study; 1,000 primary and 1,000 permanent bitewing radiographs; 5,704 lesions.	YOLOv5x for caries detection in primary and permanent teeth.	Annotated radiographic lesions; 80/10/10 data split.	Primary teeth: F1 0.79, AUC 0.87. Permanent teeth: F1 0.76, AUC 0.76.	Unclear (QUADAS-2)
Uzel et al. (2025) [18]	Türkiye	Retrospective study; 2,000 pediatric panoramic radiographs; 140 annotated (71 positive, 69 negative); independent validation set of 20.	AI-based segmentation and classification of a pediatric dental finding on panoramic images.	Annotations by three calibrated pediatric dentists (κ = 0.92).	Segmentation precision 100%, recall 38%, F1 55%; classification accuracy 100% on internal/external datasets. Independent validation sample was small.	Unclear (QUADAS-2)
van Nistelrooij et al. (2025) [20]	Netherlands/Brazil	3D dental-impression dataset; 716 digital impressions from 351 patients, including primary/mixed dentition.	Automated tooth detection, segmentation and labeling in 3D intraoral scans.	Reference tooth annotations; five-fold cross-validation.	Tooth-detection F1 0.996, segmentation Dice 0.969, and labeling macro-F1-score 0.989; processing time under two seconds.	Unclear (QUADAS-2)
Yelkenci et al. (2026) [15]	Türkiye	Retrospective model-development study; 200 pediatric panoramic images.	YOLOv8n-OBB for molar detection and maturity/angulation estimation to support space-closure decision-making after first permanent molar extraction.	Anatomical annotations and a decision matrix; no validation against actual post-extraction outcomes.	mAP@0.5 0.983; F1 0.978 for second molars and 0.904 for third molars. Detection performance should not be interpreted as prognostic accuracy.	High (PROBAST+AI)

TABLE 1: Characteristics of studies included in the systematic review

AI: artificial intelligence, ROC-AUC: area under the receiver operating characteristic curve, CBCT: cone-beam computed tomography, CNN: convolutional neural network, F1: F1-score, ICDAS: International Caries Detection and Assessment System, IoU: intersection over union, LLMs: large language models, mAP: mean average precision, OPG: orthopantomogram, PR-AUC: area under the precision–recall curve, QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2, JBI: Joanna Briggs Institute, PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for AI, YOLO: You Only Look Once, HOG: histogram of oriented gradients, RT-DETR: real-time detection transformer, 1D-CNN: one-dimensional convolutional neural network, LSTM: long short-term memory, BERT: bidirectional encoder representations from transformers, OBB: oriented bounding box, 3D: three-dimensional

Risk-of-Bias and Methodological Quality Assessment

Risk of bias was assessed using instruments appropriate to the methodological design and intended application of each included study. QUADAS-2 was applied to 14 diagnostic or image-classification studies, the JBI Critical Appraisal Checklist for Analytical Cross-Sectional Studies was used for Naik et al. [9], and PROBAST+AI was applied to the prognostic model investigated by Yelkenci et al. [15]. As these instruments evaluate different methodological domains, their judgments were interpreted separately and were not combined into a single numerical score.

QUADAS-2 assessment: Among the 14 studies assessed using QUADAS-2, five were judged to have a high overall risk of bias: de Brito et al. [3], Fadilah et al. [4], Çetin et al. [23], Jaiswal et al. [19], and Özüdoğru et al. [5]. The remaining nine studies were classified as having an unclear overall risk of bias: Azhari et al. [17], Alangari and AlShenaifi [2], Gülşen et al. [21], Herawati et al. [24], Karamüftüoğlu et al. [12], Pham and Al-Hebshi [22], Güçlü et al. [16], Uzel et al. [18], and van Nistelrooij et al. [20]. No study was unequivocally judged to be at low risk of bias across all assessed domains. The domain-level judgments and supporting rationales are presented in Table *2*.

Study	Patient selection	Index test	Reference standard	Flow and timing	Overall
Azhari et al. (2023) [17]	Unclear	Low	Unclear	Unclear	Unclear
de Brito et al. (2025) [3]	High	Unclear	Low	Unclear	High
Alangari and AlShenaifi (2026) [2]	Unclear	Low	Unclear.	Low	Unclear
Fadilah et al. (2025) [4]	High	Low	Low	Unclear	High
Çetin et al. (2026) [23]	High	Low .	High	Unclear	High
Gülşen et al. (2026) [21]	Unclear	Low	Low	Unclear	Unclear
Herawati et al. (2026) [24]	Unclear	Unclear	Unclear	Low	Unclear
Jaiswal et al. (2023) [19]	High	Unclear .	Unclear	Unclear	High
Karamüftüoğlu et al. (2025) [12]	Unclear	Low	Unclear	Unclear	Unclear
Özüdoğru et al. (2025) [5]	High .	Low	Low	Unclear	High
Pham and Al-Hebshi (2026) [22]	Unclear	Unclear	Unclear	Unclear	Unclear
Güçlü et al. (2025) [16]	Unclear	Low	Unclear	Unclear	Unclear
Uzel et al. (2025) [18]	Unclear	Low	Low	Unclear	Unclear
van Nistelrooij et al. (2025) [20]	Unclear	Low	Low	Unclear	Unclear

TABLE 2: QUADAS-2 domain-level judgments for diagnostic/classification studies

QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2

Concerns regarding patient selection were common across the included diagnostic studies. Several investigations used retrospective image collections, selected subsets from larger datasets, or did not adequately describe consecutive or random recruitment. These limitations raised concerns about selection and spectrum bias and may have limited the representativeness of the study populations. Such concerns were particularly evident in de Brito et al. [3], Fadilah et al. [4], Çetin et al. [23], Jaiswal et al. [19], and Özüdoğru et al. [5].

The index-test domain was generally better described, as most studies specified the AI architecture, model-development process, and performance evaluation. Nevertheless, reporting was insufficient in several investigations to establish whether thresholds were prespecified, whether assessors were blinded to reference labels, and whether data partitioning occurred at the participant level rather than the image level. These uncertainties were particularly relevant to studies involving multiple images from individual participants.

