Journal of Dentistry 147 (2024) 105104

Contents lists available at ScienceDirect

Journal of Dentistry

journal homepage: www.elsevier.com/locate/jdent

Review article

Artificial intelligence for detecting periapical radiolucencies: A systematic review and meta-analysis

Utku Pul [a], Falk Schwendicke [b,*]

[a] University for Digital Technologies in Medicine and Dentistry, Wiltz, Luxembourg
[b] Conservative Dentistry and Periodontology, LMU Klinikum, Goethestr. 70, Munich 80336, Germany

ARTICLE INFO

Keywords:
AI
Accuracy
Deep learning
Endodontology
Image analysis
Radiographs

ABSTRACT

Objectives: Dentists' diagnostic accuracy in detecting periapical radiolucency varies considerably. This systematic review and meta-analysis aimed to investigate the accuracy of artificial intelligence (AI) for detecting periapical radiolucency.
Data: Studies reporting diagnostic accuracy and utilizing AI for periapical radiolucency detection, published until November 2023, were eligible for inclusion. Meta-analysis was conducted using the online MetaDTA Tool to calculate pooled sensitivity and specificity. Risk of bias was evaluated using QUADAS-2.
Sources: A comprehensive search was conducted in PubMed/MEDLINE, ScienceDirect, and Institute of Electrical and Electronics Engineers (IEEE) Xplore databases. Studies reporting diagnostic accuracy and utilizing AI tools for periapical radiolucency detection, published until November 2023, were eligible for inclusion.
Study selection: We identified 210 articles, of which 24 met the criteria for inclusion in the review. All but one study used one type of convolutional neural network. The body of evidence comes with an overall unclear to high risk of bias and several applicability concerns. Four of the twenty-four studies were included in a meta-analysis. AI showed a pooled sensitivity and specificity of 0.94 (95 % CI = 0.90–0.96) and 0.96 (95 % CI = 0.91–0.98), respectively.
Conclusions: AI demonstrated high specificity and sensitivity for detecting periapical radiolucencies. However, the current landscape suggests a need for diverse study designs beyond traditional diagnostic accuracy studies. Prospective real-life randomized controlled trials using heterogeneous data are needed to demonstrate the true value of AI.
Clinical significance: Artificial intelligence tools seem to have the potential to support detecting periapical radiolucencies on imagery. Notably, nearly all studies did not test fully fledged software systems but measured the mere accuracy of AI models in diagnostic accuracy studies. The true value of currently available AI-based software for lesion detection on both 2D and 3D radiographs remains uncertain.

1. Introduction

Endodontology is a dental specialty that deals with problems related to the root canal system. Non-surgical root canal treatment is commonly used to treat diseases related to the pulp and tissues surrounding the root of teeth. Accurately diagnosing the specific condition affecting the pulp and periapical tissues is crucial for successful treatment. Failure to do so can lead to pain and negatively impact the overall treatment plan [1].

A thorough endodontic examination includes a dental and medical history, clinical evaluation, and radiologic assessment. The latter allows the detection of periapical radiolucencies, indicating an inflammatory response to a bacterial load due to infected necrotized pulp tissue, an unsatisfactory root canal treatment, a cancerous lesion, a cystic lesion, or a manifestation of a systemic disease [2]. The prevalence of periapical radiolucencies has been determined at 5 % and 6.40 % in two meta-analyses conducted on 300,861 and 679,414 teeth, respectively [3,4].

Periapical radiolucencies can be assessed using two-dimensional (2D) and three-dimensional (3D) imagery. While 3D data like cone beam computed tomography (CBCT) provide more robust and accurate assessments [5], guidelines discourage routine 3D assessments for periapical diagnostic purposes [6] due to the significantly higher radiation

* Corresponding author.
E-mail address: Falk.schwendicke@med.uni-muenchen.de (F. Schwendicke).

https://doi.org/10.1016/j.jdent.2024.105104
Received 20 April 2024; Received in revised form 24 May 2024; Accepted 27 May 2024
Available online 6 June 2024

dose required and the time, expertise, and costs associated with obtaining and interpreting 3D data [7].

Artificial Intelligence (AI) refers to the ability of a computer or a computer-controlled system to exhibit behaviors similar to humans, perform tasks such as logical reasoning, motion, speech, and sound perception, and carry out various activities in a manner resembling intelligent beings. An increasing wealth of studies focus on AI to detect periapical radiolucencies in 2D and 3D imagery [8]. At the same time, the consistency of findings and the robustness of the overall body of evidence is not well established. The present review systematically identified and appraised studies on AI for detecting periapical radiolucencies and quantitively synthesized the obtained accuracy data using meta-analysis.

2. Methods

Reporting of this study follows the PRISMA checklist [9].

2.1. Studies and search

Our search and inclusion and exclusion criteria were determined using the PICOS scheme. The Population was patients receiving dental radiographs, the Intervention was AI used to detect periapical radiolucencies, and the Control was a reference test established by clinicians, histological analysis, or CBCT. Our Outcome looked at accuracy or its derivates, like the area under the curve. Only diagnostic accuracy studies were included.

Two researchers conducted the systematic search independently and duplicative. Their agreement was assessed using Cohen's kappa. In case of a disagreement, a third reviewer was consulted as a tiebreaker. The following electronic databases were searched: Medline via PubMed, Institute of Electrical and Electronics Engineers (IEEE) Xplore, and ScienceDirect. The search strategy can be seen in the Appendix.

Accompanying the electronic search, a manual search was conducted on the following resources: reference lists of the included papers and identified reviews (cross-referencing), three journals (01/2000 to 11/2023), namely the Journal of Endodontics, the International Endodontic Journal, and the Journal of Dental Research.

2.2. Data extraction

Details on the study design and the results of the included studies were extracted into a spreadsheet, encompassing information about the authors, publication year, study aims, datasets used for training and validation, information on labeling, information on preprocessing and augmentation, the specific AI algorithm employed, the outcome in any form of accuracy, and comparison with dentists or any other standard of care, if available.

2.3. Risk of bias

The risk of bias was determined by two independent examiners using Quadas-2 [10], a tool for the Quality Assessment of Diagnostic Accuracy Studies. Studies were evaluated for patient selection, index test, reference test, flow, and timing. Applicability concerns for patient selection, index, and reference tests were assessed. Disagreements were resolved by discussion.

2.4. Meta-analysis

Meta-analysis was conducted using MetaDTA (v2.0, Shinyapps, RStudio, Boston, USA) [11,12], an online tool for meta-analysis of diagnostic accuracy studies. Studies were included if the number of true positive, true negative, false positive, and false negative diagnostic cases were provided. Pooled sensitivity and specificity of the included studies with 95 % confidence intervals were determined using random-effects modeling, assuming individual study estimates to vary but to come from a joint underlying distribution with an unstructured between-study covariance matrix [13,14]. We also generated hierarchical summary receiver operating characteristic (HSROC) curves, including summary points, confidence, and predictive regions, estimated as described elsewhere [15], via generalized linear mixed effect modeling using the glmer function in the R-package lme4 [13,14,16].

