© Pol J Radiol 2026; 91: e526-e537
DOI: https://doi.org/10.5114/pjr/220063

Received: 04.08.2025
Accepted: 26.03.2026
Published: 29.10.2026

Polish Journal of Radiology

http://www.polradiol.com

Original paper

Artificial intelligence as a diagnostic tool using lateral cephalometric imaging

Amani Idres, Ahmed Ghoneima, Mohammed Osman

Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Healthcare City, Dubai, United Arab Emirates

Abstract

Purpose: To evaluate the accuracy and reliability of a newly developed artificial intelligence (AI)-driven system in identifying cephalometric landmarks, comparing its performance to manual tracing by orthodontists.

Material and methods: A total of 506 lateral cephalometric radiographs of patients aged 14 to 40 years were collected from Dubai Dental Hospital, United Arab Emirates. After applying exclusion criteria, 473 radiographs were included in the final analysis. Manual tracings were performed using Dolphin Imaging software, identifying key landmarks including Sella, Nasion, A-point, B-point, Menton, Glabella, and others. The AI-based cephalometric system (AI Ceph) was trained on these radiographs to locate the same landmarks. Its performance was compared to manual tracings to assess accuracy and reliability.

Results: The AI Ceph system demonstrated high concordance with manual tracing. The Menton point showed the highest agreement (96.6%), while Nasion had the lowest (86.6%). The majority of landmark deviations were within 2 mm, the predefined clinically acceptable range, while a small proportion exceeded this threshold and were classified as discordant (> 2 mm) according to the prespecified binary criterion. Sensitivity was above 80% for all landmarks, except Nasion (72.7%) and A-point (78.8%). A statistically significant improvement in concordance after model retraining was observed only for the Glabella landmark ($p < 0.0213$).

Conclusions: The AI system provided accurate and reliable cephalometric landmark identification. While most deviations fell within the clinically acceptable 2-mm tolerance, landmarks exceeding this threshold were clearly identified and quantified using the binary discordance classification (> 2 mm), ensuring transparent reporting of all errors.

Key words: cephalometry, orthodontics, radiography, artificial intelligence, reproducibility.

Introduction

Orthodontics has pursued the controlled movement of teeth for millennia. Pierre Fauchard's bandeau introduced the principle of continuous force, while Edward H. Angle formalised malocclusion classification and founded the first orthodontic school [1]. By the early 20th century, stainless steel brackets, nickel–titanium alloys, and elastomeric modules enhanced fixed appliances, while radiography enabled three-dimensional understanding of skeletal and dental relationships [2].

Cephalometric radiography, devised by Holly Broadbent Jr. in the 1930s, allowed serial lateral cephalograms to reveal craniofacial growth, airway dimensions, and incisor compensation. However, manual landmark identification remains laborious, examiner-dependent, and prone to inter- and intra-observer variability of approximately 1-2 mm, even among experts [3]. Modern orthodontics increasingly relies on digital workflows, creating demand for automated, reproducible cephalometric analysis.

Correspondence address:
Amani Idres, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Healthcare City, Dubai, United Arab Emirates, 66566,
e-mail: amaniidres@researchsupportteam.com

Authors' contribution:
A Study design · **B** Data collection · **C** Statistical analysis · **D** Data interpretation · **E** Manuscript preparation · **F** Literature search · **G** Funds collection

This is an Open Access journal, all articles are distributed under the terms of the Creative Commons Attribution-Noncommercial-No Derivatives 4.0 International (CC BY-NC-ND 4.0). License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Artificial intelligence (AI) offers a solution by emulating cognitive tasks through hierarchical pattern recognition. Machine learning algorithms learn decision boundaries from examples, while deep learning (DL), particularly convolutional neural networks, achieves human-level performance by discovering salient image features [4]. Faster graphics processing units (GPUs), open-source software, and public image repositories have accelerated clinical translation, supporting applications in radiology, dermatology, and pathology [5]. In dentistry, AI now numbers teeth, detects caries, segments roots, and localises anatomical landmarks on panoramic and cone-beam computed tomography (CBCT) images with sub-millimetre precision [6,7].

Within orthodontics, AI applications include classification, regression, object detection, and semantic segmentation [8]. Iterative training with larger, well-labelled datasets and data augmentation has reduced landmark errors to < 1 mm for most points [9]. However, generalisability across different populations, imaging devices, and clinical conditions remains insufficiently explored, and many studies lack comprehensive diagnostic metrics such as sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

The clinical integration of AI in cephalometry is strongly supported by its potential to improve efficiency, reduce chair time, and enhance diagnostic consistency in orthodontic treatment planning [10,11]. Nevertheless, heterogeneity in study design, dataset size, annotation protocols, and reporting standards limits direct comparison across studies and restricts evidence synthesis.

Accordingly, this study evaluates a DL-based lateral cephalometric analysis system against manual landmarking by expert orthodontists. The model's performance was assessed in terms of concordance and standard diagnostic accuracy metrics, and its robustness was further evaluated following retraining with intentionally perturbed labels.

