



Accuracy of automated cephalometric landmark detection with artificial intelligence: a systematic review

Precisión de la detección automatizada de puntos cefalométricos con inteligencia artificial: revisión sistemática

Melissa Castillo Jimenez¹  

Maylon Felix Guerra Reyes¹  

¹ Universidad de Ciencias Médicas de La Habana. La Habana, Cuba.

Submitted: 24-03-2025 | Revised: 29-05-2025 | Accepted: 29-09-2025 | Published: 14-10-2025

Corresponding Author: Maylon Felix Guerra Reyes 

ABSTRACT

Introduction: Cephalometric analysis depends on manual landmark identification, which is time-consuming and observer-dependent. Artificial intelligence has been proposed to automate this step, but reported accuracy varies. **Methods:** Following PRISMA 2020, PubMed, OpenAlex, and Crossref were searched for studies published from 2018 to 2025 that evaluated artificial intelligence models for cephalometric landmark detection against manual annotation. Risk of bias was assessed with QUADAS-2. **Results:** Forty-three studies were included, most using convolutional neural networks on lateral cephalograms. Mean radial error ranged from 0.62 mm to 2.67 mm across studies, and success detection rate at a 2 mm threshold ranged from 26.8% to 97.3%. **Conclusions:** Artificial intelligence detects cephalometric landmarks with clinically acceptable mean radial error in most studies, but success detection rate at the 2 mm threshold remains heterogeneous and varies with landmark and dataset. External validation on clinical cohorts is still scarce.

Keywords: artificial intelligence; cephalometry; anatomic landmarks; deep learning; diagnostic imaging; systematic review.

RESUMEN

Introducción: El análisis cefalométrico depende de la identificación manual de puntos, que es laboriosa y dependiente del observador. Se ha propuesto la inteligencia artificial para automatizarla, pero la precisión reportada es variable. **Métodos:** Siguiendo PRISMA 2020, se buscó en PubMed, OpenAlex y Crossref estudios publicados entre 2018 y 2025 que evaluaran modelos de inteligencia artificial para la detección de puntos cefalométricos frente a anotación manual. El riesgo de sesgo se evaluó con QUADAS-2. **Resultados:** Se incluyeron cuarenta y tres estudios, la mayoría con redes neuronales convolucionales en cefalogramas laterales. El error radial medio osciló entre 0,62 mm y 2,67 mm, y la tasa de detección exitosa a un umbral de 2 mm entre 26,8% y 97,3%. **Conclusiones:** La inteligencia artificial detecta puntos cefalométricos con un error radial medio clínicamente aceptable en la mayoría de los estudios, pero la tasa de detección exitosa a 2 mm sigue siendo heterogénea y varía según el punto y el conjunto de datos. La validación externa en cohortes clínicas sigue siendo escasa.

Palabras clave: inteligencia artificial; cefalometría; puntos anatómicos; aprendizaje profundo; diagnóstico por imagen; revisión sistemática.

INTRODUCTION

Cephalometric analysis is a cornerstone of orthodontic and orthognathic diagnosis, and its first step, the identification of anatomical landmarks on a lateral or posteroanterior cephalogram, is the most tedious and error-prone part of the workflow. Manual landmarking requires a trained operator to locate each point, and the time and inter-examiner variability it introduces are well documented. The measurements that follow, the angles and distances of the skeletal classification, are only as reliable as the points on which they are anchored, so an error in a single landmark propagates into the diagnosis. Artificial intelligence, and deep learning in particular, has been proposed to automate this step, and the number of primary studies has grown rapidly since 2018.^(1,2)

The clinical stakes of this step are higher than they might appear. Cephalometric measurements are used to decide whether a patient needs orthodontic treatment, whether surgery is indicated, and how the treatment should be planned, and a misclassified skeletal pattern can change all three. The manual workflow is therefore not only slow but also a source of diagnostic risk, and the prospect of replacing it with a reproducible automatic procedure is the main motivation behind the large body of work reviewed here.

The appeal of automation is concrete. Landmark identification is repeated on every patient, the task is well defined, and large annotated datasets now exist, which makes it an ideal target for supervised learning. If a model can locate landmarks with an error no larger than the variability between two human examiners, it can remove a bottleneck and standardize the analysis across clinicians.⁽³⁾

The task has a further advantage as a benchmark. A public dataset released for a landmark detection challenge in 2015, containing several hundred lateral cephalograms with 19 annotated landmarks, has become the de facto standard on which most models are compared, and it provides a common yardstick that is unusual in dental artificial intelligence.⁽²⁾ The landmark set is also standardized by clinical convention, with well-established definitions for the sella, nasion, and the other points that anchor the common cephalometric measurements. These properties make cephalometric landmark detection one of the most measurable and comparable tasks in dental imaging, which is why it has attracted so many primary studies and why a focused synthesis is feasible.

At the same time, the comparability is only partial. The public benchmark uses a fixed set of 19 landmarks, whereas clinical tracing uses a larger and variable set, and many studies evaluated on private datasets with landmarks of their own choosing. The number of landmarks, and specifically the inclusion of the difficult mandibular points, therefore differs across studies, and this difference, more than any algorithmic detail, drives the spread in the reported summary statistics. This distinction between a common benchmark and a heterogeneous clinical task is central to interpreting the results that follow.

Three questions summarize what a clinician needs to know about this technology. First, what mean radial error do the models achieve, that is, how far from the manual reference does the automatic point fall on average. Second, what success detection rate do they achieve at the clinically meaningful thresholds of 2, 3, and 4 mm. Third, which architectures, datasets, and reference standards explain the differences among studies. These three questions organize the present review, and each is answered in the corresponding subsection of the results and the discussion.

The aim of this systematic review was to synthesize the accuracy, expressed as mean radial error and success detection rate, of artificial intelligence models for automated cephalometric landmark identification. The review deliberately separates the distance error from the threshold-based detection rate, because these two measures convey different information about whether a model is fit for clinical use.

METHODS

Protocol and Eligibility criteria

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines.⁽⁴⁾

The review considered original research articles published between 2018 and 2025 that investigated the development, application, or performance of artificial intelligence (AI), machine learning (ML), or deep learning (DL) algorithms for the automated identification or localization of anatomical landmarks in two-dimensional cephalometric radiographs. To qualify for inclusion,

studies were required to provide quantitative evidence of model performance using a defined reference standard. Eligible performance indicators included mean radial error (MRE), expressed in millimeters, successful detection rate (SDR) at specified distance thresholds, point-to-point localization error, or other quantitative measures of detection accuracy.

Investigations based primarily on cone-beam computed tomography, volumetric imaging, or three-dimensional anatomical reconstructions were excluded. Review articles, editorials, and publications lacking numerical assessments of landmark detection performance were also considered ineligible.

Information sources and search strategy

A structured literature search using three bibliographic sources: PubMed, OpenAlex, and Crossref. The search was restricted to publications appearing from January 2018 through March 2025. Search expressions incorporated combinations of controlled concepts and free-text terminology associated with artificial intelligence, automated cephalometric analysis, and anatomical landmark recognition.