3. Results

3.1. Search and included studies

The electronic searches yielded 210 records (Fig. 1). No additional articles were included through the manual search. After removing duplicate and irrelevant titles, 43 records remained. Of these, five articles could not be obtained; fourteen were excluded based on the full-text review, resulting in 24 articles being included. Cohen's kappa between reviewers was 0.84, showing almost perfect agreement. The list of excluded studies and the reason for their exclusion is in Appendix Table S3.

Table 1 below lists all studies with study features. Table S2 in the Appendix shows the outcomes of the included studies.

Studies differed in their aims, imaging method, dataset size, reference standard, model structure, and performance measurements. Most studies used a type of convolutional neural network (CNN) [17,18,21,22,24–38,40]. Ten studies performed segmentation [17,18,25,26,29,30,32,35,37,40], 11 studies classification [19–21,23,24,29,31,34,36,38,39] and six object detection [21,22,27,28,33,38]. Two combined object detection and classification [29,41] and one classification and segmentation [35].

Twelve studies used periapical radiographs as test data [17,18,20–22,26,27,31–34,36], seven studies panoramic radiographs [23,24,28,30,37,41] and six CBCT [19,25,29,35,39,40]. Twenty-one studies [17–25,28–37,39,40] labeled the reference set the same way they labeled the test data. Fifteen studies [17,18,20–24,28,30–34,36,37] relied on 2D radiological diagnosis to create a reference test. Six out of the twenty-one used a 3D radiological method, CBCT [19,25,29,35,39,40]. One study [27] chose to label the reference set in CBCT, while another imaging method was used on the test set. Only one study [38] supported radiological diagnoses with histopathology and clinical tests to establish the reference test. One study [26] did not specify how labeling is done.

Five studies [17,19,21,27,29] employed only one dentist to establish the reference test, and 13 studies [18,23,24,30–38,40] two or more dentists. In six studies [20,22,25,26,28,39], who or how many experts established the reference tests remained unclear.

Reported performance measurements were highly heterogeneous. Sensitivity and specificity were most common, while only four studies [25,31,33,35] reported true positive, true negative, false positive, and false negative diagnostic cases.

3.2. Bias analysis of included studies

The risk of bias is displayed in Table 2 and Fig. 2.

If a study showed a low risk of bias in 2 or more areas without any high risk of bias, it was graded as low risk. If a study showed a high risk of bias in any area, it was graded as having a generally high risk of bias. Studies other than these two groups were graded as unclear risk of bias. Out of 24 articles, seven showed a low risk of bias [17,24,25,27–29,38]; eight had an unclear risk of bias [21–23,30–32,37], and eight had a high risk of bias [18–20,26,34–36,40].

For the patient selection domain, studies had generally unclear to high risk of bias, mainly because the methodology for selecting positive and negative image cases was often not adequately described. Without clear criteria or a detailed process for how these cases were chosen, it was difficult to ascertain if the samples truly represented the broader

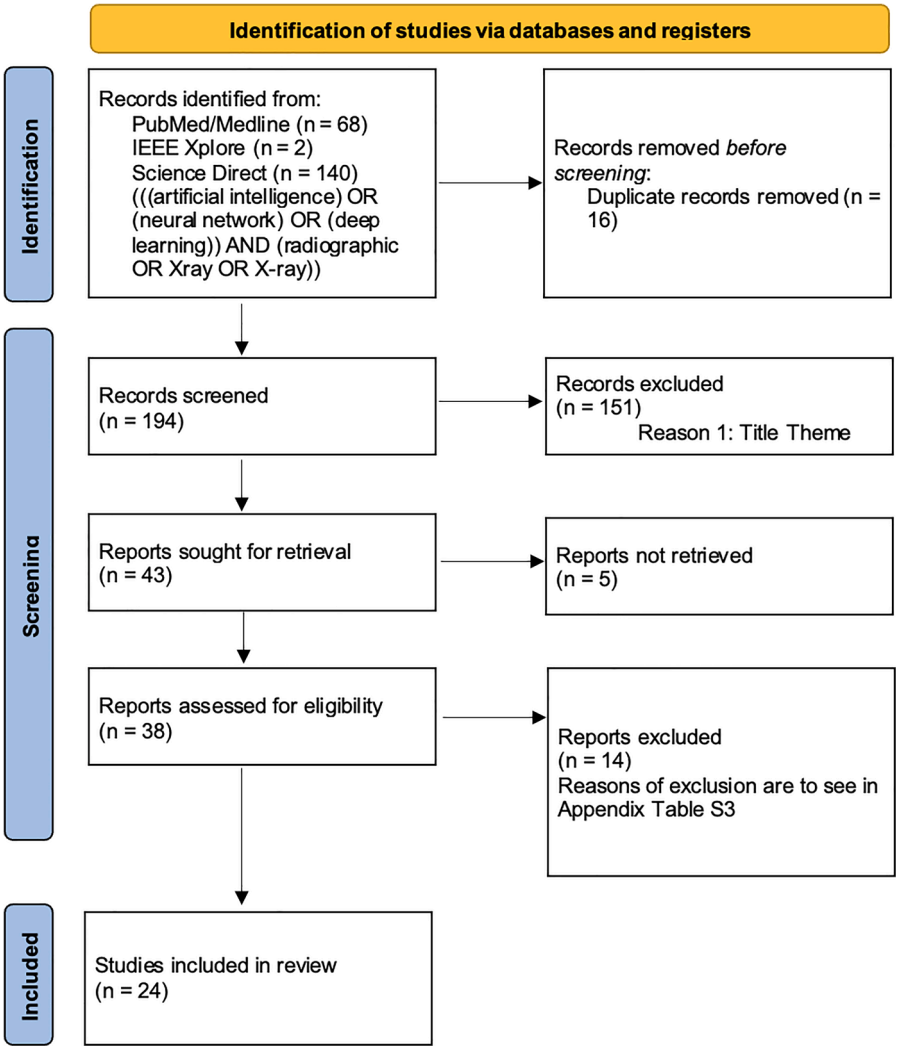

Fig. 1. Selection process.

population. This lack of detail raises questions about the generalizability of the study results.

In the index test domain, the risk of bias was mainly evaluated as low to unclear. The primary issue here was the examiners' lack of blinding for the reference test when establishing the index tests, which could obviously lead to biased results.

The reference standard domain frequently showed an unclear risk of bias, mainly as it was not substantiated whether the reference tests were sufficiently reliable, valid, and precise to detect the condition of interest.

The flow and timing domain also typically had an unclear risk of bias because many studies did not report the chronological order in which the index and reference tests were performed. Ideally, these tests should be performed independently (see above) but in timely juxtaposition to avoid changes in the target condition (e.g., a CBCT to establish the reference test should not have been taken months after the panoramic used for the index test). Notably, this may not always be relevant (e.g., when both tests are established on the same image).

The patient selection domain generally posed an unclear risk of applicability problems when considering the studies' relevance to real-world settings. Due to uncertainty about their representativeness, it is uncertain whether the selected patient populations in specific studies can be replicated in future studies.

The index test domain posed a low risk of applicability issues. The architectural and modeling parameters used in the studies were largely replicable and well-described, meaning that other researchers or practitioners could use similar methods and expect comparable results.

