BMC Oral Health

SYSTEMATIC REVIEW
Open Access

Deep learning for tooth detection and segmentation in panoramic radiographs: a systematic review and meta-analysis

M. Bonfanti-Gris[1], A. Herrera[1], M. P. Salido Rodríguez-Manzaneque[1*], F. Martínez-Rus[1] and G. Pradíes[1]

Abstract

Background This systematic review and meta-analysis aimed to summarize and evaluate the available information regarding the performance of deep learning methods for tooth detection and segmentation in orthopantomographies.

Material and methods Electronic databases (Medline, Embase and Cochrane) were searched up to September 2023 for relevant observational studies and both, randomized and controlled clinical trials. Two reviewers independently conducted the study selection, data extraction, and quality assessments. GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) assessment was adopted for collective grading of the overall body of evidence. From the 2,207 records identified, 20 studies were included in the analysis. Meta-analysis was conducted for the comparison of mesiodens detection and segmentation ($n=6$) using sensitivity and specificity as the two main diagnostic parameters. A graphical summary of the analysis was also plotted and a Hierarchical Summary Receiver Operating Characteristic curve, prediction region, summary point, and confidence region were illustrated.

Results The included studies quantitative analysis showed pooled sensitivity, specificity, positive LR, negative LR, and diagnostic odds ratio of 0.92 (95% confidence interval [CI], 0.84–0.96), 0.94 (95% CI, 0.89–0.97), 15.7 (95% CI, 7.6–32.2), 0.08 (95% CI, 0.04–0.18), and 186 (95% CI, 44–793), respectively. A graphical summary of the meta-analysis was plotted based on sensitivity and specificity. Hierarchical Summary Receiver Operating Characteristic curves showed a positive correlation between logit-transformed sensitivity and specificity ($r=0.886$).

Conclusions Based on the results of the meta-analysis and GRADE assessment, a moderate recommendation is advised to dental operators when relying on AI-based tools for tooth detection and segmentation in panoramic radiographs**.**

Keywords Orthopantomography, Panoramic radiograph, Artificial intelligence, Convolutional neural network, Deep learning, Dentistry

Introduction

Panoramic radiographs (OPGs) are essential diagnostic tools used in dentistry, as they help clinicians outline treatment plans based on a greater initial diagnosis. However, their diagnostic accuracy is influenced by experts' overall expertise and fatigue [1].

Artificial Intelligence (AI) aims to create computer-based systems (Convolutional Neural Networks, CNN) that provide human-like decision-making solutions [2].

*Correspondence:
M. P. Salido Rodríguez-Manzaneque
mpsalido@ucm.es
[1] Department of Conservative and Prosthetic Dentistry, Faculty of Dentistry, Universidad Complutense de Madrid, Plaza Ramón y Cajal S/N, Madrid 28040, Spain

AI is known to be divided into Machine Learning (ML) and Deep Learning (DL) depending on its pre-processing steps for learning purposes [1, 3]. To date, both subsets have been applied for radiographic analysis and interpretation – which falls into Computer Vision AI's subset—although this last one has achieved better outcomes [1, 4]. Overall, a typical DL model consists of three main components: the input layer, which receives raw data; hidden layers, which perform computations and extract hierarchical features; and the output layer, which provides the final predictions or classifications. These multiple layers enable deep learning models to automatically learn complex patterns from large datasets, making them highly effective for tasks such as image recognition, natural language processing, and medical diagnostics [2].

AI has been reported to be used for periodontal bone-loss detection [5], dental implant classification [6], or even caries, periapical pathology and mesiodens identification in both 2D and 3D images [7, 8]. However, accurate pathology detection and dental structure association remain somehow challenging—especially in OPGs. Thus, AI-based systems must be trained to identify both dental and maxillofacial structures accurately [9] for which Object detection (OD) and segmentation (OS) tasks play a crucial role, especially when finding supernumeraries. Yet, their overall performance can be hindered due to the anatomical boundary challenges that OPGs pose [1].

CNNs like ResNet, Fast/Faster Region, Visual Geometry Group (VGG) and You Only Look Once (YOLO) have been reported to be successful in both OD and OS, but recent studies have described that combining them may enhance their overall outcomes. Also, it is important to consider that, up to date, these networks have been broadly employed using different training and validation methodologies and different sample sizes—so the performance of AI systems might be both over and underrated [4].

A systematic review by Umer et al. emphasized the need for standardized methodology and performance metrics in DL models for OD. They questioned the use of DL models for this purpose, but the evolving research landscape may have altered this perspective [1].

Therefore, this systematic review aimed to summarize and evaluate the available information regarding the performance of Deep Learning AI-based systems for tooth detection and segmentation in panoramic radiographs, compared to the reference human-executed diagnosis.

Methods

This systematic review was conducted following the PRISMA DTA guidelines [10]. The review protocol was registered and allocated with the identification number CRD42023453081 PROSPERO (University of York, National Institute for Health Research, United Kingdom).

Literature search

Two electronic literature searches were conducted in three databases during December 2022 and September 2023—PubMed (Medline), Embase, and Cochrane. A hand search of specific references was also performed. The search keywords were based on the Medical Subject Headings (MeSH) terms specified in Additional File 1. Keywords referencing all dental radiological images were used to obtain the largest number of studies where panoramic radiographs might have also been involved in the methodological process. All of the selected study lists were manually reviewed for cross-references.

Inclusion criteria (PIRD)

Both observational studies and randomized and controlled clinical trials (RCT/CCT) published in English and other languages up to January 2013 were screened based on the PIRD criteria:

- Population: Panoramic Radiographs, without exclusion of patient age range or sex.
- Index Test: Deep Learning models employing teeth identification tasks – object detection, segmentation, and labeling.
- Reference Standard: Ground Truth or Reference Test established by experts.
- Diagnostic Accuracy: Performance indicators based on object or pixel level, including Dice Coefficient, F1-Score, Area Under Curve (AUC), Sensitivity (S), Specificity (E), Positive Predictive Value (PPV), Negative Predictive Value (NPV), Precision (P), Recall (R), Average Precision (AP), Mean Average Precision (mAP) and Average Recall (AR). The diagnostic performance of Deep Learning models for tooth identification and labeling was set as the primary outcome. Two secondary outcomes were also considered: AI-based software's capacity to detect A) mesiodens and B) impacted mandibular third molars.

Studies registered as protocols only, literature and scoping reviews and unpublished manuscripts were not considered. Also, studies without sufficient details on the AI method used or on the data used for validation were excluded.

Both title and abstract and full-text screenings were performed by two independent operators (A.H and M.B-G). If disagreement was found at any stage of the selection process, it was solved through consensus with a third reviewer (G.P). The inter-reviewer reliability (percentage of agreement and kappa correlation coefficient) for title and abstract screening and full-text analysis was conducted.

Data extraction and analysis

Two different reviewers individually collected data from each of the 26 studies included in the present systematic review (A.H and M.B-G). A third reviewer revised data extraction for discrepancies and disagreements (G.P). The following data items were extracted: Author, Date of Publication, Country, Objective, Study Design, AI Technique, Convolutional Neural Network, Sample Size, Dataset, Data Augmentation, Validation Method, Reference/Standard Comparison (including experience years and total number of operators), Outcome Measurement, and Main Study Outcomes.

When data were incomplete or missing, corresponding authors were contacted for clarification. If the information could not be obtained in time, this was excluded until notification was available. If the results of a given study were published more than once, only the most recent one was included. Also, if more than one neural network was employed in the same study, only the results of the most accurate ones were used for analysis. In case a study included performance evaluation of a CNN for several variables, only the ones considering OD and OS for dental structures were considered.

When referring to data analysis, the main information regarding Accuracy (A), Sensitivity (S), Specificity (E), Precision (P), Recall (R), Mean Average Precision and Recall (mAP and mAR) and F1 Score were sought. Nevertheless, performance measurements like Dice Coefficient, PPV, NPV, AUC and confusion matrix results were also considered if available. If a study only did not show the results for the main aforementioned but did report other metrics with which to calculate them, this was carried out by the authors in the present study.

Risk of bias assessment

This process involved two reviewers (A.H and M.B-G), with discrepancies resolved by a third assessor (G.P). QUADAS-2 checklist was used to assess the risk of bias and methodological quality, focusing on patient selection, index test, reference standard, and flow of patients [11]. In our case, data clarification questions were added [12] (Additional File 2). Overall, the checklist questions were answered as positively (+), negatively (-), unclear (NR), and occasionally, as NA (Not Applicable). Based on the final results, studies were rated for high (HR), low (LR), or unclear risk (UR) of bias.

Overall, high risk for patient selection resulted from limited dataset information. Index test was categorized as high risk if test reproducibility was poor, or model construction details were lacking. Reference Standard risk was also high when an unclear definition was provided or when just one operator established ground truth.

The GRADE approach was also used to assess the evidence certainty and to determine the strength of clinical practice recommendations. Ratings were categorized as either very low, low, moderate, or high risk of certainty of evidence [13, 14]. Publication bias was addressed by performing Egger's test for Diagnostic Odds Ratio variable.