For PubMed, three complementary search strategies were implemented. The first combined artificial intelligence-related terms (“artificial intelligence”, “deep learning”, and “machine learning”) with cephalometric terminology (“cephalometric landmark” and “cephalometric analysis”) and localization-related expressions (detection, identification, and localization). The second focused on deep learning applications in cephalometric landmark identification and incorporated performance-related terms such as “mean radial error”, “detection rate”, and accuracy. The third targeted convolutional neural network approaches applied to cephalograms and landmark detection.

Corresponding search expressions were adapted to the retrieval characteristics and syntax of OpenAlex and Crossref. Each query was configured to retrieve a maximum of 50 records per database search. Retrieved references were subsequently consolidated, and duplicate entries were identified and removed through comparisons of digital object identifiers and publication titles.

Study selection process

Study selection followed a sequential screening procedure comprising an initial examination of titles and abstracts and a subsequent assessment of potentially eligible publications. Records that clearly failed to satisfy the predefined eligibility criteria were excluded during preliminary screening, whereas publications with insufficient information to determine eligibility were retained for additional examination.

Full-text documents were consulted whenever accessible. For publications without available full text, eligibility decisions and relevant methodological information were established using the information reported in their abstracts. The screening process was undertaken by the author with support from an AI-assisted tool, followed by verification of the resulting decisions. Independent duplicate screening by two reviewers was not performed.

Data collection and extraction

Relevant methodological and performance characteristics were systematically collected from each eligible publication. Extracted information included the first author's name, publication year, country of origin, computational model or network architecture, dataset size, reference standard, and quantitative indicators of landmark localization accuracy.

Particular attention was given to mean radial error, successful detection rates, and the spatial thresholds used to define accurate landmark identification. Numerical results were transcribed directly from the original publications without mathematical transformation or recalculation. Where performance estimates were presented for individual anatomical landmarks or multiple detection thresholds, the corresponding values were preserved separately to maintain the level of detail provided by the original investigators.

Localization errors expressed in pixels were retained in their original measurement units rather than converted into millimeters. Similarly, studies reporting alternative outcomes, including angular deviations, were documented according to their original metrics without attempting to derive unreported spatial errors. Whenever available, the distance threshold associated with each successful detection rate was explicitly recorded.

Risk of bias and applicability assessment

Methodological quality was examined using the Quality Assessment of Diagnostic Accuracy

Studies 2 (QUADAS-2) framework, which addresses four domains: patient selection, index test, reference standard, and flow and timing.⁽⁵⁾ Potential sources of bias and concerns regarding applicability were evaluated qualitatively, taking into consideration the characteristics of the datasets, model development procedures, annotation methods, and validation strategies.

Regarding patient selection, particular attention was directed toward the frequent use of publicly available benchmark datasets or images obtained from individual institutions, as these sampling strategies may restrict the generalizability of model performance across different clinical populations. Within the index test domain, the reproducible nature of computational predictions was considered alongside inconsistencies in the reporting of image preprocessing, data augmentation, and model implementation procedures.

For the reference standard, potential limitations were associated with reliance on manually identified cephalometric landmarks, particularly when interobserver or intraobserver variability was insufficiently documented. Assessment of flow and timing focused on the independence of training and evaluation datasets, including possible overlap in data sources across investigations. The limited availability of independent external validation was also considered when interpreting the applicability of reported findings.

Data synthesis and analysis

The evidence was integrated through a structured narrative synthesis rather than a quantitative meta-analysis. This approach was selected because substantial methodological and clinical heterogeneity was anticipated across the included investigations, particularly in relation to computational architectures, imaging datasets, anatomical landmark definitions, and performance evaluation procedures.

Direct statistical aggregation was considered inappropriate because the number and anatomical distribution of evaluated landmarks differed considerably among studies. For example, combining results from models assessing 13 landmarks with those evaluating 61 landmarks could produce summary estimates that fail to reflect differences in task complexity. Furthermore, certain anatomical structures, including gonion and condylion, present greater localization challenges and were not consistently represented across datasets.

Differences in the reporting and interpretation of mean radial error and successful detection rate further limited the comparability of study outcomes, particularly when detection thresholds and landmark configurations varied. Consequently, results were organized according to the principal research questions and summarized descriptively, emphasizing variations in model performance, landmark-specific accuracy, and methodological characteristics rather than generating pooled effect estimates.

RESULTS

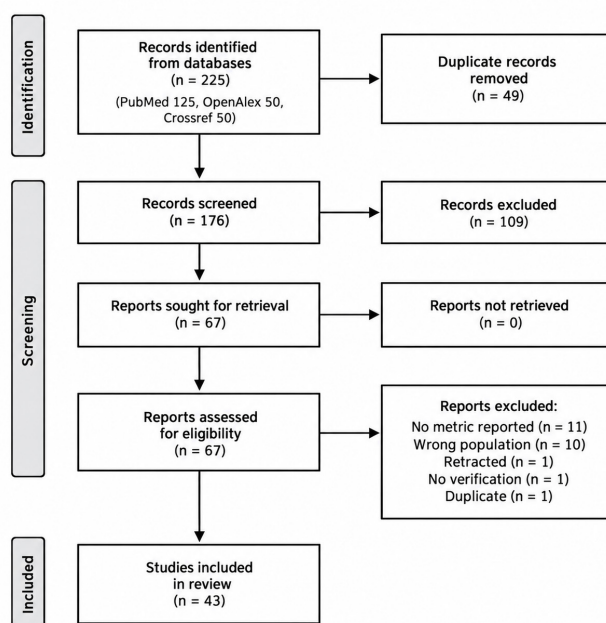
Study selection

The searches retrieved 225 records, of which 49 duplicates were removed and 176 unique records were screened (Figure 1). Title and abstract screening excluded 109 records, leaving 67 reports assessed for eligibility. Of these, 24 were excluded: 11 because no quantitative accuracy metric was reported, 10 because the outcome was not landmark detection, and the remainder for a retracted report, a duplicate, a non-cephalogram modality, or an unverifiable reference. Forty-three studies were included⁽⁴³⁾.

Characteristics of included studies

The included studies spanned 2019 to 2025, with the largest number published in 2023 (12 studies) and 2025 (10 studies) (Table 1). Lateral cephalograms were the most common input, with a smaller number of studies using posteroanterior cephalograms⁽⁵⁾. Convolutional neural networks dominated the architectures, ranging from U-Net and its variants^(6,7) to cascaded and two-stage networks,⁽⁸⁻¹⁰⁾ object detectors,^(11,12) and coarse-to-fine regression networks.^(13,14) Within the convolutional family, several design choices recurred. Cascaded networks first localize a region of interest and then refine the landmark within it, which improves the difficult gonion and condylion,^(8,15,16) whereas stacked regression networks predict all landmarks jointly and then refine them iteratively.⁽¹⁷⁾ A minority used transformer-based models,^(18,19) machine learning on hand-crafted features such as histograms of oriented gradients,^(20,21) or reinforcement learning.⁽²²⁾

Figure 1. PRISMA 2020 flow diagram of study selection.