Concerns regarding the reference standard were also identified. Several studies relied on expert-generated annotations without sufficiently detailed reporting of independent assessment, consensus procedures, adjudication, or blinding. The most notable concern was in Çetin et al. [23], where extraction indications were inferred from Demirjian developmental stages rather than verified against independent clinical decisions or actual post-extraction outcomes. Conversely, studies such as de Brito et al. [3], Fadilah et al. [4], Özüdoğru et al. [5], and Uzel et al. [18] provided clearer descriptions of specialist annotation, calibration, or consensus procedures.

The flow-and-timing domain was frequently classified as unclear because some studies did not fully report image-level exclusions, failed or incomplete cases, repeated images, or the relationship between image acquisition and reference assessment. In particular, the possibility of images from the same participant being allocated across different data partitions required clarification in some investigations, including Fadilah et al. [4] and van Nistelrooij et al. [20].

JBI assessment: Naik et al. [9] was assessed separately using the JBI Critical Appraisal Checklist for Analytical Cross-Sectional Studies and was judged to have substantial methodological limitations. Although the study described the pediatric population, standardized clinical scenarios, and dentist-generated treatment plans followed by expert review, the sample consisted of only 10 children. Potential influences such as case complexity, prompt formulation, and clinician experience were not comprehensively addressed. Furthermore, although the study reported no statistically significant difference between ChatGPT-generated and clinician-generated assessments ($p = 0.926$), the small exploratory sample was insufficient to establish equivalence or non-inferiority. The item-level appraisal and supporting rationale are provided in Table *3*.

JBI item	Judgment	Rationale
1. Were the inclusion criteria clearly defined?	Yes	The target population and clinical scenario were described, although the pilot sample was very small.
2. Were the study subjects and setting described in detail?	Yes	The study described ten cooperative children aged 6–12 years and a standardized case/scenario approach.
3. Was the exposure measured in a valid and reliable way?	Yes	ChatGPT 3.5 was used with standardized case information; reproducibility across prompts/sessions remains a concern.
4. Were objective, standard criteria used for measurement of the condition?	Unclear	The clinical scenarios and diagnostic reference process were described, but full operational criteria for each diagnosis/plan were not sufficiently clear.
5. Were confounding factors identified?	Unclear	Potential influences such as case complexity, prompt wording, and clinician experience were not comprehensively handled.
6. Were strategies to deal with confounding factors stated?	Unclear	No robust adjustment or design strategy addressing these potential influences was evident.
7. Were the outcomes measured in a valid and reliable way?	Unclear	Dentist review and blinded senior validation were reported, but the small number of cases and limited detail on reliability restrict confidence.
8. Was appropriate statistical analysis used?	Unclear	A non-significant p-value (p = 0.926) was reported; the study was not designed or powered to establish equivalence/non-inferiority.

TABLE 3: JBI Critical Appraisal Checklist for Analytical Cross-Sectional Studies

JBI: Joanna Briggs Institute

Naik et al. (2025) [9]

PROBAST+AI assessment: Yelkenci et al. [15] was assessed using PROBAST+AI and was judged to have a high overall risk of bias. Concerns were identified across all four domains: participants, predictors, outcome, and analysis. The study used a relatively small dataset of 200 pediatric panoramic radiographs, raising concerns about sample representativeness. Additionally, second-molar maturity was approximated using bounding-box height, which represents an indirect measure rather than a validated developmental predictor. Most importantly, spontaneous space-closure categories were assigned using a predefined decision matrix and were not validated against observed post-extraction outcomes in a longitudinal patient cohort. The analysis was also limited to internal data partitioning without external or prospective clinical validation. Consequently, the reported mAP@0.5 of 0.983 primarily reflected anatomical detection performance rather than validated prognostic accuracy for spontaneous space closure. The domain-level judgments and supporting explanations are presented in Table *4*.

Domain	Judgment	Rationale
Participants	High	The model was developed/evaluated using 200 pediatric panoramic images; representativeness and participant selection are limited/unclear.
Predictors	High	Second-molar maturity was approximated using bounding-box height, which is an indirect proxy rather than a validated developmental measure.
Outcome	High	Clinical success/space-closure categories were assigned through a decision matrix and were not validated against observed post-extraction outcomes in a longitudinal cohort.
Analysis	High	Evaluation was internal and based on a relatively small image set; calibration, optimism correction, and external validation were not established.

TABLE 4: PROBAST+AI assessment

PROBAST+AI: Prediction Model Risk of Bias Assessment Tool for AI

Yelkenci et al. (2026) [15]

Overall, the methodological appraisal identified recurring limitations, including retrospective or selectively sampled datasets, small or single-center populations, incomplete reporting of reference-standard procedures, uncertainty about participant-level data partitioning, and limited external or prospective

validation. These limitations were considered when interpreting the reported AI performance and its potential applicability to routine pediatric dental practice.

Discussion

This systematic review evaluated 16 studies investigating AI applications relevant to pediatric dental diagnosis, treatment planning, and clinical decision-making. The findings demonstrate promising technical performance across several applications, particularly automated caries detection, tooth identification, segmentation, and assessment of developmental abnormalities. However, an important finding was the discrepancy between the number of studies demonstrating technical model performance and the limited number evaluating actual clinical decision-making or patient outcomes. Most included investigations assessed intermediate tasks, such as image classification, object detection, and anatomical segmentation, rather than whether AI-assisted recommendations changed treatment selection, improved treatment appropriateness, or led to better clinical outcomes. Therefore, although AI shows considerable potential as a diagnostic support tool in pediatric dentistry, its effectiveness as a treatment-planning intervention remains insufficiently established.

The findings of this review are consistent with those of Rokhshad et al., who identified promising AI performance across several pediatric dental applications, including early childhood caries detection, tooth identification, oral health assessment, and supernumerary tooth identification [25]. Their systematic review and meta-analysis included 33 studies and reported accuracy estimates ranging from 60% to 99% across different applications. Nevertheless, the authors concluded that the available evidence was insufficient to establish clinical effectiveness. Similarly, a more recent umbrella review by Garg et al. [13] identified promising diagnostic performance across existing systematic reviews of AI in pediatric dentistry, while highlighting methodological weaknesses, limited risk-of-bias assessment, and the need for stronger clinical validation. The present review complements this broader evidence by examining not only diagnostic applications but also studies that attempt to connect AI-generated findings with treatment-planning decisions. An important observation is that the included evidence is unevenly distributed across clinical applications. Most studies concentrated on image-based detection, classification, and segmentation, whereas comparatively fewer investigated complete treatment-planning processes or patient-level outcomes. Consequently, although AI has demonstrated encouraging technical performance in several pediatric dental tasks, its role in supporting comprehensive clinical decision-making remains less established.