Data analysis

The pooled diagnostic performance values were also estimated for detection and segmentation as a meta-analysis. The estimates were presented with 95% confidence intervals (CIs). Heterogeneity between the studies was examined by the I^2 index, which ranges from 0.0% to 100.0%, and a *P*-value less than 0.10 was considered significant. Forest plots were utilized to describe the results of the meta-analysis. A hierarchical logistic regression model was used for meta-analyses on diagnostic accuracy. Only studies that reported true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in their test sets were included in the meta-analyses. The metandi command in STATA software version 17 (StataCorp, College Station, Texas) was used to calculate the pooled version of the following parameters.

Results

Search

Figure 1 depicts the flow chart summarizing the results of the selection process. The search strategy rendered a total of 1717 titles after the removal of duplicated records. From these, a final number of 20 articles were finally included for data extraction. Kappa Index for title and abstract screening and full-text analysis was 0.926 and 0.872, respectively (IC =95%). Overall reviewer agreement was 98% for first screening and 97.8% for full-text analysis.

Description of the included studies

From the final twenty studies that were included in the present systematic review, fourteen evaluated tooth detection of all classes, and six focused on mesiodens detection on panoramic radiographs. Characteristics of the included studies are summarized in Table 1. In terms of study design, nineteen studies performed a development and validation research and only one carried out a validation model with an already pre-trained neural network [15]. Despite all studies utilizing deep learning methods, two of them utilized this approach combined with collaborative learning [4] or other optimization techniques [16]. Most of them used a combination of several architectures to approach their research question. Only four studies stook with just

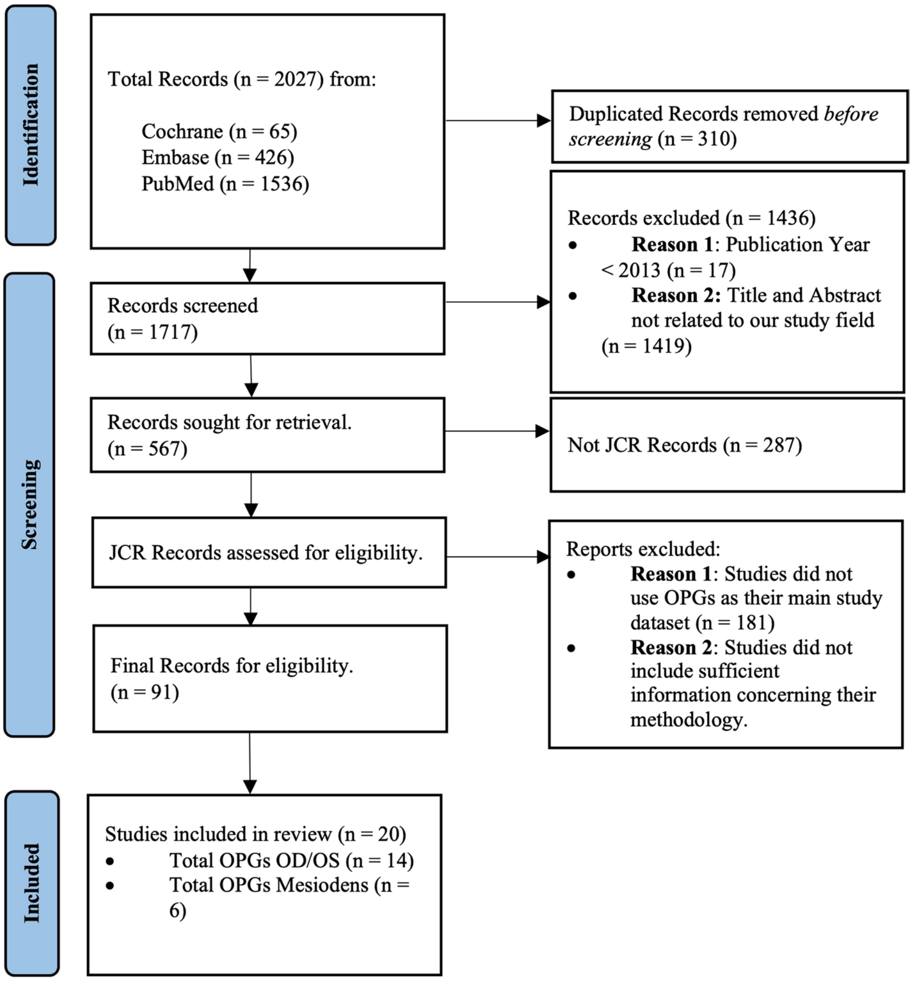

Identification
Screening
Included
Total Records (n = 2027) from:
Cochrane (n = 65)
Embase (n = 426)
PubMed (n = 1536)
Duplicated Records removed *before screening* (n = 310)
Records screened (n = 1717)
Records excluded (n = 1436)
Reason 1: Publication Year < 2013 (n = 17)
Reason 2: Title and Abstract not related to our study field (n = 1419)
Records sought for retrieval. (n = 567)
Not JCR Records (n = 287)
JCR Records assessed for eligibility.
Reports excluded:
Reason 1: Studies did not use OPGs as their main study dataset (n = 181)
Reason 2: Studies did not include sufficient information concerning their methodology.
Final Records for eligibility. (n = 91)
Studies included in review (n = 20)
Total OPGs OD/OS (n = 14)
Total OPGs Mesiodens (n = 6)

Fig. 1 PRISMA 2020 flow diagram: selection of studies evaluating DL-CNN for object detection and segmentation in panoramic radiographs

one neural network [17–21]. Overall, the most repeated CNN architecture families used were Faster R-CNN [7], ResNet [6], VGG-16 [5], YOLO [5] and Mask R-CNN [3]. Other utilized neural networks include Inception V2 [2], UNet [2], Inception V3 [1], AlexNet [1], DetectNet [1], SqueezeNet [1], DMLNet [1], DeepLabv3 [1], SwinNet [1] and EfficientNet [1]. Validation procedures included cross-validation [3], and both independent [1] and split sample techniques [15].

As for the sample size, the majority of the included studies counted a total of 500–1500 OPGs for all Training, Validation and Testing Phases. Only one study used 50 OPGs, which was the smaller sample reported. Two studies used both sample sizes between 1500 and 2500 images, and over 2500. To decrease overfitting probability, 9 studies used data augmentation to increase the amount of training data based on modifications made to the images included in this set.

The majority of the included titles reported their results in terms of precision (P), sensibility (S), recall (R), specificity (E) and F1 Score. Nevertheless, it was also common to encounter results expressed in mean average precision (mAP), mean average recall (mAR), false discovery rate (FDR), or false negative rate (FNR).

Risk of bias and certainty of evidence assessment

Egger's Test was performed to rate the publication bias for Diagnostic Odds Ratio, (p value $=0.025$) suggesting asymmetry in the funnel plot and, thus, publication bias (Fig. 2). Quality assessment results are shown in Additional File 3 and Table 2. Overall, 6 studies showed HR of bias and the "Reference Standard" domain appeared to be the most debatable. Applicability concerns were also evaluated, resulting in a majority of studies accounting to be appropriate for the matter (Fig. 3).

Table 1 All included studies general characteristics

Title	Date	Country	Objective	Study Design	AI Technique	CNN	Sample Size (OPGs)	Datase (OPGs)	Data Augmentation	Validation Method	Reference Standard (OPs, YE)	Outome Measurement	Main Outcomes
(Vinayahalingam et al. 2021) [25]	2021	Netherlands	To develop a CNN that can detect, segment and label teeth and dental treatments on OPGs	D and V	Deep Learning	Mask R-CNN and ResNet-50 Architectures	2000	TrS = 1600; VS = 200; TeS = 200	No	Split Sample Validation	Expert Annotation and Labelling (3, > 10)	P, F1 Score, and R	OD: *P* = 0.997; F1 Score = 0.992; *R* = 0.989 OS: *P* = 0.971; F1 Score = 0.952; *R* = 0.937 OL: *P* = 0.975; F1 Score = 0.970; *R* = 0.965
(Bonfanti-Gris et al. 2022) [15]	2022	Spain	To evaluate the reliability of a web-based AI CNN for object detection and labelling on OPGs	V	Deep Learning	Faster R-CNN and VGG-16 Architectures	300	TeS = 300	No	-	Expert Annotation and Labelling (2, NS)	S, E, PPV, NPV, AUC ROC Curve	OD: S = 0.693 OL: S = 0.500 E = 0.500 **Error Types Available*
(Tuzoff et al. 2019) [27]	2019	Canada	To propose a novel CNN that detects and labels dental structures present on OPGs	D and V	Deep Learning	Faster R-CNN Architecture, VGG-16	1574	TrS = 1352; TeS = 222	Yes	Split Sample Validation	Expert Annotation and Labelling (5, NS)	Detection: S and P; Labeling: S and E	OD: S = 0.9941; *P* = 0.9945; OL: S = 0.9800; E = 0.9994
(H.-R. Choi et al. 2022) [22]	2022	Korea	To propose a deep learning system that automatically detects teeth and dental treatment on OPGs for potential human identification	D and V	Deep Learning	EfficientDet-D3 + EfficientNet-B3	1638	TrS = 983; VS = 328; TeS = 327	Yes	Split Sample Validation	Expert Annotation and Labelling (1, NS)	IoU AP and AR	OD: AP = 0.991; AR = 0.996
(Mahdi et al. 2020) [16]	2020	Japan	To propose an automatic teeth recognition model	D and V	Deep Learning + Optimization Technique	Faster R-CNN Architecture, (1) ResNet-101 + (2) ResNet-105	1000	TrS = 900; TeS = 100	No	Split Sample Validation	Expert Annotation and Labelling (NS, NS)	IoU AP, AR and F1 Score	OD: (1) mAP = 0.955, mAR = 0.990; F1-Score = 0.978. (2) mAP = 0.930, mAR = 0.966, F1 Score = 0.965