The reference standard domain had an unclear risk of applicability problems. In nearly all the studies, the reference tests were based on experts' opinions rather than objective measures. Since experts' opinions can vary widely, the validity and reliability of the reference standard are doubtful. Moreover, the strategy to uniform the varying opinions was not always clear and is generally not scientifically substantiated (instead, different kinds of "common practices" have been established, while for none a robust scientific justification is available).

3.3. *Meta-analysis*

Figs. 3 and 4 show the specificity and sensitivity of each study in a forest plot, respectively; the pooled specificity (95 % CI) was 0.96 (0.91, 0.98); the pooled sensitivity was 0.94 (0.90, 0.96).

Fig. 5 displays the HSROC and the computed 5 % confidence and 95 % predictive regions. HSROC provides a comprehensive method to evaluate diagnostic test accuracy across multiple studies. It considers the variability between studies and accounts for the potential correlation between sensitivity and specificity, offering a more robust summary of diagnostic performance. This is particularly useful when there is heterogeneity in study results, as the HSROC allows for the inclusion of random effects to provide a summary estimate.

Table 1
Features of all included studies.

Author	Year	Description	Aim of Study	Used Imaging Method	Data Size (Training/ Test)	Labeling on	Labeling Done by	Data Augmentation	Model Structure	Performance Measurements and Outcomes	Comparison with Dentists other than Labelers
Ari, T. et al. [17].	2022	Automatic Feature Segmentation in Dental Periapical Radiographs	Segmentation	Periapical Radiographs	292 (266/26)	Periapical Radiographs	1 Oral Radiologist with 12 Years of Experience	No	CNN	Sensitivity: 0.92 Precision: 0.85 F1-Score:: 0.86	No
Bayrakdar, I. S. et al. [18].	2022	U-Net for Apical Lesion Segmentation on Panoramic Radiographs	Segmentation	Periapical Radiographs	470 (380/47)	Periapical Radiographs	3 Oral Radiologists with At Least 3 Years of Experience	Yes	CNN	Sensitivity:0.92 Precision: 0.84 F1-Score: 0.88	No
Calazans, M. A.A. et al. [19].	2022	Automatic Classification for Periapical Lesions in CBCT	Classification	CBCT	885 (na)	CBCT	1 Oral Radiologist with 10 Years of Experience	Yes	Siamese Concatenated Network a CNN	Accuracy: 0.70 Sensitivity: 0.64 Precision: 0.76 Specificity: 0.76 F1-Score: 0.70	No
Caputo, B. et al. [20].	2000	Analysis of Periapical Lesions using Statistical Textural Features	Classification	Periapical Radiographs	108 (30/78) 108 (50/58)	Periapical Radiographs	na	na	A three-layers, Feedforward, Backpropagating Neural Network a CNN	TP Rate: 0.69 FP Rate: 0.09	No
Chen, H. et al. [21].	2021	Dental Disease Detection on Periapical Radiographs Based on Deep Convolutional Neural Networks	Object Detection and Classification	Periapical Radiographs	2900 (na)	Periapical Radiographs	1 Dentist with More Than 5 Years of Clinical Experience	na	CNN	IoU: 0.69 Precision: 0.52 Sensitivity: 0.52	No
Chuo, Y. et al. [22].	2022	A High-Accuracy Detection System: Based on Transfer Learning for Apical Lesions on Periapical Radiograph	Object Detection	Periapical Radiographs	760 (662/98)	Periapical Radiographs	na	Yes	4 different CNNs	Accuracy AlexNet: 0.96 ResNet101: 0.95 ResNet50: 0.94 GoogleNet: 0.88	No
Ekert, T. et al. [23].	2019	Deep Learning for Radiographic Detection of Apical Lesions	Classification	Panoramic Radiographs	2579 (2238/341)	Panoramic Radiographs	A Majority Vote of 6 Independent, Experienced Dentists	Yes	na	AUC: 0.85 Sensitivity: 0.65 Specificity: 0.87 PPV: 0.49 NPV: 0.93	No
Endres, M.G. et al. [24].	2020	Development of a Deep Learning Algorithm for Periapical Disease Detection in Dental Radiographs	Classification	Panoramic Radiographs	3099 (2902/102)	Panoramic Radiographs	Four OMF Surgeons with Experiences Ranging from 5 to 20 Years	na	CNN	PPV: 0.67 F1-Score: 0.58 AP: 0.60	Yes
Ezhov, M. et al. [25]	2020	Clinically Applicable Artificial Intelligence System for Diagnosis of CBCT	Segmentation	CBCT	2800 (na)	CBCT	Dental and OMF Radiologists	No	CNN	Sensitivity: 1.00 Specificity: 0.84	Yes
Fatima, A. et al. [26].	2023	Deep Learning-Based Multiclass Instance Segmentation for Dental Lesion Detection	Segmentation	Periapical Radiographs	534 (453/81)	na	Experienced Radiologists and Dentists	Yes	CNN	mAP: 0.85 Sensitivity: 0.89 Precision: 0.86 F1-Score: 0.89 mIoU: 0.71	No
Hamdan, M. H. et al. [27].	2022	Deep Learning for Detecting Apical Radiolucencies on Periapical Radiographs	Object Detection	Periapical Radiographs	184 (na/130)	CBCT	1 Oral-maxillofacial Radiologist with 10 Years of Experience	Yes	CNN	AFROC-AUC: 0.89 Specificity: 0.73 Sensitivity: 0.93	Yes

(*continued on next page*)

Table 1 (*continued*)