Dental caries detection represented one of the most frequently investigated applications in the included studies. Alangari and AlShenaifi [2], de Brito et al. [3], Fadilah et al. [4], Azhari et al. [17], and Güçlü et al. [16] evaluated different deep-learning architectures to identify or classify carious lesions using intraoral photographs, bitewing radiographs, or other dental radiographic images. Although several models showed encouraging performance, results varied by imaging modality, lesion category, and evaluation metric.

For example, Alangari and AlShenaifi [2] reported a balanced accuracy of 0.929 using ResNet-18 for pediatric caries classification from intraoral photographs. De Brito et al. [3] reported an accuracy of 91% and a precision of 98% using YOLOv8 for caries detection in young permanent molars. In contrast, Fadilah et al. [4] reported an mAP of 45.8% for YOLOv8x, with an overall recall of 41.1%, demonstrating that performance may be substantially lower for particular datasets and detection tasks. Similarly, Azhari et al. [17] observed considerable variation in segmentation performance across lesion categories, while Güçlü et al. [16] reported differences in classification performance between primary and permanent teeth. These differences suggest that the effectiveness of AI-based caries detection depends not only on the selected architecture but also on image characteristics, lesion definition and severity, and the quality of reference annotations.

These findings are consistent with the systematic review and meta-analysis conducted by Abbott et al. [26], which included 45 studies evaluating AI platforms for dental caries detection. Reported accuracy varied widely, from 41.5% to 98.6%, while the quantitative synthesis of seven studies demonstrated a pooled sensitivity of 76%, a specificity of 91%, and an AUC of 0.92. The authors also identified substantial heterogeneity across AI platforms and imaging modalities. This variation is comparable to the differences observed among the caries-detection studies included in the present review. More recently, Karamüftüoğlu et al. [12] reported pooled sensitivity, specificity, and AUC estimates of 0.91, 0.97, and 0.98, respectively, for radiographic caries detection in their systematic review and meta-analysis specifically addressing pediatric dentistry. Although these pooled estimates are encouraging, differences in study eligibility, included datasets, reference standards, and analytical approaches mean that they should not be directly compared with the individual performance estimates in the present review.

Collectively, the evidence supports the potential of AI-assisted caries detection to identify lesions and improve the consistency of image interpretation, although differences in lesion severity, imaging modality, and reference standards limit direct comparisons between models. Further validation using standardized diagnostic criteria and clinically representative populations is required to establish the reliability of these AI models in routine pediatric dental practice.

Another important application identified in this review was the automated identification, classification, and segmentation of dental structures. Jaiswal et al. [19] evaluated convolutional neural network architectures for distinguishing deciduous and permanent teeth, with EfficientNet-B0 and EfficientNet-B3 achieving 98% accuracy. Özüdoğru et al. [5] investigated nnU-Net v2 for tooth detection and numbering in CBCT images,

reporting a numbering accuracy of 0.99 and a Dice similarity coefficient of approximately 0.81. Similarly, van Nistelrooij et al. [20] reported high performance in automated tooth detection, segmentation, and labeling using digital dental impressions, with a detection F1-score of 0.996 and a segmentation Dice similarity coefficient of 0.969. Uzel et al. [18], however, demonstrated that high classification accuracy does not necessarily translate into equally reliable segmentation, reporting a segmentation recall of 38% and an F1-score of 55%.

These findings indicate that AI can perform well in tasks involving the recognition and localization of relatively well-defined anatomical structures. However, differences in the imaging modalities and segmentation targets limit direct comparison between studies. Panoramic radiographs and CBCT images provide two-dimensional and three-dimensional representations, respectively, while digital impressions capture dental surface morphology without the same radiographic information. Consequently, model performance in one modality cannot automatically be extrapolated to another.

Gülşen et al. [21] extended these applications to detect ectopic canines and molars during the mixed dentition stage using a YOLOv8-based model. The overall mAP@0.5 was 0.756, with better detection performance for ectopic canines than for ectopic molars. The study also reported excellent interobserver agreement between expert assessors. These findings suggest that AI may help identify developmental and eruption abnormalities that require early clinical assessment. However, the study primarily evaluated detection of ectopic eruption on panoramic radiographs and did not establish the model's ability to predict subsequent eruption patterns or the need for specific interventions. The broader pediatric dentistry literature supports the potential of AI in automated tooth identification and developmental assessment. Rokhshad et al. identified tooth numbering, mesiodens detection, and caries assessment among the most frequently investigated applications in pediatric dentistry [25]. Garg et al. similarly reported that computer-vision approaches, particularly those involving panoramic radiographs, represented a substantial proportion of the existing pediatric AI literature [13]. The present findings are consistent with this pattern and suggest that image-based AI applications currently represent a more developed area of pediatric dental research than comprehensive AI-assisted treatment planning.

Although most included studies focused on diagnostic or image-analysis tasks, several investigated applications more directly related to treatment planning. Çetin et al. [23] evaluated a machine-learning approach for identifying the developmental timing criterion relevant to first permanent molar extraction. The model achieved 77.78% accuracy and an AUC of 0.77. However, the classification was based on Demirjian developmental stages rather than actual clinical extraction decisions or subsequent treatment outcomes. Therefore, the reported performance should be interpreted as the ability to classify a radiographic developmental criterion rather than the accuracy of comprehensive extraction planning.

Karamüftüoğlu et al. [12] investigated deep-learning models to detect furcation lesions in primary molars and link radiographic findings to potential treatment recommendations. The best-performing RT-DETR-X model achieved an mAP@0.5 of 0.434, indicating limitations in lesion-detection performance. Furthermore, although the model linked radiographic categories with treatment recommendations, the appropriateness of those recommendations was not established through subsequent clinical outcomes. Decisions regarding pulp therapy or extraction require consideration of clinical symptoms, restorability, pathological findings, the child's overall dental condition, and other patient-specific factors that may not be represented in a radiographic image alone. Yelkenci et al. [15] developed a model for assessing second- and third-molar characteristics in relation to spontaneous space closure following first permanent molar extraction. Although the model achieved an mAP@0.5 of 0.983 for anatomical detection, this value should not be interpreted as prognostic accuracy for spontaneous space closure. The subsequent prognostic classifications were generated using a predefined decision matrix and were not validated against observed post-extraction outcomes in an independent longitudinal cohort. Consequently, the study demonstrates the feasibility of automated radiographic assessment but does not establish the accuracy of the proposed prognostic application. These findings highlight an important distinction between identifying anatomical or radiographic features and determining the most appropriate treatment for an individual child. Pediatric dental treatment planning frequently involves multiple interacting considerations, including symptoms, clinical examination findings, developmental status, caries risk, behavioral cooperation, treatment feasibility, and parental preferences. An algorithm that accurately identifies a radiographic feature may provide useful information but may not be sufficient to independently recommend appropriate treatment.

