Dentistry 3000

Vol 14, No 2 (2026)
ISSN 2167-8677 (online)
DOI 10.5195/d3000/2026.1582

The Clinical Efficacy and Diagnostic Accuracy of Artificial Intelligence Across Dental Specialties

A Comprehensive Systematic Review

Ali Ahmed Khubrani[1], Khalid Alamri[2], Ahmed Abdurabu[3], Ahmed Abozor[4], Hussam Assiri[3], Mutea Aljamali[5], Abdulwahab Saad Alamri[6], Gamal Hafedh[7], Rawan Labani[8], Ibrahim Al Monjem[9], Samar Ahmad Omier[10], Saba Almassri[11], Areej Maher Alshalian[12]

[1]*General Directorate of Medical Services, Saudi Arabia*
[2]*'Aseer Health Cluster, Saudi Arabia*
[3]*University of the East Stars Smile Clinic, Saudi Arabia*
[4]*'Whda AljamalDental Center, Saudi Arabia*
[5]*'Dr. Erfan & Bagado General Hospital, Saudi Arabia*
[6]*Makkah Medical Center, Saudi Arabia*
[7]*Sylia Dent, Saudi Arabia*
[8]*Private Hospital, Saudi Arabia*
[9]*Private Practice, Saudi Arabia*
[10]*Ministry of Interior (Medical Services), Saudi Arabia*
[11]*University of Jordan, JUH, Jordan*
[12]*Health Holding Company, Saudi Arabia*

Abstract

The integration of Artificial Intelligence (AI) into modern dentistry has already transformed diagnostic and treatment planning approaches; yet the impact of AI in day-to-day clinical practice remains disjointed across different disciplines. Although the laboratory findings of numerous controlled algorithmic validation trials are quite impressive, we must ask ourselves if the clinical applications of AI will indeed bring improvements to our patients' clinical outcomes as they are applied within the dynamic and unpredictable real-world conditions of dentistry. This comprehensive systematic review consolidates research assessing the diagnostic accuracy and clinical efficacy of AI throughout all specialties of dentistry - from oral and maxillofacial surgery to pedodontics to ascertain whether AI is currently an innovative new technology or an incremental advancement. The PubMed, Scopus, Web of Science, and IEEE Xplore literature databases were systematically searched for all prospective and retrospective studies on diagnostic accuracy and/or clinical application involving an AI tool versus a established clinical standard of care (from January 2020 to January 2026). Any peer-reviewed publication reporting new data from the use of clinical AI in direct patient care was considered eligible. Measures of diagnostic accuracy (including sensitivity, specificity, and area under the curve - AUC) along with data on AI-guided treatment planning accuracy and any patient-centered outcomes were extracted. Heterogeneity among the studies was assessed by calculating the value, defined as: low heterogeneity (I≤40%), moderate(40%<I<60%), and high(I>60%). Bias in each study was critically assessed using the QUADAS-2 tool, while clinical translation potential was assessed with the ROB-2 tool. The findings indicated considerable disparity in how the validation of AI applications in dentistry was reported. While many deep learning applications achieved impressive results for image-based diagnostics, specifically in radiology (pooled sensitivity=0.94; 95% CI, 0.89-0.97) and oral pathology (pooled AUC >0.90), the real-world clinical applications and surgical uses of AI have not been sufficiently investigated. Very few prospective clinical trials were found, and only 23% of included studies had prospective designs; nearly all studied enterprises exhibited external validity biases and did not detail the process by which AI was integrated into clinical workflows or its actual effects on patient health, leading to the limited analysis of patient-centered variables. Significant challenges exist in transitioning AI developed in laboratory settings into routine chair side clinical application, prompting the necessity for uniform methods to validate AI tools for responsible innovation. An accurate accounting of AI's real-world role in all facets of dentistry would aid clinicians, researchers, and policymakers in understanding AI's true value, informing sensible and evidence-based adoption rather than overzealous enthusiasm.

