BMC Oral Health

SYSTEMATIC REVIEW

Open Access

Performance of artificial intelligence on cervical vertebral maturation assessment: a systematic review and meta-analysis

Termeh Sarrafan Sadeghi[1], Seyed AmirHossein Ourang[1], Fatemeh Sohrabniya[1], Soroush Sadr[2], Parnian Shobeiri[3] and Saeed Reza Motamedian[1,4*]

Abstract

Background Artificial intelligence (AI) methods, including machine learning and deep learning, are increasingly applied in orthodontics for tasks like assessing skeletal maturity. Accurate timing of treatment is crucial, but traditional methods such as cervical vertebral maturation (CVM) staging have limitations due to observer variability and complexity. AI has the potential to automate CVM assessment, enhancing reliability and user-friendliness**.** This systematic review and meta-analysis aimed to evaluate the overall performance of artificial intelligence (AI) models in assessing cervical vertebrae maturation (CVM) in radiographs, when compared to clinicians.

Methods Electronic databases of Medline (via PubMed), Google Scholar, Scopus, Embase, IEEE ArXiv and MedRxiv were searched for publications after 2010, without any limitation on language. In the present review, we included studies that reported AI models' performance on CVM assessment. Quality assessment was done using Quality assessment and diagnostic accuracy Tool-2 (QUADAS-2). Quantitative analysis was conducted using hierarchical logistic regression for meta-analysis on diagnostic accuracy. Subgroup analysis was conducted on different AI subsets (Deep learning, and Machine learning).

Results A total of 1606 studies were screened of which 25 studies were included. The performance of the models was acceptable. However, it varied based on the methods employed. Eight studies had a low risk of bias in all domains. Twelve studies were included in the meta-analysis and their pooled values for sensitivity, specificity, positive and negative likelihood ratios, and diagnostic odds ratio (DOR) were calculated for each cervical stage (CS). The most accurate CVM evaluation was observed for CS1, boasting a sensitivity of 0.87, a specificity of 0.97, and a DOR of 213. Conversely, CS3 exhibited the lowest performance with a sensitivity of 0.64, and a specificity of 0.96, yet maintaining a DOR of 32.

Conclusion AI has demonstrated encouraging outcomes in CVM assessment, achieving notable accuracy.

Keywords Artificial intelligence, Growth and development, Cervical vertebrae, Orthodontics, Computer algorithm

*Correspondence:
Saeed Reza Motamedian
drmotamedian@gmail.com
Full list of author information is available at the end of the article

Background

Artificial intelligence (AI) refers to the ability of machines to perform tasks that typically require human intelligence, such as learning, problem-solving, and decision-making [1]. Within AI, machine learning (ML) represents a key subset, enabling systems to learn from data and improve performance over time without explicit programming [2]. Building on the foundation of ML, deep learning (DL) leverages artificial neural networks (ANNs), particularly deep neural networks with multiple layers. These layers process data in increasingly abstract ways and enable models to automatically learn complex patterns and representations [3]. In contrast, rule-based AI employs explicitly defined rules crafted by human experts to solve problems. While this approach does not rely on data-driven insights, it demands significant expertise and is challenging to maintain or adapt as datasets evolve [4]. Statistical Modeling is the process of analyzing data using statistical methods to provide insight into the importance and relationships between independent and dependent variables [5]. DL eliminates the need for manual feature engineering by learning hierarchical representations of data through multiple layers [6]. Traditional AI methods, such as symbolic reasoning and rule-based systems, rely on predefined rules and logical operations to process inputs and generate outputs. These systems are explicitly programmed by human experts to perform specific tasks. Unlike rule-based AI, DL can discover patterns and representations directly from data without human intervention [3].

Deep learning has shown significant advancements in orthodontics by automating diagnosis, treatment planning, and monitoring outcomes [7, 8]. It can be utilized for: cephalometric landmark detection [9], diagnosis of malocclusion and skeletal patterns [10], detection and segmentation of teeth from dental images [11] and prediction of orthodontic treatment outcomes [12].

Timing is a fundamental element of orthodontic treatment. A patient's skeletal parameters change in the sagittal, transversal, and vertical planes due to growth and development [13]. The timing of treatment onset may be as important as selecting a specific treatment protocol. Treatment procedures should be initiated at the appropriate developmental stage for the patient to achieve the most favorable outcome with the most minor potential morbidity [14]. As in the case of growth modification treatment for mandible deficiency, it is recommended that this treatment should be started before the growth spurt [15].

Several biological indicators can be used to determine skeletal maturity, including the height of the person [16], the hand and wrist development [17], the eruption and development of the teeth [18], menarche or the transformation of the voice [19], and the cervical vertebrae maturation (CVM) [20]. In the past, radiographs taken of the wrist were the gold standard for measuring growth periods [13]. The use of hand-wrist radiographs in an orthodontic practice is discouraged since it requires additional radiographic examinations. CVM staging has been increasing in popularity among orthodontists since it can be assessed on lateral cephalograms required for orthodontic diagnosis. In lateral cephalograms, CVM is assessed based on morphological changes of the second, third, and fourth vertebrae (C2, C3, and C4) [14, 21]. However, a suboptimal intraobserver agreement exists with the CVM degree method [22, 23]. According to some studies, the CVM method is not reliable or reproducible since observers do not agree with the results. In this regard, clinicians who lack technical knowledge and experience may have difficulty using the CVM method [24].

In addition to not being user-friendly, the CVM method requires experienced practitioners to implement it. The computerized cephalometric analysis does not incorporate the conventional visual assessment of CVM stages [25]. If we can diagnose the developmental stage fully automatically and with few errors, we can overcome these limitations and utilize it as a diagnostic clinical aid in orthodontic practice.

A considerable degree of dynamic development is observed in the field of AI in CVM assessment, resulting in substantial variations in methods and results among studies. Additionally, the robustness of the body of evidence remains uncertain. Hence, we reviewed and appraised studies using AI (DL, ML, statistical modeling and rule-based AI) for CVM assessment and compared their performance measures, including accuracy, specificity, and sensitivity.

Methods

Protocol

The study protocol was registered at PROSPERO (https://www.crd.york.ac.uk/prospero/; trial registration number: PROSPERO CRD42022374636) and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of diagnostic test accuracy (PRISMA-DTA) guidelines [26]. Specifically, the review question was based on PICO (P = population, I = intervention, C = comparison, and O = outcome): "What is the diagnostic performance (O) of AI (I) for CVM assessment via radiographs (P) compared to clinician's diagnosis (C)?".

Inclusion criteria

The following inclusion criteria were utilized for selecting the studies:

P: Studies employed AI approaches on radiography for CVM analysis.
I, C: AI models for classification tasks, compared with a reference test
O: Any kind of model performance evaluation

The exclusion criteria were: Non-human studies, studies that used AI for conditions not related to the CVM, studies that did not mention details of the study samples, studies without a clear definition of the AI model, studies that did not report quantitative outcomes for AI-based CVM staging, review articles, conference papers and studies with 5-stage classification for CVM.

Information sources and search

An electronic literature search was conducted in December 2024 using the following databases: Medline (via PubMed), Google Scholar, Scopus, Embase, IEEE, ArXiv and MedRxiv. No language restrictions were applied. The specific search strategies utilized for each database, along with the number of records retrieved, are detailed in Table 1. Reference lists of eligible articles were handsearched to identify any relevant studies missed by the database searches. Retrieved citations were imported into EndNote X9 (Clarivate Analytics, Philadelphia, PA, USA) to facilitate removal of duplicate records, study screening, and overall management of the assembled literature.

Selection of sources of evidence

Title and abstract screening were performed independently by two reviewers (T.S.S, F.S) against predefined inclusion and exclusion criteria. In cases where the information provided in the title and abstract was insufficient for making an eligibility determination, the full text of the article was retrieved for review. Any disagreements between reviewers regarding study inclusion were resolved through consensus discussion with a third reviewer (S.S).

Data charting process and items

Two reviewers (T.S.S, SAH.O) independently completed the data charting process using a standardized form to extract key parameters from each included study. Extracted items comprised: bibliographic details (name of authors and the year of publication, data modality, data size (train/valid/test if available), inclusion and exclusion criteria, type of AI subset, labeling procedure, preprocessing procedure, augmentation, model structure, hardware, performance measurement, and outcome (Table 2). Any discrepancies in data charting between the two reviewers were resolved through discussion and consensus with a third reviewer (S.S).

Table 1 Search queries and results in each database

Data Base	Search Query	Results	Date
PubMed (via Medline)	(automat* OR "Artificial Intelligence" OR "artificial intelligence"[MeSH Terms] OR "deep learning" OR "deep learning"[Mesh Terms] OR "machine learning" OR "machine learning"[MeSH Terms] OR "convolutional neural network" OR "convolutional neural network"[MeSH Terms] OR "artificial neural network" OR "artificial neural network"[MeSH Terms] OR "neural network" OR "computer vision" OR "Image processing" OR "Computer-Assisted" OR CNN OR ANN OR DL OR ML OR AI) AND ((cervical AND vertebra* AND stag*) OR "skeletal maturation")	279	December 2024
Google Scholar	(("artificial intelligence" OR "deep learning" OR "machine learning" OR "image processing" OR "neural network" OR "convolutional neural network" OR "artificial neural network") AND ("cervical vertebral") AND ("skeletal maturation"))	252	December 2024
Embase	("artificial intelligence" OR AI OR "deep learning" OR DL OR "machine learning" OR ML OR "image processing" OR "neural network" OR NN OR " Convolutional neural network" OR CNN) AND "cervical vertebral" AND ("stage" OR "maturation")	38	December 2024
Scopus	ALL(("artificial intelligence" OR AI OR "deep learning" OR DL OR "machine learning" OR ML OR "artificial neural network" OR ANN OR "convolutional neural network" OR CNN OR "image processing" OR "neural network" OR NN) AND "cervical vertebral" AND ("stage" OR "maturation"))	662	December 2024
ARxiv	all="cervical vertebra" OR CVM OR "vertebral maturation" OR "cervical maturation" OR "skeletal maturation"; AND all="artificial intelligence" OR "deep learning" OR "machine learning" OR "artificial neural network" "computer vision" OR "convolutional neural network" OR "neural network" OR AI OR ML OR DL OR NN OR CNN OR ANN	28	December 2024
medRxiv	""cervical vertebral" AND ("deep learning" OR "machine learning" OR "neural network") AND ("stage" OR "maturation")"	165	December 2024
IEEE	(automat* OR "Artificial Intelligence" OR AI OR "deep learning" OR DL OR "machine learning" OR ML OR "convolutional neural network" OR CNN OR "artificial neural network" OR ANN OR "neural network" OR NN OR "computer vision" OR "Image processing" OR Computer-Assisted) AND ((cervical AND vertebra* AND stag*) OR "skeletal maturation" OR "bone age")	182	December 2024