Author	Year	Description	Aim of Study	Used Imaging Method	Data Size (Training/ Test)	Labeling on	Labeling Done by	Data Augmentation	Model Structure	Performance Measurements and Outcomes	Comparison with Dentists other than Labelers
Kim, C. et al. [28].	2022	Tooth-Related Disease Detection System Based on Panoramic Images and Optimization Through Automation	Object Detection	Panoramic Radiographs	10,000 (na)	Panoramic Radiographs	Radiologists with 20-Year Experience	na	R-CNN ResNet Inception	Precision: 0.82 Sensitivity: 0.95 Specificity: 0.89	No
Kirnbauer, B. et al. [29].	2022	Automatic Detection of Periapical Osteolytic Lesions on CBCT Using Convolutional Neuronal Networks	Classification and Segmentation	CBCT	144 (2128 ROI)	CBCT	1 Oral Surgeon	No	CNN	Sensitivity: 0.97 Specificity: 0.88 TP rate: 0.97 FN rate: 0.03 TN rate: 0.88 FP rate: 0.12	No
Krois, J. et al. [30].	2021	Generalizability of Deep Learning Models for Dental Image Analysis	Segmentation	Panoramic Radiographs	1300 (1000/300) 650 (500/150) 650 (500/150)	Panoramic Radiographs	4 Specialists 1 Expert for Validation	Yes	CNN	F1-Score: 0.54 Sensitivity: 0.48 Precision: 0.64 Specificity: 1.00	Yes
Li, C.W. et al. [31].	2021	Detection of Dental Apical Lesions Using CNNs on Periapical Radiograph	Classification	Periapical Radiographs	460 (322/138)	Periapical Radiographs	3 Dentists	Yes	CNN	Accuracy: 0.93 Specificity: 0.90 Sensitivity 0.95 Precision: 0.92	No
Moidu, N. et al. [32].	2022	Deep Learning for Categorization of Endodontic Lesions Along the Periapical Index	Segmentation	Periapical Radiographs	1950 (1250/250)	Periapical Radiographs	3 Endodontists	Yes	CNN	Sensitivity: 0.92 Sensitivity: 0.76 Precision: 0.86 F1-Score: 0.89 Matthews Coefficient: 0.71	No
Ngoc, V. et al. [33].	2021	Periapical Lesion Diagnosis Support System Based on X-ray Images Using Machine Learning	Object Detection	Periapical Radiographs	1130 (1000/130)	Periapical Radiographs	2 Experienced Endodontists	na	CNN	Sensitivity: 0.89 Specificity: 0.98 Accuracy: 0.96	No
Sajad, M. et al. [34].	2019	Automatic Lesion Detection in Periapical X-rays	Classification	Periapical Radiographs	534 (453/81)	Periapical Radiographs	1 Radiologist and 1 Dentist	Yes	CNN for extraction K-nearest Neighbor Learning Support Vector Machine for classification	Accuracy: 0.79	No
Setzer, F. et al. [35].	2020	Computer-aided Detection of Periapical Lesions in CBCT	Segmentation	CBCT	20 (16/4)	CBCT	1 Radiologist, 1 Endodontist, 1 Oral Radiology Fellow	No	CNN	Sensitivity: 0.93 Specificity: 0.88 PPV: 0.87 NPV: 0.93 DICE: 0.67	No
Shafi, I. et al. [36].	2023	Apical Lesion Detection Using Deep Learning and the Internet of Things	Classification	Periapical Radiographs	534 (453/81)	Periapical Radiographs	2 Dentists with More Than 10 Years of Experience	Yes	CNN for extraction K-nearest Neighbor Learning Support Vector Machine for classification	Accuracy: 0.98 Precision: 0.76 Sensitivity: 0.75 F1 Score: 0.75	No
Song, I. S. et al. [37].	2022	Deep learning-based Apical Lesion	Segmentation	Panoramic Radiographs	1000 (800/100)	Panoramic Radiographs	3 Oral and Maxillofacial Radiologists with	Yes	CNN	Precision: 0.74 Sensitivity: 0.74 F1-Score: 0.74	No

(*continued on next page*)

Table 1 (*continued*)

Author	Year	Description	Aim of Study	Used Imaging Method	Data Size (Training/ Test)	Labeling on	Labeling Done by	Data Augmentation	Model Structure	Performance Measurements and Outcomes	Comparison with Dentists other than Labelers
		segmentation from Panoramic Radiographs					More Than 10 Years of Experience				
Ver Berne, J. et al. [38].	2023	A Deep Learning Approach for Radiological Detection and Classification of Radicular Cysts and Periapical Granulomas	Object Detection and Classification	Panoramic Radiographs	(na) 585/67	Pathology Results, Clinical and Radiological Information	2 Authors Radiologically Supported by Pathology and Clinical Results	Yes	2 CNNs	Specificity: 0.95 Sensitivity: 1.00 AUC for MobileNetV2: 0.88 Average Precision for YOLOV3: 0.74	No
Yilmaz, E. et al. [39].	2017	Computer-aided Diagnosis of Periapical Cyst and Keratocystic Odontogenic Tumor on CBCT	Classification	CBCT Axial Slides	6397 (na/ na)	CBCT	na	na	Support Vector Machine	Accuracy: 0.94 F1-Score: 0.94	No
Zheng, Z. et al. [40].	2020	Anatomically Constrained Deep Learning for Automating Dental CBCT Segmentation and Lesion Detection	Segmentation	CBCT	20 (15/5)	CBCT	3 Dentists, Semi-automated Segmentation with ITK-SNAP	Yes	2 CNNs	Precision: 0.90 Sensitivity: 0.84	No

AFROC-AUC: alternative free-response receiver operating characteristic- area under the curve, AP: average precision, AUC: area under the curve, CBCT: cone-beam computed tomography, CNN: convolutional neural network, FN: false negative, FP: false positive, IoU: intersection over union, mAP: mean average precision, mIoU: mean intersection over union, na: not available, NPV: negative predictive value, OMF: oral maxillo-facial, PPV: positive predictive value, R-CNN: region-CNN, ResNet: residual network, ROI: region of interest, TN: true negative, TP: true positive.

Table 2
Risk of bias according to the QUADAS-2 tool [10].

	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING
Ari, T. et al. [17].	Low	Low	Low	Low
Bayrakdar, I.S. et al. [18].	High	High	High	Unclear
Calazans, M.A. A. et al. [19].	High	Unclear	High	Unclear
Caputo, B. et al. [20].	High	High	Low	Low
Chen, H. et al. [21].	Unclear	Low	Unclear	Unclear
Chuo, Y. et al. [22].	Unclear	Unclear	Unclear	Unclear
Ekert, T. et al. [23].	Unclear	Unclear	Low	Unclear
Endres, M.G. et al. [24].	Low	Low	Low	Low
Ezhov, M. et al. [25].	Unclear	Low	Low	Unclear
Fatima, A. et al. [26].	High	High	High	Unclear
Hamdan, M.H. et al. [27].	Unclear	Low	Low	Unclear
Kim, C. et al. [28].	Low	Low	Unclear	Unclear
Kirnbauer, B. et al. [29].	Unclear	Low	Low	Unclear
Krois, J. et al. [30].	Low	Unclear	Unclear	Unclear
Li, C.W. et al. [31].	Unclear	Unclear	Unclear	Unclear
Moidu, N. et al. [32].	Unclear	Low	Unclear	Unclear
Ngoc, V. et al. [33].	Low	Unclear	Unclear	Unclear
Sajad, M. et al. [34].	High	Unclear	Unclear	Unclear
Setzer, F. et al. [35].	High	Unclear	Unclear	Unclear
Shafi, I. et al. [36].	High	Unclear	Unclear	Unclear
Song, I. S. et al. [37].	Unclear	Low	Unclear	Unclear
Ver Berne, J. et al. [38].	Unclear	Low	Low	Unclear
Yilmaz E. et al. [39].	High	Low	High	Unclear
Zheng, Z. et al. [40].	High	Unclear	Unclear	Unclear

4. Discussion

In this systematic review and meta-analysis, we assessed the diagnostic accuracy of AI, mainly deep neural networks (nearly all CNNs), in detecting periapical radiolucency on different radiological imaging modalities. We identified 24 studies using different image modalities, modeling tasks, and setups, with an overall unclear or high risk of bias and a range of applicability concerns. The quantitative synthesis of four studies, reporting data in sufficient detail to allow meta-analysis, confirmed that AI has high sensitivity and specificity. Notably, given the paucity of comparable data, conducting further subgroup or stratified analysis and yield estimates for specific image modalities or modeling tasks was impossible. Comparisons of the AI against the current standard of care, unaided dentists, were extremely scarce.