The application of LLMs represents a relatively new area of pediatric dental research. Naik et al. [9] evaluated diagnoses and treatment plans generated by ChatGPT 3.5 in a small pilot study involving 10 children and compared the outputs with assessments from pediatric dentists. Although no statistically significant difference was identified between the assessments ($p = 0.926$), this result should not be interpreted as evidence of equivalence or non-inferiority. The small sample, exploratory design, and dependence on standardized clinical scenarios limit the generalizability of the findings to routine clinical practice.

Herawati et al. [24] compared multimodal LLMs for orthodontic assessment using standardized intraoral photographs from children. Agreement with the orthodontist varied considerably across the evaluated models, with kappa values ranging from −0.15 to 0.63 and AUC values from 0.56 to 0.85. These findings suggest that the performance of generative AI systems may depend substantially on the model architecture and the clinical task being evaluated. Furthermore, assessment based on photographs alone may not provide the clinical and radiographic information required for comprehensive orthodontic treatment planning.

Pham and Al-Hebshi [22] investigated a text-driven framework in which GPT-4 generated structured descriptions from panoramic radiographs, followed by classification using 1D-CNN, LSTM, and BERT architectures. The 1D-CNN achieved 84% accuracy in the reported holdout evaluation, whereas BERT and LSTM demonstrated lower accuracy. Importantly, these results represent the performance of downstream classification models using generated text rather than the independent diagnostic accuracy of GPT-4.

The broader dental literature also identifies considerable variability in LLM performance. Alhazmi et al., in a systematic review of 31 studies, identified concerns about inaccurate responses, insufficient source attribution, and fabricated references, concluding that LLMs may be better regarded as supplementary educational tools rather than autonomous clinical decision-making systems [27]. A separate systematic review and meta-analysis of LLM-generated responses to dental patient inquiries reported an overall pooled accuracy of approximately 81.9%, but also identified limitations in the clinical acceptability and reliability of responses to complex questions [28]. These findings are consistent with the variability observed in the LLM studies included in the present review. Accordingly, generative AI may have potential applications in assisting clinicians with information synthesis, structured documentation, and preliminary decision support. Nevertheless, its use in pediatric dentistry requires particular caution because inaccurate recommendations may affect developing dentition and long-term oral health. Human oversight and verification against established clinical guidelines and patient-specific findings remain essential.

A major finding of this review was the considerable heterogeneity in reported AI performance. The included studies differed in their clinical objectives, imaging modalities, sample sizes, AI architectures, reference standards, and validation strategies. Consequently, the reported performance measures were not directly comparable, and quantitative pooling was not considered appropriate. The use of different imaging modalities is an important source of variation. Panoramic radiographs, bitewing radiographs, CBCT images, intraoral photographs, and digital dental impressions provide different types and levels of anatomical information. Furthermore, the complexity of the diagnostic task may influence model performance. Identifying a tooth or segmenting an anatomical structure is fundamentally different from detecting an early carious lesion or generating a treatment recommendation that requires integrating multiple clinical variables.

Differences in dataset composition and class distribution may also affect reported results. Studies involving selected or balanced datasets may produce performance estimates that differ from those obtained in populations reflecting the actual prevalence and spectrum of disease. Similarly, the number and experience of annotators, the criteria used to establish ground truth, and the methods used to handle uncertain cases may influence the apparent accuracy of an AI system. These issues are particularly important when interpreting models developed using retrospective image collections. The findings of Abbott et al. [26] and Karamüftüoğlu et al. [12] further demonstrate that heterogeneity is a recurring challenge in dental AI research, particularly in caries detection [27,28]. Therefore, the absence of a meta-analysis in the present review should not be considered a methodological shortcoming. Given the differences in clinical tasks and outcome measures, a pooled estimate could have been misleading.

The methodological appraisal identified important limitations across the included studies. Among the 14 diagnostic or classification studies assessed using QUADAS-2, five were judged to have a high overall risk of bias, while nine had an unclear overall risk of bias. No study was considered unequivocally at low overall risk across all assessed domains. The separate JBI assessment of Naik et al. [9] identified substantial methodological concerns, particularly related to the very small sample and exploratory design. Similarly, the PROBAST+AI assessment of Yelkenci et al. [15] identified a high overall risk of bias in the prognostic component, particularly because the outcome was not validated against observed post-extraction space-closure outcomes and external validation was absent.

Patient selection and dataset representativeness were recurring concerns. Several investigations relied on retrospective image collections or selected subsets of available radiographs. Such approaches may introduce selection bias and limit the representation of the full clinical spectrum encountered in pediatric dental practice. Small datasets may also result in unstable estimates of model performance, particularly when multiple diagnostic categories or uncommon abnormalities are evaluated.

The validity of reference standards was another important methodological consideration. Although several studies used specialist annotations, calibration procedures, or expert consensus, the independence and reproducibility of these assessments were not consistently reported. Furthermore, a radiographic label or predefined developmental criterion does not necessarily represent a clinically verified treatment decision. This distinction is particularly relevant to studies assessing extraction indications and prognostic treatment outcomes.

Validation strategies also varied across studies. Patient-level data partitioning, as reported in studies such as Gülşen et al. [21], is important because it reduces the risk of information leakage between training and testing datasets. However, an internal test set does not establish that a model will perform similarly in a different hospital, geographic population, or imaging environment. Independent external validation is therefore essential before clinical implementation.