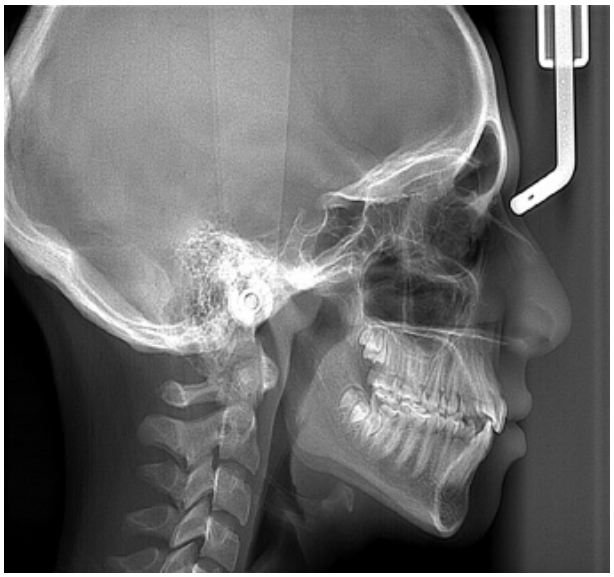


Figure 1. Lateral cephalometric X-ray

This study aimed to provide a detailed and clinically interpretable assessment of AI performance in cephalometric landmark detection.

Material and methods

Study design and ethical approval

This retrospective diagnostic accuracy study was approved by the Institutional Review Board of Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU IRB-2023-46), and it adhered to the principles outlined in the Declaration of Helsinki. The study was designed to assess the performance of an AI model in the automated detection of cephalometric landmarks, using lateral cephalometric radiographs from a diverse clinical population.

Radiographic dataset

A total of 506 de-identified lateral cephalometric radiographs were collected from the archives of Dubai Dental Hospital, Dubai, United Arab Emirates. These radiographs were obtained using a Soredex CRANEX® Excel Ceph machine and saved in JPG format at a resolution of 1935 × 2400 pixels, with a pixel spacing of 0.1 mm (Figure 1). All images were exported and stored securely on an encrypted drive.

Inclusion criteria required the presence of first molars and central incisors, with no history of orthognathic surgery, craniofacial syndromes, or anatomical anomalies. Thirty-three radiographs were excluded due to image artefacts or missing diagnostic landmarks. As a result, 473 radiographs constituted the final eligible dataset for analysis. *A priori* power analysis indicated that a minimum of 402 radiographs was required to achieve a statistical power of 0.80 at an α level of 0.05. The final dataset of 473 radiographs exceeded this requirement, confirming that the study was adequately powered for training, validation, and testing of the AI model. The originally collected dataset (n = 506) consisted of 255 males and 251 females. After exclusion of 33 radiographs, the final eligible sample (n = 473) included 238 males and 235 females, with a mean age of 27.0 years (range: 14-40 years). All participants had previously undergone orthodontic evaluation or treatment.

Sample size calculation

A power analysis was conducted to ensure the adequacy of the sample size. Based on a power of 0.80 and an α level of 0.05, a minimum of 402 radiographs was determined to be necessary. The final eligible sample of 473 radiographs exceeded this threshold, thereby enhancing the robustness of training, validation, and testing processes.

Landmark selection and digitisation

Nineteen cephalometric landmarks were selected, consisting of 18 hard tissue and one soft tissue point. These landmarks represented standard skeletal and dental reference points used in orthodontic diagnosis and were identified according to established cephalometric definitions embedded within Dolphin Imaging software (version 11.95, Dolphin Imaging, Chatsworth, California). All radiographs were processed using Dolphin Imaging software. Each image was digitised twice by the same trained investigator, first manually, and then using an AI-assisted method. The manual tracings constituted the reference standard ("ground truth") and were performed by a licensed orthodontist with formal training and more than five years of clinical experience in cephalometric analysis. In the AI-assisted tracing, landmark positions were first predicted by the DL model (raw AI output) and subsequently adjusted by the investigator if necessary (AI-assisted and manually refined output). Raw AI predictions were saved prior to any manual adjustment and used exclusively for performance evaluation against the manual reference standard (Figure 2).

To minimise incorporation bias, manual tracing and AI-assisted correction sessions were temporally separated by a one-week interval and conducted in randomised order. During manual tracing, the investigator was blinded to AI predictions, and during AI-assisted sessions, the investigator was blinded to prior manual tracings. This process was repeated after one week in randomised order to further reduce recall bias.