Table 2 Summary of the included studies

Author, year	Data modality	Data set size (train/valid/test)	Inclusion and exclusion criteria (if any)	AI type	Labeling procedure	Pre-processing	Augmentation	Model structure	Performance measurements	Outcome
Akay, G., et al. 2023 [27]	Cephalograms	588 (447/141)	Inclusion: patients between 8–22 years, clear C2, C3 and C4, images with no artifacts and distortions	Deep learning	By two dentomaxilofacial radiologists, final decision but agreement among observers	Cropping and labeling, size reduction	NA	CNN	Kappa coefficient,	0.88
									Precision (per class)	(0.47–0.82)
									Recall (per class)	(0.37–0.74)
									F1 score (per class)	(0.44–0.76)
									Accuracy	0.57
Khazaei M. et al. 2023 [28]	Cephalograms	1846 (1477/-/369)	Inclusion: age between 5–18 years, clear C2, C3 and C4, no trauma or surgery in head and neck area, no orthodontic treatment, no medical condition affecting bone development, no systemic disease, no growth delay, no craniofacial anomalies, no growth disorders and no growth hormone therapy	Deep learning	By two orthodontists	Cropping	Translation, rotation, zoom, intensity shift, normalization	ConvNext-Base-296	Accuracy	0.82
									F-score	0.81
								Efficient-NetB3-386	Accuracy	0.81
									F-score	0.80
								DenseNet-121	Accuracy	0.80
									F-score	0.80
								DenseNet-169	Accuracy	0.67
									F-score	0.66
								VGG-16	Accuracy	0.75
									F-score	0.74
								VGG-19	Accuracy	0.68
									F-score	0.67
								ResNet-101	Accuracy	0.65
									F-score	0.65
								ResNet50	Accuracy	0.65
									F-score	0.65
Li H., et al. 2022 [29]	Cephalograms	10,200 (7111/1544/1545)	Inclusion: No congenital or acquired malformation of the cervical vertebrae, no trauma or operation in the head and neck area, no disorder affecting bone development, no systemic disease, no growth and development retardation, no congenital acquired malformations in the head and neck region, clear C2, C3, and C4	Deep learning	By two orthodontists in case of disagreement the third orthodontist was consulted	Automatic ROI extraction using YOLOv3 and shape recognition network	NA	ConvNet	Precision (per class)	(0.57–0.85)
									Recall (per class)	(0.63–0.81)
									F1 score (per class)	(0.60–0.81)
									Accuracy	0.70
									AUC	0.94
									ICC	0.94

Table 2 (continued)

Author, year	Data modality	Data set size (train/ valid/test)	Inclusion and exclusion criteria (if any)	AI type	Labeling procedure	Pre-processing	Augmentation	Model structure	Performance measurements	Outcome
Makaremi, M., et al. (2019) [30]	Cephalograms	600 (300/ 200/ 100) and 900 and 1900	NA	Deep learning	By a radiologic technician	Cropping, Sobel filtering, entropy filter	NA	CNN	Precision (per class) 6 layer	(0.59–0.99)
									Recall (per class) 6 layers	(0.67–0.99)
									F1 score (per class) 6 layers	(0.74–0.92)
									Recall (per class) 7 layers	(0.67–0.99)
									Precision (per class) 7 layer	(0.59–0.99)
									F1 score (per class) 7 layers	(0.74–0.92)
Zhou, J., et al. 2021 [21]	Cephalograms	1080 (980/ -/ 100)	Inclusion: clear contour of c2, c3, c4, 6–22 years exclusion: congenital disease	Deep learning	By two examiners in case of disagreement the third examiner was consulted	Cropping, extracting and crafting the features (measurement between landmarks)	NA	CNN	ICC	0.98
									Accuracy	0.71
Kim, E.-G et al. 2021 [31]	Cephalograms	600 (fivefold cross validation)	Inclusion: 6–18 years	Deep learning	By two specialists	Automated ROI extraction using U-net	Rotation, horizontal and vertical flip, changes in brightness, saturation, contrast and hue	CNN	Accuracy	0.62
Makaremi, M., et al. (2020) [32]	Cephalograms	600 (300/ 200/ 100) and (200/200/200)	NA	Deep learning	By an expert	Cropping, Sobel filtering	NA	CNN	Accuracy	0.90
Kok, H., et al. 2020 [33]	Cephalograms	360 (fivefold cross validation)	Inclusion: 8–17 years	Deep learning, Machine learning	By an orthodontist	Extracting and crafting the features (measurement between landmarks)	NA	ANN	Accuracy	0.94
									Precision (per class)	(0.83–1.0)
									Recall (per class)	(0.83–1.0)
									F1 score (per class)	(0.83–0.97)
									Kappa coefficient	0.95
								Naïve Bayes model	Accuracy	0.68
			Exclusion: disease preventing bone development, systemic diseases and syndromes, growth and development retardation, an anomaly with prevention of craniofacial growth, endocrine disorders or malnutrition, long-term infectious disease						Kappa coefficient	0.61
									Precision (per class)	(0.25–1.0)
									Recall (per class)	(0.05–1.0)
									F1 score (per class)	(0.08–0.90)
Kok, H., et al. 2021 [34]	Cephalograms	419 patients (293/ 63/63)	Inclusion: 8–17 years	Deep learning	By an experienced researcher	Extracting and crafting the features (measurement between landmarks)	NA	ANN	Accuracy	0.94
									Sensitivity (per class)	(0.88–1.0)
									Specificity (per class)	(0.97–1.0)
									F1 score (per class)	(0.90–1.0)

Table 2 (continued)

Author, year	Data modality	Data set size (train/ valid/test)	Inclusion and exclusion criteria (if any)	AI type	Labeling procedure	Pre-processing	Augmentation	Model structure	Performance measurements	Outcome
Amasya, H., et al. 2020 [35]	Cephalograms	647 (498/ -/ 149)	Inclusion: no congenital or acquired malformation of the cervical vertebrae, proper visualization of C2, C3, C4 and C5, age between 10 and 30	Deep learning, Machine learning	By a software and two radiologists	Extracting and crafting the features (measurement between landmarks)	NA	ANN	Agreement	0.86
									Kappa coefficient (wk)	0.92
								LR	Agreement	0.78
									Kappa coefficient (wk)	0.86
								SVM	Agreement	0.81
									Kappa coefficient (wk)	0.87
								RF	Agreement	0.82
									Kappa coefficient (wk)	0.90
								DT	Agreement	0.85
									Kappa coefficient (wk)	0.92
Amasya H. et al. 2020 [36]	Cephalograms	647	Inclusionage between 10 and 30, no congenital or acquired malformation of the cervical vertebrae, good quality of C2, C3, C4 and C5 Exclusion: current orthodontic treatment, permanent incisors or first molars missing, erupted or supernumerary teeth overlapping incisor apex, obvious skeletal asymmetry	Deep learning	By a software and two radiologists	NA	NA	ANN	Agreement with observers	0.58
Mohammad-Rahimi, H., et al. 2022 [37]	Cephalograms	890 (692/ 99/ 99)	Inclusion: cephalograms with visible c2 to c4 exclusion: images of patients wearing items, non-standard images, low quality images	Deep learning	By two orthodontists	Cropping	Random cropping, random color jitter, random affine, random gaussian noise	ResNet 101	Accuracy,	0.61
									Precision (per class)	(0.25–0.88)
									Recall (per class)	(0.33–0.78)
									F1 score (per class)	(0.29–0.82)
Liao, N., et al. 2022 [38]	Cephalograms	900 (fivefold cross-validation)	Inclusion: 7–25 years	Deep learning	By three orthodontics and radiologists	Cropping	Random horizontal flipping, color jittering, random rotation	Resnet 50- iCVM	Accuracy (CVM-900, CVM-900-subset)	(0.69, 0.84)
									Kappa coefficient (CVM-900, CVM-900-subset)	(0.94, 0.96)
									MAE (CVM-900, CVM-900-subset)	(0.33,0.16)

Table 2 (continued)

Author, year	Data modality	Data set size (train/valid/test)	Inclusion and exclusion criteria (if any)	AI type	Labeling procedure	Pre-processing	Augmentation	Model structure	Performance measurements	Outcome
Li, H., et al. 2022 [39]	Cephalograms	6079 (4255/912/ 912)	Inclusion: complete medical record, qualified cephalograms, age less than 18 years old	Deep learning	By two experienced orthodontists in case of disagreement the third orthodontist was consulted	Cropping	Random translation, random rotation, adaptive histogram equalization	Resnet 152	Kappa coefficient	0.82
									AUC	0.93
									Accuracy	0.67
									Precision (per class)	(0.52–0.77)
									Recall (per class)	(0.52–0.84)
									F1 score (per class)	(0.52–0.81)
								VGG16	Kappa coefficient	0.79
									AUC	0.92
									Accuracy	0.61
								GoogLeNet	Kappa coefficient	0.81
									AUC	0.92
									Accuracy	0.64
			exclusion: syndromes, metabolic disease, special drugs, disease affecting growth and development					DenseNet161	Kappa coefficient	0.81
									AUC	0.92
									Accuracy	0.64
Seo, H., et al. 2021 [24]	Cephalograms	600 (480/ -/120)	Inclusion: 6–19 years	Deep learning	By a radiologist	Cropping	Rotation, horizontal and vertical translation, horizontal and vertical scaling	Inception-Resnet v2	Accuracy	0.94 ± 0.018
									Precision	0.84 ± 0.064
									Recall	0.84 ± 0.061
									F1 score	0.84 ± 0.051
								ResNet-18	Accuracy	0.92 ± 0.025
									Precision	0.80 ± 0.094
									Recall	0.80 ± 0.065
									F1 score	0.80 ± 0.074
								MobileNet-v2	Accuracy	0.91 ± 0.022
									Precision	0.77 ± 0.111
									Recall	0.77 ± 0.040
									F1 score	0.77 ± 0.070
								ResNet-50	Accuracy	0.92 ± 0.025
									Precision	0.80 ± 0.096
									Recall	0.80 ± 0.068
									F1 score	0.80 ± 0.075
								ResNet-101	Accuracy	0.93 ± 0.020
									Precision	0.82 ± 0.113
									Recall	0.83 ± 0.096
									F1 score	0.82 ± 0.054
								Inception-v3	Accuracy	0.93 ± 0.027
									Precision	0.82 ± 0.119
									Recall	0.83 ± 0.100
									F1 score	0.82 ± 0.082