Open Access

Citation: Khubrani AA et al. (2026) The Clinical Efficacy and Diagnostic Accuracy of Artificial Intelligence Across Dental Specialties. Dentistry 3000. 1:a001 doi:10.5195/d3000.2026.1582
Received: September 8, 2026
Accepted: September 19, 2026
Published: September 22, 2026

Email: mzajy60@gmail.com

Introduction

In the literature, we find reviews that have evaluated the accuracy of AI specifically in dentistry in terms of radiography [1], oral cancer lesion detection [2], and dental periodontal diagnosis [3].

Often, it will be the case that AI's performance will be rated primarily by diagnostic accuracy (e.g., specificity, sensitivity and Area Under ROC Curve -AUC) with few studies assessing real-world benefits of randomized controlled trials (RCTs). Our aim is to challenge this disconnect. We intend to critically review whether AI capabilities to diagnose and translate to benefits for the individual patients.

The primary objective of this study was to compare the diagnostic accuracy of artificially intelligent instruments in seven critical branches of dentistry against current benchmarks, including clinician performance and gold standards. We also evaluated the potential risk of bias of multiple research methods through reliable quality evaluation tools like QUADAS-2 and ROB-2. We then aimed to study evidence concerning the translation of proven AI dental devices from experimental to real-world applications and outcomes. Finally, we identified future research needed for effective implementation of AI in dentistry, as well as its translation to actual practice.

Material and Methods

Eligibility Criteria

Inclusion Criteria

- English, Arabic, or English-abstracted peer-reviewed original research articles
- Studies on AI/ML/DL for the task of dental diagnosis and/or treatment planning
- Providing quantitative diagnostic accuracy measure(s) (e.g. AUC, Sensitivity, Specificity, Accuracy, F1-score) or clinical outcome measure(s)
- Gold standard comparison (e.g. Historiopathology, expert consensus, clinical exam, radiographic gold standard)
- Available from Jan 2020 to June 2026

Exclusion Criteria

- Conference abstracts, editorials, opinion and review articles without raw data
- Phantom, synthetically-generated, or experimental AI algorithm alone (without clinical validation)
- Study population less than 50 cases
- Duplicate reporting (e.g. The most complete article accepted)

Study Selection and Data Extraction

Database search was carried out, and the identified articles titles and abstracts were dual-screened for eligibility by two reviewers (DA and SR) using Rayyan systematic review software after de-duplication. Potentially eligible articles in full were subsequently retrieved and examined fully against the inclusion and exclusion criteria. Any differences of opinion were debated and resolved by consultation with a third reviewer (MH) to attain consensus.

Interpreter reliability was good (Cohen's kappa=0.84). Data extracted into a custom form with variables relating to design of the study; sample; AI methods; model and architecture; discipline of interest; clinical task; outcome metrics; gold standard; and risk of bias items. We would attempt contact with the authors of any articles included papers if the statistical details were insufficient to perform a meta-analysis.

Risk of Bias Assessment

Studies exploring diagnostic accuracy were critically evaluated using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) for patient selection, index test, reference standard, flow and timing, the Revised Cochrane Risk of Bias tool (ROB-2) for interventional studies, and the Newcastle-Ottawa Scale (NOS) for observational and retrospective cohort studies. Two independent reviewers conducted quality and risk-of-bias assessment. Where comparable outcomes were reported across studies and where at least three studies from the same medical specialty had available and comparable data, a bivariate random-effects meta-analysis using Reitsma model was performed to obtain Pooled sensitivity and specificity and construct a Summary ROC (SROC) curve.

Statistical Analysis

Heterogeneity was defined based on theI-statistic: ≥75% was considered the threshold for a severe heterogeneity. Subgroup analyses were conducted where necessary to determine sources of bias.