Table 1 (continued)

Title	Date	Country	Objective	Study Design	AI Technique	CNN	Sample Size (OPGs)	Datase (OPGs)	Data Augmentation	Validation Method	Reference Standard (OPs, YE)	Outome Measurement	Main Outcomes
(Yüksel et al. 2021) [17]	2021	Turkey	To develop a Deep Learning framework that detects and labels dental treatments and structures on OPGs	D and V	Deep Learning	YOLO Architecture	1005	TrS = 855; TeS = 150	Yes	Split Sample Validation	Student Annotation and Labelling, Expert Validation (1 OP, NS)	IoU AP and AR	OD: mAP (0.5–0.95) = 0.477, MAX score 89.4% OL mAR (0.5–0.95) = 0.564, MAX score 59%
(Chandrashekar et al. 2022) [4]	2022	USA	To propose a collaborative learning approach to supplement existing object recognition and segmentation algorithms on OPGs	D and V	Deep Learning, *Collaborative Learning (CL)*	OD: YOLO-5 (1) and Faster R-CNN (2) Architectures; OS: Mask R-CNN (3) and U-Net(4) Architectures	1500	(OD, 1 + 2) TrS = 750; VS = 150 and TeS = 100 *(CL) TeS = 150* (OS, 3 + 4) TrS = 193; VS = 83; TeS = 1224 *(CL) TeS = 150*	Yes	Independent	NS	A, P, R, F1 Score and mAP	OD (1) A = 0.995; F1 Score = 0.998 and mAP = 0.995. (2) A = 0.910, F1 Score = 0.900 and mAP = 0.910 *(CL) A = 0.9844, F1 Score = 0.987 and mAP = 0.977* OS (3) A = 0.960, F1 Score = 0.980 and mAP = 0.950. (4) A = 0.9697, F1 Score = 0.936 and mAP = 0.920. *(CL) A = 987, F1 Score = 0.988 and mAP = 0.973*
(Sheng et al. 2022 Oct 14) [18]	2022	China	To evaluate the tooth segmentation capacity of a DL-CNN in OPGs using a reduced dataset	D and V	Deep Learning	SwinNet Architecture	100	TrS = 90; TeS = 10	No	Split Sample Validation	Expert Annotation and Labelling (NS, > 5)	A, mIoU and F1 Score	OS A = 0.885; mIoU = 0.468; F1 Score = 0.637

Table 1 (continued)

Title	Date	Country	Objective	Study Design	AI Technique	CNN	Sample Size (OPGs)	Datase (OPGs)	Data Augmentation	Validation Method	Reference Standard (OPs, YE)	Outome Measurement	Main Outcomes
(Leite et al. 2021) [9]	2021	Belgium	To evaluate the performance of a new AI-driven tool for tooth detection and segmentation on OPGs	D and V	Deep Learning	Deeplab-v3 Architecture + ResNet-101	153	TrS = 70; VS = 18; TeS = 65	Yes	Split Sample Validation	Expert Annotation and Labelling (1, > 20)	(OD) S and P (OS) IoU, P, R and F1 Score	OD S = 0.989, *P* = 0.996 OS Mean Values: *P* = 0.958; *R* = 0.975: IoU = 0.936; F1 Score = 0.966
(Lee et al. 2020) [19]	2020	Korea	To evaluate a fully DL- CNN for automated tooth segmentation using OPGs	D and V	Deep Learning	Mask R-CNN Architecture	50	TrS = 30; VS = 10; TeS = 10	Yes	Split Sample Validation	Expert Annotation and Labelling (1, 5)	mIoU, F1 Score, P, R, and Visual Analysis	OS Mean IoU = 0.877; F1 Score = 0.875; *P* = 0.858; *R* = 0.893
(Kılıc et al. 2021) [23]	2021	Turkey	To evaluate the use of DL approach for automated detection and numbering of deciduous teeth using OPGs	D and V	Deep Learning	Faster R-CNN Inception v2 Architecture	421	TrS = 329; VS = 46; TeS = 46	No	Split Sample Validation	Expert Annotation and Labelling (1, > 10)	S, P and F1 Score	S = 0.9804; *P* = 0.9571; F1 Score = 0.9686
(Estai et al. 2022) [28]	2022	Australia	To evaluate teeth detection and classification capacity of an automated CNN using OPGs	D and V	Deep Learning	U-Net, Faster R-CNN and VGG-16	591	TrS = 531; VS/TeS = 60	No	Cross-Validation	Expert Annotation and Labelling (3, > 10)	(OD) P and R (OL) P, R, E, A and F1 Score,)	OD *P* = 0.992, *R* = 0.994 OL *P* = 0.980, *R* = 0.980, E = 0.999, A = 0.998 and F1 Score = 0.980
(Bilgir et al. 2021) [24]	2021	Poland	To verify the diagnostic performance of an AI-based system to detect and number dental structures on OPGs	D and V	Deep Learning	Faster R-CNN Inception v2 Architecture	2482	TrS = 1985; VS = 249; TeS = 248	No	Split Sample Validation	Expert Annotation and Labelling (3, > 7)	S, P, F1 Score, FDR and FNR	S = 0.955; *P* = 0.9652; F1 Score = 0.9696; FDR = 0.034 and FNR = 0.044

Table 1 (continued)

Title	Date	Country	Objective	Study Design	AI Technique	CNN	Sample Size (OPGs)	Datase (OPGs)	Data Augmentation	Validation Method	Reference Standard (OPs, YE)	Outome Measurement	Main Outcomes
(Kaya et al. 2022)15/1/24 12:22:00 [20]	2022	Turkey	To evaluate the performance of a deep learning system for automated tooth detection and numbering on pediatric OPGs	D and V	Deep Learning	YOLO v4 Architecture	4545	TrS = 4045; VS/TeS = 500	No	Split Sample Validation	Expert Annotation and Labelling (NS, NS)	P, R, F1 Score, AP, AR	OD mAP = 0.922 and mAR = 0.944 OL mAP = 0.889, mAR = 0.901; F1 Score = 0.911
(Dai et al. 2023) [29]	2023	China	To develop a fully automatic method for mesiodens localization on OPGs	D and V	Deep Learning	DMLnet based don YOLOv5 Architecture	850	TrS = 655; VS = 85; TeS = 100	No	Split Sample Validation	Expert Annotation and Labelling (2, NS)	A, P, S, E and mAP	A = 0.94, S = 0.95, E = 0.93, *P* = 0.93, mAP = 0.99
(Ha et al. 2021) [21]	2021	South Korea	To develop an artificial intelligence model that can detect mesiodens on OPGs of various dentition groups	D and V	Deep Learning	YOLOv3 Architecture	860	TrS = 551; VS = 61; TeS = 248	No	Split Sample Validation	Expert Annotation and Labelling (1, > 20)	A, S and E	A = 0.930; S = 0.915 and E = 0.943
(Ahn et al. 2021) [31]	2021	South Korea	To develop and evaluate the performance of deep-learning models to automatically detect and classify mesiodens in primary or mixed dentition on OPGs	D and V	Deep Learning	ResNet-18 (1), ResNet-101 (2), Inception ResNet-V2 (3) and SqueezeNet (4) Architectures	1100	TrS = 800; VS = 200; TeS = 100	Yes	Cross-Validation	Expert Annotation and Labelling (1, NS)	A, P, R, F1 Score, and mAP	(1) A = 0.914, *P* = 0.883, *R* = 0.958, F1 Score = 0.918 (2) A = 0.927, *P* = 0.911, *R* = 0.948, F1 Score = 0.928 (3) A = 0.924, *P* = 0.916, *R* = 0.934, F1 Score = 0.925 (4) A = 0.833, *P* = 0.779, *R* = 0.960, F1 Score = 0.855
(Kim et al. 2022) [26]	2022	South Korea	To develop and evaluate the performance of a deep-learning model that automatically detects mesiodens in OPGs of growing children	D and V	Deep Learning	DeepLabV3Plus and Inception-ResNet-v2 Architectures	988	TrS = 790; VS/Tes = 198	Yes	Cross-Validation	Expert Annotation and Labelling (1, NS)	A, P, R and F1 Score	A, P, R and F1 Score = 0.971

Table 1 (continued)