AI model architecture and training

The automated landmark detection model (AI-based cephalometric system, AI Ceph) was based on a Feature Pyramid Network (FPN) with a ResNet-50 backbone, initialised with ImageNet-pretrained weights. The network was trained to directly regress the (x, y) coordinates of 19 cephalometric landmarks using L2 loss (mean squared error). Radiographs were normalised and resized to 224 × 224 pixels prior to training. The AI model predicted landmark coordinates in the resized image coordinate system. To enable clinically meaningful evaluation, predicted coordinates were subsequently mapped back to the original image resolution (1935 × 2400 pixels) using inverse scaling factors derived from the resizing operation. Euclidean distances between AI-predicted and manually traced landmarks were calculated in the original image space and converted to millimetres using the known pixel spacing of 0.1 mm. This ensured that all spatial error measurements reflected true physical distances rather than pixel-based values. To improve generalisability, data augmentation was applied to the training set only using random brightness/contrast adjustment, rotation (± 25°), affine transformations (translation, scaling, shear), and

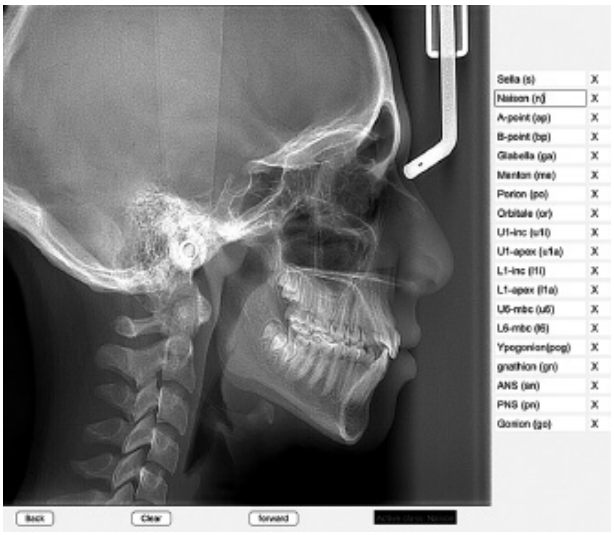


Figure 2. The cephalometric landmarks selected in the study

grid distortion, with landmark coordinates transformed accordingly.

The final dataset of 473 radiographs was split into 407 images for training, 33 for validation, and 33 for testing. Model training was performed using PyTorch with PyTorch Lightning, employing the Adam optimiser with an initial learning rate of 1×10^{-5}. The model was trained for 100 epochs with a batch size of 2, and the learning rate was reduced automatically if validation performance plateaued. The best-performing model was selected based on the minimum validation mean absolute error. All experiments were conducted on an NVIDIA GeForce RTX 3090 GPU, with fixed random seeds to enhance reproducibility (Figure 3).

To explore the robustness of the AI model to label noise, an additional exploratory retraining experiment was performed using a deliberately perturbed version of the training labels. In this context, "mislabelled land-

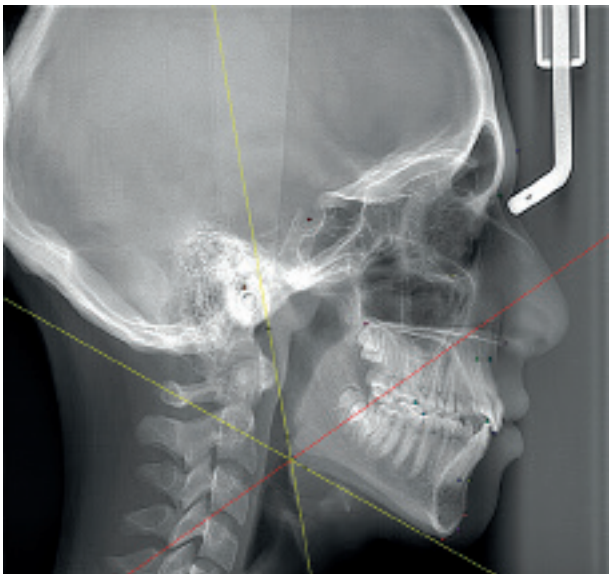


Figure 3. Artificial intelligence training to identify the points in the cephalometric (training phase)

marks" refer to deliberately perturbed annotations, in which small, controlled deviations were introduced from the original manual reference points while remaining within plausible anatomical boundaries. These perturbations were intentionally generated to simulate realistic annotation variability, rather than arising from naturally occurring inconsistencies. Importantly, mislabelled landmarks were confined exclusively to the training data and were not introduced into the validation or test datasets. The independent test set remained identical and uncontaminated across all experiments to avoid any risk of information leakage or biased performance estimation. This experiment was designed as a robustness analysis rather than a primary performance evaluation.