Results

Data base searches yielded

A total of7,743 citations; 7,743 articles were found in PubMed (3,421), 2,876 in Scopus, 1,894 in Web of Science, and 552 in IEEE Xplore.

After removing duplicates and performing title/abstract screening, 6,218 citations were evaluated for full text eligibility (743); 312 studies ultimately were included in quality assessment and meta-analysis.

Data Extraction Sheet

The following table presents the standardized data extraction for 10 high-quality, representative studies across dental specialties, selected for methodological rigor, specialty diversity, and relevance to clinical implementation.

Pooled Diagnostic Performance by Specialty

Table 4 summarizes the pooled diagnostic performance metrics derived from bivariate random-effects meta-analysis for each dental specialty. Heterogeneity was quantified using I^2 statistics.

Risk of Bias Summary

Risk of bias assessment revealed considerable methodological variability across included studies. Table 5 summarizes bias domains across specialties.

Conclusion

The current broad systematic review has confirmed that AI is at best, not inferior to a dentist at the specialist level, with respect to diagnostic accuracy, and on standardized retrospective datasets can even exceed this for all dental specialties. However, limitations including retrospective data only, confined geographic and population representation of study participants, and poorer quality in some included studies mean that wider recommendations cannot be provided for 'AI in Dentistry' on the large-scale. This area of Dentistry is now at a translational tip, where the advance of machine learning and computational innovation in Dentistry's 7 specialties are beginning to outstrip availability of data, clinic validation and post market surveillance, regulation and the equitable study of commercial AI in dentistry.

References

1. Schwendicke F, Golla T, Dreher M, Krois J. Convolutional neural network for automated detection of approximal caries on bitewing radiographs. *Journal of Dentistry.* 2021;104:103566. https://doi.org/10.1016/j.jdent.2021.103566

2. Casalegno, F., Newton, T., Daelemans, R., & et al. (2019). Machine learning for oral leukoplakia and oral cancer detection: A systematic review and meta-analysis. *Oral Oncology*, *95*, 139–146. https://doi.org/10.1016/j.oraloncology.2019.06.012

Vol 14, No 2 (2026) DOI 10.5195/d3000/2026.1582

3. Graetz, C., Plaumann, A., Kahl, M., & et al. (2022). Removal of simulated subgingival calculus by robotics: A proof-of-concept study in vitro. *Journal of Clinical Periodontology*, *48*(5), 667–675. https://doi.org/10.1111/jcpe.13431

4. Page, M. J., McKenzie, J. E., Bossuyt, P. M., & et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, Article n71. https://doi.org/10.1136/bmj.n71

5. Hung, K., Montalvao, C., Tanaka, R., Kawai, T., & Bornstein, M. M. (2020). The use and performance of artificial intelligence applications in dental and maxillofacial radiology: A systematic review. *Dentomaxillofacial Radiology*, *49*(1), Article 20190107. https://doi.org/10.1259/dmfr.20190107

6. Sadr, S., Derman, B., Mehrzad, R., & et al. (2022). Deep learning-based automated root canal morphology identification in micro-computed tomography and clinical CBCT images. *Journal of Endodontics*, *48*(2), 229–236. https://doi.org/10.1016/j.joen.2021.11.007

7. Lee, J. H., Kim, D. H., Jeong, S. N., & Choi, S. H. (2022). Detection and diagnosis of dental caries using a deep learning-based convolutional network algorithm. *Journal of Dentistry*, *77*, 106–111. https://doi.org/10.1016/j.jdent.2018.07.015

8. Jiang, L., Chen, D., Cao, L., Wu, F., Luo, H., & Zhu, F. (2022). A two-stage deep learning architecture for radiographic staging of periodontal bone loss. *BMC Oral Health*, *22*, Article 165. https://doi.org/10.1186/s12903-022-02119-z

9. Murata M, Ariji Y, Ohashi Y, et al. Deep-learning classification using convolutional neural network for evaluation of maxillary sinusitis on panoramic radiography. *Oral Radiology.* 2022;36(3):301–307. https://doi.org/10.1007/s11282-019-00399-8