Title	Date	Country	Objective	Study Design	AI Technique	CNN	Sample Size (OPGs)	Datase (OPGs)	Data Augmentation	Validation Method	Reference Standard (OPs, YE)	Outome Measurement	Main Outcomes
(Aljabri et al. 2022) [32]	2022	Saudi Arabia	To develop Deep Learning models that classify the type of canine impaction	D and V	Deep Learning	DenseNet (1), VGG-16 (2), Inception V3 (3) and ResNet-50 (4) Architectures	EXP 1 = 416 and EXP 2 = 268	EXP 1 TrS = 332; VS/TeS = 84; EXP 2 TrS = 214; VS/TeS = 54	Yes	Split Sample Validation	Expert Annotation and Labelling (NS, NS)	A, P, R, E and F1 Score	EXP 1 (1) A = 0.797; *P* = 0.782, *R* = 0.6219, E = 0.9178, F1 Score = 0.693; (2) A = 0.654; *P* = 0.100, *R* = 0.016, E = 0.9305, F1 Score = 0.028 (3) A = 0.809; *P* = 0.714, *R* = 0.454, E = 0.935, F1 Score = 0.555 (4) A = 0.761; *P* = 0.824, *R* = 0.778, E = 0.921, F1 Score = 0.800 EXP 2 (1) A = 0.685; *P* = 0.782, *R* = 0.6674, E = 0.814, F1 Score = 0.720 (2) A = 0.574; *P* = 0.493, *R* = 0.333, E = 0.656, F1 Score = 0.398 (3) A = 0.925; *P* = 0.935, *R* = 0.935, E = 0.913, F1 Score = 0.935 (4) A = 0.870; *P* = 0.856, *R* = 0.908, E = 0.847, F1 Score = 0.881

Table 1 (continued)

Title	Date	Country	Objective	Study Design	AI Technique	CNN	Sample Size (OPGs)	Datase (OPGs)	Data Augmentation	Validation Method	Reference Standard (OPs, YE)	Outome Measurement	Main Outcomes
(Kuwada et al. 2020) [30]	2020	Japan	To verify and compare the performance of 3 deep learning systems for classifying maxillary impacted supernumerary teeth in patients with fully erupted incisors through OPGs	D and V	Deep Learning	AlexNet (1), VGG-16 (2) and DetectNet (3)	550	TrS = 400; VS = 100; TeS = 50	No	Split Sample Validation	Expert Annotation and Labelling (2, > 30)	A, S, E, ROC Curve	(1) S = 0.840; E = 0.960; A = 0.900; AUC = 0.900 (2) S = 0.440; E = 0.600; A = 0.520; AUC = 0.520 (3) S = 0.920; E = 1.000; A = 0.960; AUC = 0.960

D Development, *V* Validation, *OPs* Operators, *YE* Years of Experience, *TrS* Training Set, *VS* Validation Set, *TeS* Testing Set, *NS* Not Specified, *A* Accuracy, *P* Precision, *R* Recall, *IoU* Intersection over Union, *mAP* Mean Average Precision, *mAR* Mean Average Recall, *FDR* False Discovery Rate, *FNR* False Negative Rate, *OD* Object Detection, *OS* Object Segmentation, *OL* Object Labeling

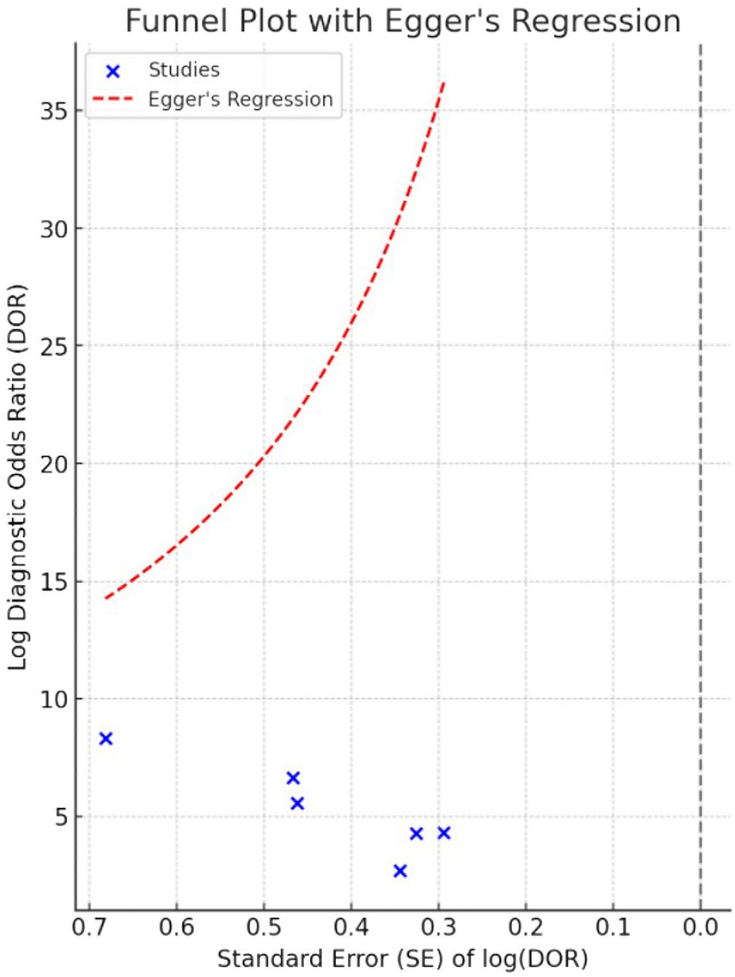


Fig. 2 Egger's Test results rating the publication bias for diagnostic odds ratio (*p* value = 0.025)

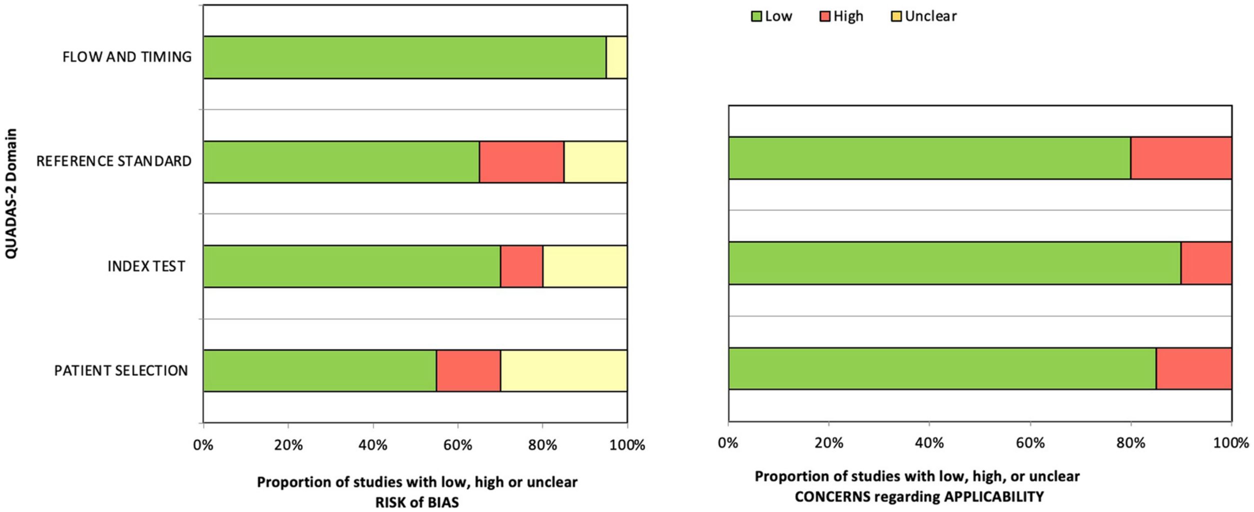


Fig. 3 Applicability concerns resulting fro the risk of bias assessment

Table 2 Overall risk of bias assessment results for each QUADAS-2 tool domain

TITLES	QUADAS-2 Domain			
	Patient Selection	Index Test	Reference Standard	Flow and Timing
Vinayahalingam et al. 2021	🟢	🟡	🟢	🟢
Bonfanti-Gris et al. 2022	🟢	🟢	🟢	🟢
Tuzoff et al. 2019	🟢	🟢	🟡	🟢
Choi et al. 2022	🟢	🟢	🟢	🟢
Mahdi et al. 2020	🟡	🟢	🔴	🟢
Yüksel et al. 2021	🟢	🟢	🟢	🟢
Chandrashekar et al. 2022	🟡	🟢	🔴	🟡
Sheng et al. 2022	🟢	🔴	🟢	🟢
Leite et al. 2021	🟢	🟢	🟡	🟢
Lee et al. 2020	🟢	🟡	🟢	🟢
Kilic et al 2021	🔴	🟢	🔴	🟢
Estai et al. 2022	🔴	🟢	🟢	🟢
Bilgir et al. 2021	🟡	🟢	🟢	🟢
Kaya et al. 2022	🔴	🔴	🔴	🟢
Dai et al. 2023	🟡	🟢	🟢	🟢
Ha et al. 2021	🟡	🟡	🟢	🟢
Ahn et al. 2021	🟢	🟢	🟢	🟢
Kim et al. 2022	🟢	🟡	🟢	🟢
Aljabri et al. 2022	🟢	🟢	🟡	🟢
Kuwada et al. 2020	🟡	🟢	🟢	🟢

Colors indicate low (green), high (red) or unclear (yellow) risk of bias

Assessment of the certainty of evidence was also conducted, and results based on risk of bias, inconsistency, indirectness, imprecision, and publication bias can be analyzed in Additional File 4. Overall, for outcome [1] authors detected a moderate risk of bias and inconsistency due to unclear description of reference standard and non-blinded evaluators to results of both alternative test and reference standard and unexplained sensitivity, specificity, and likelihood ratios, respectively.