The diagnostic accuracy of dentists for detecting periapical radiolucency on both 2D and 3D radiographs has been reported to range between 53 and 90 % [43–46]. As the reliability of clinical pulpal vitality tests is questioned [47], supporting unreliable clinical tests with inconsistent radiological diagnoses makes diagnosing pulpal and periapical health highly subjective. A wide range of deep neural networks

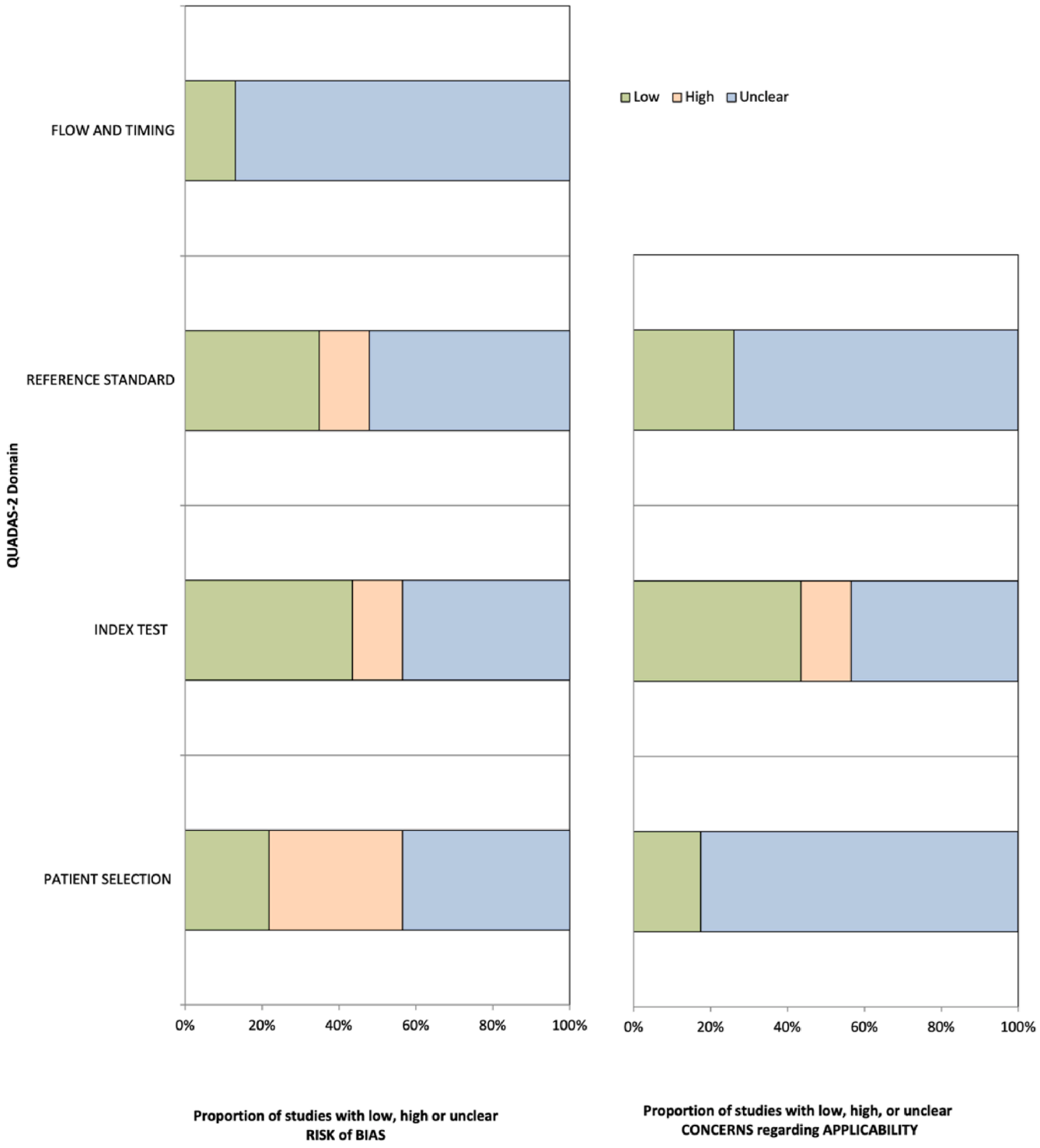

Fig. 2. Bias analysis and applicability concerns of included studies.

have been employed to assist dentists in this task, with networks becoming deeper over time. Moreover, the range of tasks also increased, with earlier studies focusing on classification, while more recent studies also employ segmentation or object detection as well as combinations of those. Notably, some studies attempted to transform segmentation or object detection tasks into classification outcomes, which yield metrics interpretable for clinicians. In our meta-analysis, we included such classification outcomes if they came from object detection and segmentation studies. Future object detection and segmentation studies should consider possible data synthesis and should provide their findings on both pixel and tooth levels, respectively. This would allow gauging the clinical usefulness of a developed AI (it remains unclear if pixel classification accuracy of 50, 60, or 70 % is useful or not). Generally, our review calls for standardization in reporting and outcomes to allow critical appraisal, comparison between studies, and synthesis.

Another element introducing heterogeneity and risk of bias was the reference test against which the performance of the AI was constituted (and on which the AI was trained). Five studies [17,19,21,27,29] employed only one dentist to establish this reference test, while in six

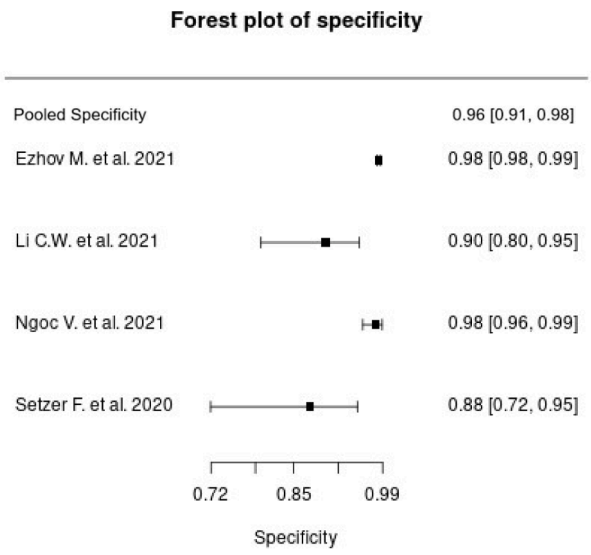

Fig. 3. Forest plot of specificity: mean and 95 % confidence intervals of specificity values are provided; studies are ordered alphabetically.

