Journal of Taibah University Medical Sciences (2026) 21(4), 683–697

Taibah University

Journal of Taibah University Medical Sciences

www.sciencedirect.com

Review Article

Artificial intelligence for skeletal classification in orthodontics: A systematic review and meta-analysis of lateral cephalometric studies

Mahmood Dashti, DDS [a,b,*], Farshad Khosraviani, DDS [c], Sara Esmaeili, DDS [d], Niloofar Ghadimi, DDS, MSc [e], Kaan Orhan, DDS, MSc [f], Atieh Meyari, DDS, MSc [g], Zohaib Khurshid, DDS, MSc [h,i,j] and Falk Schwendicke, DDS, MDPH [k]

[a] *Dentofacial Deformities Research Center, Research Institute of Dental Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
[b] *Department of Artificial Intelligence Engineering, Graduate School of Natural and Applied Sciences, Istinye University, Istanbul, Türkiye*
[c] *UCLA School of Dentistry, CA, USA*
[d] *School of Dentistry, Başkent University, Ankara, Türkiye*
[e] *Dental Materials Research Center, TeMS.C., School of Dentistry, Azad University of Medical Sciences, Tehran, Iran*
[f] *Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Ankara University, Ankara, Turkey*
[g] *Department of Esthetic & Restorative Dentistry, School of Dentistry, Tehran Azad University of Medical Sciences, Tehran, Iran*
[h] *Department of Prosthodontics and Dental Implantology, College of Dentistry, King Faisal University, Al-Ahsa, KSA*
[i] *Department of Anatomy, Faculty of Dentistry, Center of Artificial Intelligence and Innovation (CAII), Center of Excellence for Dental Stem Cell Biology, Chulalongkorn University, Bangkok, Thailand*
[j] *Department of Oral and Maxillofacial Surgery, Biruni University, Turkey*
[k] *Conservative Dentistry, Periodontology and Digital Dentistry, University Hospital of the Ludwig-Maximilians-University, Munich, Germany*

Received 11 December 2025; revised 14 May 2026; accepted 11 June 2026; Available online 1 July 2026

الملخص

أهداف البحث: تجميع الأدلة المتعلقة بالأداء التشخيصي للذكاء الاصطناعي في التصنيف الهيكلي السهمي باستخدام الصور الشعاعية السيفالومترية الجانبية للرأس، وتحديد الفجوات المنهجية التي قد تؤثر في تطبيقه سريريًا.

طرق البحث: وفقًا لإرشادات PRISMA، جرى البحث في قواعد بيانات PubMed وEmbase وScopus وWeb of Science حتى 13 يوليو 2025. شملت الدراسات المؤهلة تلك التي طبّقت الذكاء الاصطناعي على الصور السيفالومترية الجانبية لتصنيف العلاقات الهيكلية إلى الصنف الأول والثاني والثالث، مع الإبلاغ عن مؤشرات الأداء التشخيصي. واستُبعدت النماذج غير المعتمدة على الصور، والدراسات التي اقتصرت على تحديد النقاط المرجعية، وتقييمات البرمجيات التجارية، والتقارير غير الأصلية. جرى تقييم خطر التحيز باستخدام أداة QUADAS-2. كما أُجريت تحليلات تلوية استكشافية باستخدام نموذج التأثيرات العشوائية بطريقة الاحتمالية العظمى المقيدة REML، مع إجراء تحليلات فرعية حسب الصنف الهيكلي. ونظرًا لمحدودية عدد الدراسات المشمولة، ووجود تباين ملحوظ بينها، وتعدد النتائج على مستوى النماذج المستندة إلى مجموعات بيانات مشتركة، فُسرت التقديرات المجمعة بحذر.

النتائج: جرى فحص 91 سجلًا، وخضع 12 منها للمراجعة بالنص الكامل، واستوفت أربع دراسات معايير الاشتمال، بينما أُدرجت ثلاث دراسات في التحليل الكمي المجمع، بإجمالي حجم عينة اختبار فعّال بلغ 2743، وتسعة أذرع للنماذج. واستُبعدت دراسة واحدة من التحليل المجمع بسبب عدم قابليتها للمقارنة، إذ استخدمت مدخلات مشتركة من الصور الخلفية الأمامية والجانبية، واعتمدت على الإبلاغ عن أفضل تقسيم تدريبي. بلغ الأداء المجمع للتصنيف متعدد الفئات: حساسية قدرها 87.0% بفاصل ثقة 95%: 85.6–88.4، ونوعية قدرها 91.1% بفاصل ثقة 95%: 87.3–94.9، ودقة كلية قدرها 85.3% بفاصل ثقة 95%: 79.9–90.7، ومساحة تحت منحنى ROC قدرها 0.94 بفاصل ثقة 95%: 0.90–0.98. وقد أظهرت نماذج DenseNet أداءً أفضل عمومًا، في حين سجّل نموذج المحوّلات Swin-T الأداء الأضعف. وأشار تقييم QUADAS-2 إلى وجود خطر تحيز منخفض إلى متوسط، مع ملاحظات تتعلق بالاعتماد على عينات من مركز واحد وغياب التحقق الخارجي.

* Corresponding address: Department of Artificial Intelligence Engineering, Graduate School of Natural and Applied Sciences. University: Istinye University, Istanbul, Türkiye.
E-mail: dashti.mahmood72@gmail.com (M. Dashti)
Peer review under responsibility of Taibah University.

Production and hosting by Elsevier

1658-3612 © 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). https://doi.org/10.1016/j.jtumed.2026.06.003

الاستنتاجات: يُظهر الذكاء الاصطناعي أداءً تشخيصيًا واعدًا في التصنيف الهيكلي السهمي اعتمادًا على الصور السيفالومترية الجانبية للرأس، إلا أن قاعدة الأدلة ما تزال محدودة على مستوى الدراسات، ومتباينة، وتعتمد غالبًا على دراسات أحادية المركز. لذلك، ينبغي تفسير التقديرات المجمعة بوصفها تقديرات استكشافية، وليست تقديرات نهائية للأداء السريري.

الكلمات المفتاحية: الذكاء الاصطناعي؛ القياسات السيفالومترية؛ التعلم العميق؛ التشخيص؛ تقويم الأسنان

Abstract

Objective: To synthesize the available evidence on the diagnostic performance of artificial intelligence for sagittal skeletal classification using lateral cephalometric radiographs and to identify methodological limitations affecting clinical translation.

Methods: PubMed, Embase, Scopus, and Web of Science were searched up to July 13, 2025, following PRISMA guidelines. Eligible studies used artificial intelligence to classify skeletal Class I, II, and III relationships from lateral cephalograms and reported diagnostic performance metrics. Non-image-based models, landmarking-only studies, commercial software evaluations, and non-original reports were excluded. Risk of bias was assessed using QUADAS-2. Exploratory random-effects meta-analyses using restricted maximum likelihood were performed, with subgroup analyses by skeletal class. Because of the small number of eligible studies, heterogeneity, and multiple model-level results from shared data sets, pooled estimates were interpreted cautiously.

Results: Ninety-one records were screened, 12 underwent full-text review, four met the inclusion criteria, and three were included in quantitative synthesis, comprising an effective test sample of 2,743 across nine model arms. One study was excluded from pooling because it used combined posteroanterior and lateral inputs with best-fold reporting, limiting comparability. The pooled multiclass performance showed sensitivity of 87.0%, specificity of 91.1%, accuracy of 85.3%, and area under the receiver operating characteristic curve of 0.94. DenseNet-based models generally showed the strongest performance, whereas the Swin-T transformer model showed the weakest performance. QUADAS-2 indicated low to moderate risk of bias, with major concerns related to single-center sampling and lack of external validation.

Conclusions: Artificial intelligence shows promising diagnostic performance for sagittal skeletal classification from lateral cephalograms. However, the current evidence remains limited, heterogeneous, and mainly single center. Therefore, pooled estimates should be considered exploratory rather than definitive indicators of clinical performance.