Table 2 (continued)

Author, year	Data modality	Data set size (train/valid/test)	Inclusion and exclusion criteria (if any)	AI type	Labeling procedure	Pre-processing	Augmentation	Model structure	Performance measurements	Outcome
Atici, S. F., et al. 2023 [40]	Cephalograms	1012 (823/-/189)	Inclusion: clear and visible C2, C3 and C4, exclusion: abnormalities of head and neck, low image quality	Deep learning	By an orthodontist	Segmentation and cropping	Random translation, rotation and auto contrast	Aggregate net	Accuracy	(male:0.75, female:0.82)
									Intra-examiner agreement (wk)	0.95
									Inter-examiner agreement (wk)	0.90
Atici, SF., et al. 2022 [25]	Cephalograms	1018 (761/ -/ 257)	Inclusion: age between 4 and 29, adequate quality, clear c2/c3/c4, exclusion: poor quality, head and neck malformation	Deep learning, Machine learning	By an expert Orthodontist Scientist and by an oral and maxillofacial surgeon	Automatic ROI extraction by Aggregate channel features object detector	Not needed	CNN	Accuracy	0.84
									Intra-examiner agreement (wk)	0.95
									Inter-examiner agreement	0.90
									Recall (per class)	(0.52–0.77)
									Precision (per class)	(0.55–0.78)
									F1 score (per class)	(0.55–0.76)
								MobileNet V2	Accuracy (with directional filters)	0.69
								ResNet101	Accuracy (with directional filters)	0.68
								Xception	Accuracy (with directional filters)	0.71
								SVM	Accuracy (with directional filters)	0.60
Radwan, M., et al. 2022 [41]	Cephalograms	1501 (1201/150/150)	Inclusion: patients between 7–25 years exclusion: artifacts, incomplete C2, C3 or C4, syndromes affecting maxillofacial, incorrect head position	Deep learning	By an orthodontic resident	Automatic ROI extraction using U-net	NA	Alex-net	ICC	0.97
									Kappa coefficient	0.87 ± 0.027
									Accuracy (per class)	(0.80–0.91)
									Sensitivity (per class)	(0.45–0.98)
									Specificity (per class)	(0.75–0.94)
									F1 score (per class)	(0.57–0.90)
Xie, L., et al. 2021 [42]	CBCT	231	Inclusion: no history of systemic or physiological disorders, no history of trauma or surgery in the dentofacial region and reliable CBCT scans, female, 7–17 years old	Machine learning	By three orthodontists	Reorientation, MPR mode, extracting and crafting the features (measurement between landmarks)	NA	LR	Accuracy	0.87
									AUC	0.94

Table 2 (continued)

Author, year	Data modality	Data set size (train/valid/test)	Inclusion and exclusion criteria (if any)	AI type	Labeling procedure	Pre-processing	Augmentation	Model structure	Performance measurements	Outcome
Kok, H., et al. 2019 [13]	Cephalograms	300 (fivefold cross validation)	Inclusion: 8–17 years,balance quality,clear c2/c3/c4, no trauma, operation, congenital or acquired malformations in the head and neck area, no history of orthodontic treatment, no disorder interposing with bone development, no systemic disease or growth and development retardation	Deep learning, Machine learning	By an orthodontist	Extracting and crafting the features (measurement between landmarks)	NA	DT	Classification accuracy (per class)	(0.83–0.99)
									AUC (per class)	(0.71–0.98)
									F1 score (per class)	(0.42–0.97)
									Precision (per class)	(0.40–0.97)
									Recall (per class)	(0.40–0.98)
								kNN	Classification accuracy (per class)	(0.81–0.92)
									AUC (per class)	(0.81–0.95)
									F1 score (per class)	(0.44–0.82)
									Precision (per class)	(0.48–0.78)
									Recall (per class)	(0.38–0.86)
								SVM	Classification accuracy (per class)	(0.88–0.95)
									AUC (per class)	(0.90–0.99)
									F1 score (per class)	(0.50–0.91)
									Precision (per class)	(0.51–0.84)
									Recall (per class)	(0.50–0.98)
								RF	Classification accuracy (per class)	(0.83–0.97)
									AUC (per class)	(0.84–0.99)
									F1 score (per class)	(0.39–0.95)
									Precision (per class)	(0.40–0.91)
									Recall (per class)	(0.38–0.98)
								Neural network	Classification accuracy (per class)	(0.85–0.97)
									AUC (per class)	(0.90–0.99)
									F1 score (per class)	(0.48–0.95)
									Precision (per class)	(0.47–0.93)
									Recall (per class)	(0.50–0.97)
								Naïve Bayes	Classification accuracy (per class)	(0.95–0.83)
									AUC (per class)	(0.85–0.98)
									F1 score (per class)	(0.38–0.88)
									Precision (per class)	(0.44–0.92)
									Recall (per class)	(0.33–0.85)
								LR	Classification accuracy (per class)	(0.81–0.90)
									AUC (per class)	(0.81–0.96)
									F1 score (per class)	(0.25–0.75)
									Precision (per class)	(0.36–0.75)
									Recall (per class)	(0.19–0.98)

Table 2 (continued)

Author, year	Data modality	Data set size (train/valid/test)	Inclusion and exclusion criteria (if any)	AI type	Labeling procedure	Pre-processing	Augmentation	Model structure	Performance measurements	Outcome
Sokic, E., et al. 2012 [43]	Cephalograms	211	Inclusion: 8–16 years	Machine learning	By orthodontists	Prescaling, bilinear projective transformation, special markers, extracting and crafting the features (measurement between landmarks)	NA	Fuzzy C means	Accuracy	0.70
Xie, L., et al. 2022 [44]	CBCT	709 (447/-/262)	Inclusion: 7–19 years, no history of systemic or physiological syndromes, no history of trauma or surgery in the dentofacial area, dependable and suitable CBCT scans	Statistical modeling	By three orthodontists	Reorientation, MPR mode, extracting and crafting the features (measurement between landmarks)	NA	LR	Agreement percentage Kappa coefficient AUC ICC (range)	0.88 0.90 0.96 (0.94–0.99)
Yang, Y. M., et al. (2014) [45]	CBCT	121	Inclusion: 6–18 years exclusion: cleft lip and/or palate, trauma, or syndromes	Statistical modeling	By an investigator	NA	NA	Regression models	R 2 (Female-male)	(0.84–0.9)
Baptisa, R. S., et al. 2012 [46]	Cephalograms	188 (tenfold cross validation)	NA	Machine learning	By specialist in orthodontics and radiology and a specialist in orthodontics and then by an examiner using a software	Extracting and crafting the features (measurement between landmarks), cropping	NA	Naïve bayes 1	Kappa coefficient Accuracy	0.99 ± 0.019 0.90
Feng, X., et al. (2021) [47]	Cephalograms and CBCT	60	Inclusion: 8–16 years, in the age of growth and development; Exclusion: unclear CBCT, incomplete C2 and C4, history of craniofacial deformity, syndrome affecting the shape of the cervical spine, intense systemic STDs	Rule-based AI	By a researcher with three years of experience in CVM assessment	Otsu's method, three-dimensional least squares method, superpixel segmentation, And marking the selected points automatically with morphological algorithm and manual method, extracting and crafting the features (measurement between landmarks)	NA	Decision Tree	Kappa coefficient Gamma value	0.87 0.99

NA Not assigned, *CNN* Convolutional neural network, *ICC* Intraclass correlation coefficient, *ROI* Region of interest, *ANN* Artificial neural network, *MAE* Mean Absolute Error, *CVM* Cervical vertebrae maturation, *AUC* Area under curve, *WK* Weighted kappa, *MPR* Multiplanar reformation, *CBCT* Cone beam computed tomography, *STD* Sexually transmitted diseases, *AI* Artificial intelligence, *LR* Likelihood ratio, *DT* Decision tree, *RF* Random forest, *SVM* Support vector machine, *KNN* K-nearest neighbors

Critical appraisal of individual sources of evidence

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool [48] was utilized independently by 2 reviewers (T.S.S, SAH.O) to evaluate the risk of bias. It encompasses four domains: patient selection, index test, reference standard, and flow and timing. It also addresses applicability concerns in three areas: patient selection, index test, and reference standard. Low risk of bias in patient selection is linked to clear description of patient selection methods, avoidance of inappropriate exclusions, such as difficult-to-diagnose cases or outliers excluded without a defined detection method, prevention of data leakage (e.g., overlap between training and testing datasets). In the index test domain, blinding of the reference standard to the results of the index test, transparent reporting of the test threshold, adequate information on model development and test reproducibility will lead to low risk of bias. For the reference standard, low risk of bias arises from Use of sensitive reference standards, such as evaluation by multiple examiners or robust diagnostic methods, blinding of the reference standard to the index test results. The flow and timing domain evaluates the consistency of reference standards across samples and the intervals between index tests and reference standards. Finally, the tool assesses applicability by evaluating how well the study, including its dataset, and AI model addresses the research question in the context of the intended clinical use.