Several studies evaluated commercially available software, including WebCeph, CefBot, Planmeca Romexis, and CranioCatch.(23-26)

The reference standard was manual annotation in nearly all studies, whether by a single expert, a calibrated pair, or a consensus. The public ISBI 2015 benchmark, which contains 400 cephalograms with 19 landmarks, was the most common evaluation dataset,(6,27-30)and many studies supplemented it with private or institutional datasets. The reliance on a single benchmark has a subtle consequence: the benchmark is reused across so many studies that it becomes progressively less independent, because each new publication is informed by the performance of its predecessors on the same images. The private datasets differed in size and in the landmarks they annotated, from a few dozen skulls with lead-sphere references(31)to several thousand clinical cephalograms.(10,14) Sample sizes ranged from fewer than 100 radiographs(31,32) to more than 9,000.(10)Only a minority reported validation on a cohort independent of the training distribution, and only one study reported a physical phantom as the reference.(31)

The studies also differed in whether they validated against human variability. Some compared the model with the reference annotators and reported the model error relative to the inter-examiner difference,(11,32) which is the comparison that matters clinically, whereas others reported only the model error against a single annotation, which leaves the clinical significance of the error unresolved. The studies that made the comparison generally found that the model error was within or near the range of human variability, which is the standard against which any replacement technology should be judged.

Risk of bias

Across QUADAS-2 domains, the reference standard was the principal source of concern because it was manual annotation in nearly all studies, and the variability among the annotators who defined the reference is itself on the order of the model error being measured. Patient and image selection raised further concerns because many studies evaluated on the same public benchmark, which raises the risk of overlap between training and evaluation across studies, and few reported external validation. Flow and timing were generally acceptable, but blinding and the handling of unidentifiable landmarks were reported inconsistently. The overall risk of bias was therefore moderate to high, and the applicability of the benchmark-based estimates to routine clinical images is uncertain.

Table 1. Characteristics of the 43 included studies evaluating artificial intelligence for cephalometric landmark detection.

First author, year	Country	Architecture	Sample	Reference standard	Notes
Bagdy-Balint 2025	NR	cascaded CNN	1600 train / 78 test	manual examiners	angular errors 0.05-1.86 deg; proportional error 3.14%
Bu 2025	China	CNN	NR	NR	accuracy 90.8% (ISBI 2015)
Guo 2025	Macau	diffusion data generation + detection	NR	NR	SDR improved 6.5% with generated data
Han 2025	NR	CNN	1019 cephalograms	NR	22 landmarks
Kim 2025	South Korea	CNN	3004 train/val; 224 test	human standard	13 soft tissue landmarks
Laitenberger 2025	NR	Vision Transformer	NR	NR	improvement >2 mm MRE vs state of the art
Neeraja 2025	NR	CephTransXnet (CNN + Swin Transformer)	ISBI 2015 + private	NR	SDR at 2 mm
Silva 2025	Brazil	CefBot / WebCeph	60 radiographs	two examiners	accuracy 90% (CefBot) / 95% (WebCeph); population-dependent
Sumon 2025	NR	Self-CephaloNet (self-ONN + HRNetV2)	ISBI 2015 + PKU	NR	
Tageldin 2025	Egypt	WebCeph	112	expert orthodontist (OnyxCeph)	coordinate differences -0.86 to 3.15 mm; less precise in cleft patients
Gonca 2024	Turkey	FARNet	1431 PA (1177/117/137)	manual annotation	PA cephalograms
Han 2024	South Korea	cascaded CNN	2903 PA (2150/376/377)	two orthodontists	midline landmarks, PA
Hori 2024	Japan	CNN (6 conv + 2 affine)	833 train; 179 test	NR	SNA/SNB difference <1 deg
Jiao 2024	NR	Mask R-CNN	400 (320/80)	manual (19 landmarks)	detection rate 98.3%; 43.4% <1 mm; 85.7% <4 mm
Mehra 2024	India	VGG-16	ISBI 2015	manual	
Rashmi 2024	NR	Detectron2 / YOLOv8-pose	1692 + 400	NR	accuracy 85.9% (Detectron2) / 72.9% (YOLOv8)
Rashmi 2024	India	UNet/UNetPlusPlus + ensemble	400 public + 700 private	NR	accuracy 79.9% to 95.4%
Silva 2024	NR	CefBot / WebCeph	10 skulls	lead-sphere skulls	ICC CefBot >=0.994; WebCeph >=0.787
Bao 2023	China	Planmeca Romexis 6.2	85 patients; 19 landmarks	Dolphin Imaging	commercial software
Chen 2023	NR	U-Net + Monte Carlo dropout	ISBI 2015	NR	
Hong 2023	NR	DQN / DDQN (reinforcement learning)	ISBI 2015 + 500	NR	
Jiang 2023	NR	two-stage CNN	9870 cephalograms	manual annotation	measurement classification accuracy 89.3%
Lee 2023	South Korea	CNN	1032 PA / 82 test	two experts	PA cephalograms
Neeraja 2023	India	CephXNet (CNN + SE)	ISBI 2015 + 100 private	NR	classification accuracy 97.7%
Panesar 2023	NR	deep learning software	30 cephalograms; 31 landmarks	human examiners	MAE 1.57 mm/deg; ICC 0.97
Popova 2023	Germany	CNN	430 train / 460 test	NR	MRE by landmark, values not in abstract
Rashmi 2023	NR	HOG + LightGBM	ISBI + Dental Ceph	NR	mean accuracy 78.2%
Rashmi 2023	NR	LGBM ensemble (HOG)	300 lateral	NR	
Yang 2023	South Korea	Ceph-Net (stacked regression + attention)	NR	NR	superior MRE/SDR, values not stated

First author, year	Country	Architecture	Sample	Reference standard	Notes
Zhao 2023	NR	MS-YOLOv3	NR	NR	
Hong 2022	South Korea	CNN	3188 (3004/184)	human gold standard	20 landmarks, serial lateral
Le 2022	NR	DACFL	1193 train; 100 test	NR	
Ugurlu 2022	Turkey	CNN (CranioCatch)	1620 (1360/140/180)	labeled dataset	
Yao 2022	NR	CNN (global + local)	512 (312/100/100)	labeled dataset	37 landmarks
Yoon 2022	NR	UNet + EfficientNetB0	600 (500 train)	ground truth vs examiners	16 landmarks
Bulatova 2021	NR	YOLOv3	110 cephalograms	Dolphin Imaging	no statistical difference in 12/16 points
Kim 2021	South Korea	multi-stage CNN	430 (345/85)	one examiner	PA from CBCT
Kim 2021	South Korea	2-stage CNN (ROI + detection)	950 lateral	two examiners	
Song 2021	Japan	coarse-to-fine CNN	ISBI 2015	NR	
Takeda 2021	NR	deep CNN vs random forest	400 PA (320 train)	two orthodontists	R2 = 0.715 (CNN)
Zeng 2021	NR	3-stage cascaded CNN	NR	NR	benchmark, values not in abstract
Song 2020	China	ResNet50 + ROI	ISBI 2015 + Shandong	NR	SDR/SCR reported, values not in abstract
Nishimoto 2019	NR	CNN (coordinate regression)	219 (153/66)	manual	mean prediction error 17.02 px

Note: NR = not reported; CNN = convolutional neural network; PA = posteroanterior.