Keywords: Artificial intelligence; Cephalometry; Deep learning; Diagnosis; Orthodontics

© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

The demand for orthodontic treatment remains high worldwide due to the significant prevalence of malocclusion and dentoskeletal discrepancies.[1,2] Malocclusion is an esthetic concern but also affects oral function, psychosocial well-being, and overall quality of life.[3,4] The classification of skeletal patterns has become a standard diagnostic procedure in orthodontics and serves as the basis for formulating appropriate treatment strategies.[5]

Cephalometric analysis has a central role in orthodontic diagnosis and treatment planning by providing essential anatomical measurements that facilitate morphometric quantification of craniofacial growth and the evaluation of spatial relationships between hard and soft dentomaxillofacial structures. This process is indispensable for accurate diagnosis, individualized treatment planning, and assessing treatment outcomes.[6,7] Since its introduction by Broadbent and Hofrath in the 1930s, cephalometric radiography has remained a fundamental diagnostic tool in orthodontics.[8,9] Precisely identifying an individual's skeletal class, particularly the sagittal relationship between the maxilla and mandible, is crucial for achieving stable, long-term orthodontic results.[10] Determining an anteroposterior discrepancy is also critical for distinguishing sagittal from vertical skeletal anomalies, ensuring that treatment is directed appropriately in the correct dimension.[11] Skeletal Class I represents an ideal anteroposterior relationship between the jaws, whereas skeletal Class II is characterized by a retrognathic mandible and/or a prognathic maxilla, and Class III involves a prognathic mandible and/or a retrognathic maxilla. These sagittal disharmonies are observed in up to one-quarter of the population, although their prevalence varies significantly among ethnic groups.[12,13]

In recent years, artificial intelligence (AI) technologies, particularly convolutional neural networks (CNNs), have exhibited promise in applications to medical imaging and diagnostics.[14–16] These methods have increasingly been applied in orthodontics, especially cephalometric analysis, where both traditional machine-learning approaches[17,18] and advanced deep learning models[19,20] are used to improve the reliability, consistency, and accuracy of landmark identification and skeletal classification in both two-dimensional and three-dimensional contexts. However, most previous studies were based on relatively small or single-center data sets, thereby limiting the generalizability of their findings.[21]

Previous syntheses of evidence have examined the application of AI in dental imaging and cephalometric workflows, but their focus differed from that of the present review. Reviews of the use of deep learning and CNNs in dental image diagnostics have demonstrated promising performance across several dental imaging tasks, as well as highlighting the heterogeneity of data sets, model architectures, validation strategies, and reporting quality.[15,16] In cephalometry, systematic reviews and meta-analyses have mainly focused on automated landmark detection and the tracing accuracy, showing that AI can localize cephalometric landmarks with clinically useful performance, although errors can vary depending on the landmark, image quality, and study design.[7,20] A related review of mandibular growth prediction

also emphasized that previous AI studies remain limited by issues such as small data sets, retrospective designs, and insufficient external validation.[21] Therefore, although these syntheses support the potential use of AI in orthodontic imaging, they have not directly shown whether image-based AI can classify sagittal skeletal Class I, II, and III patterns from lateral cephalometric radiographs as a diagnostic endpoint. Thus, in the present review, we addressed a narrower and clinically distinct question by focusing on skeletal classification as an endpoint rather than cephalometric preprocessing or landmark identification.

Several recent studies have explored the potential of AI for skeletal classification using lateral cephalometric radiographs. Yu et al.[22] first demonstrated the feasibility of using CNNs for end-to-end skeletal diagnosis and reported accuracies above 90% across sagittal and vertical classifications. Jeong et al.[23] extended this work by showing that cranio-spinal features visible on cephalograms could be employed for skeletal class prediction, achieving high accuracy even when mandibular structures were masked. Li et al.[24] compared multiple deep learning architectures and confirmed that advanced networks such as DenseNet performed better than shallower CNNs (e.g., VGG16 and GoogLeNet) for differentiating Class I, II, and III skeletal patterns. More recently, Nan et al.[25] provided the first pediatric-focused data set, and applied DenseNet to children's cephalograms and reported accuracies above 90%. In addition, Midlej et al. investigated machine learning approaches using cephalometric parameters in Arab populations, achieving very high accuracy in Class II versus III classification[26] and reliable prediction of Class I versus II patterns.[27] Overall, these studies highlight the promise of using AI for skeletal diagnosis, but also the heterogeneity of methods, populations, and outcomes, necessitating systematic synthesis and quantitative evaluation.

AI has been widely studied in dental imaging and cephalometric analysis, but these studies mainly focused on landmark detection, automated tracing, measurement reproducibility, or broader diagnostic applications rather than direct skeletal Class I, II, and III classification from lateral cephalometric radiographs.[7,21,23] This distinction is clinically important because skeletal classification is a diagnostic endpoint that directly informs orthodontic treatment planning. In addition, these primary studies differed substantially in terms of their data set composition, imaging inputs, model architecture, and reporting of diagnostic metrics, making it difficult to judge the consistency and clinical relevance of current evidence.

The present systematic review and meta-analysis aims to comprehensively synthesize the current evidence and quantitatively assess the diagnostic performance of image-based AI models for skeletal class classification using lateral cephalometric radiographs.

Methods

Protocol and reporting

This systematic review and meta-analysis was conducted in accordance with established methodological standards for diagnostic accuracy reviews and reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement.[28,29]

Search strategy

Comprehensive electronic search was performed based on PubMed, Embase, Scopus, Web of Science, and an additional Scopus secondary search from database inception to July 13, 2025. The search combined terms related to AI, machine learning, deep learning, cephalometry, and skeletal classification. The search terms were selected to align closely with the predefined review question and to maintain specificity for lateral cephalometric radiographs; therefore, broader terms that could capture non-target imaging modalities were not prioritized (see Table 1).

Eligibility criteria

Studies were included if they met all of the following criteria:

1. Human participants undergoing orthodontic assessment with lateral cephalometric radiographs were included.
2. A lateral cephalometric radiograph was used as a primary imaging input. In this review, primary imaging input refers to the lateral cephalogram being directly analyzed by the AI model for skeletal classification, rather than only using it to manually extract cephalometric measurements before model development. Models could use lateral cephalograms alone or in combination with additional demographic or imaging inputs, provided that the lateral cephalogram was directly analyzed by the model.
3. An AI, machine-learning, or deep-learning model was applied to classify sagittal skeletal Class I, Class II, and/or Class III patterns.
4. At least one diagnostic performance metric relevant to skeletal classification was reported, including the accuracy, sensitivity, specificity, precision, F1-score, or area under the receiver operating characteristic curve (AUC).
5. Original peer-reviewed research articles.

Studies were excluded if met any of the following criteria:

1. Imaging modalities without lateral cephalograms were used, such as panoramic radiographs, cone-beam computed tomography alone, or synthetic images.
2. They focused only on cephalometric landmark detection, measurement reproducibility, head posture analysis, cervical maturation, craniofacial growth prediction, or forensic applications without skeletal Class I, II, or III classification as an outcome.
3. Machine learning was conducted based only on manually derived cephalometric parameters, without direct image analysis of lateral cephalometric radiographs.
4. Commercial AI-assisted cephalometric tracing software was evaluated without reporting skeletal classification outcomes.
5. Reviews, case reports, conference abstracts, theses, or reports without sufficient diagnostic performance data.

Table 1: Summary of search queries used for different databases.

Database	Keyword	Result
PubMed	(("Machine Learning"[Mesh] OR "Deep Learning"[Mesh] OR "Supervised Machine Learning"[Mesh] OR "Unsupervised Machine Learning"[Mesh] OR "Neural Networks, Computer"[Mesh] OR "artificial intelligence") AND ("lateral cephalometry" OR "lateral cephalometric")) AND ("skeletal" OR "skeletal classification" OR "class I″ OR "class II" OR "class III")	16
Embase	('machine learning'/exp OR 'machine learning' OR 'deep learning'/exp OR 'deep learning' OR 'supervised machine learning'/exp OR 'supervised machine learning' OR 'unsupervised machine learning'/exp OR 'unsupervised machine learning' OR 'artificial neural network'/exp OR 'artificial neural network' OR 'artificial intelligence') AND (' 'lateral cephalometry' OR 'lateral cephalometric') AND ('skeletal' OR 'skeletal classification' OR 'class i' OR 'class ii' OR 'class iii')	18
Scopus	(TITLE-ABS-KEY ("Machine Learning") OR TITLE-ABS-KEY ("Deep Learning") OR TITLE-ABS-KEY ("Supervised Machine Learning") OR TITLE-ABS-KEY ("Unsupervised Machine Learning") OR TITLE-ABS-KEY ("Neural Network") OR TITLE-ABS-KEY ("artificial intelligence")) AND (TITLE-ABS-KEY ("lateral cephalometry") OR TITLE-ABS-KEY ("lateral cephalometric")) AND (TITLE-ABS-KEY ("skeletal") OR TITLE-ABS-KEY ("skeletal classification") OR TITLE-ABS-KEY ("class I″) OR TITLE-ABS-KEY ("class II") OR TITLE-ABS-KEY ("class III"))	31
Scopus secondary	(TITLE-ABS-KEY ("Machine Learning") OR TITLE-ABS-KEY ("Deep Learning") OR TITLE-ABS-KEY ("Supervised Machine Learning") OR TITLE-ABS-KEY ("Unsupervised Machine Learning") OR TITLE-ABS-KEY ("Neural Network") OR TITLE-ABS-KEY ("artificial intelligence")) AND (TITLE-ABS-KEY ("lateral cephalometry") OR TITLE-ABS-KEY ("lateral cephalometric")) AND (TITLE-ABS-KEY ("skeletal") OR TITLE-ABS-KEY ("skeletal classification") OR TITLE-ABS-KEY ("class I″) OR TITLE-ABS-KEY ("class II") OR TITLE-ABS-KEY ("class III"))	1
Web of Science	((TS=("Machine Learning"[Mesh] OR "Deep Learning"[Mesh] OR "Supervised Machine Learning"[Mesh] OR "Unsupervised Machine Learning"[Mesh] OR "Neural Networks, Computer"[Mesh] OR "artificial intelligence")) AND TS=("lateral cephalometry" OR "lateral cephalometric")) AND TS=("skeletal" OR "skeletal classification" OR "class I″ OR "class II" OR "class III")	25

Study question and PICO

The review question was: "In lateral cephalometric radiographs, how accurately can AI-based algorithms classify skeletal Classes I, II, and III compared with ground truth labels from conventional cephalometric analysis."