Synthesis of results and meta-analysis

A hierarchical logistic regression model was utilized for the meta-analysis of diagnostic test accuracy. Inclusion criteria required studies to provide sufficient raw data to extract true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each diagnostic class under evaluation. This enabled the calculation of pooled estimates and 95% confidence intervals for sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR-), and diagnostic odds ratio (DOR) across studies. Subgroup analyses were conducted for different AI techniques and different image modalities to quantify differences in diagnostic performance between artificial intelligence methods. The 12 studies included in the meta-analysis reported the values for TN, TP, FP, and FN for each of the six CVM stages. In most of the included studies, these metrics were presented in the form of a confusion matrix. Studies that did not provide complete reporting of these metrics or failed to report metrics for all six stages were excluded from the meta-analysis. Publication bias was evaluated through visual inspection of Deek's funnel plot asymmetry and Egger's regression test. The primary meta-analytic findings were visualized through forest plots, hierarchical summary receiver operating characteristic (HSROC) curves, and Deek's funnel plots. All statistical analyses were executed using the metadta, metandi, and midas commands in STATA 16 (StataCorp LLC, College Station, TX).

The reliability and validity of the evidence from the studies compiled in the meta-analysis, covering various imaging techniques and tasks, were evaluated through the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) framework, as outlined by the GRADE Working Group (https://www.gradeworkinggroup.org).

Results

Study selection and characteristics

The database searches yielded 656 records, of which 31 were retrieved for full-text review following title and abstract screening. After examining the full texts, 6 articles were excluded for reasons detailed in Supplementary Table 1. Thus, a total of 25 studies satisfied the inclusion criteria.; The number of included studies rose over the observed period. Also, the data types diversified over time (Fig. 1).

The included studies utilized two main image modalities: cephalograms ($n=22$) [13, 21, 24, 25, 27, 35] and cone beam computed tomography (CBCT) scans ($n=4$) [42, 44, 45, 47], as summarized in Table 1. The majority of studies ($n=22$) established ground truth labels via evaluation by clinical experts. Specifically, the reference standard was defined by one expert ($n=8$ studies) [13, 24, 32–34, 40, 41, 47], two experts (n=8) [25, 27, 28, 31, 35–37, 46], or three or more experts (n=5) [29, 38, 39, 42, 44] and one study [43] did not report the number of experts involved. Three studies employed a combination of clinical experts and software analysis to determine labels [35, 36, 46]. A total of 55 AI models were utilized, with DL being the most common approach (n=19) [13, 21, 24, 25, 27–41] analysis to determine labels analysis to determine labels, followed by ML (n=7) [13, 25, 33, 35, 42, 43, 46], statistical modeling (n=2) [44, 45], and rule-based AI (n=1) [47]. Among DL technologies, CNNs stood out as the predominant model (n=14) [21, 24, 25, 27–32, 37–41], including ResNet architectures (n=6)[24, 25, 28, 37–39]. The most utilized ML techniques were Naïve Bayes [13, 33, 46], and support vector machines [13, 25, 35], each applied in three studies and Logistic Regression applied in five studies [13, 35, 42, 44, 45]. Logistic regression models were the primary focus in statistical modeling (n=2) [44, 45]. Augmentation techniques, such as rotation and translation, were implemented in seven DL studies [24, 28, 31, 37–40] to increase the size of the training data. Feature extraction, using landmark coordinate measurements, was performed in 10 studies [13, 21, 33–35, 42–44, 46, 47].

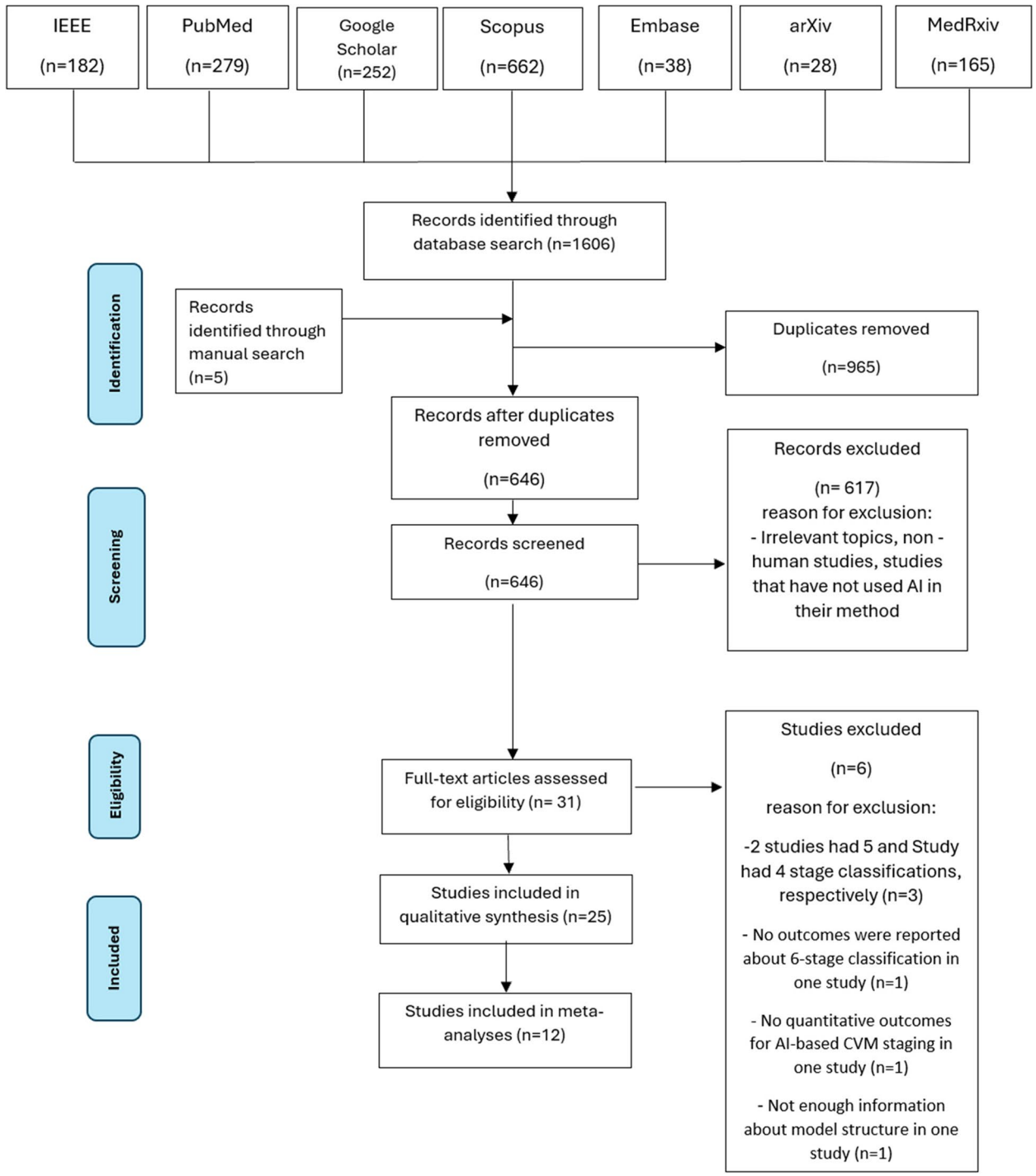

Fig. 1 Prisma flowchart

Additionally, the automation of region of interest (ROI) detection was carried out in four studies [25, 29, 31, 41], with methods like U-Net being used in two instances [31, 41] to delineate crucial anatomical areas from the images.

Performance metrics most reported for DL studies included accuracy, kappa coefficient, precision, recall, and F1 score. Additional measures such as mean absolute error, area under the receiver operating characteristic curve, sensitivity, and specificity were also occasionally

utilized. The statistical modeling studies reported a wider range of metrics encompassing agreement percentage, R squared, predictability, as well as some of the aforementioned measures. The machine learning studies focused on accuracy and area under the curve, while the single rule-based AI study used kappa coefficients and Goodman-Kruskal gamma correlation. The detailed description of each metric is presented in supplementary Table 2.

Risk of bias and applicability

Quality assessment identified 8 studies [21, 27, 35, 36, 38, 39, 42, 44] as having low risk of bias and concerns regarding applicability across all domains. The greatest issues were found in the reference standard domain, with 13 studies [13, 24, 25, 28, 30, 32–34, 40, 41, 43, 45, 47] deemed at high or unclear risk of bias and 17 studies [13, 24, 25, 28–34, 37, 40, 41, 43, 45–47] harboring applicability concerns. Performance of the included studies for assessed domains are summarized in Table 3.

Results of individual studies

The accuracy of DL for cephalograms varied widely from 0.57 to 0.95, the Kappa coefficient from 0.8 to 0.96, precision from 0.25-0.99, recall from 0.33-0.99, and F1 score from 0.29-1.0. The highest reported accuracy was 0.95 using ANN on 360 cephalograms [33] and the lowest accuracy was 0.57 using CNN on 588 cephalograms [27]. The highest reported precision was 0.99 using CNN [33] and the lowest was 0.25 using ResNet [37]. Also, the highest reported sensitivity, specificity, and F1 score was 100% using ANN [34].

The reported accuracy of ML ranged from 0.60 to 0.99. The highest reported accuracy was 0.99 using Decision tree on the 300 cephalograms [13] and the lowest was 0.60 using SVM on the 1018 cephalograms [25].

All studies conducting statistical modeling used CBCTs [44, 45]. The reported kappa coefficient was 0.90 using logistic regression on CBCT [44]. The reported R2 was 0.84 for females and 0.90 for males using regression models on CBCT [45].

The only study using rule-based AI was conducted on cephalograms and CBCTs. The reported Kappa coefficient and Gamma value were 0.87 and 0.99 [47].

Synthesis of results

The meta-analysis included 12 studies (28 models/datasets) [13, 24, 25, 27, 29–31, 33, 37–40]. For each CS the pooled sensitivity, specificity, diagnostic odds ratio (DOR), positive likelihood ratio (LR+), negative likelihood ratio (LR-) were computed separately. The summary points for sensitivity were based on a range of estimates across the studies included in the analysis, with values ranging from 0.64 for CS3 (95% confidence interval [CI], 0.57–0.71) to 0.87 for CS1 (95% CI, 0.81–0.91). The pooled estimate of specificity ranged from 0.94 for CS4 (95% CI, 0.92–0.95) to 0.97 for CS1 (95% CI, 0.95–0.98).