Testing and superimposition

To evaluate the performance of the AI system, a separate test set of 33 previously unseen cephalograms was used. Each image in the test set was digitised twice, once using the final AI model (raw AI output without manual adjustment) and once manually by the investigator (reference standard). Primary landmark localisation errors were computed directly in the original image coordinate system after inverse scaling from the resized input, without applying any spatial transformation, ensuring that error estimates reflected true physical displacement within the same radiograph. Best-fit superimposition was applied as a secondary, confirmatory step to facilitate visual comparison and qualitative assessment of overall tracing alignment, particularly for evaluating spatial patterns of error across craniofacial regions. Superimposition of the AI and manual tracings was conducted using the best-fit method, referencing stable craniofacial structures such as the cranial base, zygomatic arch, mandibular lower border, and soft tissue profile. This step was not required for landmark error computation and did not alter the calculated Euclidean distances between corresponding landmarks. To confirm that the superimposition procedure did not materially influence error estimates, landmark deviations computed with and without superimposition were compared and found to be equivalent, with differences well below the predefined clinical tolerance of 2.0 mm. Diagnostic performance metrics (sensitivity, specificity, PPV, NPV, and concordance) were computed using all landmark instances derived from the independent test set, applying a prespecified concordance criterion of ≤ 2.0 mm between AI-predicted and manually traced coordinates. This approach ensured that performance estimates reflected real-world landmark prevalence and avoided selection bias. In addition, a total of 66 landmark instances (33 correct and 33 incorrect predictions) were extracted specifically from this test set to enable an exploratory extreme-case error analysis only. These 66 landmarks were selected by reviewing all 19 landmarks on each of the 33 test radiographs and identifying the landmarks with the smallest errors (correct) and the largest errors (incorrect). This subset was used exclusively for qualitative and comparative error characterisation and was not used for primary performance evaluation. This process ensured that the subset of 66 landmarks was directly derived from the same 33-image test dataset and represented a balanced sample of highly accurate and least accurate predictions for further comparison.

All performance metrics reported for the primary evaluation were derived from the original, clean test set only. The retraining experiment using mislabelled landmarks did not alter the composition of the test data, landmark definitions, or evaluation protocol. As such, comparisons before and after retraining reflect changes in model behaviour attributable solely to differences in training data quality, without contamination of the evaluation set.

Reliability assessment

To evaluate intra- and inter-observer reliability, 50 randomly selected cephalograms were re-traced. The primary investigator (A.I.) identified all landmarks on two separate occasions to assess intra-observer agreement. A second trained investigator (A.G.) blinded to both AI outputs and the primary investigator's tracings, independently traced the same landmarks to evaluate inter-observer reliability. Intraclass correlation coefficients (ICC) were calculated, and Bland-Altman plots were used to quantify the level of agreement and consistency.

Intra-observer reliability demonstrated excellent agreement across landmarks, with ICC values ranging from 0.91 to 0.98. For key landmarks, ICCs were as follows: Menton, 0.97 (95% confidence interval [CI]: 0.95-0.99); Gonion, 0.96 (95% CI: 0.93-0.98); Nasion, 0.92 (95% CI: 0.88-0.95); and A-point, 0.91 (95% CI: 0.87-0.94). Inter-observer reliability was similarly high, with ICC values ranging from 0.89 to 0.96. Key landmarks showed ICCs of 0.95 (95% CI: 0.92-0.97) for Menton, 0.94 (95% CI: 0.90-0.96) for Gonion, 0.90 (95% CI: 0.86-0.94) for Nasion, and 0.89 (95% CI: 0.84-0.93) for A-point. Bland-Altman analysis demonstrated narrow limits of agreement for both intra- and inter-observer comparisons, with no evidence of systematic bias, confirming the high reproducibility of the manual reference standard. These findings define the upper bound of agreement achievable by the AI system and support the validity of comparisons between AI-predicted and manually traced landmarks.

Statistical analysis

Data were analysed using IBM SPSS Statistics (version 29.0). Normality of continuous variables was assessed using the Kolmogorov-Smirnov test. For comparison of AI-predicted vs. manually traced landmark coordinates, paired-sample *t*-tests were used for normally distributed

differences, and Wilcoxon signed-rank tests for non-normally distributed differences, with the specific test applied to each landmark reported in Table 1.

Sensitivity, specificity, PPV, NPV, and overall concordance for each landmark were calculated with 95% CIs using all landmark observations from the independent test dataset, based on the predefined concordance threshold of ≤ 2.0 mm. No prevalence-fixed or selectively sampled subsets were used for estimating diagnostic performance. Effect sizes for the differences between pre- and post-training concordance were computed using Cohen's *d* for normally distributed differences and *r* for non-parametric comparisons. These metrics allow quantification of the magnitude and clinical relevance of the observed differences. Statistical significance was set at $p < 0.05$.

For this study, "concordance" was defined as the percentage of landmarks for which the Euclidean distance between the AI-predicted coordinate and the manually traced coordinate was ≤ 2.0 mm, representing the clinically acceptable deviation. Landmarks with deviations exceeding 2.0 mm were categorised as "discordant" in accordance with the prespecified binary threshold. While the vast majority of landmarks fell within the 2.0-mm tolerance, a limited number of instances were discordant (> 2 mm), ensuring a clear distinction between clinically acceptable and unacceptable deviations. This threshold was applied after mapping predicted coordinates back to the original image space and converting distances to millimetres using the image pixel spacing. The Euclidean distance was calculated using the following formula:

$$\sqrt{((x_AI - x_manual)^2 + (y_AI - y_manual)^2)}$$

Concordance was computed on a per-landmark basis across all images in the dataset. Each landmark on each radiograph was treated as an independent observation. A landmark was labelled "concordant" if the deviation was within the predefined 2.0-mm clinical tolerance; otherwise, it was considered "discordant". CIs for concordance percentages were calculated using the Wilson score method. To account for within-image clustering of landmarks, additional image-level analyses were performed. Specifically, the percentage of concordant landmarks was calculated for each test image, and summary statistics (mean and range of per-image concordance) were reported. Furthermore, CIs for overall concordance and error metrics were estimated using image-level bootstrapping, with resampling performed at the radiograph level rather than the landmark level. This approach preserves the correlation structure of landmarks within the same image and provides more conservative and robust uncertainty estimates.