The population (P) consisted of patients undergoing orthodontic assessment with lateral cephalometric radiographs. The intervention (I) was applying AI, machine learning, or deep learning models for skeletal classification. The comparison (C) denoted the reference skeletal class assigned according to standard cephalometric diagnostic criteria. The outcome (O) was the diagnostic performance quantified as accuracy, sensitivity, specificity, or AUC.

Study selection

All records identified through database searches were imported into a reference manager and duplicates were removed. Two reviewers (NG and FK) independently screened titles and abstracts to exclude clearly irrelevant articles. The full texts of potentially eligible studies were then assessed in detail against the inclusion and exclusion criteria. Disagreements were resolved by discussion, and a third reviewer (MD) acted as arbiter when necessary. The study selection process is summarized in the PRISMA flow diagram shown in Figure 1.

Data extraction

For each of the studies included, two reviewers independently extracted data using a standardized data extraction form developed for this review. The extracted variables included author, year, country, study population, data set size and splits, skeletal classification scheme, reference standard, AI model architecture, training and validation strategy, and reported diagnostic performance metrics (accuracy, sensitivity, specificity, AUC, and positive predictive value (PPV) and F1 score when available). For studies that involved k-fold cross-validation, the total number of unique participants was recorded and the mean performance across folds was extracted. The extraction form was reviewed by the study team before full extraction. Discrepancies were resolved by discussion and by consultation with a third reviewer (MD) when required. A formal inter-reviewer agreement statistic was not calculated.

Risk of bias assessment

Risk of bias and applicability concerns were assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool,[30] which evaluates four domains for risk of bias (patient selection, index test, reference standard, and flow and timing) and three domains for applicability (patient selection, index test, reference standard). Two reviewers (NG and FK) independently rated each study,

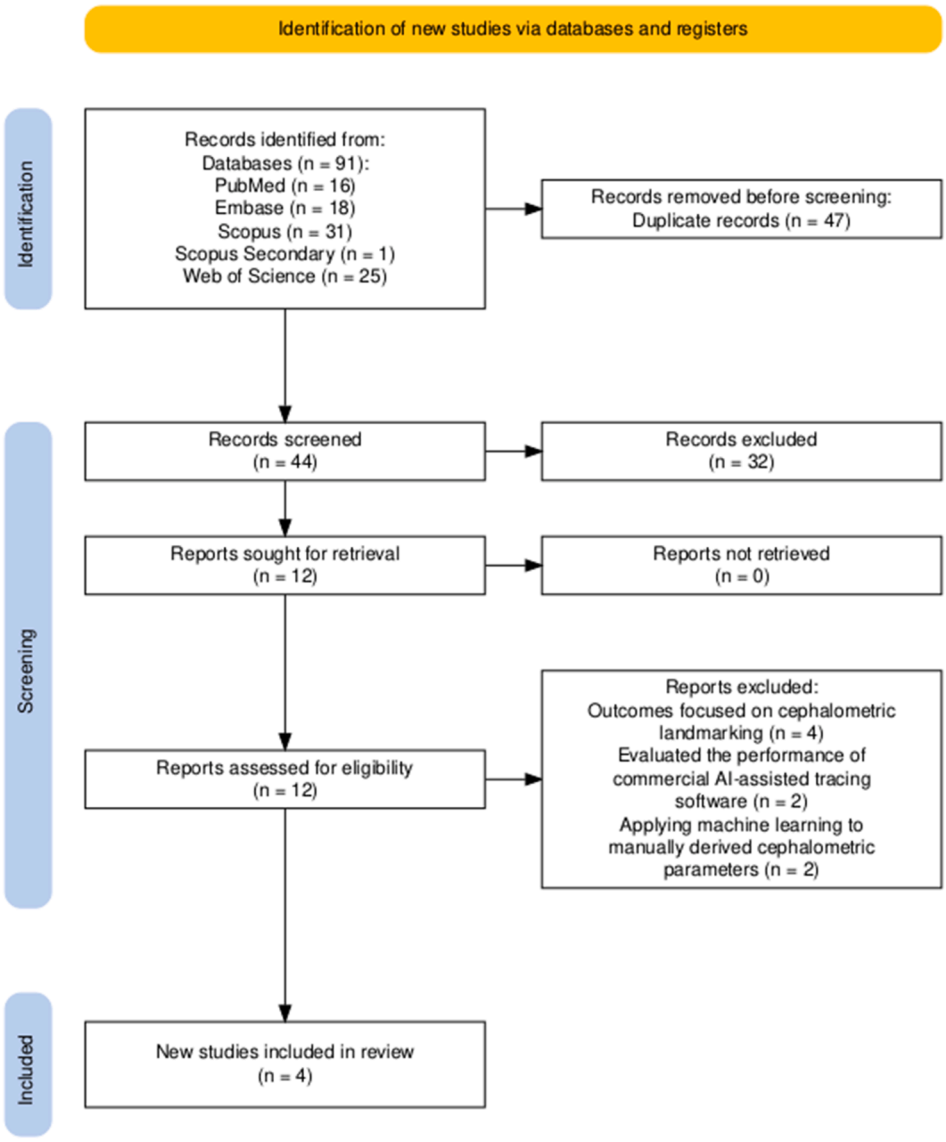


Figure 1: PRISMA flow diagram for selected studies.

and any disagreements were resolved by consulting with a third reviewer (MD). The QUADAS-2 assessment results are summarized in Table 4.

Statistical analysis

Quantitative synthesis was conducted for multiclass classification studies that reported extractable summary metrics, including accuracy, sensitivity, specificity, and AUC. Binary classification outcomes were analyzed narratively. The included studies reported heterogeneous multiclass outcome structures and did not provide sufficiently comparable paired diagnostic data to support robust hierarchical modeling of sensitivity and specificity, so hierarchical summary receiver operating characteristic and bivariate diagnostic meta-analysis models were not used. Instead, the reported summary metrics were pooled separately after logit transformation and then back-transformed for interpretability. Therefore, these pooled estimates should be interpreted as exploratory summary measures rather than definitive hierarchical diagnostic accuracy estimates. Several of the studies reported evaluations of more than one AI architecture based on the same underlying data set. Random-effects models with restricted maximum likelihood were used because heterogeneity of the data sets, imaging inputs, model architectures, and evaluation protocols was expected. Heterogeneity was quantified using I^2 and τ^2.[31]

These model-level results were retained for descriptive comparison of architectures, but they did not represent independent study populations. Accordingly, the pooled quantitative results may have been affected by within-study dependence and should be interpreted with caution. All quantitative analyses, including metric transformation, random-effects pooling, and heterogeneity estimation, were performed in Stata 17 (StataCorp, College Station, TX, USA).

Certainty of evidence

An overall certainty-of-evidence assessment was not performed. Given the very small number of diagnostic accuracy studies included, substantial methodological heterogeneity across data sets, inputs, and evaluation designs, and the dependence introduced by multiple model architectures reported within individual studies, we considered that a formal certainty framework would be of limited interpretive value in this review. This is acknowledged as a limitation of the present synthesis.

Only a small number of studies were included but these studies evaluated multiple AI architectures, facilitating exploratory comparisons of model performance; however, several models were assessed within shared data sets, and thus did not represent independent study populations.

Data synthesis

The primary outcomes were accuracy, sensitivity, specificity, and AUC. Subgroup analyses were performed for skeletal Class I, II, and III where possible. Narrative synthesis was provided for studies that could not be pooled due to methodological incomparability.

Results

Study selection

In total, 91 articles were retrieved from the five databases searched. After removing 47 duplicates, 44 unique records remained. Title and abstract screening excluded 32 articles that did not address skeletal classification using AI. Twelve articles were selected for full-text review. Among these 12 articles, four were excluded because their outcomes focused on cephalometric landmarking or measurement reproducibility,[32–35] two were excluded because they evaluated the performance of commercial AI-assisted tracing software rather than skeletal classification models,[36,37] and two studies were excluded that applied machine learning to manually derived cephalometric parameters rather than directly analyzing lateral cephalometric radiographs[26,27] as the focus of this review was image-based AI. Ultimately, four studies met the eligibility criteria and were included in the systematic review,[22–25] and three studies were selected for the meta-analysis.[22,24,25]

Among the four studies included, three contributed to the quantitative synthesis. The study by Jeong et al.[23] was retained in the qualitative review but excluded from the meta-analysis because it used combined posteroanterior and lateral imaging and reported the best-performing fold rather than averaged cross-validation results, thereby limiting comparability with the remaining studies. The PRISMA flow chart in Figure 1 illustrates the selection process.[29]

Study characteristics

The characteristics of the studies are summarized in Table 2. Extracted data covered details such as authorship, year of publication, country of origin, study population, data set size, reference standards used for skeletal classification, AI architecture or machine learning model applied, and reported diagnostic performance outcomes, including accuracy, sensitivity, specificity, and AUC (see Table 3).