Pooled diagnostic odds ratio (DOR) was between 32 for CS3 (95% CI, 22-47) to 213 for CS1 (95% CI, 115-394). Other results of the meta-analysis are summarized in Table 4.

The plot (Fig. 2) depicted the visualization of various elements, including the Hierarchical Summary Receiver Operating Characteristic (HSROC) curve, prediction region, summary point, and confidence region. The HSROC model is a combination of sensitivity and specificity estimates from individual studies. The beta parameter was computed for each CS (from 1 to 6), and the values were −0.02, −0.06, −0.15, −0.24, −0.83, and 0.68, respectively. This indicates that there is no significant skewness in the diagnostic odds ratio for all CSs except for CS5 and CS6 ($p < 0.05$).

No significant publication bias was detected across all studies ($p > 0.05$). In analyzing heterogeneity across the included studies, the inconsistency index (I^2) was found to be over 97%, indicating that nearly all observed variability in outcomes is due to genuine heterogeneity. Subgroup analysis was done for machine learning and deep learning tasks (Table 5).

The summary points for sensitivity were based on a range of estimates across the studies included in the analysis, with values ranging from 0.67 for CS3 (95% CI, 0.60-0.73) to 0.84 for CS1 (95% CI, 0.76-0.89) for deep learning and from 0.52 for CS4 (95% CI, 0.38-0.65) to 0.93 (95% CI, 0.86-0.97) for machine learning models. The pooled estimate of specificity ranged from 0.94 for CS5 (95% CI, 0.92-0.95) to 0.97 for CS1 (95% CI, 0.96-0.98), and from 0.93 for CS5 (95% CI, 0.87-0.97) to 0.97 for CS2 (95% CI, 0.94-0.98). The diagnostic accuracy in deep learning and machine learning models revealed significant heterogeneity ($p < 0.0001$), with minimal impact from subgroup analyses on sensitivity or specificity. Subgroup analyses for different imaging modalities could not be conducted due to limited data availability across modalities. Publication bias was assessed through Egger's regression test and visual inspection of Deek's funnel plots generated for each CS. There was no strong statistical evidence of significant publication bias or small study effects in the analyzed dataset (Egger's test, $p > 0.05$). The Deek's funnel plots for each CS are presented in Fig. 3.

Appraisal using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework deemed the overall Certainty of Evidence for the studies compiled in the meta-analyses to be "moderate" (Table 6).

Table 3 Quality assessment in individual studies. Green represents low risk, orange represents uncertain risk and red represents high risk. If the risk is low in all domains, the study is at low risk of bias. A study is judged to be at risk of bias if there was unclear or high risk in one or more domains

Author, Date	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Time and flow	Patient selection	Index test	Reference standard
Atici, S. F., et al. 2023 [40]							
Akay, G., et al. 2023 [27]							
Radwan, M., et al. 2022 [41]							
Li, H., et al. 2022 [29]							
Xie, L., et al. 2022 [44]							
Mohammad-rahimi, H., et al. 2022 [37]							
Liao, N., et al. 2022 [38]							
Li, H., et al. 2022 [39]							
Zhou, J., et al. 2021 [21]							
Xie, L., et al. 2021 [42]							
Kim, E.-G et al. 2021 [31]							
Seo, H., et al. 2021 [24]							
Kok, H., et al. 2021 [33]							
Kok, H., et al. 2021 [34]							
Feng, X., et al. 2021 [47]							
Makaremi, M., et al. 2020 [32]							
Amasya, H., et al. 2020 [35]							
Amasya, H., et al. 2020 [36]							
Makaremi, M., et al. 2019 [30]							
Kok, H., et al. 2019 [13]							
Yang, Y. M., et al. 2014 [45]							
Sokic, E., et al. 2012 [43]							
Baptisa, R. S., et al. 2012 [46]							
Atici, SF., et al. 2022 [25]							
Khazaei, M., et al. 2023 [28]							

Atici, S. F., et al. 2023 [40], Akay, G., et al. 2023 [27], Radwan, M., et al. 2022 [41], Li, H., et al. 2022 [29], Xie, L., et al. 2022 [44], Mohammad-rahimi, H., et al. 2022 [37], Liao, N., et al. 2022 [38], Li, H., et al. 2022 [39], Zhou, J., et al. 2021 [21], Xie, L., et al. 2021 [42], Kim, E.-G et al. 2021 [31], Seo, H., et al. 2021 [24], Kok, H., et al. 2021 [33], Kok, H., et al. 2021 [34], Feng, X., et al. 2021 [47], Makaremi, M., et al. 2020 [32], Amasya, H., et al. 2020 [35], Amasya, H., et al. 2020 [36], Makaremi, M., et al. 2019 [30], Kok, H., et al. 2019 [13], Yang, Y. M., et al. 2014 [45], Sokic, E., et al. 2012 [43], Baptisa, R. S., et al. 2012 [46], Atici, SF., et al. 2022 [25], Khazaei, M., et al. 2023 [28]

Discussion

In the field of orthodontics, determining the optimal time to initiate treatment is crucial for maximizing its effectiveness. Failing to accurately identify and address orthodontic issues at the appropriate stage may necessitate surgical intervention later on to correct jaw deformities, highlighting the importance of timely intervention [30, 49, 50]. The conventional technique for determining the initiation of the orthodontics treatments is based on evaluating CVM. However, this approach has several limitations, such as its subjective nature, which will lead to a low intra-observer agreement, and an inability to detect

Table 4 Summary of the meta-analysis result

	CS1	CS2	CS3	CS4	CS5	CS6
Se	0.87 (0.81, 0.91)	0.73 (0.65, 0.79)	0.64 (0.57, 0.71)	0.71 (0.54, 0.78)	0.72 (0.64, 0.79)	0.79 (0.75, 0.83)
Sp	0.97 (0.95, 0.98)	0.95 (0.93, 0.96)	0.95 (0.93, 0.96)	0.94 (0.92, 0.95)	0.94 (0.93, 0.95)	0.97 (0.95, 0.98)
DOR	213 (115, 394)	50 (28, 89)	32 (22, 47)	38 (23, 63)	41 (26, 64)	106 (65, 174)
LR +	28.3 (18.6, 43.0)	14.4 (9.9, 20.8)	12.0 (9.3, 15.6)	11.6 (8.5, 15.8)	11.9 (9.5, 15.0)	23.0 (15.3, 34.6)
LR-	0.13 (0.09, 0.20)	0.29 (0.22, 0.38)	0.38 (0.31, 0.45)	0.30 (0.24, 0.39)	0.29 (0.22, 0.39)	0.22 (0.18, 0.26)

Se Sensitivity, *Sp* Specificity, *DOR* Diagnostic odds ratio, *LR* + Positive Likelihood Ratio, *LR*- Negative Likelihood Ratio

Numbers in the parenthesis indicate the range for each metric by 95% CI

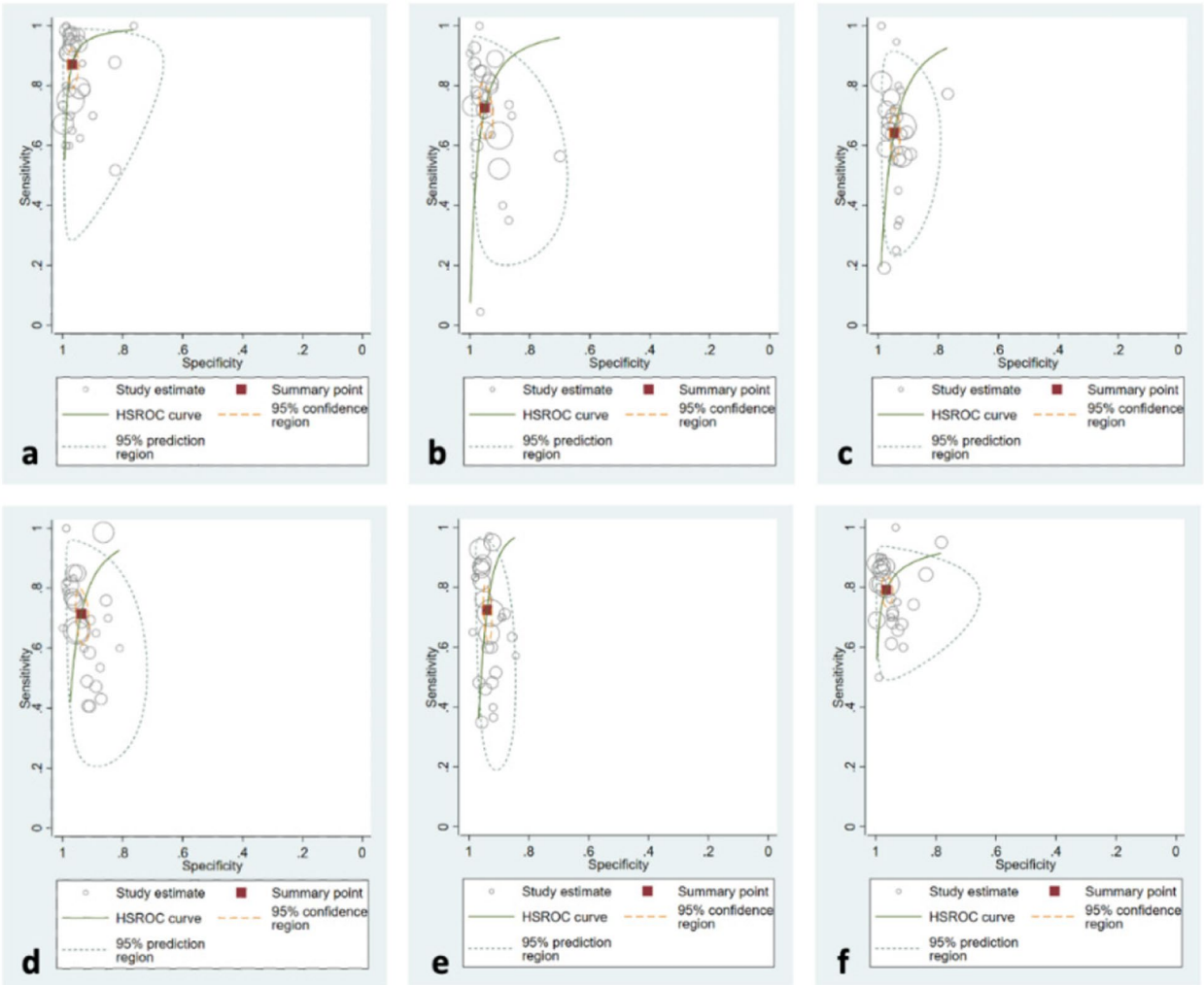

Fig. 2 Results of meta-analysis on Sensitivity and Specificity of the deep learning models for assessment of CVM, and the Hierarchical Summary Receiver Operating Characteristic (HSROC) curve. Each "study estimate" is shown as a data point, representing sensitivity, specificity, and sample size. The "Summary point" displays the pooled sensitivity and specificity from all studies. The "95% confidence region" indicates the expected location of the pooled summary point with 95% certainty, while the "95% prediction region" forecasts sensitivity and specificity ranges for future studies. Pictures "a" to "f" refer to CS1 to CS6 respectively. The beta parameter was significant in CS5($p = 0.013$) and CS6($p = 0.016$) showing heterogenicity among included studies for these stages