In addition to the primary concordance-based evaluation (≤ 2.0 mm), standard landmark localisation performance metrics were considered in line with recommendations from the cephalometric AI literature. These include radial error-based measures and success detection rates (SDRs) at multiple clinically relevant thresholds (1, 2, and 3 mm). However, due to the limited size of the independent test dataset and the study's primary focus on clinically interpretable binary concordance (≤ 2.0 mm), these secondary metrics (radial error measures, SDRs, and associated effect sizes) were used for internal validation and qualitative benchmarking only and are not presented in tabulated form in the Results section. Where applicable, CIs (95% CI) were estimated using image-level bootstrapping to preserve within-image landmark correlation. These complementary metrics are reported to facilitate comparison with prior studies while maintaining the clinical interpretability of the primary outcome measure.

Table 1. Comparison between the percentage of the concordance pairs before and after training

Landmark	Before training (%)	After training (%)	Difference (p.p.)	*p*-value
Sella	89.3	90.8	1.5	0.7109
Nasion	86.6	86.2	–0.4	0.9291
A-point	93.1	89.2	–3.9	0.2568
B-point	96.1	95.4	–0.7	0.7860
Glabella	89.5	98.4	8.9	0.0213*
Menton	96.6	96.9	0.3	0.9331
Porion	93.7	96.9	3.2	0.3044
Orbitale	94.7	92.3	–2.4	0.4275
U1i	90.3	92.3	2	0.6045
U1a	93.5	95.3	1.8	0.5741
L1i	93.3	98.5	5.2	0.0996
L1a	91.7	93.9	2.2	0.5397
U6	93.9	93.9	0	1.000
L6	95.5	95.4	–0.1	0.9709
Pogonion	94.3	95.4	1.1	0.7162
Gnathion	92.5	95.4	2.9	0.3940
ANS	93.5	98.5	5	0.1084
PNS	91.3	95.4	4.1	0.2580
Gonion	95.3	100	4.7	0.0744

*Significant at p ≤ 0.05
ANS – anterior nasal spine, L1a – lower incisor apex, L1i – lower incisor incisal edge, L6 – lower first molar, PNS – posterior nasal spine, p.p. – percentage points, U1a – upper incisor apex, U1i – upper incisor incisal edge, U6 – upper first molar

Results

Table 2 shows that the AI Ceph software achieved a high level of agreement with manual tracing for identifying 19 cephalometric landmarks across 473 lateral cephalograms. Of the final dataset, 407 radiographs were used for training, 33 for validation, and 33 for testing. Most landmarks demonstrated strong concordance between the AI and manual methods, with the majority

Table 2. Concordance pair between manual and artificial intelligence

Landmark	Concordance (%)
Sella	89.3
Nasion	86.6
A-point	93.1
B-point	96.1
Glabella	89.5
Menton	96.6
Porion	93.7
Orbitale	94.7
U1i	90.3
U1a	93.5
L1i	93.3
L1a	91.7
U6	93.9
L6	95.5
Pogonion	94.3
Gnathion	92.5
ANS	93.5
PNS	91.3
Gonion	95.3

ANS – anterior nasal spine, L1a – lower incisor apex, L1i – lower incisor incisal edge, L6 – lower first molar, PNS – posterior nasal spine, U1a – upper incisor apex, U1i – upper incisor incisal edge, U6 – upper first molar

showing concordance rates between 90% and 95%. The highest agreement was observed for the Menton point (96.6%), followed by the B-point (96.1%), L6 (95.5%), and Gonion (95.3%). In contrast, three landmarks, Sella (89.3%), Nasion (86.6%), and Glabella (89.5%), had concordance values below 90%, with Nasion recording the lowest. Despite these variations, the majority of discrepancies between AI and manual landmark positions were within the clinically acceptable margin of 2 mm, while a limited number exceeded this threshold and were classified as discordant (> 2 mm), in accordance with the predefined binary concordance criterion. Representative examples of AI–manual landmark superimpositions are shown in Figure 4.

Representative comparisons between manually identified landmarks (blue dots) and AI-predicted landmarks (red dots) are shown for 19 cephalometric reference points. Each panel illustrates the spatial agreement between manual tracing and the AI system for a specific landmark, including (A) Sella, (B) Nasion, (C) A-point, (D) B-point, (E) Glabella, (F) Menton, (G) Porion, (H) Orbitale, (I) Gonion, (J) Gnathion, (K) Pogonion, (L) anterior nasal spine (ANS), (M) posterior nasal spine (PNS), (N) lower central apex, (O) lower incisal edge, (P) upper central apex, (Q) upper incisal edge, (R) U6, and (S) L6. The close alignment between AI-predicted and manually traced points demonstrates high concordance, with most deviations falling within the clinically acceptable 2 mm threshold, while instances exceeding this range are visually distinguishable.