Mean radial error

Twenty studies reported a mean radial error (Table 2). The reported values ranged from 0.62 mm for the sella in one study, although this landmark was the easiest and the gonion reached 8.30 mm,(26)to an overall study mean of 2.67 mm in a transfer learning model.(27)Most studies reported an overall mean radial error between 0.9 and 2.3 mm.(8,10,14,25,33-37)The error was consistently larger for the gonion and condylion, which lie on curved and low-contrast structures, than for landmarks on the incisal edge or the sella, which are sharply defined.(26,38) Two studies reported error in pixels rather than millimeters, with a mean prediction error of 17 pixels(39)and 3.1 pixels,(40)which limits direct comparison.

The reference standard conditioned the error. Studies that validated against a physical phantom with lead-sphere landmarks reported agreement that was high for one commercial tool and lower for another,(31)and studies that compared the model against the variability of human examiners found that the model error was comparable to or smaller than inter-examiner variability in several cases.(11,32) Studies that reported the error per landmark, rather than as a single summary, showed that the distribution is skewed by the difficult points, and that a study-level mean conceals landmarks that fall far outside the acceptable range.(26,41)

The mean radial error also depended on the projection. On lateral cephalograms, the values clustered between 0.9 and 2.3 mm,(10,14,33,34)whereas on posteroanterior cephalograms the error was larger and more variable, from 1.26 mm for midline landmarks(9)to 2.23 mm in a study that reconstructed posteroanterior images from cone-beam data.(8) The posteroanterior projection presents additional landmarks, such as the condylion, that are harder to localize, and fewer studies have used it, so the evidence is thinner. The number of landmarks also mattered: studies that detected 37 or more landmarks reported summary errors that were pulled upward by the extra difficult points,(34)whereas studies restricted to a smaller set of easy landmarks reported lower means.(7,33) Two studies illustrate the importance of the landmark set directly. One reported a mean radial error of 0.62 mm for the sella, the easiest landmark, and 8.30 mm for the gonion, the hardest, within the same model and dataset.(26)Another reported a mean radial error of 0.94 mm for the best landmark and 3.31 mm for the worst, again within the same model.(38)This within-model range, larger than the between-model differences, shows that the choice of which landmarks to include in the summary is a first-order determinant of the reported mean, and it explains why studies of different landmark sets are not directly comparable.

Table 2. Reported accuracy of artificial intelligence models for cephalometric landmark detection.

First author, year	Mean radial error (mm)	Success detection rate
Bagdy-Balint 2025	NR	NR
Bu 2025	NR	NR
Guo 2025	NR	82.2%
Han 2025	0.97 +/- 0.52 mm	95.4% (<2 mm)
Kim 2025	1.50 +/- 0.48 mm	64.9% (<1.5 mm)
Laitenberger 2025	NR	NR
Neeraja 2025	NR	88.7% / 79.1% (2 mm)
Silva 2025	NR	NR
Sumon 2025	NR	82.3% (2 mm)
Tageldin 2025	NR	NR
Gonca 2024	0.94-3.31 mm	32.8-100%
Han 2024	1.26 mm	83.2% (<2 mm)
Hori 2024	3.1 px	NR
Jiao 2024	2.19 mm	68.3% (<2 mm)
Mehra 2024	2.67 mm	26.8% (2 mm); 59.9% (3 mm); 94.4% (4 mm)
Rashmi 2024	NR	NR
Rashmi 2024	NR	NR
Silva 2024	NR	NR
Bao 2023	2.07 +/- 1.35 mm	58.6% (2 mm); 82.0% (3 mm); 91.4% (4 mm)
Chen 2023	1.39 / 1.33 mm	79.7% (<2 mm); 97.2% (<4 mm)
Hong 2023	<2 mm	67.3% (2 mm)
Jiang 2023	0.94 +/- 0.74 mm	NR
Lee 2023	1.87 +/- 1.53 mm	67.5% (<2 mm); 91.5% (<4 mm)
Neeraja 2023	NR	88.1% (2 mm)
Panesar 2023	NR	NR
Popova 2023	NR	NR
Rashmi 2023	NR	77.1% (<2 mm)
Rashmi 2023	NR	75.8% (2 mm)
Yang 2023	NR	NR
Zhao 2023	NR	80.8% (<2 mm); 93.8% (<3 mm); 98.1% (<4 mm)
Hong 2022	1.17 mm	NR
Le 2022	1.87 +/- 2.04 mm	73.2% (2 mm)
Ugurlu 2022	0.62-8.30 mm	48.3-99.4%
Yao 2022	1.038 +/- 0.893 mm	91.9% (<1.5 mm); 97.3% (<2 mm); 100% (<4 mm)
Yoon 2022	0.80 / 1.78 mm	93.4% (<2 mm); 98.7% (<4 mm)
Bulatova 2021	NR	NR
Kim 2021	2.23 +/- 2.02 mm	60.9% (<2 mm)
Kim 2021	1.84 mm	36.1%
Song 2021	1.0-1.6 mm	NR
Takeda 2021	NR	NR
Zeng 2021	NR	NR
Song 2020	NR	NR
Nishimoto 2019	NR	NR

Note: MRE = mean radial error; SDR = success detection rate; px = pixels.

Success detection rate

Twenty-three studies reported a success detection rate at one or more distance thresholds (Table 2). At the 2 mm threshold, which is the most clinically cited, the reported values ranged from 26.8% in a transfer learning model(27) to 97.3% in a global-to-local regression network.(34) The two ends of this range correspond to very different experimental settings, and the spread should not be read as a pure ranking of model quality, because the transfer learning model evaluated on a public benchmark with a fixed landmark set, whereas the global-to-local network evaluated on a smaller set of landmarks with a refined local adjustment stage. The intermediate values are therefore the more representative guide to what a well-configured model achieves in practice. Intermediate values clustered between 60% and 90%.(6,8,9,12,18,28,29,33,35,37,42) A two-stage region-of-interest network reported a success detection rate of only 36.1% at an unspecified threshold,(16) which is among the lowest reported and illustrates how much the choice of landmark set and threshold moves the result. At the more permissive 4 mm threshold, most studies exceeded 90%.(6,7,12,25,27,34) Success detection rate was lower for posteroanterior cephalograms than for lateral ones,(38,42,43) and lower for the gonion and condylion than for midline landmarks.(26,38) The success detection rate also depended on the threshold, and this dependency is informative: a model that detects 60% of landmarks within 2 mm may detect more than 90% within 4 mm, and the choice of threshold therefore changes whether the model is judged acceptable.(6,25,27) Studies that reported only a single threshold or that omitted the threshold altogether made comparison difficult.(16,22,44)

The success detection rate and the mean radial error are complementary rather than redundant. A model can have a low mean radial error because most landmarks are located very precisely, yet a low success detection rate because a few landmarks fall just outside the threshold, and the reverse is also possible. Reporting both therefore gives a fuller picture than either alone, but only twenty studies reported both, and the remainder reported one or the other.(33,37) The inconsistency in reporting is itself a barrier to synthesis, and it means that the ranges reported here for each measure draw on partly different sets of studies.