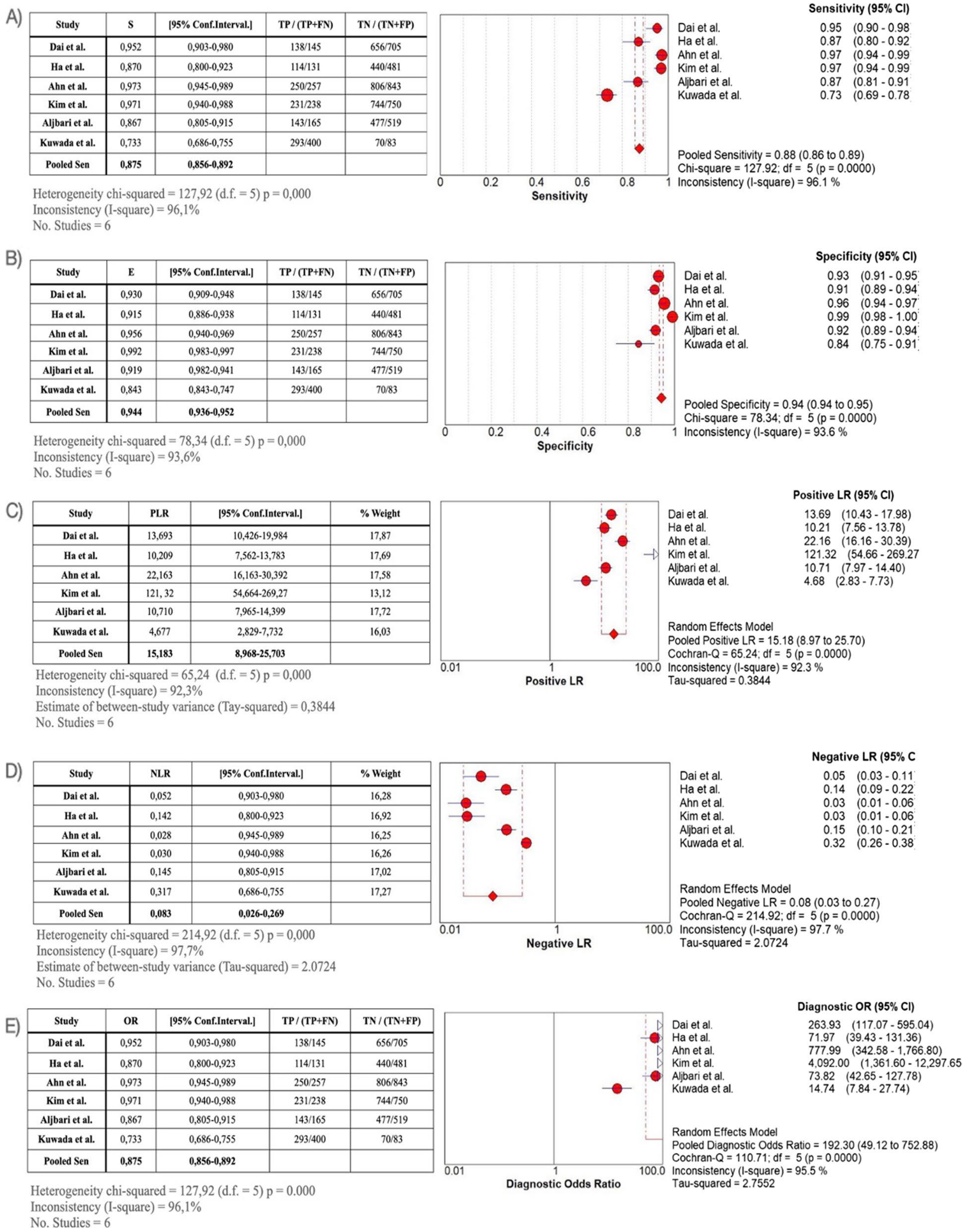
If a study was found to have a high risk of bias in a single QUADAS-2 domain, it was automatically classified as HR. If a study showed to have both a single domain with unclear risk of bias or a drawing situation between low risk and unclear risk of bias, the study was classified as UR. Only studies with all domains clear of bias were categorized as LR.

Meta-analysis

Six studies were included in the meta-analysis [21, 26, 29–32] used 14 different models (YOLOv5 Architecture, ResNet-18, ResNet-101, Inception ResNet-101, SqueezeNet, DeepLabV3Plus, Inception-ResNet-V2, DenseNet, VGG, Inception V3, ResNet-50, AlexNet, VGG-16, and DetectNet). Therefore, it was included as 6 separate studies in the meta-analysis, increasing the total number of models to 14. Pooled sensitivity, specificity, positive LR, negative LR, and diagnostic odds ratio of included studies were 0.92 (95% confidence interval [CI], 0.84–0.96), 0.94 (95% CI, 0.89–0.97), 15.7 (95% CI, 7.6–32.2), 0.08 (95% CI, 0.04–0.18), and 186 (95% CI, 44–793) respectively (Fig. 4).

A graphical summary of the meta-analysis was plotted based on sensitivity and specificity. Hierarchical Summary Receiver Operating Characteristic curve, prediction region, summary point, and confidence region were illustrated in the plot. Hierarchical Summary Receiver Operating Characteristic curve model combines study-specific estimates of sensitivity and specificity. There was a positive correlation between logit-transformed sensitivity and specificity ($r = -0.886$). The beta parameter was significant ($P = 0.019$), indicating heterogeneity (Fig. 5).

Based on the six studies, the highest sensitivity was predicted to be in Ahn et al., and Kim et al., with 97%, and the highest specificity in Kim et al., with 99%.



A)

Study	S	[95% Conf.Interval.]	TP / (TP+FN)	TN / (TN+FP)
Dai et al.	0,952	0,903-0,980	138/145	656/705
Ha et al.	0,870	0,800-0,923	114/131	440/481
Ahn et al.	0,973	0,945-0,989	250/257	806/843
Kim et al.	0,971	0,940-0,988	231/238	744/750
Aljbari et al.	0,867	0,805-0,915	143/165	477/519
Kuwada et al.	0,733	0,686-0,755	293/400	70/83
Pooled Sen	0,875	0,856-0,892		

Heterogeneity chi-squared = 127,92 (d.f. = 5) p = 0,000
Inconsistency (I-square) = 96,1%
No. Studies = 6

Sensitivity (95% CI)
Dai et al. 0.95 (0.90 - 0.98
Ha et al. 0.87 (0.80 - 0.92
Ahn et al. 0.97 (0.94 - 0.99
Kim et al. 0.97 (0.94 - 0.99
Aljbari et al. 0.87 (0.81 - 0.91
Kuwada et al. 0.73 (0.69 - 0.78
Pooled Sensitivity = 0.88 (0.86 to 0.89)
Chi-square = 127.92; df = 5 (p = 0.0000)
Inconsistency (I-square) = 96.1 %
0
0.2
0.4
0.6
0.8
1
Sensitivity

B)

Study	E	[95% Conf.Interval.]	TP / (TP+FN)	TN / (TN+FP)
Dai et al.	0,930	0,909-0,948	138/145	656/705
Ha et al.	0,915	0,886-0,938	114/131	440/481
Ahn et al.	0,956	0,940-0,969	250/257	806/843
Kim et al.	0,992	0,983-0,997	231/238	744/750
Aljbari et al.	0,919	0,982-0,941	143/165	477/519
Kuwada et al.	0,843	0,843-0,747	293/400	70/83
Pooled Sen	0,944	0,936-0,952		

Heterogeneity chi-squared = 78,34 (d.f. = 5) p = 0,000
Inconsistency (I-square) = 93,6%
No. Studies = 6

Specificity (95% CI)
Dai et al. 0.93 (0.91 - 0.95
Ha et al. 0.91 (0.89 - 0.94
Ahn et al. 0.96 (0.94 - 0.97
Kim et al. 0.99 (0.98 - 1.00
Aljbari et al. 0.92 (0.89 - 0.94
Kuwada et al. 0.84 (0.75 - 0.91
Pooled Specificity = 0.94 (0.94 to 0.95)
Chi-square = 78.34; df = 5 (p = 0.0000)
Inconsistency (I-square) = 93.6 %
0
0.2
0.4
0.6
0.8
1
Specificity

C)

Study	PLR	[95% Conf.Interval.]	% Weight
Dai et al.	13,693	10,426-19,984	17,87
Ha et al.	10,209	7,562-13,783	17,69
Ahn et al.	22,163	16,163-30,392	17,58
Kim et al.	121, 32	54,664-269,27	13,12
Aljbari et al.	10,710	7,965-14,399	17,72
Kuwada et al.	4,677	2,829-7,732	16,03
Pooled Sen	15,183	8,968-25,703	