These concerns are consistent with the broader literature on AI model development and evaluation. The TRIPOD+AI statement emphasizes transparent reporting of model development, validation, performance, and applicability, including providing sufficient information to assess the reliability and generalizability of

prediction models [29]. Similarly, the DECIDE-AI guideline highlights the importance of early clinical evaluation, safety assessment, human factors, and the interaction between AI systems and healthcare professionals [30]. These recommendations are particularly relevant to pediatric dentistry, where clinical decisions often require interpretation of imaging findings alongside behavioral, developmental, and patient-specific information.

The findings suggest that current evidence supports using AI primarily as an adjunct to pediatric dental diagnostic assessment rather than as an autonomous treatment-planning system. Image-based applications may assist with lesion identification, tooth localization, segmentation, and recognition of developmental abnormalities. However, their integration into clinical decision-making requires validation beyond technical performance, including assessment of recommendation appropriateness, clinical safety, and patient-related outcomes. Future research should therefore move beyond retrospective image classification toward prospective evaluation of AI-assisted clinical workflows.

Future research should prioritize multicenter prospective studies involving diverse pediatric populations and independent external validation datasets. Studies should compare AI-assisted assessment with conventional clinical evaluation and determine whether integrating AI improves diagnostic accuracy, treatment appropriateness, referral decisions, or patient-related outcomes. For prognostic applications, models should be evaluated against actual longitudinal outcomes rather than surrogate classifications or predefined decision matrices. Research involving LLMs should also use standardized clinical scenarios, clearly specified model versions, reproducible prompts, and independent expert evaluation. The performance of these systems should be assessed across different levels of case complexity and with access to relevant clinical and radiographic information. Additionally, future investigations should evaluate whether AI-generated recommendations are safe, explainable, and consistent with established pediatric dental guidelines.

This systematic review has several limitations that should be considered when interpreting its findings. First, although 16 studies were included, the evidence remains heterogeneous in terms of study designs, sample sizes, AI technologies, clinical applications, imaging modalities, and outcome measures, preventing meaningful quantitative synthesis or meta-analysis. Second, most included studies evaluated diagnostic or image-analysis tasks rather than complete treatment-planning processes, limiting the extent to which the findings can be used to infer clinical effectiveness. Third, several investigations relied on retrospective or selectively sampled datasets, relatively small populations, or standardized clinical scenarios, potentially restricting generalizability to routine pediatric dental practice. Furthermore, reference standards and validation approaches varied, and independent external or prospective clinical validation was limited. The risk-of-bias assessments also identified methodological concerns to consider when interpreting the reported performance estimates. Additionally, inter-reviewer agreement was not formally quantified using Cohen's kappa, which may limit the assessment of consistency during study selection and risk-of-bias evaluation. Although the review protocol was developed before starting the review, it was not prospectively registered in PROSPERO or another systematic review registry. Finally, differences in geographical settings, imaging protocols, and clinical practices may affect the transferability of the evaluated models to other pediatric dental populations. These limitations highlight the need for larger, multicenter prospective studies using standardized outcome measures, independent external validation, and direct evaluation of AI-assisted clinical decisions and patient outcomes.

Conclusions

AI shows promise for supporting pediatric dental practice, particularly in automated caries detection, tooth identification, image segmentation, and assessment of developmental and structural abnormalities. The included studies reported encouraging performance across several deep-learning architectures and image-analysis applications, suggesting that AI-based systems may assist clinicians in diagnostic assessment and selected treatment-planning decisions. However, the available evidence remains insufficient to establish that these technologies improve treatment success, clinical decision-making, or patient outcomes. Considerable variation in study designs, dataset characteristics, reference standards, and validation procedures, together with the methodological limitations identified in the risk-of-bias assessments, restricts the generalizability of the reported findings. Furthermore, high technical accuracy in image detection or classification should not be interpreted as evidence of validated clinical or prognostic effectiveness. Future research should prioritize large-scale, multicenter prospective studies, independent external validation, uniform diagnostic reference standards, and direct comparisons with experienced pediatric dental clinicians. Studies should also evaluate whether AI-assisted decisions translate into better treatment selection, improved clinical outcomes, and greater efficiency in routine practice. Until such evidence becomes available, AI should be considered an adjunct to, rather than a replacement for, professional clinical judgment in pediatric dentistry.

Appendices

Database	Search strategy/terms
PubMed/MEDLINE	("Artificial Intelligence"[MeSH Terms] OR "artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "convolutional neural network*" OR AI) AND ("Pediatric Dentistry"[MeSH Terms] OR "pediatric dentistry" OR pedodontic* OR child* OR children OR infant* OR "early childhood caries" OR "primary teeth" OR "pediatric dental") AND ("treatment planning" OR "clinical decision support" OR "treatment outcome*" OR diagnos* OR "risk prediction" OR "caries detection")
ScienceDirect	("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural networks") AND ("pediatric dentistry" OR "paediatric dentistry" OR "pediatric dental" OR "paediatric dental") AND ("primary dentition" OR "mixed dentition" OR "deciduous teeth" OR "primary teeth") AND ("dental treatment planning" OR "orthodontic treatment planning" OR "pulp therapy" OR "dental caries detection")
Cochrane CENTRAL	("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "artificial neural network*" OR "convolutional neural network*" OR "computer-assisted diagnosis" OR "automated diagnosis" OR "predictive model*" OR "intelligent system*") AND ("pediatric dentistry" OR "paediatric dentistry" OR "pediatric dental" OR "paediatric dental" OR "children" OR "child" OR "preschool children" OR "primary dentition" OR "mixed dentition") AND ("treatment planning" OR "clinical decision making" OR "treatment decision*" OR "treatment recommendation*" OR "treatment outcome*" OR "treatment success" OR "diagnostic accuracy" OR "clinical management" OR "risk prediction" OR "dental diagnosis")
Google Scholar	"artificial intelligence" pediatric dentistry treatment planning; "machine learning" children dentistry diagnosis; "deep learning" pediatric dental treatment; AI clinical decision support pediatric dentistry

TABLE 5: Electronic database search strategy

AI: artificial intelligence, MeSH: Medical Subject Headings, MEDLINE: Medical Literature Analysis and Retrieval System Online, CENTRAL: Cochrane Central Register of Controlled Trials

No.	Primary reason for exclusion	Number of studies
1	No pediatric-specific population or separately extractable pediatric results	4
2	Evaluated general AI knowledge, chatbot responses, or parental information rather than patient-specific diagnosis or treatment planning	5
3	AI application outside the review's defined scope (e.g., forensic age estimation or non-eligible prediction methodology)	7
4	Insufficient methodological information (conference abstract)	2
	Total excluded	18