Quantitative synthesis

Quantitative synthesis was possible for only three studies. Due to the small number of studies, substantial methodological heterogeneity, and inclusion of multiple model-level results from shared data sets, the resulting pooled estimates should be interpreted as exploratory summaries rather than definitive diagnostic accuracy estimates. For studies that reported explicit data set splits, the number of test subjects was extracted from the dedicated test set. For example, Yu et al.[22] reported sagittal classification test cohorts with 351 subjects in Class I, 228 in Class II, and 186 in Class III, totaling 765 cephalograms. Similarly, Li et al.[24] divided their data set into 70% for training, 15% for validation, and 15% for testing, corresponding to 365 cephalograms in the test set. By contrast, in studies that applied k-fold cross-validation,[25] the total data set size was treated as the effective number of test cases since each subject was included once in a test fold and the reported diagnostic metrics represented the mean across folds. For the study by Li et al.,[24] the average sensitivity, precision, and F1-scores were calculated by taking the means of the values across skeletal Classes I, II, and III, providing a single representative metric for inclusion in the meta-analysis.

Sensitivity: The overall pooled sensitivity was 87.02% (95% confidence interval (CI): 85.64–88.41%). Substantial heterogeneity was observed ($\tau^2 = 3.21$, $I^2 = 99.46\%$, $H^2 = 184.18$; $p < 0.001$). Class-specific exploratory summaries had sensitivities of 85.50% (95% CI: 80.82–90.17%; $I^2 = 69.21\%$) for Class I, 89.13% (95% CI: 86.52–91.74%; $I^2 = 0\%$) for Class II, and 91.20% (95% CI: 88.84–93.57%; $I^2 = 0\%$) for Class III. DenseNet-161 had the highest sensitivity among the model-level results at 89.76% (95% CI: 86.64–92.88%) (Figures 2 and 3). These findings should be interpreted with caution because the pooled results were derived from only three studies and several model-level entries originated from shared data sets.

Specificity: The pooled specificity was 91.08% (95% CI: 87.30–94.87%) and there was substantial heterogeneity ($\tau^2 = 24.37$, $I^2 = 96.83\%$, $H^2 = 31.56$; $p < 0.001$). Among the model-level results, DenseNet-121 achieved the highest specificity of 94.87% (95% CI: 91.81–97.93%), whereas Swin-T showed the lowest value of 80.04% (95% CI: 77.24–82.84%) (Figure 4). These pooled estimates should be interpreted as exploratory given the limited study-level evidence base.

Accuracy: The overall pooled accuracy was 85.25% (95% CI: 79.85–90.66%) and there was substantial heterogeneity ($\tau^2 = 65.66$, $I^2 = 96.83\%$, $H^2 = 31.55$; $p < 0.001$). DenseNet-121 had the highest accuracy among the model-level results at 90.50% (95% CI: 87.34–93.66%), whereas Swin-T had the lowest at 62.30% (95% CI: 57.07–67.53%) (Figure 5).

Table 2: Summary characteristics for each study.

Author (year)	Country	Data set test size	Architecture model	Accuracy (%)	Sensitivity/ recall (%)	Specificity (%)	Precision (PPV) (%)	F1-score/dice coefficient (%)	AUC
Jeong et al. (2021) (23)	South Korea	832	DenseNet-121	90.43	NR	NR	NR	NR	NR
Li et al. (2022) (24)	China	365	DenseNet-161	89.58	Avg: 89.76 Class I: 91.0 (85.0–94.0) Class II: 88.0 (81.0–93.0) Class III: 90.0 (83.0–94.0)	NR	Avg: 89.21 Class I: 83.0 (77.0–88.0) Class II: 93.0 (86.0–96.0) Class III: 93.0 (86.0–96.0)	Avg: 89.49 Class I: 87.0 (81.0–91.0) Class II: 90.0 (83.0–94.0) Class III: 92.0 (86.0–96.0)	Avg: 0.977 Class I: 0.955 Class II: 0.981 Class III: 0.993
			ResNet-152	89.04	Avg: 89.0 Class I: 89.0 (83.0–93.0) Class II: 88.0 (80.0–92.0) Class III: 90.0 (84.0–95.0)	NR	Avg: 89.25 Class I: 83.0 (75.0–87.0) Class II: 89.0 (82.0–94.0) Class III: 97.0 (92.0–99.0)	Avg: 89.13 Class I: 86.0 (79.0–90.0) Class II: 88.0 (81.0–93.0) Class III: 94.0 (88.0–97.0)	Avg: 0.974 Class I: 0.950 Class II: 0.979 Class III: 0.994
			VGG16	88.76	Avg: 88.48 Class I: 86.0 (79.0–91.0) Class II: 90.0 (83.0–94.0) Class III: 90.0 (84.0–95.0)	NR	Avg: 88.69 Class I: 84.0 (78.0–89.0) Class II: 88.0 (81.0–93.0) Class III: 95.0 (90.0–98.0)	Avg: 88.74 Class I: 85.0 (78.0–90.0) Class II: 89.0 (82.0–93.0) Class III: 93.0 (87.0–96.0)	Avg: 0.973 Class I: 0.951 Class II: 0.977 Class III: 0.993
			GoogleNet	87.94	Avg: 87.78 Class I: 80.0 (73.0–86.0) Class II: 91.0 (84.0–95.0) Class III: 94.0 (88.0–97.0)	NR	Avg: 87.96 Class I: 87.0 (80.0–92.0) Class II: 84.0 (77.0–90.0) Class III: 93.0 (87.0–96.0)	Avg: 88.0 Class I: 83.0 (76.0–89.0) Class II: 88.0 (80.0–92.0) Class III: 94.0 (87.0–96.0)	Avg: 0.972 Class I: 0.949 Class II: 0.976 Class III: 0.992
Nan et al. (2023) (25)	China	1613	Swin-T	62.30 ± 2.67	58.11 ± 2.74	80.04 ± 1.43	NR	NR	0.7881 ± 0.0154
			ResNet-101	85.49 ± 1.36	84.86 ± 1.62	92.96 ± 0.76			0.9557 ± 0.0051
			DenseNet-121	85.49 ± 0.89	83.99 ± 0.81	92.44 ± 0.40			0.9680 ± 0.0039
			ConvNeXt-T	86.30 ± 1.31	85.86 ± 1.23	93.14 ± 0.64			0.9600 ± 0.0039
Yu et al. (2020) (22)	South Korea	765	DenseNet-121	Avg: 90.50 (87.00–93.30) Class I: 85.75 (81.65–89.24) Class II: 92.31 (89.01–94.87) Class III: 93.45 (90.33–95.80)	Avg: 85.75 (78.21–91.39) Class I: 78.63 (70.09–85.67) Class II: 88.03 (80.74–93.30) Class III: 90.60 (83.80–95.21)	Avg: 92.88 (88.84–95.76) Class I: 89.32 (84.63–92.97) Class II: 94.44 (90.69–97.01) Class III: 94.87 (91.21–97.32)	NR	NR	Avg: 0.938 Class I: 0.889 Class II: 0.950 Class III: 0.967

NR: not reported.

Table 3: Exploratory pooled diagnostic performance metrics for skeletal classification AI models.

	Effect size (95% CI)	Class I	Class II	Class III
Sensitivity	87.02 (85.64–88.41) I^2: 99.46%	85.50 (80.82–90.17) I^2: 69.21%	89.13 (86.52–91.74) I^2: 0%	91.20 (88.84–93.57) I^2: 0%
Specificity	91.08 (87.30–94.87) I^2: 96.83%	–	–	–
Accuracy	85.25 (79.85–90.66) I^2: 96.83%	–	–	–
AUC	0.94 (0.90–0.98) I^2: 98.19%	0.94 (0.90–0.97) I^2: 0%	0.97 (0.95–1.00) I^2: 0%	0.99 (0.98–1.00) I^2: 0%

Table 4: Bias assessment for the selected studies.

Authors (year)	Risk of bias				Applicability concerns		
	Patients' selection	Index test	Reference standard	Flow and timing	Patients' selection	Index test	Reference standard
Jeong et al. (2021) (23)	Unclear	High	Low	Low	Unclear	Unclear	Low
Li et al. (2022) (24)	Unclear	Low	Low	Low	Low	Low	Low
Nan et al. (2023) (25)	Unclear	Unclear	Low	Low	Low	Low	Low
Yu et al. (2020) (22)	Unclear	Low	Low	Low	Low	Low	Low

Given the small number of contributing studies and the dependence among some model-level results, these findings should be interpreted with caution.

AUC: The overall pooled AUC was 0.94 (95% CI: 0.90–0.98), suggesting the promising discriminatory capacity of the studies included. However, significant heterogeneity was observed ($I^2 = 98.19\%$; $p < 0.001$). Exploratory subgroup summaries obtained pooled AUCs of 0.94 (95% CI: 0.90–0.97) for Class I, 0.97 (95% CI: 0.95–1.00) for Class II, and 0.99 (95% CI: 0.98–1.00) for Class III. Among the individual model-level results, DenseNet-161 had the highest AUC at 0.98 (95% CI: 0.92–1.00) (Figures 6 and 7). These subgroup findings should be interpreted with caution given the very limited number of contributing studies and the lack of independence between some model-level results.