Heterogeneity chi-squared = 65,24 (d.f. = 5) p = 0,000
Inconsistency (I-square) = 92,3%
Estimate of between-study variance (Tay-squared) = 0,3844
No. Studies = 6

Positive LR (95% CI)
Dai et al. 13.69 (10.43 - 17.98)
Ha et al. 10.21 (7.56 - 13.78)
Ahn et al. 22.16 (16.16 - 30.39)
Kim et al. 121.32 (54.66 - 269.27
Aljbari et al. 10.71 (7.97 - 14.40)
Kuwada et al. 4.68 (2.83 - 7.73)
Random Effects Model
Pooled Positive LR = 15.18 (8.97 to 25.70)
Cochran-Q = 65.24; df = 5 (p = 0.0000)
Inconsistency (I-square) = 92.3 %
Tau-squared = 0.3844
0.01
1
100.0
Positive LR

D)

Study	NLR	[95% Conf.Interval.]	% Weight
Dai et al.	0,052	0,903-0,980	16,28
Ha et al.	0,142	0,800-0,923	16,92
Ahn et al.	0,028	0,945-0,989	16,25
Kim et al.	0,030	0,940-0,988	16,26
Aljbari et al.	0,145	0,805-0,915	17,02
Kuwada et al.	0,317	0,686-0,755	17,27
Pooled Sen	0,083	0,026-0,269	

Heterogeneity chi-squared = 214,92 (d.f. = 5) p = 0,000
Inconsistency (I-square) = 97,7%
Estimate of between-study variance (Tau-squared) = 2.0724
No. Studies = 6

Negative LR (95% C
Dai et al. 0.05 (0.03 - 0.11
Ha et al. 0.14 (0.09 - 0.22
Ahn et al. 0.03 (0.01 - 0.06
Kim et al. 0.03 (0.01 - 0.06
Aljbari et al. 0.15 (0.10 - 0.21
Kuwada et al. 0.32 (0.26 - 0.38
Random Effects Model
Pooled Negative LR = 0.08 (0.03 to 0.27)
Cochran-Q = 214.92; df = 5 (p = 0.0000)
Inconsistency (I-square) = 97.7 %
Tau-squared = 2.0724
0.01
1
100.0
Negative LR

E)

Study	OR	[95% Conf.Interval.]	TP / (TP+FN)	TN / (TN+FP)
Dai et al.	0,952	0,903-0,980	138/145	656/705
Ha et al.	0,870	0,800-0,923	114/131	440/481
Ahn et al.	0,973	0,945-0,989	250/257	806/843
Kim et al.	0,971	0,940-0,988	231/238	744/750
Aljbari et al.	0,867	0,805-0,915	143/165	477/519
Kuwada et al.	0,733	0,686-0,755	293/400	70/83
Pooled Sen	0,875	0,856-0,892		

Heterogeneity chi-squared = 127,92 (d.f. = 5) p = 0.000
Inconsistency (I-square) = 96,1%
No. Studies = 6

Diagnostic OR (95% CI)
Dai et al. 263.93 (117.07 - 595.04)
Ha et al. 71.97 (39.43 - 131.36)
Ahn et al. 777.99 (342.58 - 1,766.80)
Kim et al. 4,092.00 (1,361.60 - 12,297.65
Aljbari et al. 73.82 (42.65 - 127.78)
Kuwada et al. 14.74 (7.84 - 27.74)
Random Effects Model
Pooled Diagnostic Odds Ratio = 192.30 (49.12 to 752.88)
Cochran-Q = 110.71; df = 5 (p = 0.0000)
Inconsistency (I-square) = 95.5 %
Tau-squared = 2.7552
0.01
1
100.0
Diagnostic Odds Ratio

Fig. 4 Summary performance estimates for (**A**) Sensitivity (S), (**B**) Specificity (E), (**C**) Positive Likelihood Ratio (PLR), (**D**) Negative Likelihood Ratio (NLR) and (**E**) Odds Ratio (OR)

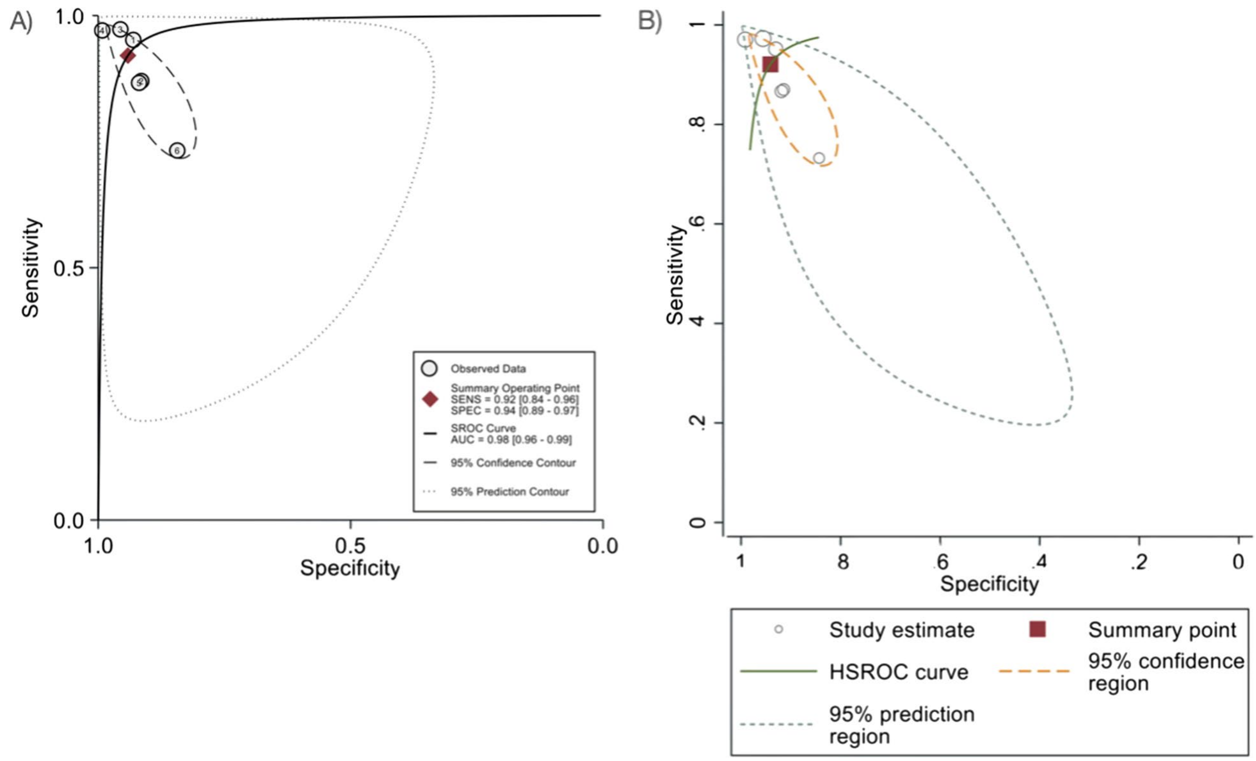


Fig. 5 **A** Summary ROC curve with confidence and prediction regions around mean operating sensitivity and specificity point, and (**B**) Results of meta-analysis on Sensitivity and Specificity of the deep learning models for tooth detection and segmentation in panoramic radiographs and the Hierarchical Summary Receiver Operating Characteristic (HSROC) curve. *"Study estimate" represents the estimates of each of the included studies, which includes sensitivity, specificity, and sample size (shown by the size of diameter). "Summary point" represents an estimated summary of pooled sensitivity and specificity of the included studies. "95% confidence region" represents the area where the actual pooled summary point was anticipated to be. "95% prediction region" represents a forecasted confidence region for the sensitivity and specificity of future studies*

Overall, based on the included studies the sensitivity was established as 88% and specificity as 94%.

Discussion

Artificial Intelligence-based models are increasingly being explored in dentistry to enhance diagnostic accuracy, particularly while using orthopantomographies – which are fundamental for pathology identification. Nevertheless, these images frequently present superimpositions and deformations that pose unavoidable challenges for training neural networks to perform tasks like object detection (OD) and segmentation (OS). Thus, although human diagnostic capabilities might surpass computer-based ones, it is thought that deep learning models could be a potential solution to refine them bearing in mind inherent limitations – especially when considering inexperienced operators. However, the diagnostic power of these models should be interpreted with caution, as biases and performance variability across different architectures and datasets can impact their generalizability [9].

When referring to image interpretation and Computer Vision (CV), it is crucial to discern between different tasks concerning object identification. This way, OD and OS are the predominant terms used in contemporary literature – with OS further categorized into Semantic Segmentation (SS) and Instance Segmentation (IS). SS involves creating pixel-level boundaries around objects of a specific category (*e.g.* dental structures segmented individually but all of them categorized as tooth), while IS distinguishes individual instances within a category (*e.g.* dental structures segmented individually being categorized as tooth but also specifying their tooth number) [1].