TABLE 6: Summary of reasons for full-text exclusion

AI: artificial intelligence

Sr. No.	Study (year)	Task	Model	Modality	n	Sens.	Spec.	Acc./AUC	Other metrics	95% CI	AI vs. clinici
1	Alangari and AlShenaifi (2026) [2]	Caries classification	(C) ResNet-18/EfficientNet-B0/MobileNetV3-Small (same order in all metric columns)	Intraoral photos	435	(C) 1.00/0.93/0.98	(C) 0.86/0.90/0.81	(C) Acc 0.955/0.924/0.924; AUC 0.956/0.978/0.952	Bal. acc 0.929/0.919/0.894; weighted F1 0.954/0.925/0.920; PR-AUC 0.952/0.990/0.976; Grad-CAM	Bootstrap (ROC-AUC, PR-AUC; values NR in text)	None (betwe model benchmarkin only)
2	Fadilah et al. (2025) [4]	Caries detection	YOLOv8x	Intraoral photos	3,221	82.31-96.45% (D1-D6)	77.10-99.33% (D1-D6)	mAP 45.8% (no Acc/AUC)	Precision 72.6%; recall 41.1%; IoU	NR	(C) Mann-Whitney U - diagnostic classification performance, vs. dentists: p

											0.05 (NS); examination t p < 0.001 (AI faster)
3	de Brito et al. (2025) [3]	Caries detection + classification	YOLOv8 + EfficientNet-B0	Bitewing	500 radiographs	(C) NR (previously derived ≈92.4% removed; not explicitly reported, TN unavailable)	NR (TN not available)	YOLO 91%; EffNet-B0 89%	YOLO precision 98%; inter-examiner κ 0.88	κ only: 0.85-0.90 (overall)	GEE only (surface/age/ no AI-vs-examiner test
4	Çetin et al. (2026) [23]	Extraction indication	Gabor-HOG-SVM	Panoramic	176 (88/88)	0.81 (indication recall)	(C) 0.74 - review-calculated (TN 20/FP 7); not author-reported	Acc 77.78%; AUC 0.77	Macro F1 0.78; comparators 48.5-60.6%	NR	Post hoc vs. chance p < 0. no paired AI-v clinician test
5	Azhari et al. (2023) [17]	Caries segmentation	Ensemble U-Net (ResNet50/ResNeXt101/Inception-ResNet-v2)	Bitewing	771 (adult 554/pediatric 217)	NR	NR	NR	(C) IoU, adult: no caries 0.98/primary 0.23/moderate 0.19/advanced 0.51. Pediatric: no caries 0.97/primary 0.08/moderate 0.17/advanced 0.25. Encoder means (adult): ResNet50 0.456/ResNeXt101 0.476/Inception-ResNet-v2 0.434	NR	Advanced vs. primary caries detection: p < 0.05
6	Gülşen et al. (2026) [21]	Ectopic eruption (canine/molar)	YOLOv8x-seg	Panoramic	1,059 (1,632 canine + 245 molar labels)	0.771/0.650	NR	NR	(C) mAP@0.5 = 0.756 overall; canine 0.793/molar 0.719 (tied to YOLOv8x-seg, threshold 0.5); F1 0.778/0.722; Dice 0.768/0.757	NR (κ 0.83/0.93)	None (stated limitation)
7	Herawati et al. (2026) [24]	Malocclusion (LLM)	ChatGPT 5.2 Pro/Gemini 3 Fast/Claude 4.5 Sonnet/Copilot	Intraoral photos	(C) 50 children (participant level); 250 photographs (image level)	NR (radar only)	NR (radar only)	AUC ranges: alignment 0.41-0.85; sagittal 0.47-0.70; vertical 0.44-0.55; transverse 0.48-0.61; arch 0.41-0.68	κ: −0.15 to 0.63 by domain	NR	Descriptive o (no p-values)
8	Jaiswal et al. (2023) [19]	Tooth-type classification (deciduous vs. permanent)	27 CNNs (VGG/ResNet/Inception/MobileNet/DenseNet/NASNet/EfficientNet)	Panoramic	620	(C) 0.98 - recall (= sensitivity) for EfficientNet-B0 and EfficientNet-B3; range 0.93-0.98 across all models	NR (no confusion matrix published)	(C) Acc 0.98 (EfficientNet-B0/B3); ResNet50 0.93 (lowest); others 0.95-0.97	Precision 0.98; F1 0.98 (EfficientNet-B0/B3)	NR	None (descrip ranking only)
9	Karamüftüoğlu et al. (2025) [12]	Furcation lesion + treatment decision	YOLOv12x/RT-DETR-L/RT-DETR-X	Panoramic	387	(C) Recall - RT-DETR-X 0.483 (best); YOLOv12x and RT-DETR-L not numerically stated in text	NR	mAP@0.5 - RT-DETR-X 0.434; YOLOv12x 0.397; RT-DETR-L 0.326	(C) Precision - YOLOv12x 0.442 (highest); RT-DETR-X 0.440; per-class metrics NR (reporting gap)	NR	None
10	Naik et al. (2025) [9]	Diagnosis + treatment planning (LLM)	ChatGPT 3.5	Text vignettes	10 children	NR (case-level accuracy only)	NR	Acc 80% (8/10); pediatricians 80%, 80%, 90%, 90%	κ vs. ChatGPT: 1.00, 1.00, 0.615, 0.615	κ p-values only (0.002, 0.100, 0.035,	Fisher's exac = 0.926 (NS)