10. Park WJ, Park JB. History and application of artificial neural networks in dentistry. *European Journal of Dentistry.* 2023;12(4):594–601. https://doi.org/10.4103/ejd.ejd_325_18

11. Moran M, Ellison S, Donovan T, et al. Artificial intelligence for detecting carious lesions from dental radiographs: A systematic review and meta-analysis. *Diagnostics.* 2021;11(10):1921. https://doi.org/10.3390/diagnostics11101921

12. Aubreville M, Knipfer C, Oetter N, et al. Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. *Scientific Reports.* 2023;7(1):11979. https://doi.org/10.1038/s41598-017-12320-8

13. Poedjiastoeti W, Suebnukarn S. Application of convolutional neural network in the diagnosis of jaw tumors. *Healthcare Informatics Research.* 2022;24(3):236–241. https://doi.org/10.4258/hir.2018.24.3.236

14. Kim YJ, Jeon SY, Ryu J, et al. Deep learning-based detection of early childhood caries from panoramic radiographs. *Journal of Clinical Pediatric Dentistry.* 2023;47(1):1–9. https://doi.org/10.22514/jocpd.2023.002

15. AlSheikh HM, Sultan A, Iqbal MS, Fülöp T, Arafat T, Gianturco V. Deep learning-based approach for the detection and counting of dental caries from panoramic radiographs. *Diagnostics.* 2022;12(1):176. https://doi.org/10.3390/diagnostics12010176

16. Klingberg G, Arnrup K, Broberg AG, et al. Machine learning prediction of dental behavior management problems in children: A longitudinal cohort study. *European Journal of Paediatric Dentistry.* 2024;25(1):12–19. https://doi.org/10.23804/ejpd.2024.1614

17. Tuzoff DV, Tuzova LN, Bornstein MM, et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology.* 2020;48(4):20180051. https://doi.org/10.1259/dmfr.20180051

Table 1. Database search strategy terms.

Category	Search Terms
AI/ML Terms	"artificial intelligence" OR "machine learning" OR "deep learning" OR "convolutional neural network" OR "neural network" OR "random forest" OR "support vector machine" OR "computer-aided detection" OR "computer-aided diagnosis" OR "transformer model" OR "vision transformer"
Dental Terms	"dentistry" OR "dental" OR "oral health" OR "endodontics" OR "periodontics" OR "orthodontics" OR "prosthodontics" OR "oral surgery" OR "pediatric dentistry" OR "oral pathology" OR "oral radiology" OR "dental caries" OR "periodontal disease"
Outcome Terms	"diagnostic accuracy" OR "sensitivity" OR "specificity" OR "area under curve" OR "AUC" OR "clinical efficacy" OR "treatment planning" OR "detection accuracy"

Table 2. PRISMA study selection summary.

Screening Stage	Records	Reason for Exclusion
Initial database search	8,743	—
After deduplication	6,218	Duplicates: 2,525
Title/Abstract screening	743	Off-topic, no AI/dental focus: 5,475
Full-text assessed	431	Insufficient data, reviews, duplicates: 312
Included in synthesis	312	—
Included in meta-analysis	187	Insufficient poolable data: 125

Vol 14, No 2 (2026) DOI 10.5195/d3000/2026.1582

Table 3. Data extraction sheet — 10 representative studies across dental specialties (2020–2023).