A further reporting inconsistency concerns the threshold itself. The 2 mm threshold is the one cited most in the literature and is the one used by the original benchmark, but some studies reported only 1 mm, 1.5 mm, 3 mm, or 4 mm, and a few omitted the threshold entirely.(36,44) This matters because the success detection rate rises steeply as the threshold loosens, and a value quoted without its threshold is uninterpretable. The studies that reported the full set of thresholds showed this steepness clearly: one model moved from 26.8% at 2 mm to 94.4% at 4 mm,(27) and another from 79.7% at 2 mm to 97.2% at 4 mm.(6) A single threshold therefore tells only part of the story, and the full threshold curve is what a clinician actually needs to set expectations.

Beyond mean radial error and success detection rate, a subset of studies reported different measures. Some reported a detection rate, defined as the proportion of landmarks that the model found at all regardless of distance,(37) which is a weaker criterion than the success detection rate. Others reported classification accuracy for the presence of landmarks or for derived measurement categories,(10,29,45-47) and still others reported angular or proportional error in the derived measurements.(40,48) These measures answer different questions and are not comparable to the distance-based metrics, but they are informative in their own right, because they show that the downstream measurements inherit the landmark error and remain clinically usable when the anchor landmarks are reliable.

A small number of studies approached the task through alternative paradigms that are worth recording for completeness. One used reinforcement learning, in which an agent learns to move toward each landmark by maximizing a reward, and reported a success detection rate of 67.3% at 2 mm.(22) Another augmented the training data with a generative diffusion model and reported a measurable improvement in the success detection rate.(44) A third framed the problem as a segmentation task over heatmaps and reported a mean radial error around 1.3 to 1.4 mm.(6) These alternative formulations converge on similar accuracy, which again suggests that the formulation matters less than the data and the landmarks, and that the field has largely reached the performance available from the current benchmark rather than from any particular architectural choice.

Sources of variability

The variability among studies has identifiable sources. The landmark set differed, and studies that included the difficult gonion and condylion reported worse summary statistics than those restricted to easier landmarks. The dataset differed, and models evaluated on the public ISBI 2015 benchmark(6,13,34) reported systematically different values from those evaluated on private clinical datasets.(10,14,33) The metric differed, because some studies reported mean radial error, others

success detection rate, and still others classification accuracy(45-47) or angular and proportional error,(48) which are not interchangeable. Data augmentation, including generative diffusion models,(44) and self-supervised or reinforcement learning approaches,(22) further modulated the reported accuracy.

The commercial tools merit separate note, because they are the ones most likely to reach clinical use and they showed distinct behavior. WebCeph and CefBot performed well in the populations in which they were developed but degraded when applied to a different ethnic population,(23) and WebCeph was less precise than manual tracing in patients with cleft lip and palate.(24) A commercial software evaluated against a computer-assisted manual reference achieved a mean radial error of 2.07 mm and a success detection rate of 58.6% at 2 mm,(25) which is within the range of the research models but not at the top of it. These observations suggest that commercial performance should not be assumed from the research literature, and that each product requires evaluation in the intended population.

The field also evolved over the review period, and the direction of travel is informative. The earliest studies used small datasets and reported error in pixels rather than millimeters,(39) which made clinical interpretation difficult. The more recent studies used larger datasets, reported the success detection rate at standardized thresholds, and increasingly evaluated commercial software.(23-25) This maturation of reporting, more than any specific architectural advance, is what has made the literature increasingly interpretable, and it suggests that the field is converging on the conventions that would permit a future meta-analysis at the landmark level.

DISCUSSION

The answer to the first question, the mean radial error, is that artificial intelligence models locate cephalometric landmarks within roughly 1 to 2 mm of a manual reference in most studies. This is within the range of inter-examiner variability reported for manual landmarking and is therefore clinically encouraging.(3) The answer to the second question, the success detection rate at 2 mm, is that performance is heterogeneous, from below 30% to above 95%, and depends strongly on which landmarks are included and which dataset is used. The answer to the third question is that the architecture matters less than the dataset and the landmark set, because convolutional networks, transformers, and even hand-crafted-feature models converge to similar accuracy when evaluated on the same benchmark.(6,18,20)

The first answer deserves a qualification. A mean radial error near 1 mm is encouraging only if the error is uniform, and the per-landmark reports show that it is not. The gonion and condylion, which anchor the mandibular plane and the ramus measurements, are the points that matter most for skeletal classification and are also the points that the models locate least reliably.(26,38) A clinician who relies on an automatic gonion that is occasionally off by several millimeters will propagate that error into the mandibular plane angle and, from there, into the skeletal diagnosis. The mean radial error therefore understates the clinical risk for the very measurements that orthodontists weigh most heavily.

These findings align with and refine the earlier syntheses. A review of artificial intelligence in orthodontic diagnosis reported that landmark detection was among the most mature applications, with high accuracy but limited clinical translation.(1) A comprehensive review of cephalometric analysis with artificial intelligence reached the same conclusion.(2) A narrative review of artificial intelligence in cephalometric analysis emphasized the promise while noting the absence of external validation.(3) A recent systematic review with meta-analysis of the SNA, SNB, ANB, and WITS measures concluded that automatic detection is reliable for these derived measurements.(49) The present review adds the observation that the summary statistics conceal a wide spread at the 2 mm threshold, which is the threshold that determines whether a landmark is usable for a measurement.

The apparent contradiction between the favorable meta-analytic conclusion for the derived measurements(49) and the wide spread observed here is resolved by noting what each measures. The SNA, SNB, ANB, and WITS measures are anchored on landmarks that the models locate reliably, whereas the difficult gonion and condylion do not enter those measures and therefore do not drag down their pooled accuracy. This means that a model can be entirely adequate for the specific measurements that a meta-analysis pools while still failing on the landmarks needed for a full skeletal analysis. The two conclusions are both correct, and they are reconciled by the same per-landmark perspective that organizes this review.

For the reader who must decide whether to adopt a tool, the practical summary is as follows.

The mean radial error is acceptable in most studies, the success detection rate at 2 mm is good but not uniformly high, and the weakest points are the gonion and condylion, which are also the points that anchor the skeletal classification. A tool can therefore be adopted with confidence for a reviewed first pass, but its output should be checked point by point, with particular attention to the mandibular landmarks, and it should be validated in the local patient population before it is trusted for unattended measurement.

The clinical consequence of this spread is that a mean radial error below 2 mm does not guarantee that every landmark is usable. A model with a mean radial error of 1.5 mm may still place the gonion 8 mm away, and the angular and linear measurements that depend on that point will then be wrong.(26,38) Clinicians adopting these tools should therefore verify the landmarks that anchor the measurements they actually use, rather than relying on a summary error. The commercial tools showed the same behavior, with population-dependent accuracy and weaker performance in patients with craniofacial anomalies.(23,24)

The workflow that follows from these findings is a reviewed first pass rather than an unattended replacement. The model places the landmarks, and the clinician then inspects and corrects the few that are misplaced, which converts a fully manual task into a mostly automatic one. This use preserves the time savings while keeping the clinician accountable for the measurements that drive the treatment plan. Several of the included studies explicitly framed their models in this way, comparing the time required for manual tracing with the time required to review and correct automatic output, and reported meaningful reductions.(25,32) The question for adoption is therefore not whether the model is perfect, but whether the combined error of the model and the reviewing clinician is acceptable relative to manual tracing alone.