										0.035)	
11	Özüdoğru et al. (2025) [5]	Tooth segmentation + FDI numbering	nnU-Net v2	CBCT	49 volumes	(C) Tooth-numbering recall 0.84	NR	(C) Tooth-numbering Acc 0.99; per-tooth AUC 0.28-0.98 (voxel-overlap discrimination, not overall summary)	(C) Segmentation: mean Dice 0.81 (0.22-0.95); mean IoU 0.81 (0.18-0.95); 95% HD 1.93; precision 0.86	NR	None
12	Pham and Al-Hebshi (2026) [22]	Caries vs. periapical (text pipeline)	1D-CNN/LSTM/BERT (vs. SqueezeNet/GoogLeNet/AlexNet)	Panoramic → text descriptions	58 (29/29)	NR (per class in Table 3; not in extracted text)	NR (in Table 3)	(C) Acc - 1D-CNN 84%; BERT 77%; LSTM 57%; image CNNs lower (exact values in Table 3)	5-fold CV; Tables 4-5; word clouds; training curves	(C) Reported CV-derived CIs (Tables 4-5; numeric values not in extracted text); no review-calculated substitutes	None (model-model only)
13	Güçlü et al. (2026) [16]	Interproximal caries segmentation	YOLOv5x-Seg	Bitewing	2,000 radiographs; 5,704 lesions	0.78/0.75 (primary/permanent)	NR (not quantifiable - enriched dataset; TN not exhaustively annotated)	ROC-AUC 0.87/0.76	Precision 0.80/0.78; F1 0.79/0.76	NR	None (descrip primary vs. permanent)
14	van Nistelrooij et al. (2025) [20]	Tooth segmentation + FDI numbering	ToothInstanceNet	Intraoral scans	716 scans/8,321 teeth	NR (segmentation/labeling task)	NR	Detection F1 0.996; Dice 0.969; macro-F1 0.989	macro-IoU 0.909 (mixed)/0.879 (3DTeethSeg22); 3DTeethSeg22 score 0.9764; <2 s/scan	NR (CV point estimates only)	Benchmark v algorithms (b score 0.9764 0.9539); no A clinician test
15	Yelkenci et al. (2026) [15]	M1 extraction prognosis	YOLOv8n-OBB	Panoramic	200 radiographs/800 quadrants	NR	NR	mAP@0.5 0.983	AP M2 0.991/M3 0.974; F1 0.978/0.904; MCC; 12.4 ms/scan; prognostic decision matrix	NR	None (descriptive; human comparison; prognosis not validated aga clinical outco
16	Uzel et al. (2025) [18]	Supernumerary teeth detection	YOLOv8-seg (segmentation) + YOLOv8x-cls (classification)	Panoramic	2,000 screened; 140 annotated	(C) Segmentation: 0.38; Classification: 1.00	(C) Segmentation: NR (no TN denominator); Classification: 1.00	(C) Segmentation: Acc NR; Classification: Acc 100%; no AUC for either task	(C) Segmentation: precision 1.00, F1 0.55. Classification: precision 1.00, F1 1.00	NR	McNemar's te > 0.05 (classification model ≈ expe inter-rater κ = 0.92

TABLE 7: Detailed AI performance metrics and validation characteristics of included studies

AI: artificial intelligence, Acc: accuracy, AP: average precision, AUC: area under the curve, Bal. acc: balanced accuracy, BERT: Bidirectional Encoder Representations from Transformers, CBCT: cone-beam computed tomography, CI: confidence interval, CNN: convolutional neural network, CV: cross-validation, EffNet: EfficientNet, F1: F1-score, FDI: Fédération Dentaire Internationale, GEE: generalized estimating equations, Grad-CAM: Gradient-weighted Class Activation Mapping, HD: Hausdorff distance, HOG: histogram of oriented gradients, IoU: intersection over union, κ: kappa statistic, LLM: large language model, LSTM: long short-term memory, mAP: mean average precision, MCC: Matthews correlation coefficient, NASNet: Neural Architecture Search Network, NR: not reported, NS: not significant, PR-AUC: area under the precision-recall curve, ResNet: residual network, ROC-AUC: area under the receiver operating characteristic curve, RT-DETR: real-time detection transformer, Sens.: sensitivity, Spec.: specificity, SVM: support vector machine, TN: true negative, VGG: Visual Geometry Group, YOLO: You Only Look Once, 1D-CNN: one-dimensional convolutional neural network, 3D: three-dimensional, M1: first molar, M2: second molar, M3: third molar, FP: false positive, OBB: oriented bounding box

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the

work.

Acquisition, analysis, or interpretation of data: Laila Alanaz, Lama Zayed A. Alzahrani, Ahmed Fahad Alzahrani, Walaa Ali Altarouti, Fatema Hussain Eisa AlSoltan, Waad M. Alahmadi, Areej O. Yousif Abduldaiem

Drafting of the manuscript: Laila Alanaz, Lama Zayed A. Alzahrani, Ahmed Fahad Alzahrani, Fatema Hussain Eisa AlSoltan, Waad M. Alahmadi, Hayfaa Al-Shayea

Concept and design: Lama Zayed A. Alzahrani, Ahmed Fahad Alzahrani, Walaa Ali Altarouti, Fatema Hussain Eisa AlSoltan, Jenan Ehsan Ali Alali, Hayfaa Al-Shayea

Critical review of the manuscript for important intellectual content: Lama Zayed A. Alzahrani, Ahmed Fahad Alzahrani, Walaa Ali Altarouti, Jenan Ehsan Ali Alali, Areej O. Yousif Abduldaiem