Table 5 Summary of the sub-group analysis

		CS1	CS2	CS3	CS4	CS5	CS6
Se	**Deep learning**	**0.84 (0.76–0.89)**	**0.73 (0.63–0.80)**	**0.67 (0.60–0.73)**	**0.77 (0.70–0.82)**	**0.77 (0.70–0.83)**	**0.81 (0.76–0.84)**
	Machine learning	0.93 (0.86–0.97)	0.72 (0.56–0.85)	0.57 (0.43–0.69)	0.52 (0.38–0.65)	0.53 (0.37–0.69)	0.74 (0.66–0.82)
Sp	Deep learning	0.97 (0.96–0.98)	0.94 (0.92–0.96)	0.95 (0.93–0.96)	0.94 (0.92–0.96)	0.94 (0.92–0.95)	0.97 (0.96–0.98)
	Machine learning	0.95 (0.89–0.97)	0.97 (0.94–0.98)	0.94 (0.90–0.96)	0.93 (0.88–0.95)	0.95 (0.92–0.96)	0.93 (0.87–0.97)

Se Sensitivity, *Sp* Specificity

Numbers in the parenthesis indicate the range for each metric by 95% CI

subtle changes. AI can address these shortcomings by providing precise and unbiased analysis of digital images of cervical vertebrae, minimizing the variation between examiners, and enabling more accurate tracking of treatment progress [24, 30]. The present study aimed to review the role of AI in CVM assessment and how most of the models show great promise due to their high accuracy in this task.

The AI models developed exhibited varying degrees of performance across different metrics. For instance, sensitivity ranged widely from 0.45 to 1, while specificity varied from 0.75 to 1. These results suggest that, while the models may not deliver optimal performance in the intricate task of CVM evaluation, the outcome of AI-based CVM staging models depends on the nature of the applied model and can excel humans in some instances. Previous studies showed low inter-examiner agreement reports which were 0.50-0.74 [13, 37].

To enhance the model's accuracy, both the input image and the subject's chronological age were utilized in Atici et al. study [40]. Recognizing the divergent growth rates between male and female patients, the dataset was segregated by gender. This stratification aimed to optimize the model's efficiency by incorporating chronological age as a variable. This approach is consistent with findings from Kim et al. [51], who observed improved model accuracy when incorporating both chronological age and gender into the input. This methodological choice underscores the significance of demographic factors in refining the predictive performance of models designed for assessing developmental stages.

Our meta-analysis demonstrated that AI models exhibited superior performance in classifying CS1 compared to other stages. This superior accuracy can be attributed to the distinct morphology of CS1, where C1, C2, and C3 are characterized by flat lower borders. The absence of concavities or other morphological changes in CS1 simplifies its detection, making it more straightforward for both AI models and human examiners to identify. On the other hand, the overall performance of detecting CS3 was lower than in other stages. This might be related to the difficulties of detecting CS3 compared to other stages, primarily due to its inherent morphological overlap with adjacent stages like CS2 and CS4. This overlap can blur the distinction between the stages, leading to potential misclassifications. Additionally, the variability in the progression of the third vertebra from being rectangular to square introduces further inconsistencies, making CS3 a particularly intricate stage to identify with high precision [14, 49].

A key distinction emerges between deep learning models designed for visual data interpretation, CNNs, versus those built for structured data analysis, such as ANNs and traditional machine learning algorithms. Among the models reviewed, 33 leveraged CNNs to process and interpret visual data directly. In contrast, 4 studies utilized ANNs for analyzing structured data, while others employed traditional algorithms reliant on structured inputs. CNNs streamline workflows by eliminating the need for manual feature extraction and anatomical measurements, automating the laborious processes required by conventional methods. This automation saves clinicians valuable time per evaluation. Additionally, algorithmic feature extraction standardizes the analytical process, enhancing diagnostic reliability and consistency compared to manual measurements prone to subjectivity and human error. By minimizing evaluator discrepancies in assessments of skeletal maturation stages, deep learning models like CNNs can better assist clinicians in cervical vertebral maturation evaluations than methods reliant on structured data inputs. Artificial Intelligence has emerged as a transformative tool in orthodontics, particularly in the analysis of 2D cephalograms and 3D CBCT images. The latest deep learning methods have enabled automated cephalometric analysis, offering precise and efficient identification of landmarks, which is crucial for diagnosis and treatment planning [52]. Innovations such as personal computer-based cephalometric

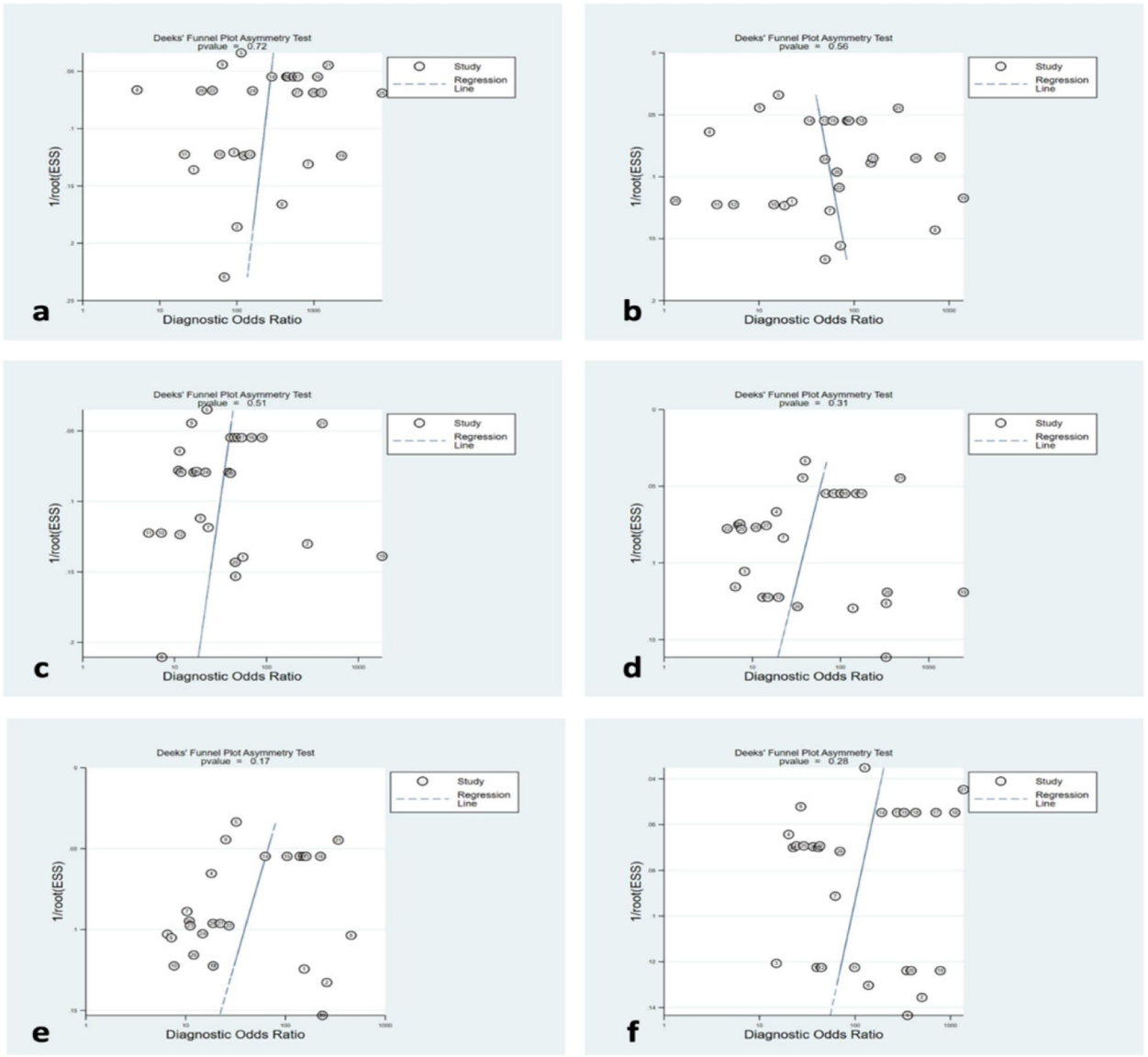

Fig. 3 Funnel plots for assessing publication bias. The funnel plots are used to visually assess publication bias across multiple studies. Each plot displays the log diagnostic odds ratio on the x-axis and the inverse standard error (SE) on the y-axis. The funnel plot is typically symmetrical in the absence of bias, with higher precision studies (smaller SE) clustering near the top and lower precision studies (larger SE) scattering toward the bottom. The open circles represent individual studies. The solid line indicates the regression line. Asymmetry in the distribution of studies relative to the regression line may suggest publication bias. Individual Plots: Plot a, b, c, d, e, and f all visually show studies distributed across the funnel. The p-value of Deeks' Funnel Plot Asymmetry Test, indicated on each plot, provides statistical evidence for asymmetry. If $p > 0.05$, there is no significant asymmetry, suggesting a low likelihood of publication bias. If $p < 0.05$, significant asymmetry is present, indicating a potential risk of publication bias

landmark detection utilizing online cephalograms further enhance accessibility and accuracy in orthodontic evaluations [53]. In 3D imaging, deep learning algorithms facilitate multiclass CBCT image segmentation [54] and automatic detection and segmentation of the pharyngeal airway, contributing to improved assessments of airway-related orthodontic conditions [55]. Additionally, deep learning-based integrated tooth models, combining intraoral scans and CBCT data, provide accurate 3D evaluations of root positions during orthodontic treatment, ensuring better outcomes and precise treatment adjustments [56].