Summary of findings

Across the primary pooled metrics, DenseNet-based models generally had the strongest performance among the model-level results. DenseNet-161 achieved the highest sensitivity at 89.76% (95% CI: 86.64–92.88%), DenseNet-121 had the highest specificity at 94.87% (95% CI: 91.81–97.93%) and accuracy at 90.50% (95% CI: 87.34–93.66%), and DenseNet-161 had the highest AUC at 0.98 (95% CI: 0.92–1.00), whereas Swin-T had the lowest overall performance. Secondary metrics, such as the PPV and F1-score when reported, had a similar general pattern, as shown in Table 2. These comparisons should be interpreted with caution because the evidence base was limited and several model-level results originated from shared data sets.

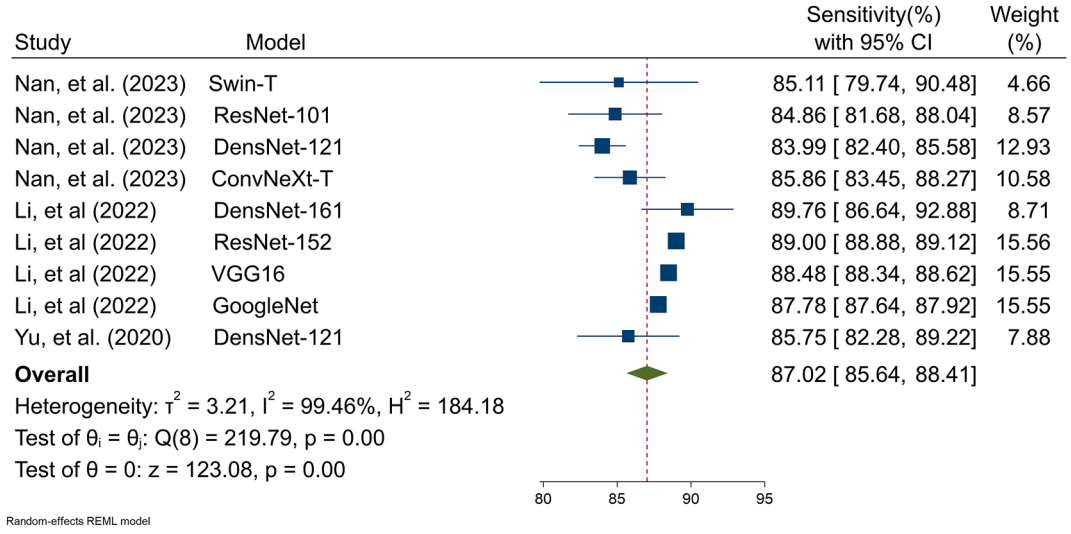


Figure 2: Exploratory forest plot of pooled sensitivity for skeletal classification models across studies. (Note: multiple architectures from the same data set are shown as model-level summaries within studies, and these are not independent observations.)

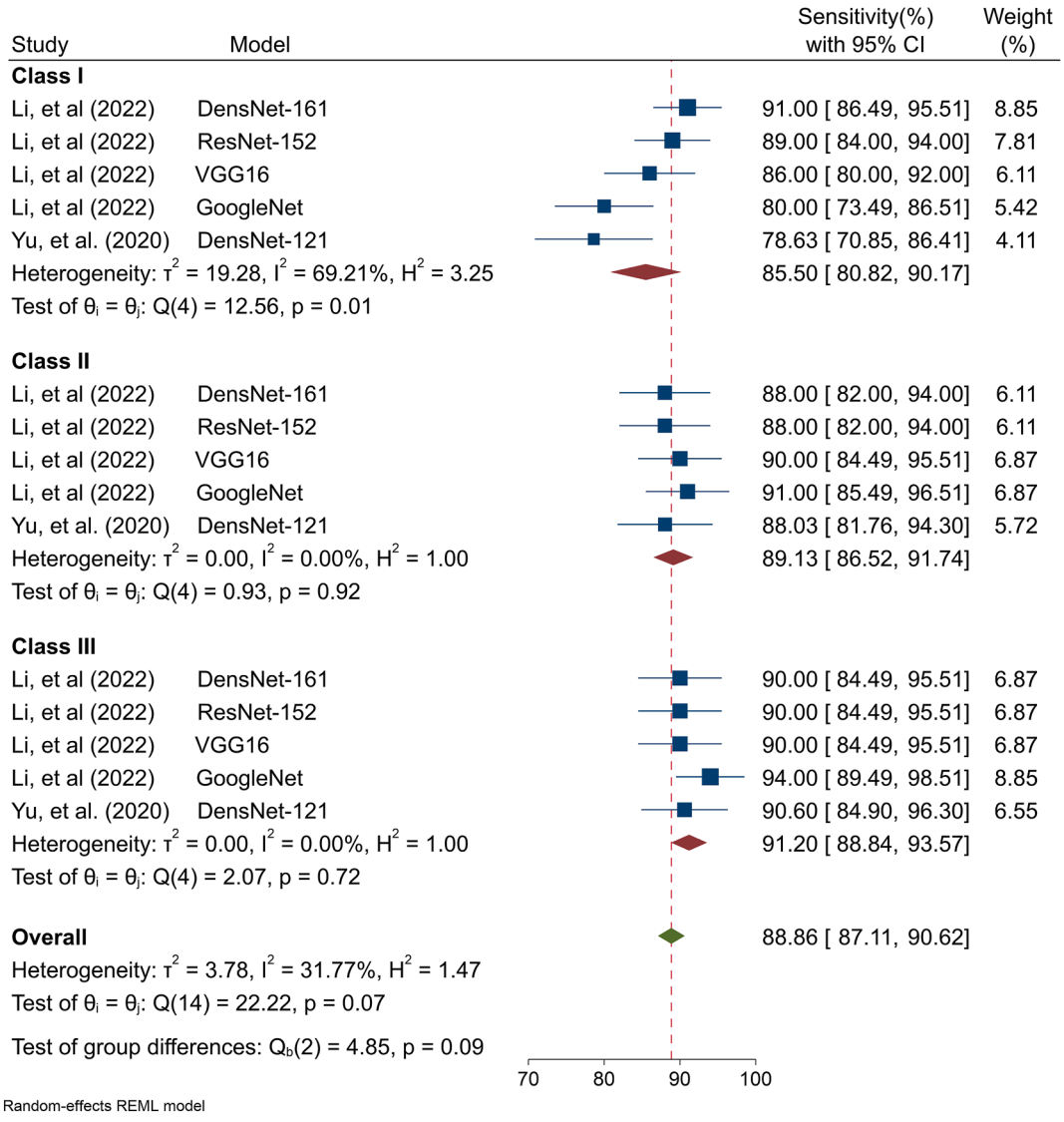


Figure 3: Exploratory subgroup forest plot showing pooled sensitivities for skeletal Class I, II, and III classifications. (Note: multiple architectures from the same data set are shown as model-level summaries within studies, and these are not independent observations.)

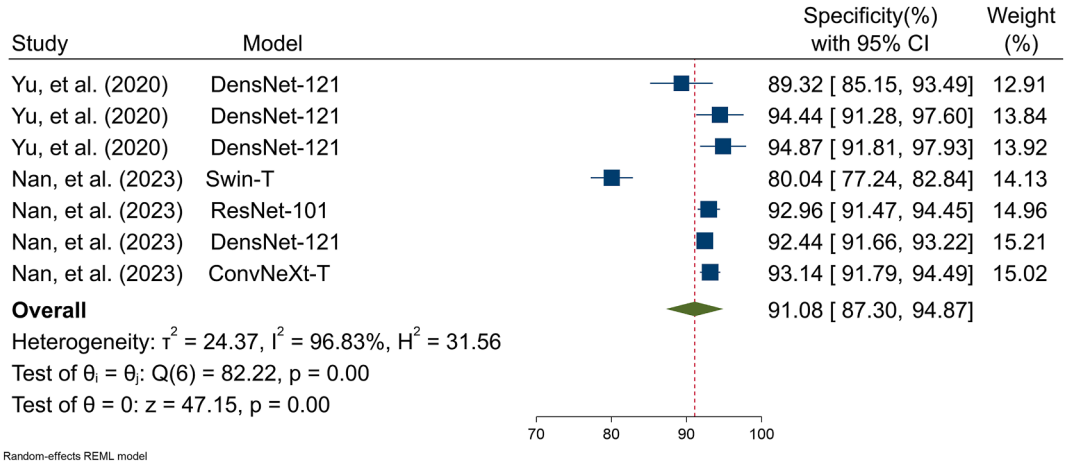


Figure 4: Exploratory forest plot of pooled specificity for skeletal classification models across studies. (Note: multiple architectures from the same data set are shown as model-level summaries within studies, and these are not independent observations.)