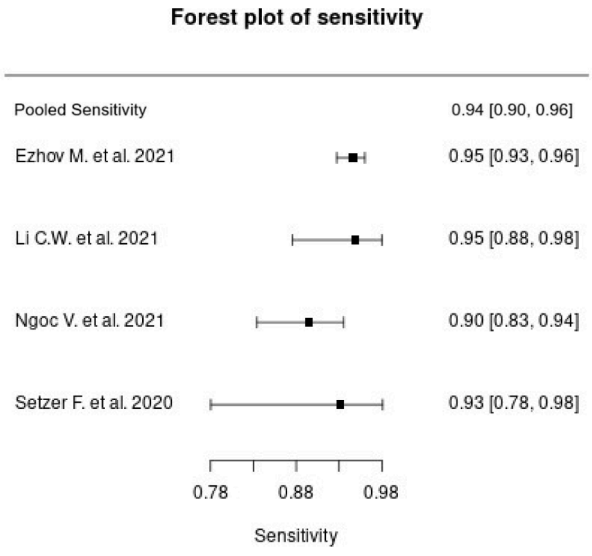

Fig. 4. Forest plot of sensitivity: mean and 95 % confidence intervals of sensitivity values are provided; studies are ordered alphabetically.

studies [20,22,25,26,28,39], it was unclear who or how many experts established it. This is highly relevant, as models trained and tested on data from only one annotator will not exceed that annotator's performance, which is why guidelines recommend multiple annotators assessing each image independently. Notably, once this is concluded, it remains unclear how best to establish one label from the multitude of inputs. Ekert et al. [23] have demonstrated the impact of different schemes in unifying multiple annotations into one reference test. Moreover, the reference test was mainly established on 2D radiographs – which are known to allow only limited accuracy for detecting periapical lesions; 3D radiographs [19,25,29,35,39,40] and histological assessment [48] or a combination of imagery and clinical tests [38] was scarce, likely as most studies were conducted retrospectively, where the availability of further data sources (like 3D images or tests) is limited.

Another problem frequently encountered was possible overfitting and, consequently, limited generalizability. Overfitting occurs if training and validation datasets are small and homogeneous. To limit overfitting, augmentation strategies like cropping or rotating, etc., are applied while still relying on data originating from the same source. For most studies, it remains impossible to gauge the full extent of overfitting, as no testing on truly independent datasets was performed. Using data from centers in Germany and India, Krois et al. [30] demonstrated the lack of generalizability of their developed model to detect apical lesions on panoramic radiographs; further research into this direction is needed.

The overall body of evidence comes with an overall unclear or high risk of bias and several applicability concerns, according to QUADAS-2 [10]. QUADAS-2 assesses the risk of bias and applicability across four domains: patient selection, index test, reference standard, and flow and timing. Each domain is evaluated for bias, and the first three are assessed for applicability concerns. This tool helped us identify methodological weaknesses, such as unclear patient selection and lack of blinding in index tests. Using QUADAS-2, we provided a transparent assessment of study quality, highlighting strengths and limitations to inform the reliability of our meta-analysis results and advance understanding of diagnostic test performance.

None of the studies was performed on a fully random sample of patients (which is generally true for studies involving radiographs, though, as these are usually taken not for surveillance purposes but on the basis of a medical justification in a clinical setting). The reference test and the limited heterogeneity in test data have been discussed. Generally, limited reporting detail led to unclear classification of risk of bias and applicability concerns for many studies. Future studies should aim to better adhere to reporting guidelines in the field [49]. Overall, our confidence in any conclusions drawn from this review needs to be limited.

On the basis of this review, a range of research gaps can be identified. First, future studies should focus less on applying different (novel) architectures on the existing small datasets and more on the impact of methodological aspects (like establishing the reference test and heterogeneity in data) on model performance and generalizability. Larger multi-centric datasets are needed for this purpose. These may only be established by international consortia, which could then focus on developing ways to benchmark models against one (or several) representative, systematically, and reliably annotated datasets. The ITU/WHO Focus Group AI For Health (soon WHO/ITU/WIPO Global Initiative AI For Health) is one such consortium. Notably, these consortia will require large-scale heterogeneous datasets; accessing and pooling such sets of sensitive data come with certain restrictions, i.e. implementing data sharing agreements in compliance with local jurisdiction and sufficient de-anonymization.

Second, the true impact of deep learning on clinicians, clinical care, and health services should be determined. The assessment of commercially available AI systems seems warranted, as it remains unclear how dentists use these systems, how their usage impacts diagnostic processes and decision-making, and how patients benefit (or not) from this new technology. Lastly, such assessments need different study designs beyond the prevailing diagnostic accuracy studies. Randomized real-world trials and nested health economic and health behavioral assessments may be needed.

AI systems designed to detect periapical lesions have the potential to impact clinical practice significantly. With AI-supported detection, dentists can more accurately identify lesions, especially in cases where the lesions are small or ambiguous. Early and accurate detection can lead to timely interventions, improving treatment outcomes and preserving dental structures. Additionally, AI can standardize the diagnostic process, reducing variability between practitioners and ensuring consistent care for patients. Notably, AI-aided users may also show a higher number of diagnostic or therapeutic interventions given a potentially lower specificity than unaided users, which may then come with detrimental health and economic outcomes [50].

AI tools can also serve as educational resources, enhancing diagnostic skills and confidence among dental students and less experienced dentists. Integrating AI into daily practice can streamline workflows, saving dentists time and allowing them to focus more on patient care.