To date, two reviews have been conducted in this area, each providing valuable insights into the application of

Table 6 Results of grading of recommendations assessment, development and evaluation (GRADE)

Number of models/actual cervical stages	Study design	Factors that may decrease certainty of evidence						Certainty of Evidence
		Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	Other consideration	
28 models/ 10,865 stages	Case–control	Serious[a]	Not serious	Very serious[b]	Not serious	Not detected	Very strong association[c]	⊕⊕⊕◯ Moderate

Explanations

[a] Evidence was downgraded by one level. Seventeen studies had high or unclear risk of bias

[b] Evidence was upgraded by 2 levels. Different imaging modalities, different artificial intelligence tasks, different data curation and different age and sex groups created inconsistency

[c] Evidence was upgraded by 2 levels because of "very large" effect size (DOR > 5)

AI and neural networks in cervical vertebral maturation (CVM) assessment. The review by Mathew et al. [57] focused on neural networks for cervical vertebral maturation (CVM) classification, reporting accuracy ranging from 50% to 90%, while highlighting concerns about bias and the need for standardized reference methods. Kazimierczak et al. [58] examined a broader spectrum of AI models, with accuracy ranging from 57% to 95%, emphasizing variability in performance due to differences in models, datasets, and methodologies, and calling for more robust research. Building on these, our review includes a larger number of studies, performs a meta-analysis, and conducts subgroup analyses based on AI methodologies, offering a more comprehensive evaluation of AI performance in CVM assessment.

The use of AI in CVM assessment holds significant promise; however, several limitations and challenges must be considered. To begin with, many of the studies withhold key information, as they often fail to share their datasets or provide in-depth details about their models. In this review, only two studies had public datasets [40]. On the other hand, there are two studies that only included female patients in their datasets [42, 44]. Moreover, these studies did not share any details about their dataset classes, for example, the number of samples from different age or sex groups [29, 30, 32], and some studies only shared data about age groups and no details about sex groups [24, 27, 31, 36, 38, 41]. Sharing model details is necessary for other researchers to reproduce the models and make the reported accuracy and metrics more reliable. Crucially, this absence of information could mask issues such as imbalanced, insufficient, or mislabeled datasets. Such issues might adversely affect the AI model's performance and generalization, both of which are heavily reliant on the quality and representativeness of the training dataset. Moreover, any errors in the dataset's labeling and annotation could significantly affect the accuracy of the AI models [59]. Therefore, expert labeling and annotation are crucial in training the models. Unfortunately, some studies [21, 34, 42, 45, 47] have used datasets labeled by operators, examiners, or researchers. It is rational to assume that annotations and datasets of these studies surely had less accuracy and reliability compared to studies that had been labeled by experienced orthodontics. To ensure the highest possible accuracy when using AI models for real-world decision-making problems, it is essential that datasets are labeled by experts, particularly in the sensitive task of CVM assessment. Therefore, we recommend that datasets for AI model training should be carefully curated and labeled by a panel of experts to achieve the most reliable performance.

Another significant limitation of utilizing AI for CVM assessment lies in the inherent unreliability of the prevailing gold standard for CVM evaluation which is highly subjective and largely depends on the observer's expertise and interpretation. The nuances in the cervical vertebrae transitions, which evolve gradually rather than abruptly, contribute to the low interexaminer agreement [13, 37]. Often, these transitions manifest as intermediary stages, displaying characteristics of two stages simultaneously. This inherent overlap complicates the clinician's task of determining a clear cut-off for each definitive stage, thereby reducing both the validity and reproducibility of annotations. To address this obstacle, we recommend that CVM annotations be cross-verified using auxiliary methods, such as hand-wrist radiographs, to enhance the accuracy and reliability of the assigned labels.

When utilizing AI in medicine and dentistry, it is crucial to consider ethical issues to ensure patient well-being. Informed consent is necessary, and patients should have the option to decline the use of AI in their diagnosis and treatment [21, 60]. Patient data must be kept secure, confidential, and only accessible to authorized personnel [61]. While AI demonstrates significant promise for clinical applications, AI systems employed in evaluating CVM stages must undergo rigorous testing and validation. While AI can assist in diagnosis and treatment

planning, a human clinician should always be involved, and the AI systems should be used as a tool rather than a replacement. Moreover, patients should also have access to information about how the AI system works and how it generates its predictions [62, 63].

AI technology has great potential for improving CVM assessment. Future directions include developing more sophisticated DL models to capture complex CVM features and incorporating diverse datasets from different populations and age groups. Combining AI with other clinical factors such as dental and skeletal findings could also improve accuracy and usefulness in clinical practice. Additionally, using public datasets and codes would promote reproducibility, collaboration, and unbiased data, addressing concerns about bias and ethical considerations associated with AI in medical diagnosis. Overall, these directions hold great promise for building accurate and reliable AI models for CVM assessment.

Conclusion

The use of AI in CVM assessment has shown relatively high accuracy and efficiency. Hence, it holds potential as an auxiliary tool for diagnosis and pinpointing the optimal initiation time for growth modification treatments in the future. However, challenges such as the unreliability of the accepted gold standard and the low level of agreement among clinicians need to be addressed to enhance the accuracy and reliability of AI models. In the future, further development and standardization of AI technology can significantly improve the accuracy and efficiency of CVM assessment and ultimately benefit both patients and clinicians.

Abbreviations

AI	Artificial intelligence
CVM	Cervical vertebral maturation
DOR	Diagnostic odds ratio
ML	Machine learning
DL	Deep learning
NA	Not assigned
CNN	Convolutional neural network
ICC	Intraclass correlation coefficient
ROI	Region of interest
ANN	Artificial neural network
MAE	Mean Absolute Error
AUC	Area under curve
WK	Weighted kappa
MPR	Multiplanar reformation
CBCT	Cone beam computed tomography
STD	Sexually transmitted diseases
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies
TP	True positive
TN	True negative
FP	False positive
FN	False negative
LR+	Positive likelihood ratio
LR-	Negative likelihood ratio
SE	Sensitivity
SP	Specificity
HSROC	Hierarchical summary receiver operating characteristic
GRADE	Grading of Recommendations Assessment, Development and Evaluation
LR	Logistic regression
SVM	Support vector machine
RF	Random forest
DT	Decision tree
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
p_0	Relative observed agreement among raters
p_e	Hypothetical probability of chance agreement
KNN	K-nearest neighbors

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-025-05482-9.

Supplementary Material 1.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

T.SS. and SAH.O. did Conceptualization, Writing-original draft, Methodology and Data curation. F.S. did Data curation, Writing-original draft and Methodology. S.S. did Data curation and Writing-original draft. P.S. did Formal Analysis. SR.M. did Conceptualization, Methodology, Writing-Review and Editing and Supervision.

Funding

The authors declared no potential funding for this article.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Dentofacial Deformities Research center, Research Institute of Dental sciences, Shahid Beheshti, University of Medical Sciences, Tehran, Iran. [2]Department of Endodontics, School of Dentistry, Hamadan University of Medical Sciences, Hamadan, Iran. [3]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, United States. [4]Department of Orthodontics, School of Dentistry Shahid Beheshti University of Medical Sciences, Daneshjoo Blvd, Evin, Shahid Chamran Highway, Tehran 1983963113, Iran.

Received: 8 November 2024 Accepted: 13 January 2025
Published online: 05 February 2025

References

1. Mohammad-Rahimi H, Rokhshad R, Bencharit S, Krois J, Schwendicke F. Deep learning: a primer for dentists and dental researchers. J Dent. 2023;134:104430.
2. Kühl N, Schemmer M, Goutier M, et al. Artificial intelligence and machine learning. Electron Markets. 2022;32:2235–44. https://doi.org/10.1007/s12525-022-00598-0.
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.