The economics of this workflow are favorable even when the model error is not zero. If the model places most landmarks correctly, the clinician corrects only a minority, and the time saved scales with the proportion correctly placed. The success detection rate at 2 mm is therefore a direct proxy for the time saving, because a landmark within 2 mm of the reference is typically accepted without correction, whereas one outside it must be moved. This interpretation reframes the heterogeneous success detection rates: a model at 90% saves more time than one at 60%, but both save time relative to fully manual tracing, so even the weaker models may be worth deploying as a first pass provided the clinician reviews the output.

The generalizability question remains the main unresolved issue. Most models were trained and evaluated on images from a limited set of institutions and populations, and the one study that tested a commercial tool across two populations of different ethnicity reported a measurable drop in accuracy.(23) Cephalometric morphology differs across ancestry and across patients with craniofacial anomalies, and a model trained predominantly on one population cannot be assumed to transfer to another. Until prospective, multi-population validation becomes the norm, the safe assumption is that a tool performs best on the population that resembles its training data, and that local validation is a prerequisite for clinical use.

The same caution applies to image acquisition. Cephalograms are acquired on equipment from different manufacturers and with different protocols, and the contrast and geometry of the image affect the visibility of the landmarks. A model trained on one device may transfer imperfectly to another, and the included studies rarely addressed this, because most used a single acquisition source. This is a further reason to interpret the benchmark-based estimates as an upper bound, and to expect that performance in a new clinic, with a new device and a new patient population, will be somewhat lower than the published figures until a local evaluation is performed.

A further concern specific to this literature is the heavy use of the public ISBI 2015 benchmark. When many studies evaluate on the same dataset, the risk of indirect test-set leakage increases, because researchers may tune architectures and hyperparameters against the benchmark over successive publications. The result is that benchmark-based estimates may overstate performance on unseen clinical images, and the private-dataset studies, which reported a wider range of success detection rates, are likely closer to real-world performance.(10,33) Future studies should report performance on both the benchmark and an independent clinical cohort so that the gap between the two can be quantified.

The same argument extends to the private datasets, which are also reused across publications from the same groups. When a research group publishes successive models on the same institutional dataset, the later models benefit from the group's accumulated knowledge of that dataset, and the reported improvement over the earlier models may not generalize. This is a normal limitation of a young field, but it means that the absolute performance figures, more than the relative improvements, are what a clinician should weigh, and that independent reproduction remains the strongest form of evidence available.

The strengths of this review are the systematic search across three bibliographic databases, the explicit 2018 to 2025 window, the separation of mean radial error from success detection rate, and the recording of values as reported without recalculation. The limitations mirror those of the primary studies. Full text was available for only a subset of reports, so some data were extracted from abstracts. Screening and extraction were performed by a single author with artificial intelligence assistance rather than two independent reviewers, which raises the risk of selection error and warrants a dual-reviewer design in future updates. A meta-analysis was not possible given the heterogeneous landmarks, datasets, and metrics, and a pooled estimate would be misleading because the underlying landmark sets differ. Publication bias may also inflate the observed accuracy, because studies with unfavorable results are less likely to be published, and the heavy use of the same public benchmark raises the risk of test-set leakage across studies. Finally, the review could not assess the time efficiency of the tools, which is the other half of their clinical value, because the studies reported it inconsistently.

For clinical practice, the findings support using automatic landmarking as a time-saving first pass that a clinician then reviews and corrects, rather than as an unattended replacement for manual tracing. For research, the most informative next studies would validate models prospectively on external clinical cohorts, report success detection rate at 2, 3, and 4 mm by landmark, and state the inter-examiner variability of the reference so that model error can be compared against it. The broader adoption of artificial intelligence in orthodontics depends on evidence of this kind.⁽⁵⁰⁾

For reporting, the field would benefit from a shared convention. At a minimum, studies should state the number and names of landmarks, the dataset and its provenance, the reference standard and its inter-examiner variability, the mean radial error and its standard deviation, and the success detection rate at 2, 3, and 4 mm by landmark. This would make studies directly comparable and would allow a future meta-analysis at the level of individual landmarks, which is where the clinical decisions actually reside. Until such reporting converges, readers should interpret any single summary figure with the caveat that it averages over landmarks of very different difficulty.

The research agenda that follows has three priorities. First, prospective validation across multiple institutions and populations, so that the generalizability question can be answered rather than assumed. Second, per-landmark reporting with the full threshold curve, so that the difficult points and the easy points can be distinguished. Third, head-to-head evaluation of the commercial tools against research models on a common external dataset, so that clinicians can choose among products on evidence rather than on marketing. These three steps would move the field from a large but heterogeneous literature of demonstrations toward a body of evidence on which routine clinical adoption can be justified.

Taken together, the evidence supports a measured conclusion. Automatic cephalometric landmark detection is mature enough to serve as a reviewed first pass, and it is not yet reliable enough, on the landmarks that anchor the skeletal classification, to be used without supervision. The distance between these two statements is small in engineering terms but large in clinical terms, and closing it requires exactly the kind of external, per-landmark validation that is currently the exception rather than the rule. For the clinician, the actionable message is to adopt the technology as an assistant, to verify the mandibular landmarks in particular, and to validate the tool in the local population before relying on its measurements. The evidence reviewed here supports that adoption while setting a clear expectation about where the residual risk lies. The field has reached the point where the remaining uncertainty is no longer about whether the technology works, but about whether it works for the specific patient, on the specific image, and on the specific landmark in front of the clinician. That is a question of validation, not of architecture, and it is the question that the next generation of studies should answer.

CONCLUSIONS

Artificial intelligence detects cephalometric landmarks with a mean radial error of roughly 1 to 2 mm in most studies, which is comparable to the variability between human examiners. The success detection rate at the 2 mm threshold is heterogeneous, ranging from below 30% to above 95%, and depends on the landmark set and the dataset more than on the architecture. The gonion and condylion are the least reliably located points and are also the ones that anchor the skeletal classification, so a summary error conceals the clinical risk for the measurements that matter most. The reference standard was manual annotation in nearly all studies, and risk of bias was moderate to high, chiefly because of benchmark-based evaluation and scarce external validation. Automatic landmarking should therefore be used as a reviewed first pass, and prospective external

validation with per-landmark success detection rates is required before unattended use can be recommended.