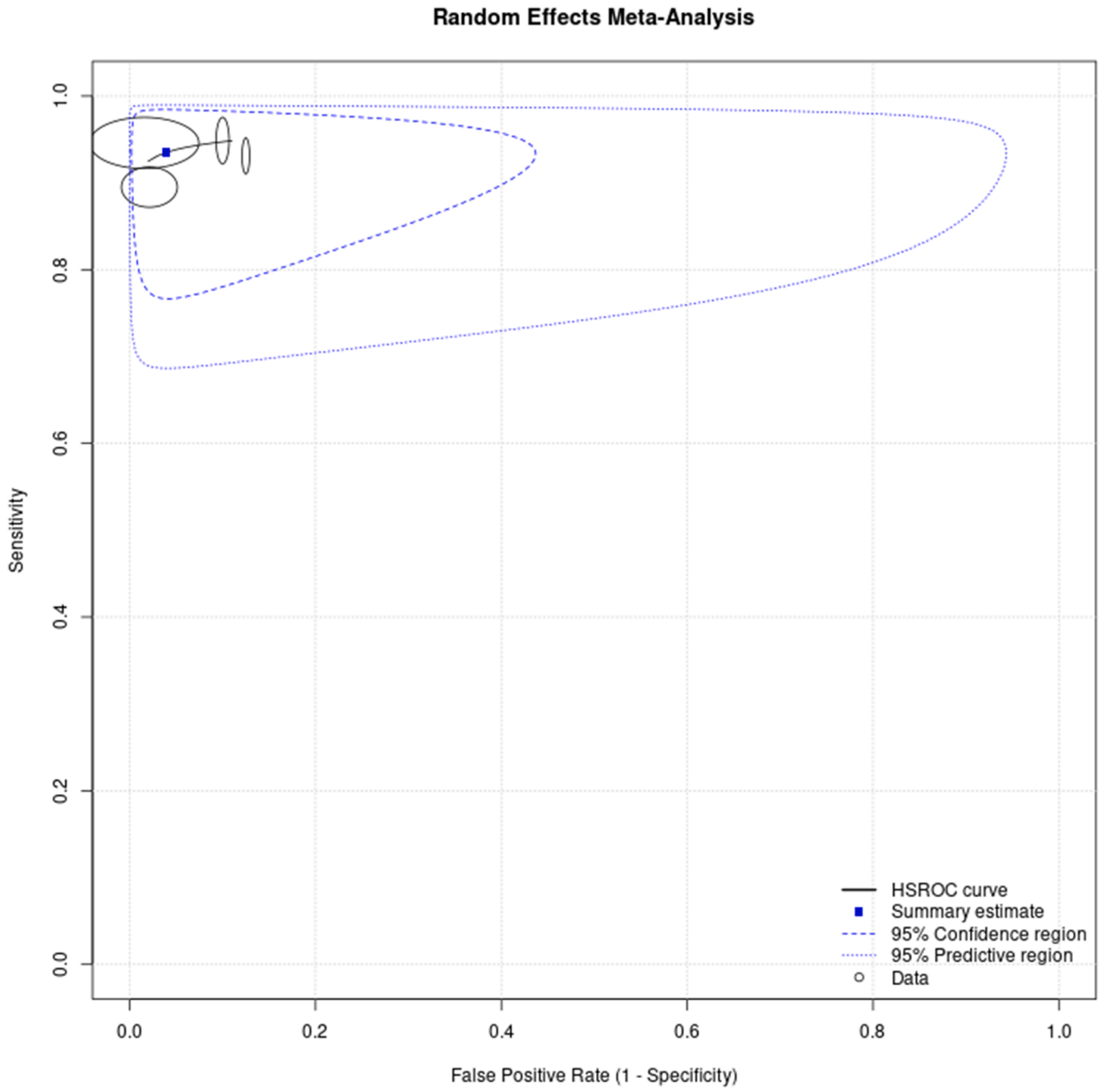

Fig. 5. Hierarchical summary receiver operating characteristic (HSROC) plot. Individual studies are shown as circles on the HSROC area; each circle's size indicates the study's weight within the provided random-effects meta-analysis. HSROC shows sensitivity and specificity median values along calculated predictive and confidence regions [42].

This study is subject to several limitations. First, the scarcity of reliable research data imposes constraints on the studies eligible for inclusion in the meta-analysis. The limited number of included studies and their relatively small sample sizes significantly curtail the statistical robustness of our analysis. Second, there is an absence of consensus regarding optimal procedures for conducting meta-analyses of diagnostic accuracy studies. We have adopted the MetaDTA Tool. This tool mandates the availability of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values from the studies. Utilizing other recommended meta-analysis methodologies [51–53] such as *metaDAS, midas*, and *metandi* may have facilitated the inclusion of a greater number of studies in the analysis, thereby enhancing its comprehensiveness and generalizability. Fourth, all included studies were retrospective in nature. While the inclusion of prospective studies would have been preferable to mitigate selection bias, such studies were not available. Last, nearly all studies employed CNNs, which constrains our ability to draw a conclusive assessment of the broader applicability of AI within this domain.

This is the second meta-analysis available assessing the accuracy of deep learning in detecting periapical lesions. A previous meta-analysis [54] showed a similarly high accuracy (0.93) but a lower specificity (0.85). Notably, the authors were less strict on including studies in their meta-analysis: One study [31] was included twice, as two models were developed – while both were tested on the same test data. Two studies [29,32] that reported only the percentage of true and false positives and negatives were included, whilst it remains unclear how accurate any transformation into total numerical values is. A recent systematic review [55] on the matter included only nine studies [18,23,24,31–33,35,56,

57], all of which were included in our review, and did not perform a meta-analysis.

5. Conclusions

AI has shown promise in accurately identifying radiographic periapical lesions, potentially assisting dentists in improving diagnostic accuracy and consistency. To fully harness this potential, future research should focus on robust validation of AI to detect periapical lesions through prospective, randomized controlled trials on diverse populations. Collaborative efforts to create large, systematically annotated datasets will help enhance model development and benchmarking. Additionally, it is essential to conduct comprehensive assessments of AI's economic and behavioral impact to understand its influence on clinical decision-making and patient outcomes. Practical integration of AI tools into dental workflows, with user-friendly interfaces, will ensure that these systems complement clinical practice without adding complexity. While challenges remain, the prospects for AI in enhancing the detection of periapical lesions are promising, offering significant benefits for both practitioners and patients.

Funding

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

CRediT authorship contribution statement

Utku Pul: Writing – original draft, Visualization, Software, Formal analysis, Data curation, Conceptualization. **Falk Schwendicke:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization.

Declaration of competing interest

FS is a cofounder of the dentalXrai Ltd., focusing on AI-based image analysis. Planning, conduct and reporting of this study was independent from any such activities.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jdent.2024.105104.