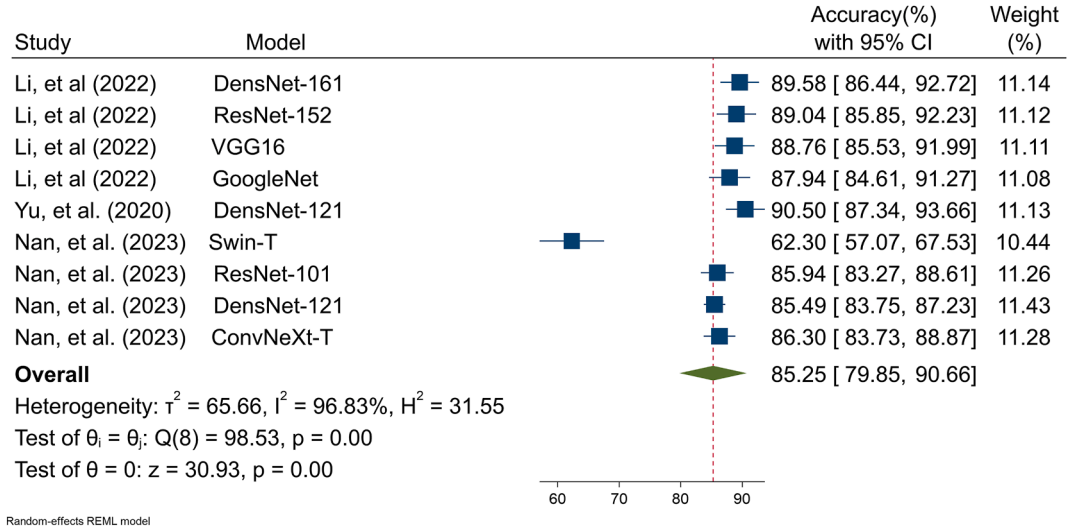


Figure 5: Exploratory forest plot of pooled accuracy for skeletal classification models across studies. (Note: multiple architectures from the same data set are shown as model-level summaries within studies, and these are not independent observations.)

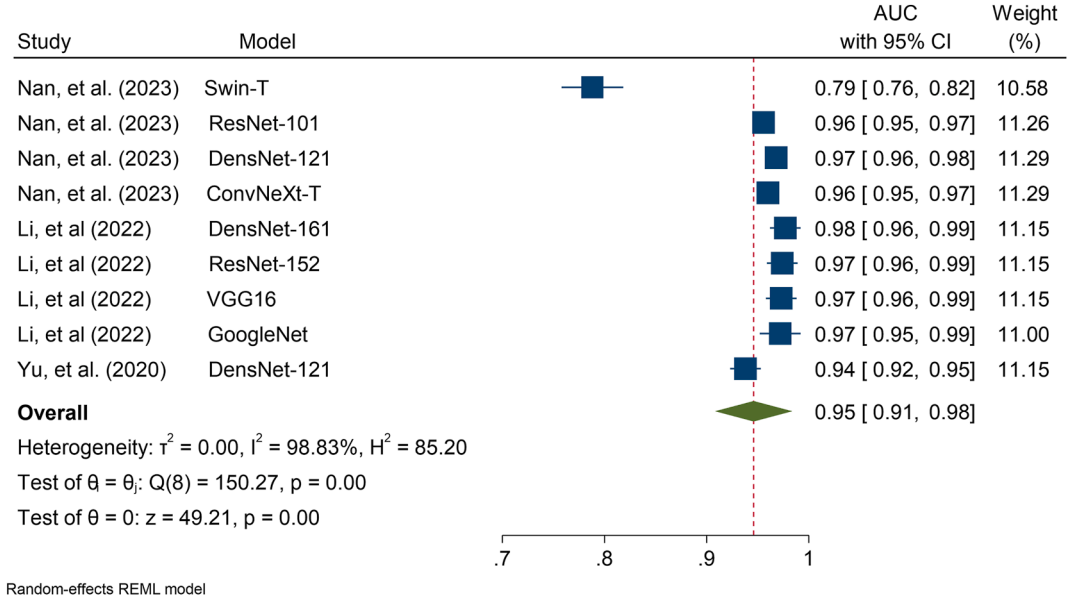


Figure 6: Exploratory forest plot of pooled AUC for skeletal classification models across studies. (Note: multiple architectures from the same data set are shown as model-level summaries within studies, and these are not independent observations.)

Risk of bias assessment

Risk of bias was evaluated using the QUADAS-2 tool across four domains (patient selection, index test, reference standard, and flow and timing), as well as three domains of applicability. Overall methodological quality among the studies included was moderate to high. The study by Li et al.[24] had the lowest risk of bias, with clearly defined data partitions (70% training, 15% validation, and 15% testing) and consistent use of a validated reference standard. The study by Yu et al.[22] also had low risk across most domains, although the retrospective single-center design and class-balancing procedure introduced some uncertainty regarding patient selection. The study by Nan et al.[25] had an unclear risk in the patient selection and index test domains because all performance metrics were derived from internal five-fold cross-validation without external validation, and borderline skeletal cases were selectively removed from the analysis. The study by Jeong et al.[23] had a high overall risk of bias due to the use of combined posteroanterior and lateral cephalograms, unclear enrollment strategy, and selective reporting of the best cross-validation fold rather than averaged results.

Applicability concerns were generally low for studies limited to lateral cephalograms,[22,24,25] which aligned with the target question addressed by this review of sagittal skeletal classification. The study by Jeong et al.[23] was excluded from quantitative synthesis because its multimodal design (posteroanterior + lateral radiographs) reduced comparability with single-projection analyses.

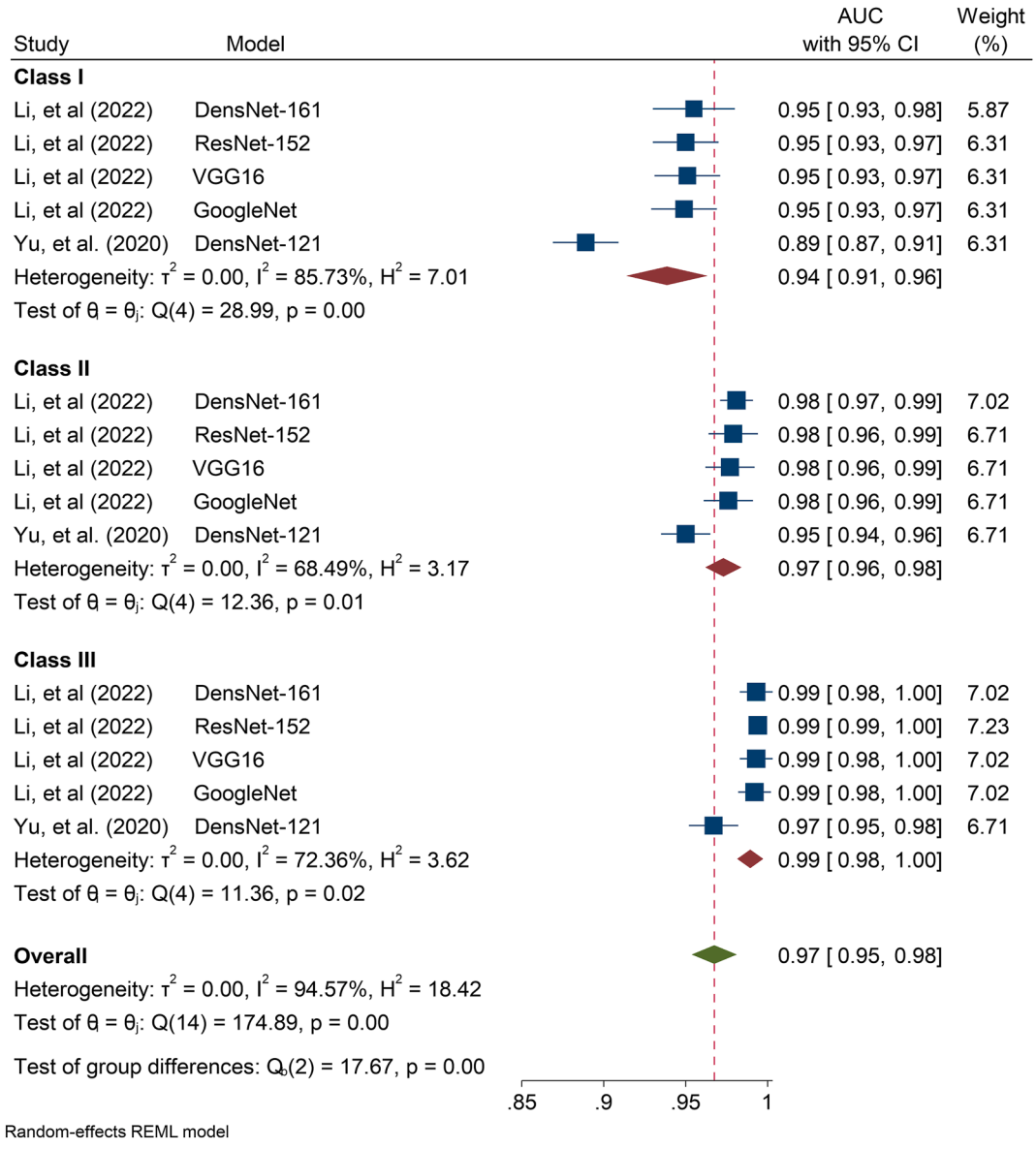


Figure 7: Exploratory subgroup forest plot showing pooled AUC for skeletal Class I, II, and III classifications. (Note: multiple architectures from the same data set are shown as model-level summaries within studies, and these are not independent observations.)