REFERENCES

1. 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. *Journal of Dental Sciences*. 2021;16(1):482-492. doi: 10.1016/j.jds.2020.05.022
2. Subramanian AK, Chen Y, Almalki A, Sivamurthy G, Kafle D. Cephalometric Analysis in Orthodontics Using Artificial Intelligence—A Comprehensive Review. *BioMed Research International*. 2022;2022(1):1880113. doi: 10.1155/2022/1880113
3. Kielczykowski M, Kamiński K, Perkowski K, Zadurska M, Czochrowska E. Application of Artificial Intelligence (AI) in a Cephalometric Analysis: A Narrative Review. *Diagnostics*. 2023;13(16):2640. doi: 10.3390/diagnostics13162640
4. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;n71. doi: 10.1136/bmj.n71
5. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(8):529-536. doi: 10.7326/0003-4819-155-8-201110180-00009
6. Chen J, Che H, Sun J, Rao Y, Wu J. An automatic cephalometric landmark detection method based on heatmap regression and Monte Carlo dropout. *Annu Int Conf IEEE Eng Med Biol Soc*. 2023;2023:1-4. doi: 10.1109/embc40787.2023.10341102
7. Yoon HJ, Kim DR, Gwon E, Kim N, Baek SH, Ahn HW, et al. Fully automated identification of cephalometric landmarks for upper airway assessment using cascaded convolutional neural networks. *Eur J Orthod*. 2022;44(1):66-77. doi: 10.1093/ejo/cjab054
8. Kim MJ, Liu Y, Oh SH, Ahn HW, Kim SH, Nelson G. Evaluation of a multi-stage convolutional neural network-based fully automated landmark identification system using cone-beam computed tomographysynthesized posteroanterior cephalometric images. *Korean J Orthod*. 2021;51(2):77-85. doi: 10.4041/kjod.2021.51.2.77
9. Han SH, Lim J, Kim JS, Cho JH, Hong M, Kim M, et al. Accuracy of posteroanterior cephalogram landmarks and measurements identification using a cascaded convolutional neural network algorithm: A multicenter study. *Korean J Orthod*. 2024;54(1):48-58. doi: 10.4041/kjod23.075
10. Jiang F, Guo Y, Yang C, Zhou Y, Lin Y, Cheng F, et al. Artificial intelligence system for automated landmark localization and analysis of cephalometry. *Dentomaxillofacial Radiology*. 2023;52(1):20220081. doi: 10.1259/dmfr.20220081
11. Bulatova G, Kusnoto B, Grace V, Tsay TP, Avenetti DM, Sanchez FJC. Assessment of automatic cephalometric landmark identification using artificial intelligence. *Orthod Craniofacial Res*. 2021;24(S2):37-42. doi: 10.1111/ocr.12542
12. Zhao C, Yuan Z, Luo S, Wang W, Ren Z, Yao X, et al. Automatic recognition of cephalometric landmarks via multi-scale sampling strategy. *Heliyon*. 2023;9(6):e17459. doi: 10.1016/j.heliyon.2023.e17459
13. SONG Y, QIAO X, IWAMOTO Y, CHEN YW, CHEN Y. An Efficient Deep Learning Based Coarse-to-Fine Cephalometric Landmark Detection Method. *IEICE Trans Inf & Syst*. 2021;E104.D (8):1359-1366. doi: 10.1587/transinf.2021edp7001
14. Hong M, Kim I, Cho JH, Kang KH, Kim M, Kim SJ, et al. Accuracy of artificial intelligence-assisted landmark identification in serial lateral cephalograms of Class III patients who underwent orthodontic treatment and two-jaw orthognathic surgery. *Korean J Orthod*. 2022;52(4):287-297. doi: 10.4041/kjod21.248
15. Zeng M, Yan Z, Liu S, Zhou Y, Qiu L. Cascaded convolutional networks for automatic cephalometric landmark detection. *Medical Image Analysis*. 2021;68:101904. doi: 10.1016/j.media.2020.101904
16. Kim YH, Lee C, Ha EG, Choi YJ, Han SS. A fully deep learning model for the automatic identification of cephalometric landmarks. *Imaging Sci Dent*. 2021;51(3):299. doi: 10.5624/isd.20210077
17. Yang S, Song ES, Lee ES, Kang SR, Yi WJ, Lee SP. Ceph-Net: automatic detection of cephalometric landmarks on scanned lateral cephalograms from children and adolescents using an attention-based stacked regression network. *BMC Oral Health*. 2023;23(1):803. doi: 10.1186/s12903-023-03452-7
18. Neeraja R, Anbarasi LJ. CephTransXnet: An attention enhanced feature fusion network leveraging neighborhood rough set approach for cephalometric landmark prediction. *Computers in Biology and Medicine*. 2025;188:109891. doi: 10.1016/j.compbiomed.2025.109891
19. Laitenberger F, Scheuer HT, Scheuer HA, Lilienthal E, You S, Friedrich RE. Cephalometric landmark detection using vision transformers with direct coordinate prediction. *Journal of Cranio-Maxillofacial Surgery*. 2025;53(9):1518-1529. doi: 10.1016/j.jcms.2025.05.021
20. S R, S S, S. Murthy P, Deshmukh S. Landmark annotation through feature combinations: a comparative study on cephalometric images with in-depth analysis of model's explainability. *Dentomaxillofacial Radiology*. 2024;53(2):115-126. doi: 10.1093/dmfr/twad011
21. Rashmi S, Srinath S, Patil K, Murthy PS, Deshmukh S. Lateral Cephalometric Landmark Annotation Using Histogram Oriented Gradients Extracted from Region of Interest Patches. *J Maxillofac Oral Surg*. 2023;22(4):806-812. doi: 10.1007/s12663-023-02025-z
22. Hong W, Kim SM, Choi J, Ahn J, Paeng JY, Kim H. Automated Cephalometric Landmark Detection Using Deep Reinforcement Learning. *Journal of Craniofacial Surgery*. 2023;34(8):2336-2342. doi: 10.1097/scs.00000000000009685
23. Silva TP, Puntigam GS, Andrade-Bortoletto MFS, Takeshita WM, Oliveira-Santos C, Freitas DQ. Population influence on cephalometric landmark identification: performance of two AI-driven software programs in Brazilian and Korean images. *BMC Oral Health*. 2025;25(1):1596. doi: 10.1186/s12903-025-06807-4