Out of the twenty studies included in the present review, only two did not evaluate OD performance [18, 19]. Among those that did, four addressed tooth labeling but did not specify performance metrics for this task [16, 22–24]. Most studies employing both OD and labeling methodologies utilized two-stage CNNs, except for three studies that used a single-stage CNN [16, 22, 24]. OS methodologies varied, as one study utilized SS [18], three used IS [9, 19, 25] and another incorporated both [4] to enhance the CNN's diagnostic performance. This was also performed by two other studies, but results were not reported [17, 26].

Establishing a reliable ground truth (GT) is crucial for evaluating deep learning models. Seven studies reported having multiple operators performing manual annotation and labeling [15, 24, 25, 27–30], while eight relied on a single clinician. Five did not specify the number of practitioners involved. Operator experience was often unreported, but varied from 5 to 30 years. Notably, studies focused on mesiodens detection with a single operator utilized Cone Beam Computed Tomography (CBCT) as GT [21, 26, 31, 32] – which may be a source of potential biases.

Sample dataset division is another key factor in model evaluation. Eleven studies correctly followed the recommended protocol of creating three independent sets of images for training (TrS), validation (VS), and testing (TeS) [4, 9, 19, 21–25, 29–31]. However, eight studies did not follow this protocol [16–18, 20, 26–28, 32], and one only included a TeS, as the study analyzed the diagnostic performance of a pre-trained and validated CNN CNN [15]. Only one study used a publicly available dataset [4]. Concerns about generalizability were raised in studies that did not include different world populations nor sets from different institutions or acquired with various X-ray machines. This way, only four studies reported using assorted testing sets [9, 18, 22, 25]. Nonetheless, studies with poor datasets tried to reduce overfitting by performing cross-validation techniques [26, 28, 31] or implementing data augmentation processes [4, 9, 17, 19, 22, 26, 27, 31, 32].

Metrics used to evaluate the performance of DL models vary within the included studies, being precision, recall and F1 Score the most frequently reported. Other pixel-based metrics such as IoU were also included. However, performance should be interpreted cautiously due to variability in dataset quality and study design. Vinayahalingam et al. published exceptional results for OD and OS, with OD precision of 0.997, recall of 0.989, and F1 Score of 0.992. While OS also achieved great results, certain limitations to this study were found, since blurred or incomplete OPGs were excluded from the dataset [25]. Similarly, Choi et al. achieved impressive average precision and recall results of 0.991 and 0.996 respectively, but the exclusion of images with primary and mixed dentition, impacted teeth or partially edentulous patients make their results not generalizable and highly biased [22].

Inconsistencies in reporting were evident when comparing studies evaluating similar tasks. When evaluating the same NN, Tuzoff et al. reported high sensitivity and precision for OD (0.9941 and 0.9945, respectively), while Bonfanti-Gris et al. reported a lower sensitivity value for this same task (0.693). Similarly, discrepancies were found for labeling task, with Tuzoff et al. reporting sensitivity and specificity of 0.980 and 0.999 and Bonfanti-Gris et al. reporting 0.500 for both [15, 27]. Other studies were also found to report poor results. Yüksel et al., observed a mean average precision of 0.477 for object detection adjusted to a threshold of 0.5–0.95. Only when this was lowered to 0.5, the model showed a maximum precision of 0.894 [17]. Nevertheless, contrarily to what was found by Bonfanti-Gris et al. with a reduced dataset, Leite et al. obtained great results both for OD and OS (S $=0.989$, $P=$ 0.996 and $P=$ 0.958, $R=$ 0.975, IoU $=0.936$ and F1 Score $=0.966$, respectively) [9]. This way, a threshold effect was observed but was not explicitly discussed, which leaves a gap of information that should be considered when applying these models in clinical settings, as decision-varying thresholds may significantly impact both sensitivity and specificity.

Results for reduced sample sizes without data augmentation techniques were also reported by Kilic et al., who achieved S= 0.9804, $P=$ 0.9571 and F1 Score $=0.9686$ outcomes for object detection and labeling [23]. Estai et al., also reported favorable outcomes for OD ($P=0.992$ and $R=$ 0,994) and labeling tasks (P, R and F1Score $=0.980$, E= 0.999 and A= 0.998) using three different NN for each of them. Nevertheless, the use of a two-set of images instead of three could have introduced a risk of bias factor [28].

Contrary to the previous, both Bilgir et al., and Kaya et al. reported remarkable results using a single CNN for OD. In the first case, authors reported high sensitivity, precision, F1 Score, False Discovery Rate and False Negative Rate, while Kaya et al. demonstrated excellent results with mAP and mAP metrics [20, 24].

Two distinct DL approaches were explored within the included studies. Mahdi et al., utilized an Optimization Technique based on Transfer Learning, showcasing positive results with CNNs like ResNet-101 and ResNet-105. Chandrashekar et al., instead, introduced a collaborative learning approach in which two DL models were integrated to obtain better results. In this case, authors compared the studied CNNs performance metrics both individually and while collaborating, obtaining – for both OD and OS – higher accuracy, F1 Score, and mAP results with the latest (> 0.973).

Finally, even studies focusing solely on OS tasks presented varying results. Sheng et al., reported accuracy values of 0.885, mean IoU of 0.468, and a F1Score of 0.637 [18]. Nevertheless, Lee et al., achieved better performance metrics while using a significantly reduced dataset and implementing data augmentation techniques – IoU $=0.877$, F1 Score $=0.875$, $P=$ 0.858, and $R=$ 0.893 [19].

When comparing the results obtained from different neural networks, the depth of the CNN should be taken into consideration, as this can affect model performance. It has been reported that deeper architectures improve accuracy but risk overfitting – especially in small datasets [4, 18]. Although data augmentation techniques might mitigate this, increasing model complexity does not always yield proportional accuracy improvements. Also, while some architectures may perform well with specific dataset sizes, others may suffer from overfitting or diminishing returns [18]. Thus, the absence of a standardized framework for selecting optimal depth and learning parameters limits the comparability and reproducibility of results [33].

Deep Learning OD and OS models have also been reported to accurately perform impacted-tooth identification. This systematic review localized six titles in which this objective was tackled by evaluating several CNNs' mesiodens identification and classification capacities. Overall, results were impressive, as Dai et al., reported A, S, E, P and mAP results of 0.94, 0.95, 0.93, 0.93, and 0.99, respectively [29]. Similarly, Ha et al., obtained outcomes from 0.915 to 0.043 for A, S, and E with a resemblant sample size dataset [21].

When comparing different CNNs, Kuwada et al. observed that DetectNet outperformed AlexNet and VGG-16, with sensitivity, specificity and accuracy values of 0.920, 1.000 and 0.960, respectively [30]. Other studies reported similar outcomes for architectures like ResNet-18, ResNet-101, Inception Resnet-V2 and SqueezeNet [31]. Variability within the outcomes was detected by Aljabri et al. while analyzing four different DL models and studying their performance on two different sample sizes experiments. Overall, worst results were observed for VGG-16 architecture [32].

Kim et al., achieved outstanding results by employing a novel OS technique to restrict the maxillary anterior region, enhancing detection accuracy for mesiodens. However, the study excluded images with distortion and blurriness, so generalizability was not ensured [34].

DL models have also been employed in dentistry for detecting ectopic eruption of maxillary first molars [35] and classifying mandibular third molar positioning [36, 37]. Object segmentation automation is crucial for digital applications, especially in 3D imagery, where manual segmentation is labor-intensive and skill-dependent. This can be particularly relevant for treatment planning, addressing intra-operative complications and planning auto-transplantations [2].

Although AI-based applications have been widely studied, clinical implications of DL models warrant further discussion. While models performed will in tooth detection and segmentation, practical challenges remain – such as the need for standardized training data, external validation and regulatory approvals before implementation in clinical practice. Additionally, model interpretability and clinicians' trust in AI-generated reports must be addressed.

Despite promising results, the systematic review highlights several limitations. First, including studies within a limited timeframe and only focusing on DL methods could be considered as liabilities.

Based on the reviewed data, future research should prioritize diverse and generalizable datasets, incorporate multicenter images to address overfitting and adopt standardized reference tests and reporting guidelines, such as STARD-AI and the CLAIM Checklist. These steps will enhance research quality, robustness and reliability in AI-based diagnostic tools for dentistry.

Conclusions

Basic AI-based object detection and segmentation tasks on OPGs are considered the pillar for the later application of this novel tool for pathology identification. If these tasks are not correctly performed, more complex milestones will not be successfully achieved by means of Deep Learning. Already published literature have demonstrated excellent performance of DL-CNN for this matter, although its clinical application remains premature due to significant limitations in study design, data generalizability and external validation. Thus, the use of AI tools in clinical practice should be approached cautiously. Future research should focus on large-scale, multicenter trials to validate findings across diverse populations and imaging modalities, addressing dataset biases and developing standardized benchmarks to enhance the generalizability and reliability of AI models in dental imaging.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-025-06349-9.

Additional file 1. Search execution strategy followed on PubMed, Embase, and Cochrane databases.

Additional file 2. QUADAS-2 original and modified (*) leading questions for critical appraisal.

Additional file 3. QUADAS-2 Tool Risk of Bias Question Results for each article included in the Systematic Review. NA = Not Applicable; NR = Does Not Refer, Unclear.