The QUADAS-2 evaluation is summarized in Table 4, where it indicated predominantly low risk in the reference standard and flow and timing domains, and only isolated high risk in the index test domain.

Overall, the studies included suggested promising diagnostic performance for the image-based AI models in sagittal skeletal classification, particularly with DenseNet-based architectures, although these findings should be interpreted with caution due to the small study-level evidence base, methodological heterogeneity, and lack of external validation.

Discussion

The results obtained in this systematic review and exploratory meta-analysis suggest that image-based AI models may achieve promising diagnostic performance at sagittal skeletal classification from lateral cephalometric radiographs. The pooled estimates indicated that the overall sensitivity, specificity, accuracy, and AUC were high. However, these findings should be interpreted with caution because the quantitative synthesis was based on only three studies, several model-level results originated from shared data sets, and substantial heterogeneity was observed across data sets, imaging inputs, model architectures, and validation approaches. Therefore, the pooled results should be viewed as exploratory summary estimates rather than definitive measures of clinical performance.

In the studies included, AI-based skeletal classification was evaluated using different model designs and input strategies. Yu et al.[22] demonstrated the feasibility of applying a CNN-based approach to lateral cephalograms for automated sagittal skeletal classification. Jeong et al.[23] extended this approach by using combined posteroanterior and lateral cephalograms, although their multimodal design and selective best-fold reporting limited comparability with

the other studies. Li et al.[24] compared several CNN architectures and found that DenseNet-161 performed strongly at sagittal craniofacial pattern classification. Nan et al.[25] further expanded the evidence base by evaluating deep-learning models in a pediatric population. Overall, these studies suggest that image-based AI can identify skeletal classification patterns from cephalometric images, but there were also important differences in the data sets, imaging inputs, model architecture, and validation design.

A consistent finding across the studies was that CNN-based architectures generally exhibited strong performance at skeletal classification. Earlier CNN architectures such as VGG and GoogLeNet have been widely used in image-classification tasks because they can learn hierarchical visual features through stacked convolutional operations.[38,39] However, deeper architectures with improved optimization mechanisms may be better suited to subtle craniofacial pattern recognition. ResNet introduced residual connections that support the training of deeper networks, while DenseNet further improved feature propagation by connecting each layer to subsequent layers.[40,41] This feature reuse may be particularly relevant in lateral cephalometric radiographs because the visual differences between skeletal Class I, II, and III patterns can be subtle, distributed across multiple craniofacial structures, and dependent on complex spatial relationships rather than a single isolated landmark.

The relatively strong performance of DenseNet-based models is also clinically plausible. Sagittal skeletal classification depends on the relationships between the maxilla, mandible, cranial base, and surrounding dentofacial structures. Therefore, CNN models that can preserve and reuse low-level and high-level image features may be well suited to identifying these patterns based on cephalometric images.[40,41] By contrast, transformer-based vision models such as Swin-T rely on self-attention mechanisms to capture broader contextual relationships, but they may require larger and more diverse data sets to achieve stable performance and generalizability.[42] This may partly explain the lower performance of Swin-T with the available pediatric data set.

The studies also demonstrated that AI-based skeletal classification is moving beyond traditional landmark-based workflows. Conventional cephalometric diagnosis usually depends on landmark identification, angular and linear measurements, and clinician interpretation. Image-based AI models may reduce the manual workload by learning diagnostic patterns directly from radiographs, which could support orthodontic workflows by improving efficiency, reducing observer variability, and providing reproducible preliminary classifications. However, these systems should currently be considered decision-support tools rather than replacements for the judgment of a clinician. Skeletal classification remains a clinical diagnosis that should be interpreted alongside facial examination, occlusal findings, growth status, patient age, and treatment objectives.

Despite the promising diagnostic metrics, the current evidence base has important limitations that restrict clinical translation. All of the included studies were retrospective and predominantly single center, thereby limiting generalizability because cephalometric images can vary according to the population characteristics, ethnicity, age distribution, radiographic equipment, acquisition protocol, head positioning, image quality, and preprocessing method. Thus, a model trained and tested within one institution may perform well on internal data but may not maintain the same accuracy when applied to images from another center. External validation was not performed in the included studies, which remains the most important barrier to determining whether these models are sufficiently robust for clinical use in the real world.

Another important issue is the dependence among the model-level results. Several studies evaluated multiple AI architectures based on the same underlying data set, which is useful for comparing model architectures within a study, but it also means that the model-level results do not represent fully independent study populations. Therefore, the pooled estimates may have been affected by within-study dependence and should not be interpreted in the same manner as pooled estimates from multiple fully independent clinical data sets. Thus, the present meta-analysis should be treated as an exploratory quantitative synthesis rather than a definitive diagnostic accuracy meta-analysis.

The substantial heterogeneity observed in the quantitative synthesis probably reflects several methodological and clinical differences among the studies, including the age groups, skeletal class distribution, data set size, imaging inputs, model architecture, validation strategy, and reporting format. Some studies used lateral cephalograms alone, whereas others included additional demographic or imaging information. Some studies used independent test sets, whereas others relied on internal cross-validation. Selective reporting, such as best-fold reporting in one study, may also have contributed to inconsistency across the evidence base. Accordingly, the pooled results should be interpreted as a broad exploratory overview of the available evidence rather than a precise estimate of expected clinical performance.

Explainability is another important consideration. CNN-based models can achieve strong performance but their decision-making process is often difficult to interpret. Visualization methods such as Grad-CAM may help identify whether a model is focusing on anatomically meaningful regions, such as the maxillomandibular complex and cranial base, rather than irrelevant image areas or acquisition-related artifacts.[43] This is particularly important for clinician trust and for detecting shortcut learning, where a model may perform well internally but rely on features that are not clinically meaningful.

The findings also highlight the need for more complete and standardized reporting in future orthodontic AI studies. High accuracy alone is not sufficient to determine whether a model is clinically useful. Future studies should report the class-specific sensitivity, specificity, precision, F1-score, and AUC because errors may have different clinical implications across skeletal Class I, II, and III patterns. For example, misclassifying a borderline Class III patient as Class I may have different treatment consequences than misclassification as Class I as Class II. Therefore, the class-specific performance should be clearly reported to allow clinicians and researchers to judge the practical value of each model. This also highlights the need for standardized methodological reporting and adherence to AI-specific reporting checklists in future studies.[44]

Future studies should also provide detailed information on reference standards, patient selection, data set partitioning, preprocessing, augmentation, model tuning, and handling of borderline skeletal cases. Borderline cases are especially important in orthodontics because skeletal classification is not always a sharply separated categorical diagnosis. Excluding borderline cases may improve apparent model performance but also reduce clinical realism. Therefore, future studies should clearly describe how skeletal classes are assigned to data sets and whether borderline or mixed-pattern cases are included or excluded.

This review has several limitations. First, the meta-analysis included only three eligible image-based AI studies, thereby limiting the statistical power and reducing the reliability of pooled estimates. Second, a formal inter-reviewer agreement statistic was not calculated during study selection and data extraction, which may have limited the ability to quantify the level of agreement between reviewers. Third, several models within individual studies were trained and tested on the same data sets, which may have introduced dependency among effect estimates derived from shared samples and modestly inflated the pooled precision. Fourth, substantial methodological heterogeneity was observed across studies, including differences in populations, imaging inputs, model architectures, validation strategies, and reporting formats. Finally, all of the studies included lacked external validation, which remains the most important barrier to determining clinical robustness in the real world.

Despite these limitations, the present review provides a focused synthesis of the available evidence on image-based AI for sagittal skeletal classification from lateral cephalometric radiographs. The findings suggest that AI, particularly CNN-based models such as DenseNet, can potentially be used as a supportive diagnostic tool in orthodontics. However, before clinical implementation, future research should prioritize multicenter data sets, independent external validation, standardized reference criteria, transparent AI reporting, and clinically meaningful evaluation. Open repositories of anonymized lateral cephalograms with consistently assigned skeletal Class I, II, and III labels would support reproducible benchmarking and fair comparisons between models. Ethical and regulatory considerations, including data privacy, transparency, clinician accountability, and safe integration into orthodontic workflows, should also be addressed before these systems are adopted in routine clinical diagnosis.

Conclusion

Evidence suggests that image-based deep-learning models can achieve promising diagnostic performance in sagittal skeletal classification from lateral cephalograms. DenseNet-based architectures obtained consistently strong performance among the available studies, but the evidence remains limited, methodologically heterogeneous, and based primarily on single-center data sets without external validation. The quantitative synthesis included only a small number of studies and several model-level results from shared data sets, so the pooled estimates should be interpreted as exploratory rather than definitive clinical performance estimates. Larger multicenter studies, standardized reporting, and independent external validation are necessary before these systems can be considered sufficiently robust for routine clinical implementation.

Source of funding

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

Conflict of interest

The authors have no conflicts of interest to declare.

Ethical approval

Ethical approval was not required because this study was a systematic review and meta-analysis based on previously published studies.

Authors contributions

MD: Conception and design of study, Acquisition of data, Drafting of article and/or critical revision, Final approval of manuscript.

FK: Conception and design of study, Final approval of manuscript, Analysis of data.