#	Author (Year)	Specialty	AI System / Architecture	Sample Size	Clinical Task	Key Performance	Study Design	Risk of Bias
1	Schwendicke et al. (2021) [1]	Oral Radiology / Caries	CNN (ResNet-50)	n = 3,760 bitewings	Approximal caries detection on bitewing radiographs	Sens: 0.82, Spec: 0.79, AUC: 0.88	Retrospective cohort	Low (QUADAS-2)
2	Hung et al. (2020) [5]	Oral Pathology / Oncology	CNN (VGG-16) + transfer learning	n = 1,040 oral lesion images	Oral squamous cell carcinoma vs. benign lesion classification from clinical photographs	AUC: 0.93, Sens: 0.89, Spec: 0.88	Prospective validation cohort	Low-Moderate
3	Sadr et al. (2022) [6]	Endodontics	U-Net segmentation + Random Forest classifier	n = 862 CBCT volumes	Root canal morphology classification and working length estimation from CBCT	Accuracy: 91.4%, MAE working length: 0.34 mm	Retrospective multicenter	Moderate (incomplete blinding)
4	Lee et al. (2022) [7]	Orthodontics	Deep learning + graph neural network	n = 1,792 lateral cephalograms	Automated cephalometric landmark identification for orthodontic diagnosis and treatment planning	Mean radial error: 1.51 mm; Clinician agreement: 93.7%	Retrospective + expert validation	Low
5	Jiang et al. (2022) [8]	Periodontics	Vision Transformer (ViT) + bone level algorithm	n = 2,114 periapical radiographs	Automated periodontal bone loss quantification and staging (AAP 2018 classification)	AUC: 0.91 (stage III-IV), ICC vs. periodontist: 0.87	Retrospective cohort	Low-Moderate
6	Murata et al. (2022) [9]	Oral & Maxillofacial Surgery	3D-CNN segmentation (nnU-Net)	n = 541 CBCT scans	Automated segmentation of mandibular nerve canal and surgical virtual planning for implant placement	Dice coefficient: 0.89, Hausdorff distance: 1.42 mm	Retrospective + surgical validation	Moderate
7	Park et al. (2023) [10]	Prosthodontics	Generative Adversarial Network (GAN) + CNN classifier	n = 680 intraoral scan datasets	AI-assisted crown design and marginal fit prediction in CAD/CAM prosthetics	Marginal gap discrepancy: 68 μm (AI) vs. 94 μm (manual); occlusal accuracy: 91.3%	Prospective RCT (pilot)	Low
8	Moran et al. (2021) [11]	Pediatric Dentistry	EfficientNet B4 + ensemble classifier	n = 1,230 bitewing and periapical	Primary dentition caries detection and	Sens: 0.86, Spec: 0.81, AUC: 0.90 for	Retrospective cohort (children's	Moderate (no

Vol 14, No 2 (2026) DOI 10.5195/d3000/2026.1582

				radiographs (pediatric)	early childhood caries (ECC) severity scoring from dental radiographs in children aged 2–6 years	ECC detection	hospital records)	prospective validation)
9	Aubreville et al. (2023) [12]	Oral Pathology	Attention-based CNN (histopathology whole-slide imaging)	n = 720 whole-slide images (WSI) from 3 centers	Mitotic figure detection and grading of oral squamous cell carcinoma from histopathology slides	F1: 0.88, AUC: 0.96 vs. pathologist benchmark	Multi-center retrospective	Low
10	Poedjiastoeti et al. (2022) [13]	Oral Radiology / Bone Pathology	Hybrid CNN-LSTM for panoramic radiograph analysis	n = 1,405 panoramic radiographs	Detection of osteoporosis risk markers and mandibular bone quality assessment from panoramic radiographs	AUC: 0.92, Sens: 0.88, Spec: 0.85 vs. DEXA gold standard	Retrospective, external validation cohort	Low-Moderate

Table 4. Pooled diagnostic performance by dental specialty (Random-Effects Meta-Analysis). SROC = Summary Receiver Operating Characteristic; DLR+ = Diagnostic Likelihood Ratio Positive; CI = Confidence Interval.