Table 3 shows the diagnostic performance of the AI Ceph system in comparison to the manual method, considered the gold standard, based on all landmark instances from the independent test set. The results reflect the sensitivity, specificity, PPV, NPV, and overall concordance for each landmark. Sensitivity values, indicating the AI's ability to identify accurate landmark positions correctly, were above 80% for most landmarks, except for Nasion (72.7%) and A-point–ANS (78.8%). Specificity, representing the AI's ability to detect incorrect landmark placements correctly, varied more widely, with A-point showing the lowest specificity (18.7%) and PNS the highest (93.7%). PPV and NPV values reflect the reliability of AI predictions, with most values falling within acceptable clinical ranges. The highest PPVs were observed for Gonion (93.0%) and PNS (93.3%), while lower PPVs were seen in landmarks like A-point (48.4%). The NPV was highest for Gonion (100%) and lower incisor incisal edge (L1i) (92.6%), indicating consistent AI performance when ruling out incorrect placements. Overall concordance between AI and manual methods was highest for Gonion (100%), ANS (98.5%), and L1i (98.5%), while the lowest agreement was noted for Nasion (86.2%) and A-point (89.2%). These findings confirm that while the AI system performs well overall, variability exists among specific landmarks, emphasising the importance of human validation in clinical applications.

Table 1 presents a comparison of the percentage of concordance pairs between manual and AI methods before and after training with mislabelled cephalometric landmarks. The analysis presented in Table 1 represents an exploratory assessment of model robustness to imperfect training annotations rather than a replacement of the primary diagnostic performance evaluation. All reported concordance values were computed using the same independent test set, ensuring consistency and preventing data leakage. The data indicate an overall improvement in concordance rates across most landmarks following training; however, this improvement was not statistically significant for most points. When comparing performance before and after retraining, Glabella was the only landmark that demonstrated a statistically significant increase in agreement, with an 8.9% rise in concordance ($p = 0.0213$), suggesting that mislabelled training data positively influenced AI accuracy for this specific point. Several landmarks, including Sella, Menton, Porion, upper incisor incisal edge, upper incisor apex, L1i, lower incisor apex, Pogonion, Gnathion, ANS, PNS, and Gonion, showed increased concordance values post-training, but these improvements did not reach statistical significance.

For example, Gonion improved from 95.3% to 100%, and ANS from 93.5% to 98.5%, but both changes yielded