Supervision: Lama Zayed A. Alzahrani, Ahmed Fahad Alzahrani

Disclosures

AI use statement: The authors used ChatGPT 5.4 as follows: language polishing, AI was used for paraphrasing, grammar correction, and language improvement. in Introduction and Discussion. The authors reviewed the AI-generated text to ensure that the original meaning was preserved. AI was used to paraphrase text in the Introduction and Discussion sections. Risk-of-bias assessments and methodological interpretations were independently conducted by the authors using the prespecified appraisal tools, without AI assistance. The authors revised or rejected AI-generated content where necessary and take full responsibility for the accuracy, integrity, and final content of the manuscript. The authors confirm that they have reviewed and verified all AI-assisted content and take full responsibility for the work. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Shahbazi S, Esmaeili S, Kavousinejad S, Younessian F, Behnaz M: Efficacy of artificial intelligence in radiographic dental age estimation of patients undergoing dental maturation: a systematic review and meta-analysis. Int Orthod. 2025, 23:101010. 10.1016/j.ortho.2025.101010
2. Alangari N, AlShenaifi N: Lightweight deep learning for automated dental caries screening from pediatric oral photographs. Diagnostics (Basel). 2026, 16:862. 10.3390/diagnostics16060862
3. de Brito DH, Dos Santos TG, Costa SG, et al.: Accuracy in diagnosing caries in young permanent molars using interproximal radiographic imaging and validation by artificial intelligence. J Clin Exp Dent. 2025, 17:648-55. 10.4317/jced.62396
4. Fadilah RP, Rikmasari R, Akbar S, Setiawan AS: Examination of new clinical dental caries in school children using real intra oral photos with artificial intelligence model YOLO-V8x. BMC Oral Health. 2025, 26:164. 10.1186/s12903-025-07486-x
5. Ozudogru S, Gulsen E, Mahyaddinova T, et al.: Artificial intelligence system for automatic tooth detection and numbering in the mixed dentition in CBCT. Eur J Paediatr Dent. 2025, 26:140-6. 10.23804/ejpd.2025.2292
6. Wang S, Liu J, Li S, He P, Zhou X, Zhao Z, Zheng L: ResNet-Transformer deep learning model-aided detection of dens evaginatus. Int J Paediatr Dent. 2025, 35:708-16. 10.1111/ipd.13282
7. Kumar M, Kumar S, Agarwal M, Yadav E, Gandi S: Predicting treatment pathways in Class II malocclusion patients using machine learning: a comparative study of four algorithms for classifying camouflage, growth modulation, and surgical decisions. Int Orthod. 2026, 24:101070. 10.1016/j.ortho.2025.101070
8. Cai J, Wang Z, Wang H, Chen Z, Yu Q, Lai Z, Xu L: Multimodal deep learning for midpalatal suture assessment in maxillary expansion. Sci Rep. 2025, 15:39723. 10.1038/s41598-025-23500-2
9. Naik S, Bhise R, Kodical SR, Chavan S, Mali SS, Shetiya N: Comparative evaluation of diagnosis and treatment plan given by pediatric dentists and generated by ChatGPT: a cross-sectional pilot study. Cureus. 2025, 17:88505. 10.7759/cureus.88505
10. Okutan AE, Sezer B: Comparative evaluation of scholargpt and ChatGPT-4 Omni in pediatric dentistry: accuracy and completeness analysis. Pediatr Dent. 2025, 47:408-48.
11. Karamüftüoğlu N, Varol EA, Bal C: Exploring artificial intelligence chatbots in pediatric fluoride education: a cross-sectional study. Sci Rep. 2025, 16:182. 10.1038/s41598-025-28857-y
12. Karamüftüoğlu N, Üçpunar BY, Birben İ, Altundağ AE, Mullaoğlu KÖ, Bal C: Artificial intelligence in pediatric dentistry: a systematic review and meta-analysis. Children (Basel). 2026, 13:152. 10.3390/children13010152
13. Garg I, Gupta S, Gupta S: Role of artificial intelligence in pediatric dentistry: an umbrella review . Int J Paediatr Dent. 2026, 36:556-66. 10.1111/ipd.70084
14. Akitomo T, Hamada M, Hamaguchi S, Kusaka S, Nomura R: Generative artificial intelligence and large language models in paediatric dentistry: a scoping review. Int Dent J. 2026, 76:109401. 10.1016/j.identj.2025.109401
15. Yelkenci A, Güven Polat G, Ciftci F, Rahebi J: Automated prognostic evaluation of first permanent molar extractions using Yolov8 with oriented bounding boxes on pediatric panoramic radiographs. Diagnostics (Basel). 2026, 16:2141. 10.3390/diagnostics16142141

16. 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, 26:202. 10.1186/s12903-025-07570-2
17. 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. Digit Health. 2023, 9:20552076231216681. 10.1177/20552076231216681
18. Uzel İ, Ghabchi B, Çoğulu D: Deep learning-based automated detection of supernumerary teeth in pediatric panoramic radiographs. PLoS One. 2025, 20:0335845. 10.1371/journal.pone.0335845
19. Jaiswal M, Sharma M, Khandnor P, et al.: Deep learning models for classification of deciduous and permanent teeth from digital panoramic images. Cureus. 2023, 15:49937. 10.7759/cureus.49937
20. van Nistelrooij N, Maia HC, Cao L, Vinayahalingam S, Loomans B, Cenci MS, Mendes FM: Automated detection and numbering of primary and permanent teeth in digital impressions of children using artificial intelligence. J Dent. 2025, 161:105976. 10.1016/j.jdent.2025.105976
21. Gülşen E, Kızılay FN, Aşar EM, et al.: Deep learning detection of ectopic canines and molars in mixed dentition. Sci Rep. 2026, 16:15667. 10.1038/s41598-026-45912-4
22. Pham TD, Al-Hebshi S: Classification of pediatric dental diseases from panoramic radiographs using natural language transformer and deep learning models. Front Artif Intell. 2026, 9:1754498. 10.3389/frai.2026.1754498
23. Çetin SG, Ertuğrul ÖF, Kavasoğlu N, Eratilla V: Artificial intelligence-based evaluation of permanent first molar extraction indications in children using panoramic radiographs. Children (Basel). 2026, 13:277. 10.3390/children13020277
24. Herawati H, Kusnoto J, Gunardi I, Wirasto A, Astoeti TE: Accuracy of orthodontic malocclusion detection using multiple AI models: a comparative study. Healthc Inform Res. 2026, 32:166-78. 10.4258/hir.2026.32.2.166
25. Rokhshad R, Banakar M, Shobeiri P, Zhang P: Artificial intelligence in early childhood caries detection and prediction: a systematic review and meta-analysis. Pediatr Dent. 2024, 46:385-94.
26. Abbott LP, Saikia A, Anthonappa RP: Artificial intelligence platforms in dental caries detection: a systematic review and meta-analysis. J Evid Based Dent Pract. 2025, 25:102077. 10.1016/j.jebdp.2024.102077
27. Alhazmi N, Alshehri A, BaHammam F, Philip M, Nadeem M, Khanagar S: Can large language models serve as reliable tools for information in dentistry? A systematic review. Int Dent J. 2025, 75:100835. 10.1016/j.identj.2025.04.015
28. Zhang R, Pan Y, Liu Y, Deng Y, Pow EH: Leveraging large language models for patient instructions in dentistry-a systematic review and meta-analysis. J Prosthodont. 2025, 10.1111/jopr.70068
29. Collins J, Hinshaw JL, Simcock E, Rosenberg MA: Radiology faculty compliance with recommended health guidelines: comparison with residents. Acad Radiol. 2009, 16:1433-42. 10.1016/j.acra.2009.06.011
30. Vasey B, Nagendran M, Campbell B, et al.: Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022, 377:070904. 10.1136/bmj-2022-070904