24. Tageldin MA, Yacout YM, Eid FY, Abdel-Haffiez SH. Accuracy of cephalometric landmark identification by artificial intelligence platform versus expert orthodontist in unilateral cleft palate patients: A retrospective study. *International Orthodontics*. 2025;23(2):100990. doi: 10.1016/j.ortho.2025.100990
25. Bao H, Zhang K, Yu C, Li H, Cao D, Shu H, et al. Evaluating the accuracy of automated cephalometric analysis based on artificial intelligence. *BMC Oral Health*. 2023;23(1):191. doi: 10.1186/s12903-023-02881-8
26. Department of Oral and Maxillofacial Radiology, Eskisehir Osmangazi University, Faculty of Dentistry, Eskisehir, Turkey, Ugurlu M. Performance of a Convolutional Neural Network- Based Artificial Intelligence Algorithm for Automatic Cephalometric Landmark Detection. *Turk J Orthod*. 2022;35(2):94-100. doi: 10.5152/turkjorthod.2022.22026
27. Mehra P, R N, Anbarasi L J, Ravi V, Al Mazroa A. VGG-16 based Deep Learning Approach for Cephalometric Landmark Detection. *TOPHJ*. 2024;17(1):e18749445338097. doi: 10.2174/0118749445338097240923063441
28. Sumon MSI, Islam KR, Hossain SA, Rafique T, Ghosh R, Hassan GS, et al. Self-CephaloNet: a two-stage novel framework using operational neural network for cephalometric analysis. *Neural Comput & Applic*. 2025;37(16):9777-9805. doi: 10.1007/s00521-025-11097-6
29. Neeraja R, Jani Anbarasi L. CephXNet: A Deep Convolutional Squeeze-and-Excitation Model for Landmark Prediction on Lateral Cephalograms. *IEEE Access*. 2023;11:90780-90800. doi: 10.1109/access.2023.3307636
30. Song Y, Qiao X, Iwamoto Y, Chen YW. Automatic Cephalometric Landmark Detection on X-ray Images Using a Deep-Learning Method. *Applied Sciences*. 2020;10(7):2547. doi: 10.3390/app10072547
31. Silva TP, Pinheiro MCR, Freitas DQ, Gaêta-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 Surgery, Oral Medicine, Oral Pathology and Oral Radiology*. 2024;137(4):431-440. doi: 10.1016/j.oooo.2024.01.011
32. Panesar S, Zhao A, Hollensbe E, Wong A, Bhamidipalli SS, Eckert G, et al. Precision and Accuracy Assessment of Cephalometric Analyses Performed by Deep Learning Artificial Intelligence with and without Human Augmentation. *Applied Sciences*. 2023;13(12):6921. doi: 10.3390/app13126921
33. Han M, Huo Z, Ren J, Zhu H, Li H, Li J, et al. Automated Landmark Detection and Lip Thickness Classification Using a Convolutional Neural Network in Lateral Cephalometric Radiographs. *Diagnostics*. 2025;15(12):1468. doi: 10.3390/diagnostics15121468
34. Yao J, Zeng W, He T, Zhou S, Zhang Y, Guo J, et al. Automatic localization of cephalometric landmarks based on convolutional neural network. *American Journal of Orthodontics and Dentofacial Orthopedics*. 2022;161(3):e250-e259. doi: 10.1016/j.ajodo.2021.09.012
35. Le VNT, Kang J, Oh IS, Kim JG, Yang YM, Lee DW. Effectiveness of Human-Artificial Intelligence Collaboration in Cephalometric Landmark Detection. *JPM*. 2022;12(3):387. doi: 10.3390/jpm12030387
36. Kim SH, Kim I, Cho JH, Kang KH, Kim M, Kim SJ, et al. Accuracy of artificial intelligence-assisted soft tissue landmark identification in serial lateral cephalograms of Class III two-jaw surgery patients. *Korean J Orthod*. 2025;55(5):392-404. doi: 10.4041/kjod24.302
37. Jiao Z, Liang Z, Liao Q, Chen S, Yang H, Hong G, et al. Deep learning for automatic detection of cephalometric landmarks on lateral cephalometric radiographs using the Mask Region-based Convolutional Neural Network: a pilot study. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*. 2024;137(5):554-562. doi: 10.1016/j.oooo.2024.02.003
38. Gonca M, Bayraktar İŞ, Çelik Ö. Does the FARNet neural network algorithm accurately identify Posteroanterior cephalometric landmarks?. *BMC Med Imaging*. 2024;24(1):294. doi: 10.1186/s12880-024-01478-z
39. Nishimoto S, Sotsuka Y, Kawai K, Ishise H, Kakibuchi M. Personal Computer-Based Cephalometric Landmark Detection With Deep Learning, Using Cephalograms on the Internet. *Journal of Craniofacial Surgery*. 2019;30(1):91-95. doi: 10.1097/scs.00000000000004901
40. HORI M, JINCHO M, HORI T, SEKINE H, KATO A, MIYAZAWA K, et al. Automatic point detection on cephalograms using convolutional neural networks: A two-step method. *Dent Mater J*. 2024;43(5):701-710. doi: 10.4012/dmj.2024-052
41. Popova T, Stocker T, Khazaei Y, Malenova Y, Wichelhaus A, Sabbagh H. Influence of growth structures and fixed appliances on automated cephalometric landmark recognition with a customized convolutional neural network. *BMC Oral Health*. 2023;23(1):274. doi: 10.1186/s12903-023-02984-2
42. Lee H, Cho JM, Ryu S, Ryu S, Chang E, Jung YS, et al. Automatic identification of posteroanterior cephalometric landmarks using a novel deep learning algorithm: a comparative study with human experts. *Sci Rep*. 2023;13(1):15506. doi: 10.1038/s41598-023-42870-z
43. Takeda S, Mine Y, Yoshimi Y, Ito S, Tanimoto K, Murayama T. Landmark annotation and mandibular lateral deviation analysis of posteroanterior cephalograms using a convolutional neural network. *Journal of Dental Sciences*. 2021;16(3):957-963. doi: 10.1016/j.jds.2020.10.012
44. Guo D, Han W, Lyu P, Zhou Y, Shen J. Towards Better Cephalometric Landmark Detection With Diffusion Data Generation. *IEEE Trans Med Imaging*. 2025;44(7):2784-2794. doi: 10.1109/tmi.2025.3557430
45. Bu WQ, Shi ZY, Tian ZQ, Han MQ, Zhang D, Wang S, et al. Accuracy comparative study of automatic landmarking and diagnostic models on lateral cephalograms. *Prog Orthod*. 2025;26(1):50. doi: 10.1186/s40510-025-00598-8
46. Rashmi S, Srinath S, Deshmukh S, Prashanth S, Patil K. Cephalometric landmark annotation using transfer learning: Detectron2 and YOLOv8 baselines on a diverse cephalometric image dataset. *Computers in Biology and Medicine*. 2024;183:109318. doi: 10.1016/j.combiomed.2024.109318

47. Rashmi S, Srinath S, Rakshitha R, Poornima BV. Ensemble learning methods with single and multi-model deep learning approaches for cephalometric landmark annotation. *Discov Artif Intell.* 2024;4(1):93. doi: 10.1007/s44163-024-00207-3
48. Bagdy-Bálint R, Szabó G, Zováthi ÖH, Zováthi BH, Somorjai Á, Köpenczei C, et al. Accuracy of automated analysis in cephalometry. *Journal of Dental Sciences.* 2025;20(2):830-843. doi: 10.1016/j.jds.2024.09.012
49. Freire RT, Mota IMBDS, Leão LDV, Calasans IDCP, Santos DDAF, Silva LND, et al. Is automatic cephalometric landmark detection with artificial intelligence reliable? - a systematic review with meta-analysis of SNA, SNB, ANB, and WITS cephalometric measures. *Rev DELOS.* 2025;18(73):e6871. doi: 10.55905/rdelosv18.n73-003
50. Hung K, Montalvao C, Tanaka R, Kawai T, Bornstein MM. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: A systematic review. *Dentomaxillofacial Radiology.* 2020;49(1):20190107. doi: 10.1259/dmfr.20190107

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Melissa Castillo Jimenez, Maylon Felix Guerra Reyes.

Writing–review and editing: Melissa Castillo Jimenez, Maylon Felix Guerra Reyes.