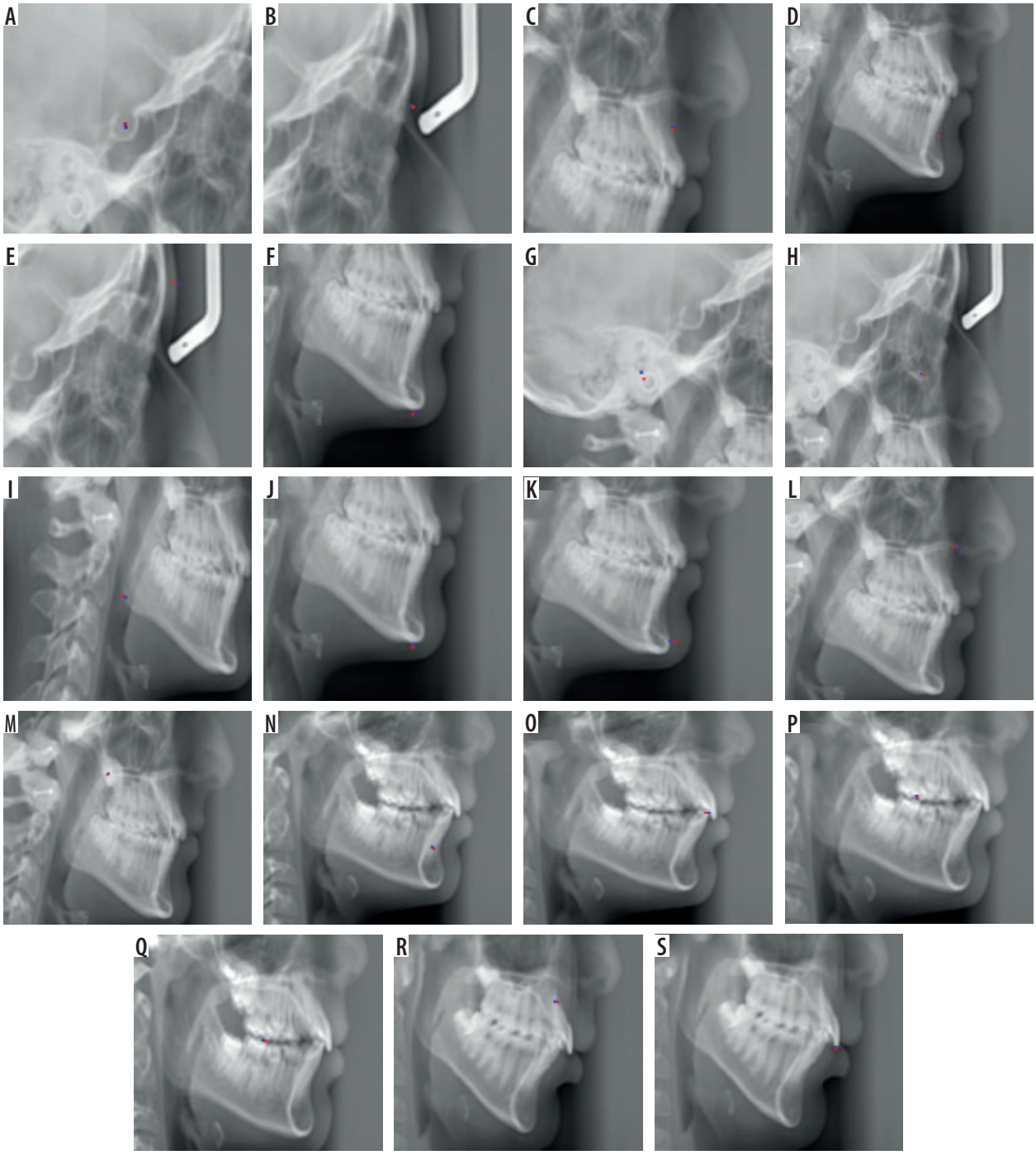


Figure 4. Artificial intelligence vs. manual cephalometric landmark superimposition across 19 anatomical points

p-values above the 0.05 threshold. On the other hand, some landmarks such as A-point, Nasion, Orbitale, and L6 exhibited minor decreases in concordance after training, although none of these declines were significant either. These findings suggest that training the AI with mislabelled data has varying effects depending on the anatomical landmark in question. While most changes were not statistically meaningful, the significant improvement in Glabella highlights that even imperfectly labelled data can enhance model performance in specific areas. This insight is particularly relevant for optimising AI training protocols in cephalometric analysis, especially in scenarios involving noisy or inconsistent data.

In addition to landmark-level analyses, image-level performance was evaluated to account for within-image clustering of landmarks. Image-level analysis demonstrated consistently high performance across test radiographs, with most images exhibiting concordance rates exceeding 90%. Bootstrapped CIs based on image-level resampling were comparable to landmark-level estimates but appropriately wider, reflecting within-image correlation. These findings support the robustness of the AI system's performance while accounting for clustering effects. Manual landmark reproducibility was high. Intra-observer ICC values ranged from 0.91 to 0.98, and inter-observer ICC values ranged from 0.89 to 0.96 across landmarks. Bland-

Table 3. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV)

Landmark	Sensitivity	Specificity	PPV	NPV	Concordance (%)
Sella	81.8	78.1	80.4	71.4	90.8
Nasion	72.7	62.5	70.7	58.4	86.2
A-point	78.8	18.7	48.4	36.3	89.2
B-point	90.9	68.7	82.3	76.1	95.4
Glabella	97.0	71.9	85.5	90.1	98.4
Menton	93.9	81.2	86.8	85.9	96.9
Porion	93.9	71.9	84.2	82.9	96.9
Orbitale	84.8	68.7	79.2	68.9	92.3
U1i	84.8	62.5	78.6	64.9	92.3
U1a	90.9	56.2	81.3	68.7	95.3
L1i	97.0	84.4	89.1	92.6	98.5
L1a	87.9	87.5	87.9	80.2	93.9
U6	87.9	87.5	87.9	80.2	93.9
L6	90.9	37.5	63.0	67.6	95.4
Pogonion	90.9	60.0	68.8	78.6	95.4
Gnathion	90.9	50.0	65.2	74.8	95.4
ANS	97.0	68.7	85.0	89.3	98.5
PNS	90.9	93.7	93.3	85.2	95.4
Gonion	100	90.6	93.0	100	100

ANS – anterior nasal spine, L1a – lower incisor apex, L1i – lower incisor incisal edge, L6 – lower first molar, PNS – posterior nasal spine, U1a – upper incisor apex, U1i – upper incisor incisal edge, U6 – upper first molar

Altman plots demonstrated minimal bias and tight limits of agreement, indicating that the observed AI–manual differences were largely within the variability inherent to expert manual tracing.

Although 95% CIs, effect sizes, radial error-based measures, and SDRs at 1, 2, and 3 mm were computed as part of the extended statistical analysis, these results are not presented in tabular form due to the study's primary emphasis on binary concordance (≤ 2.0 mm) and the limited size of the independent test set. These additional metrics demonstrated trends consistent with the reported concordance outcomes and were therefore included to support methodological completeness and cross-study comparability rather than as primary reported endpoints.