4. Sarker IH. AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. SN Comput Sci. 2022;3(2):158. https://doi.org/10.1007/s42979-022-01043-x.
5. Neale MC, Boker SM, Xie G, Maes HM. Statistical modeling. Richmond, VA: Department of Psychiatry, Virginia Commonwealth University. 1999:31.
6. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798–828.
7. Mohammad-Rahimi H, Nadimi M, Rohban MH, Shamsoddin E, Lee VY, Motamedian SR. Machine learning and orthodontics, current trends and the future opportunities: a scoping review. Am J Orthod Dentofacial Orthop. 2021;160(2):170–92. e4.
8. Rokhshad R, Mohammad-Rahimi H, Sohrabniya F, Jafari B, Shobeiri P, Tsolakis IA, et al. Deep learning for temporomandibular joint arthropathies: a systematic review and meta-analysis. J Oral Rehabil. 2024;51(8):1632–44.
9. Arık SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging. 2017;4(1):014501-.
10. Shin W, Yeom H-G, Lee GH, Yun JP, Jeong SH, Lee JH, et al. Deep learning based prediction of necessity for orthognathic surgery of skeletal malocclusion using cephalogram in Korean individuals. BMC Oral Health. 2021;21:1–7.
11. Lee J-H, Han S-S, 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.
12. Park Y, Choi J, Kim Y, Choi S, Lee J, Kim K, et al. Deep learning–based prediction of the 3D postorthodontic facial changes. J Dent Res. 2022;101(11):1372–9.
13. Kök H, Acilar AM, İzgi MS. Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics. Prog Orthod. 2019;20:1–10.
14. Baccetti T, Franchi L, McNamara Jr JA, editors. The cervical vertebral maturation (CVM) method for the assessment of optimal treatment timing in dentofacial orthopedics. United States: Seminars in Orthodontics; 2005: Elsevier.
15. Rita SN, Sadat SA. Growth modification in class II Malocclusion: a review. Update Dental College J. 2014;4(2):23–6.
16. Hunter CJ. The correlation of facial growth with body height and skeletal maturation at adolescence. Angle Orthod. 1966;36(1):44–54.
17. Hägg U, Taranger J. Skeletal stages of the hand and wrist as indicators of the pubertal growth spurt. Acta Odontol Scand. 1980;38(3):187–200.
18. Lewis AB, Garn SM. The relationship between tooth formation and other maturational factors. Angle Orthod. 1960;30(2):70–7.
19. Hägg U, Taranger J. Menarche and voice change as indicators of the pubertal growth spurt. Acta Odontol Scand. 1980;38(3):179–86.
20. Lamparski DG. Skeletal age assessment utilizing cervical vertebrae. Am J Orthod. 1975;67(4):458–9.
21. Zhou J, Zhou H, Pu L, Gao Y, Tang Z, Yang Y, et al. Development of an artificial intelligence system for the automatic evaluation of cervical vertebral maturation status. Diagnostics. 2021;11(12):2200.
22. Zhao X-G, Lin J, Jiang J-H, Wang Q, Ng SH. Validity and reliability of a method for assessment of cervical vertebral maturation. Angle Orthod. 2012;82(2):229–34.
23. Nestman TS, Marshall SD, Qian F, Holton N, Franciscus RG, Southard TE. Cervical vertebrae maturation method morphologic criteria: poor reproducibility. Am J Orthod Dentofac Orthop. 2011;140(2):182–8.
24. Seo H, Hwang J, Jeong T, Shin J. Comparison of deep learning models for cervical vertebral maturation stage classification on lateral cephalometric radiographs. J Clin Med. 2021;10(16):3591.
25. Atici SF, Ansari R, Allareddy V, Suhaym O, Cetin AE, Elnagar MH. Fully automated determination of the cervical vertebrae maturation stages using deep learning with directional filters. PLoS One. 2022;17(7):e0269198.
26. McInnes MD, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, 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):388–96.
27. Akay G, Akcayol MA, Özdem K, Güngör K. Deep convolutional neural network—The evaluation of cervical vertebrae maturation. Oral Radiol. 2023;39(4):629–38.
28. Khazaei M, Mollabashi V, Khotanlou H, Farhadian M. Automatic determination of pubertal growth spurts based on the cervical vertebral maturation staging using deep convolutional neural networks. J World Fed Orthod. 2023;12(2):56–63.
29. Li H, Li H, Yuan L, Liu C, Xiao S, Liu Z, et al. The psc-CVM assessment system: a three-stage type system for CVM assessment based on deep learning. BMC Oral Health. 2023;23(1):557.
30. Makaremi M, Lacaule C, Mohammad-Djafari A. Deep learning and artificial intelligence for the determination of the cervical vertebra maturation degree from lateral radiography. Entropy. 2019;21(12):1222.
31. Kim E-G, Oh I-S, So J-E, Kang J, Le VNT, Tak M-K, et al. Estimating cervical vertebral maturation with a lateral cephalogram using the convolutional neural network. J Clin Med. 2021;10(22):5400.
32. Makaremi M, Lacaule C, Mohammad-Djafari A, editors. Determination of the cervical vertebra maturation degree from lateral radiography. Proceedings; 2020: MDPI.
33. Kök H, İzgi MS, Acılar AM. Evaluation of the artificial neural network and Naive Bayes models trained with vertebra ratios for growth and development determination. Turk J Orthod. 2020;34(1):2.
34. Kök H, Izgi MS, Acilar AM. Determination of growth and development periods in orthodontics with artificial neural network. Orthod Craniofac Res. 2021;24:76–83.
35. Amasya H, Yildirim D, Aydogan T, Kemaloglu N, Orhan K. Cervical vertebral maturation assessment on lateral cephalometric radiographs using artificial intelligence: comparison of machine learning classifier models. Dentomaxillofac Radiol. 2020;49(5):20190441.
36. Amasya H, Cesur E, Yıldırım D, Orhan K. Validation of cervical vertebral maturation stages: artificial intelligence vs human observer visual analysis. Am J Orthod Dentofac Orthop. 2020;158(6):e173–9.
37. Mohammad-Rahimi H, Motamadian SR, Nadimi M, Hassanzadeh-Samani S, Minabi MA, Mahmoudinia E, et al. Deep learning for the classification of cervical maturation degree and pubertal growth spurts: a pilot study. Korean J Orthod. 2022;52(2):112–22.
38. Liao N, Dai J, Tang Y, Zhong Q, Mo S. ICVM: an interpretable deep learning model for CVM assessment under label uncertainty. IEEE J Biomed Health Inform. 2022;26(8):4325–34.
39. Li H, Chen Y, Wang Q, Gong X, Lei Y, Tian J, et al. Convolutional neural network-based automatic cervical vertebral maturation classification method. Dentomaxillofac Radiol. 2022;51(6):20220070.
40. Atici SF, Ansari R, Allareddy V, Suhaym O, Cetin AE, Elnagar MH. AggregateNet: A deep learning model for automated classification of cervical vertebrae maturation stages. Orthod Craniofac Res. 2023;26:111–7.
41. Radwan MT, Sin Ç, Akkaya N, Vahdettin L. Artificial intelligence-based algorithm for cervical vertebrae maturation stage assessment. Orthod Craniofac Res. 2023;26(3):349–55.
42. Xie L, Tang W, Izadikhah I, Chen X, Zhao Z, Zhao Y, et al. Intelligent quantitative assessment of skeletal maturation based on multi-stage model: a retrospective cone-beam CT study of cervical vertebrae. Oral Radiology. 2022;38(3):1–11.
43. Sokic E, Tiro A, Sokic-Begovic E, Nakas E. Semi-automatic assessment of cervical vertebral maturation stages using cephalograph images and centroid-based clustering. Acta Stomatologica Croatica. 2012;46(4):280–90.
44. Xie L, Tang W, Izadikhah I, Zhao Z, Zhao Y, Li H, et al. Development of a multi-stage model for intelligent and quantitative appraising of skeletal maturity using cervical vertebras cone-beam CT images of Chinese girls. Int J Comput Assist Radiol Surg. 2022;17(4):761–73.
45. Yang YM, Lee J, Kim YI, Cho BH, Park SB. Axial cervical vertebrae-based multivariate regression model for the estimation of skeletal-maturation status. Orthod Craniofac Res. 2014;17(3):187–96.
46. Baptista RS, Quaglio CL, Mourad LM, Hummel AD, Caetano CAC, Ortolani CLF, et al. A semi-automated method for bone age assessment using cervical vertebral maturation. Angle Orthod. 2012;82(4):658–62.
47. Feng X, Lu S, Li Y, Lin J. Establishment of an intelligent cervical vertebrae maturity assessment system based on cone beam CT data. Zhejiang Da Xue Xue Bao Yi Xue Ban = J Zhejiang University Med Sci. 2021;50(2):187–94.
48. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.

49. McNamara JA Jr, Franchi L. The cervical vertebral maturation method: a user's guide. Angle Orthod. 2018;88(2):133–43.
50. Khanagar SB, Al-Ehaideb A, Vishwanathaiah S, Maganur PC, Patil S, Naik S, et al. Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making-a systematic review. J Dent Sci. 2021;16(1):482–92.
51. Kim DW, Kim J, Kim T, Kim T, Kim YJ, Song IS, et al. Prediction of hand-wrist maturation stages based on cervical vertebrae images using artificial intelligence. Orthod Craniofac Res. 2021;24:68–75.
52. Hwang H-W, Moon J-H, Kim M-G, Donatelli RE, Lee S-J. Evaluation of automated cephalometric analysis based on the latest deep learning method. Angle Orthod. 2021;91(3):329–35.
53. Nishimoto S, Sotsuka Y, Kawai K, Ishise H, Kakibuchi M. Personal Computer-based cephalometric landmark detection with deep learning, using cephalograms on the internet. J Craniofac Surg. 2019;30(1):91–5.
54. Wang H, Minnema J, Batenburg KJ, Forouzanfar T, Hu FJ, Wu G. Multiclass CBCT image segmentation for orthodontics with deep learning. J Dent Res. 2021;100(9):943–9.
55. Sin Ç, Akkaya N, Aksoy S, Orhan K, Öz U. A deep learning algorithm proposal to automatic pharyngeal airway detection and segmentation on CBCT images. Orthod Craniofac Res. 2021;24(S2):117–23.
56. Lee S-C, Hwang H-S, Lee KC. Accuracy of deep learning-based integrated tooth models by merging intraoral scans and CBCT scans for 3D evaluation of root position during orthodontic treatment. Prog Orthod. 2022;23(1):15.
57. Mathew R, Palatinus S, Padala S, Alshehri A, Awadh W, Bhandi S, et al. Neural networks for classification of cervical vertebrae maturation: a systematic review. Angle Orthod. 2022;92(6):796–804.
58. Kazimierczak W, Jedliński M, Issa J, Kazimierczak N, Janiszewska-Olszowska J, Dyszkiewicz-Konwińska M, et al. Accuracy of artificial intelligence for cervical vertebral maturation assessment-a systematic review. J Clin Med. 2024;13(14):4047. https://doi.org/10.3390/jcm13144047.
59. Bailly A, Blanc C, Francis É, Guillotin T, Jamal F, Wakim B, et al. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. Comput Methods Programs Biomed. 2022;213:106504.
60. Brady AP, Neri E. Artificial intelligence in radiology—ethical considerations. Diagnostics. 2020;10(4):231.
61. Mörch C, Atsu S, Cai W, Li X, Madathil S, Liu X, et al. Artificial intelligence and ethics in dentistry: a scoping review. J Dent Res. 2021;100(13):1452–60.
62. Naik N, Hameed BZ, Shetty DK, Swain D, Shah M, Paul R, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? Front Surg. 2022;9:862322.
63. Farhud DD, Zokaei S. Ethical issues of artificial intelligence in medicine and healthcare. Iran J Public Health. 2021;50(11):i.

Publisher's Note