SE: Acquisition of data, Analysis of data, Drafting of article and/or critical revision, Final approval of manuscript.

NG: Conception and design of study, Analysis of data, Drafting of article and/or critical revision, Final approval of manuscript.

KO: Conception and design of study, Acquisition of data, Drafting of article and/or critical revision, Final approval of manuscript.

AM: Analysis of data, Drafting of article and/or critical revision, Final approval of manuscript.

ZK: Conception and design of study, Drafting of article and/or critical revision, Final approval of manuscript.

FS: Final approval of manuscript, Analysis of data, Drafting of article and/or critical revision.

All authors have critically reviewed and approved the final draft and are responsible for the content and similarity index of the manuscript.

Acknowledgment

None.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used a large language model-based assistant (ChatGPT-5.1, OpenAI) to help improve grammar, wording, and clarity. After using this tool, the authors carefully reviewed and edited all content and take full responsibility for the integrity and accuracy of the manuscript.

References

1. Jordan AR, Kuhr K, Frenzel Baudisch N, Kirschneck C. Prevalence of malocclusions in 8-and 9-year-old children in Germany—results of the sixth German oral health study (DMS 6). **J Orofac Orthop/Fortschritte der Kieferorthopädie 2023**; 84(Suppl 1): 1–9.
2. Bock J, Czarnota J, Hirsch C, Fuhrmann R. *Kieferorthopädischer Behandlungsbedarf einer bevölkerungsrepräsentativen adulten Probandengruppe*. Springer-Verlag; 2011.
3. Tsichlaki A, O'Brien K. Do orthodontic research outcomes reflect patient values? A systematic review of randomized controlled trials involving children. **Am J Orthod Dentofacial Orthop 2014**; 146(3): 279–285.
4. Organization WH. *Oral health surveys: basic methods*. World Health Organization; 2013.
5. Niño-Sandoval TC, Perez SVG, Gonzalez FA, Jaque RA, Infante-Contreras C. An automatic method for skeletal patterns classification using craniomaxillary variables on a Colombian population. **Forensic Sci Int 2016**; 261(159): e1–e6.
6. Lavergne J, Gasson N. Les classes II basales étudiées par les réseaux neuronaux. **L'Orthodontie Française 2008**; 79(2): 91–97.
7. Londono J, Ghasemi S, Hussain Shah A, et al. Evaluation of deep learning and convolutional neural network algorithms accuracy for detecting and predicting anatomical landmarks on 2D lateral cephalometric images: a systematic review and meta-analysis. **Saudi Dent J 2023**; 35(5): 487–497.
8. Hofrath H. Die bedeutung der röntgenfern-und abstandsaufnahme für die diagnostik der kieferanomalien. **Fortschritte der Orthodontik in Theorie und Praxis 1931**; 1(2): 232–258.
9. Broadbent BH. A new x-ray technique and its application to orthodontia. **Angle Orthod 1931**; 1(2): 45–66.
10. Graber T, Vanarsdall RL, Vig K. *Current principles and techniques*. USA: Orthodontic Book; 2005. pp. 607–611.
11. Proffit WR, Fields H, Larson B, Sarver DM. *Contemporary Orthodontics-E-Book: Contemporary Orthodontics-E-Book*. Elsevier Health Sciences; 2018.
12. Stomatologic S. Worldwide prevalence of malocclusion in the different stages of dentition: a systematic review and meta-analysis. **Eur J Paediatr Dent 2020**; 21(2): 115–122.
13. Hardy DK, Cubas YP, Orellana MF. *Prevalence of angle class III malocclusion: a systematic review and meta-analysis*; 2012.
14. Hamet P, Tremblay J. Artificial intelligence in medicine. **Metabolism 2017**; 69: S36–S40.
15. Schwendicke Fa, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. **J Dent Res 2020**; 99(7): 769–774.
16. Schwendicke F, Golla T, Dreher M, Krois J. Convolutional neural networks for dental image diagnostics: a scoping review. **J Dent 2019**; 91:103226.
17. Urschler M, Ebner T, Štern D. Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. **Med Image Anal 2018**; 43: 23–36.
18. Computerized cephalometry by game theory with shape-and appearance-based landmark refinement. In: Ibragimov B, Likar B, Pernus F, Vrtovec T, editors. *Proceedings of international symposium on biomedical imaging (ISBI)*; 2015.
19. Lee J-H, Yu H-J, Kim M-j, Kim J-W, Choi J. Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. **BMC Oral Health 2020**; 20(1): 270.
20. Schwendicke F, Chaurasia A, Arsiwala L, Lee J-H, Elhennawy K, Jost-Brinkmann P-G, et al. Deep learning for cephalometric landmark detection: systematic review and meta-analysis. **Clin Oral Invest 2021**; 25(7): 4299–4309.
21. Dashti M, Khosraviani F, Azimi T, Sehat MS, Alekajbaf E, Fahimipour A, Zare N. Predicting mandibular bone growth using artificial intelligence and machine learning: a systematic review. **Adv Stud Artif Intell Mach Learn 2024**; 4(3): 2731–2745.
22. Yu HJ, Cho SR, Kim MJ, Kim WH, Kim JW, Choi J. Automated skeletal classification with lateral cephalometry based on artificial intelligence. **J Dent Res 2020**; 99(3): 249–256.
23. Jeong SH, Yun JP, Yeom H-G, Kim HK, Kim BC. Deep-learning-based detection of cranio-spinal differences between skeletal classification using cephalometric radiography. **Diagnostics 2021**; 11(4).
24. Li H, Xu Y, Lei Y, Wang Q, Gao X. Automatic classification for sagittal craniofacial patterns based on different convolutional neural networks. **Diagnostics 2022**; 12(6).
25. Nan L, Tang M, Liang B, Mo S, Kang N, Song S, et al. Automated sagittal skeletal classification of children based on deep learning. **Diagnostics 2023**; 13(10).
26. Midlej K, Watted N, Awadi O, Masarwa S, Lone IM, Zohud O, et al. Lateral cephalometric parameters among Arab skeletal classes II and III patients and applying machine learning models. **Clin Oral Invest 2024**; 28(9): 511.
27. Midlej K, Zohud O, Lone IM, Awadi O, Masarwa S, Paddenberg-Schubert E, et al. Clustering and machine learning models of skeletal class I and II parameters of Arab orthodontic patients. **J Clin Med 2025**; 14(3).
28. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. **Ann Intern Med 2009**; 151(4): 264–269. W264.
29. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis campbell. **Syst Rev 2022**; 18:e1230.
30. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM. QUADAS-2 group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. **Ann Intern Med 2011**; 155(8): 529–536.
31. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. **BMJ 2003**; 327(7414): 557–560.
32. Alhamdani NA, Khoshnaw BM, Omer ZQ. Assessing the accuracy and reliability of fully AI-Driven cephalometric analysis in comparison to digital manual methods. **AlRafidain J Med Sci. 2025**; 8(2): 71–75.
33. Ashame LA, Youssef SM, Elagamy MN, El-Sheikh SM. An efficient hybrid 3D computer-aided cephalometric analysis for lateral cephalometric and cone-beam computed tomography (CBCT) systems. **Comput 2025**; 14(6).
34. Mercier JP, Rossi C, Sanchez IN, Renovales ID, Sahagún PMP, Templier L. Reliability and accuracy of artificial intelligence-based software for cephalometric diagnosis. A diagnostic study. **BMC Oral Health 2024**; 24(1).
35. Silva TP, Rodrigues Pinheiro MC, Freitas DQ, Gaeta-Araujo H, Oliveira-Santos C. Assessment of accuracy and reproducibility of cephalometric identification performed by 2 artificial intelligence-driven tracing applications and human examiners. **Oral Surg Oral Med Oral Pathol Oral Radiol 2024**; 137(4): 431–440.
36. Baig N, Gyasudeen KS, Bhattacharjee T, Chaudhry J, Prasad S. Comparative evaluation of commercially available AI-based cephalometric tracing programs. **BMC Oral Health 2024**; 24(1): 1241.
37. Bor S, Ciğerim S, Kotan S. Comparison of AI-assisted cephalometric analysis and orthodontist-performed digital tracing analysis. **Prog Orthod 2024**; 25(1): 41.

38. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition (Version 6). **arXiv 2014**. https://doi.org/10.48550/ARXIV.1409.1556.
39. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2015*; 2015. pp. 1–9.
40. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016*; 2016. pp. 770–778.
41. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2017*; 2017. pp. 4700–4708.
42. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision (ICCV), 2021*; 2021. pp. 10012–10022.
43. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision (ICCV), 2017*; 2017. pp. 618–626.
44. Khurshid Z, Osathanon T, Shire MA, Schwendicke F, Samaranayake L. Artificial intelligence in dentistry: a concise review of reporting checklists and guidelines. **Int Dent J 2026**; 76(1):109322.

How to cite this article: Dashti M, Khosraviani F, Esmaeili S, Ghadimi N, Orhan K, Meyari A, Khurshid Z, Schwendicke F. Artificial intelligence for skeletal classification in orthodontics: A systematic review and meta-analysis of lateral cephalometric studies. J Taibah Univ Med Sc 2026;21(4):683–697.