Additional file 4. Certainty of Evidence Assessment using GRADE Approach. VL: Very Low, L: Low, M: Moderated, H: High, A: Accuracy, P: Precision, R: Recall.

Acknowledgements

Not applicable.

Authors' contributions

Author contributions were as follows: M.B-G. contributed to conception, design, data acquisition and interpretation, and drafted and revised the manuscript; A.H., contributed to data acquisition and interpretation, and critically revised the manuscript; MP. S. contributed to conception and revision of the manuscript; F. M-R. contributed to conception and revision of the manuscript; and G.P. contributed to design, data acquisition and interpretation and critically revised the manuscript.

Funding

Not applicable.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 June 2024 Accepted: 3 June 2025
Published online: 30 July 2025

References

1. Umer F, Habib S, Adnan N. Application of deep learning in teeth identification tasks on panoramic radiographs. Dento Maxillo Facial Radiol. 2022;51(5):20210504.
2. Mureşanu S, Almăşan O, Hedeşiu M, Dioşan L, Dinu C, Jacobs R. Artificial intelligence models for clinical usage in dentistry with a focus on dentomaxillofacial CBCT: a systematic review. Oral Radiol. 2022.
3. Celik ME. Deep learning based detection tool for impacted mandibular third molar teeth. Diagn Basel Switz. 2022;12(4):942.
4. Chandrashekar G, AlQarni S, Bumann EE, Lee Y. Collaborative deep learning model for tooth segmentation and identification using panoramic radiographs. Comput Biol Med. 2022;148:105829.
5. Jiang L, Chen D, Cao Z, Wu F, Zhu H, Zhu F. A two-stage deep learning architecture for radiographic staging of periodontal bone loss. BMC Oral Health. 2022;22(1):106.
6. Sukegawa S, Yoshii K, Hara T, Matsuyama T, Yamashita K, Nakano K, et al. Multi-task deep learning model for classification of dental implant brand and treatment stage using dental panoramic radiograph images. Biomolecules. 2021;11(6):815.
7. Calazans MAA, Ferreira FABS, Alcoforado M de LMG, Santos AD, Pontual ADA, Madeiro F. Automatic classification system for periapical lesions in cone-beam computed tomography. Sensors. 2022;22(17):6481.
8. Bayrakdar IS, Orhan K, Akarsu S, Çelik Ö, Atasoy S, Pekince A, et al. Deep-learning approach for caries detection and segmentation on dental bitewing radiographs. Oral Radiol. 2022;38(4):468–79.
9. Leite AF, Gerven AV, Willems H, Beznik T, Lahoud P, Gaêta-Araujo H, et al. Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs. Clin Oral Investig. 2021;25(4):2257–67.
10. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, and the PRISMA-DTA Group, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA. 2018;319(4):388.
11. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529–36.
12. Mohammad-Rahimi H, Motamedian SR, Rohban MH, Krois J, Uribe SE, Mahmoudinia E, et al. Deep learning for caries detection: a systematic review. J Dent. 2022;122:104115.
13. Granholm A, Alhazzani W, Møller MH. Use of the GRADE approach in systematic reviews and guidelines. Br J Anaesth. 2019;123(5):554–9.
14. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336:1106.
15. Bonfanti-Gris M, Garcia-Cañas A, Alonso-Calvo R, Salido Rodriguez-Manzaneque MP, Pradies RG. Evaluation of an Artificial Intelligence web-based software to detect and classify dental structures and treatments in panoramic radiographs. J Dent. 2022;126:104301.
16. Mahdi FP, Motoki K, Kobashi S. Optimization technique combined with deep learning method for teeth recognition in dental panoramic radiographs. Sci Rep. 2020;10(1):19261.
17. Yüksel AE, Gültekin S, Simsar E, Özdemir ŞD, Gündoğar M, Tokgöz SB, et al. Dental enumeration and multiple treatment detection on panoramic X-rays using deep learning. Sci Rep. 2021;11(1):12342.
18. Sheng C, Wang L, Huang Z, Wang T, Guo Y, Hou W, et al. Transformer-based deep learning network for tooth segmentation on panoramic radiographs. J Syst Sci Complex. 2022;14:1–16.
19. Lee JH, Han SS, Kim YH, Lee C, Kim I. Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. Oral Surg Oral Med Oral Pathol Oral Radiol. 2020;129(6):635–42.
20. Kaya E, Gunec HG, Gokyay SS, Kutal S, Gulum S, Ates HF. Proposing a CNN method for primary and permanent tooth detection and enumeration on pediatric dental radiographs. J Clin Pediatr Dent. 2022;46(4):293–8.
21. Ha EG, Jeon KJ, Kim YH, Kim JY, Han SS. Automatic detection of mesiodens on panoramic radiographs using artificial intelligence. Sci Rep. 2021;11(1):23061.
22. Choi HR, Siadari TS, Kim JE, Huh KH, Yi WJ, Lee SS, et al. Automatic detection of teeth and dental treatment patterns on dental panoramic radiographs using deep neural networks. Forensic Sci Res. 2022;7(3):456–66.
23. Kılıc MC, Bayrakdar IS, Çelik Ö, Bilgir E, Orhan K, Aydın OB, et al. Artificial intelligence system for automatic deciduous tooth detection and numbering in panoramic radiographs. Dento Maxillo Facial Radiol. 2021;50(6):20200172.
24. Bilgir E, Bayrakdar İŞ, Çelik Ö, Orhan K, Akkoca F, Sağlam H, et al. An artificial ıntelligence approach to automatic tooth detection and numbering in panoramic radiographs. BMC Med Imaging. 2021;21(1):124.
25. Vinayahalingam S, Goey RS, Kempers S, Schoep J, Cherici T, Moin DA, et al. Automated chart filing on panoramic radiographs using deep learning. J Dent. 2021;115:103864.
26. Kim J, Hwang JJ, Jeong T, Cho BH, Shin J. Deep learning-based identification of mesiodens using automatic maxillary anterior region estimation in panoramic radiography of children. Dento Maxillo Facial Radiol. 2022;51(7):20210528.
27. Tuzoff DV, Tuzova LN, Bornstein MM, Krasnov AS, Kharchenko MA, Nikolenko SI, et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. Dento Maxillo Facial Radiol. 2019;48(4):20180051.
28. Estai M, Tennant M, Gebauer D, Brostek A, Vignarajan J, Mehdizadeh M, et al. Deep learning for automated detection and numbering of permanent teeth on panoramic images. Dento Maxillo Facial Radiol. 2022;51(2):20210296.
29. Dai X, Jiang X, Jing Q, Zheng J, Zhu S, Mao T, et al. A one-stage deep learning method for fully automated mesiodens localization on panoramic radiographs. Biomed Signal Process Control. 2023;80((Dai X.; Jiang X.; Zheng J.; Zhu S.; Mao T.) School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing, China). Available from: https://www.embase.com/search/results?subaction=viewrecord&id=L2020763375&from=export.
30. Kuwada C, Ariji Y, Fukuda M, Kise Y, Fujita H, Katsumata A, et al. Deep learning systems for detecting and classifying the presence of impacted supernumerary teeth in the maxillary incisor region on panoramic radiographs. Oral Surg Oral Med Oral Pathol Oral Radiol. 2020;130(4):464–9.
31. Ahn Y, Hwang JJ, Jung YH, Jeong T, Shin J. Automated mesiodens classification system using deep learning on panoramic radiographs of children. Diagn Basel Switz. 2021;11(8):1477.

32. Aljabri M, Aljameel SS, Min-Allah N, Alhuthayfi J, Alghamdi L, Alduhailan N, et al. Canine impaction classification from panoramic dental radiographic images using deep learning models. Inform Med Unlocked. 2022;30((Aljabri M., mssjabri@uqu.edu.sa) Computer Science Department, College of Computer and Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia). Available from: https://www.embase.com/search/results?subaction=viewrecord&id=L2017315174&from=export.
33. Khanagar SB, Al-ehaideb A, Maganur PC, Vishwanathaiah S, Patil S, Baeshen HA, et al. Developments, application, and performance of artificial intelligence in dentistry – a systematic review. J Dent Sci. 2021;16(1):508–22.
34. Kim YH, Jeon KJ, Lee C, Choi YJ, Jung HI, Han SS. Analysis of the mandibular canal course using unsupervised machine learning algorithm. PLoS ONE. 2021;16(11):e0260194.
35. Liu J, Liu Y, Li S, Ying S, Zheng L, Zhao Z. Artificial intelligence-aided detection of ectopic eruption of maxillary first molars based on panoramic radiographs. J Dent. 2022;125:104239.
36. Choi E, Lee S, Jeong E, Shin S, Park H, Youm S, et al. Artificial intelligence in positioning between mandibular third molar and inferior alveolar nerve on panoramic radiography. Sci Rep. 2022;12(1):2456.
37. Sukegawa S, Matsuyama T, Tanaka F, Hara T, Yoshii K, Yamashita K, et al. Evaluation of multi-task learning in deep learning-based positioning classification of mandibular third molars. Sci Rep. 2022;12(1):684.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.