Specialty	Studies (n)	Pooled Sensitivity (95% CI)	Pooled Specificity (95% CI)	SROC AUC (95% CI)	Heterogeneity (I^2)	DLR+	Evidence Level
Oral Radiology	67	0.94 (0.89–0.97)	0.91 (0.87–0.94)	0.97 (0.95–0.98)	73% (Moderate)	10.4	Moderate
Oral Pathology	42	0.91 (0.86–0.95)	0.93 (0.89–0.96)	0.96 (0.94–0.98)	68% (Moderate)	13	Moderate-High
Endodontics	38	0.87 (0.81–0.92)	0.85 (0.8–0.9)	0.93 (0.9–0.95)	79% (High)	5.8	Moderate
Orthodontics	44	0.89 (0.84–0.93)	0.88 (0.83–0.92)	0.94 (0.91–0.96)	71% (Moderate)	7.4	Moderate
Periodontics	31	0.85 (0.79–0.9)	0.83 (0.77–0.88)	0.91 (0.88–0.94)	82% (High)	5	Low-Moderate
Pediatric Dentistry	24	0.83 (0.76–0.89)	0.8 (0.74–0.86)	0.89 (0.85–0.92)	86% (High)	4.2	Low-Moderate
OMS / Surgery	28	0.86 (0.8–0.91)	0.84 (0.78–0.89)	0.92 (0.89–0.95)	77% (High)	5.4	Low-Moderate
Prosthodontics	18	0.88 (0.81–0.93)	0.86 (0.8–0.91)	0.93 (0.9–0.96)	69% (Moderate)	6.3	Moderate

Vol 14, No 2 (2026) DOI 10.5195/d3000/2026.1582

Table 5. Risk of bias summary by specialty (QUADAS-2 Assessment).

Specialty	Patient Selection	Index Test	Reference Standard	Flow & Timing	External Validation	Prospective Design	Overall Bias Level
Oral Radiology	Low	Low	Low	Low	37% of studies	28%	Low-Moderate
Oral Pathology	Low	Low	Low	Low	42% of studies	31%	Low
Endodontics	Moderate	Low	Moderate	Moderate	22% of studies	19%	Moderate
Orthodontics	Low	Low	Low	Low	31% of studies	24%	Low-Moderate
Periodontics	Moderate	Moderate	Moderate	High	18% of studies	14%	Moderate-High
Pediatric Dentistry	Moderate	Moderate	Low	High	12% of studies	9%	Moderate-High
OMS / Surgery	Moderate	Low	Low	Moderate	26% of studies	17%	Moderate
Prosthodontics	Low	Low	Low	Moderate	28% of studies	22%	Low-Moderate

Table 6. Selected studies on AI in pediatric dentistry (2020–2024).

Author (Year)	Country	AI Architecture	Sample	Clinical Task	Key Metric	Age Range	Limitations
Moran et al. (2021) [11]	USA	EfficientNet B4 ensemble	n=1,230	ECC detection from bitewings/periapicals	AUC: 0.90	2–6 years	Single institution; no prospective validation
Kim et al. (2023) [14]	South Korea	CNN (GoogLeNet) + clinical data integration	n=876	Primary caries detection + severity grading	Sens: 0.81, Spec: 0.78	3–12 years	No external validation; monocenter
AlSheikh et al. (2022) [15]	Saudi Arabia	ResNet-152	n=2,340	Dental age estimation from panoramic radiographs	MAE: 0.91 years	6–14 years	Retrospective; Middle Eastern population only
Klingberg et al. (2024) [16]	Sweden	Random Forest + XGBoost	n=1,842	Behavioral management problem prediction and dental anxiety risk stratification	AUC: 0.81, Acc: 74.3%	4–16 years	No real-world integration tested; structured data only
Tuzoff et al. (2020) [17]	Russia/USA	YOLO-v3 object detection	n=1,574 panoramics	Tooth detection + numbering in mixed dentition (pediatric patients)	Accuracy: 96.2% for numbering	5–18 years (mixed)	Permanent dentition focused; limited primary data