References

[1] A. Hilmi, et al., Efficacy of imaging techniques for the diagnosis of apical periodontitis: a systematic review, Int. Endod. J. (2023).
[2] V.S. Yu, et al., Risk score algorithm for treatment of persistent apical periodontitis, J. Dent. Res. 93 (11) (2014) 1076–1082.
[3] J.G. Pak, S. Fayazi, S.N. White, Prevalence of periapical radiolucency and root canal treatment: a systematic review of cross-sectional studies, J. Endod. 38 (9) (2012) 1170–1176.
[4] F.A. Alaidarous, et al., Prevalence of periapical radiolucency and conventional root canal treatment in adults: a systematic review of cross-sectional studies, Cureus. 15 (1) (2023) e33302.
[5] W.Y. Mao, et al., Comparison of radiographical characteristics and diagnostic accuracy of intraosseous jaw lesions on panoramic radiographs and CBCT, Dentomaxillofac. Radiol. 50 (2) (2021) 20200165.
[6] H.F. Duncan, et al., Treatment of pulpal and apical disease: the European Society of Endodontology (ESE) S3-level clinical practice guideline, Int. Endod. J. 56 (S3) (2023) 238–295.
[7] D. Donnermeyer, et al., Effectiveness of diagnosing pulpitis: a systematic review, Int. Endod. J. 56 (S3) (2023) 296-325.
[8] R. Aggarwal, et al., Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis, NPJ. Digit. Med. 4 (1) (2021) 65.
[9] J.P. Matthew, et al., PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews, BMJ 372 (2021) n160.
[10] R.A. P.F. Whiting, M.E. Westwood, S. Mallett, J.J. Deeks, J.B. Reitsma, M. M. Leeflang, J.A.C. Sterne, P.M.M. Bossuyt, QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies Ann. Intern. Med. 155 (8) (2011) 529–536.
[11] V.N. Nyaga, M. Arbyn, Metadta: a Stata command for meta-analysis and meta-regression of diagnostic test accuracy data – a tutorial, Arch. Public Health 80 (1) (2022) 95.
[12] A. Patel, et al., Graphical enhancements to summary receiver operating characteristic plots to facilitate the analysis and reporting of meta-analysis of diagnostic test accuracy data, Res. Synth. Methods 12 (1) (2021) 34–44.
[13] Partlett C., T.Y., Meta analysis of test accuracy studies in R: a summary of user-written programs and step-by-step guide to using glmer. Version 1.0. 2016.
[14] S.C. Freeman, et al., Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA, BMC Med. Res. Methodol. 19 (1) (2019) 81.
[15] H. Chu, S.R. Cole, Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach, J. Clin. Epidemiol. 59 (12) (2006), 1331-2; author reply 1332-3.
[16] Bates, D., et al., lme4: linear mixed-effects models using Eigen and S4. R package version 1. 1–7. 2014.
[17] T. Ari, et al., Automatic feature segmentation in dental periapical radiographs, Diagnostics, (Basel) 12 (12) (2022) 3081–3091.
[18] I.S. Bayrakdar, et al., A U-net approach to apical lesion segmentation on panoramic radiographs, Biomed. Res. Int. 2022 (2022) 7035367.
[19] M.A.A. Calazans, et al., Automatic classification system for periapical lesions in cone-beam computed tomography, Sensors, (Basel) 22 (17) (2022) 6481–6496.
[20] B. Caputo, G.E. Gigante, Analysis of periapical lesion using statistical textural features, Stud. Health Technol. Inform. 77 (2000) 1231–1234.
[21] H. Chen, et al., Dental disease detection on periapical radiographs based on deep convolutional neural networks, Int. J. Comput. Assist. Radiol. Surg. 16 (4) (2021) 649–661.
[22] Y. Chuo, et al., A high-accuracy detection system: based on transfer learning for apical lesions on periapical radiograph, Bioengineering. (Basel) 9 (12) (2022) 777–793.
[23] T. Ekert, et al., Deep learning for the radiographic detection of apical lesions, J. Endod. 45 (7) (2019) 917–922, e5.
[24] M.G. Endres, et al., Development of a deep learning algorithm for periapical disease detection in dental radiographs, Diagnostics, (Basel) 10 (6) (2020) 430-51.
[25] M. Ezhov, et al., Clinically applicable artificial intelligence system for dental diagnosis with CBCT, Sci. Rep. 11 (1) (2021) 15006.
[26] A. Fatima, et al., Deep learning-based multiclass instance segmentation for dental lesion detection, Healthcare (Basel) 11 (3) (2023) 347–363.
[27] M.H. Hamdan, et al., The effect of a deep-learning tool on dentists' performances in detecting apical radiolucencies on periapical radiographs, Dentomaxillofac. Radiol. 51 (7) (2022) 20220122.
[28] C. Kim, et al., Tooth-related disease detection system based on panoramic images and optimization through automation: development study, JMIR Med. Inform. 10 (10) (2022) e38640.
[29] B. Kirnbauer, et al., Automatic detection of Periapical osteolytic lesions on cone-beam computed tomography using deep convolutional neuronal networks, J. Endod. 48 (11) (2022) 1434–1440.
[30] J. Krois, et al., Generalizability of deep learning models for dental image analysis, Sci. Rep. 11 (1) (2021) 6102.
[31] C.W. Li, et al., Detection of dental apical lesions using CNNs on periapical radiograph, Sensors, (Basel) 21 (21) (2021) 7049–7066.
[32] N.P. Moidu, et al., Deep learning for categorization of endodontic lesion based on radiographic periapical index scoring system, Clin. Oral Investig. 26 (1) (2022) 651–658.
[33] V. Ngoc, et al., Periapical lesion diagnosis support system based on x-ray images using machine learning technique, World J. Dentistry 12 (2021) 189–193.
[34] Sajad, M., I. Shafi, and J. Ahmad, Automatic lesion detection in periapical x-rays. 2019. 1–6.
[35] F.C. Setzer, et al., Artificial intelligence for the computer-aided detection of periapical lesions in cone-beam computed tomographic images, J. Endod. 46 (7) (2020) 987–993.
[36] I. Shafi, et al., Teeth lesion detection using deep learning and the internet of things post-COVID-19, Sensors, (Basel) 23 (15) (2023) 6837–6857.
[37] I.S. Song, et al., Deep learning-based apical lesion segmentation from panoramic radiographs, ImAging Sci. Dent. 52 (4) (2022) 351–357.
[38] J. Ver Berne, et al., A deep learning approach for radiological detection and classification of radicular cysts and periapical granulomas, J. Dent. 135 (2023) 104581.
[39] E. Yilmaz, T. Kayikcioglu, S. Kayipmaz, Computer-aided diagnosis of periapical cyst and keratocystic odontogenic tumor on cone beam computed tomography, Comput. Methods Programs Biomed. 146 (2017) 91–100.
[40] Z. Zheng, et al., Anatomically constrained deep learning for automating dental CBCT segmentation and lesion detection, IEEE Trans. Autom. Sci. Eng. (2020) 1–12. **PP**.
[41] H. Shetty, et al., Three-dimensional semi-automated volumetric assessment of the pulp space of teeth following regenerative dental procedures, Sci. Rep. 11 (1) (2021) 21914.
[42] R.M. Harbord, et al., A unification of models for meta-analysis of diagnostic accuracy studies, Biostatistics 8 (2) (2006) 239–251.
[43] A.A. Sherwood, et al., A deep learning approach to segment and classify C-shaped canal morphologies in mandibular second molars using cone-beam computed tomography, J. Endod. 47 (12) (2021) 1907–1916.
[44] C. Kruse, et al., Diagnostic validity of periapical radiography and CBCT for assessing periapical lesions that persist after endodontic surgery, Dentomaxillofac. Radiol. 46 (7) (2017) 20170210.

[45] S.E. Stheeman, et al., Does radiographic feature recognition contribute to dentists' diagnosis of pathology? Dentomaxillofac. Radiol. 24 (3) (1995) 155–159.
[46] M. Rohlin, et al., Observer performance in the assessment of periapical pathology: a comparison of panoramic with periapical radiography, Dentomaxillofac. Radiol. 20 (3) (1991) 127–131.
[47] Z. Kong, et al., Automated periodontitis bone loss diagnosis in panoramic radiographs using a bespoke two-stage detector, Comput. Biol. Med. 152 (2023) 106374.
[48] K. Okada, et al., Noninvasive differential diagnosis of dental periapical lesions in cone-beam CT scans, Med. Phys. 42 (4) (2015) 1653–1665.
[49] V. Sounderajah, et al., Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol, BMJ Open. 11 (6) (2021) e047709.
[50] S. Mertens, et al., Artificial intelligence for caries detection: randomized trial, J. Dent. 115 (2021) 103849.
[51] J. Wang, M. Leeflang, Recommended software/packages for meta-analysis of diagnostic accuracy, J. Lab. Precis. Med. 4 (2019) 22–27.
[52] Y. Takwoingi, R.D. Riley, J.J. Deeks, Meta-analysis of diagnostic accuracy studies in mental health, Evid. Based. Ment. Health 18 (4) (2015) 103–109.
[53] J.J. Deeks, et al., Cochrane Handbook For Systematic Reviews of Diagnostic Test Accuracy, John Wiley & Sons, 2023.
[54] S. Sadr, et al., Deep learning for detection of periapical radiolucent lesions: a systematic review and meta-analysis of diagnostic test accuracy, J. Endod. 49 (3) (2023) 248–261, e3.
[55] S. Ramezanzade, et al., The efficiency of artificial intelligence methods for finding radiographic features in different endodontic treatments - a systematic review, Acta Odontol. Scand. 81 (6) (2023) 422–435.
[56] K. Orhan, et al., Evaluation of artificial intelligence for detecting impacted third molars on cone-beam computed tomography scans, J. Stomatol. Oral Maxillofac. Surg. 122 (4) (2021) 333–337.
[57] R. Pauwels, et al., Artificial intelligence for detection of periapical lesions on intraoral radiographs: comparison between convolutional neural networks and human observers, Oral Surg. Oral Med. Oral Pathol. Oral Radiol. 131 (5) (2021) 610–616.