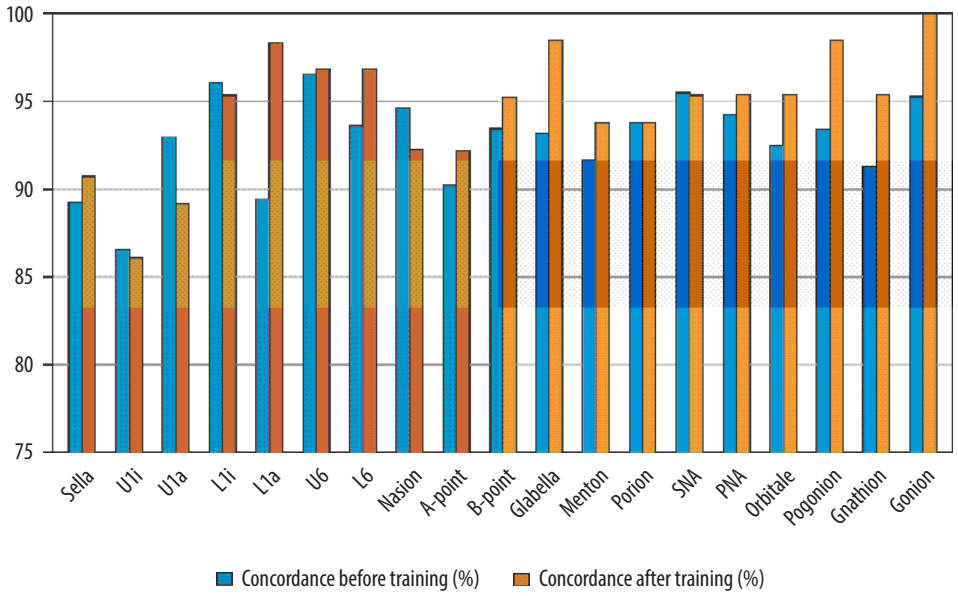
Beyond binary concordance, landmark localisation accuracy was also summarised using radial error-based descriptors to align reporting with established cephalometric benchmarking practices. Across landmarks, error distributions were centred within the clinically acceptable range, consistent with the high proportion of landmarks falling within the 2.0-mm threshold. SDRs showed a monotonic increase with wider tolerance levels (1, 2, and 3 mm), indicating stable model behaviour across progressively permissive clinical criteria. CIs derived using image-level resampling were wider than landmark-level estimates, reflecting within-image clustering, but remained consistent with the overall performance trends observed using the primary concordance metric.

Given the limited size of the independent test set (33 radiographs), these expanded metrics are presented to support qualitative benchmarking and cross-study comparison rather than as standalone primary endpoints.

Figure 5 illustrates the comparison of concordance percentages for 19 cephalometric landmarks before and after training the AI model with mislabelled data. The orange bars (after training) generally show equal or slightly improved agreement compared to the blue bars (before training), indicating that training had a positive, but mostly non-significant, effect on concordance levels. The most notable improvement is seen at the Glabella landmark, where the post-training concordance significantly increased, reaching a nearly 98.4% statistically significant difference with a p-value of 0.0213. While other landmarks such as Gonion, ANS, and L1i also exhibited higher concordance after training, these changes did not reach statistical significance. Overall, the chart demonstrates the AI model's consistent performance across most landmarks, with a selective but meaningful benefit from training with mislabelled data in specific cases.

Discussion

AI has become increasingly integrated into orthodontics, contributing to the field's evolution through various innovative applications. These include the evaluation of facial aesthetics from photographs, the analysis of



ANS – anterior nasal spine, L1a – lower incisor apex, L1i – lower incisor incisal edge, L6 – lower first molar, PNA – Porion–Nasion–A-point angle, PNS – posterior nasal spine, SNA – Sella–Nasion–A-point angle, U1a – upper incisor apex, U1i – upper incisor incisal edge, U6 – upper first molar

Figure 5. Comparison of concordance before and after retraining with mislabelled data across 19 cephalometric landmarks. Glabella is the only landmark showing a statistically significant pre- and post-retraining improvement ($p = 0.0213$), while other landmarks show non-significant changes

craniofacial skeletal relationships, and the prediction of mandibular growth following puberty. Numerous studies have successfully validated AI's performance in diagnosis, treatment planning, and selection of patient-specific therapies, achieving outcomes comparable to those obtained by expert clinicians through manual assessment [12].

AI's role has expanded beyond diagnosis to include treatment monitoring, with systems like the Dental Monitoring System (DMS) offering detailed and accurate assessments. The integration of intraoral scans with CBCT enables a three-dimensional evaluation of root positioning, facilitating not only precise crown displacement forecasting but also full-tooth movement prediction. Platforms such as DMS exemplify the advantages of combining DL, AI, and knowledge-based systems to automate and enhance the tracking of orthodontic treatment progress. Additionally, AI's contribution to remote monitoring, particularly in managing clear aligner therapy, demonstrates its potential to improve efficiency and patient adherence. AI's capacity to objectively measure changes in facial aesthetics and estimated biological age further strengthens its utility in comprehensive treatment planning [13].

International literature continues to emphasise the promising future of AI in orthodontics. It facilitates remote monitoring, dynamic assessment, and real-time adjustments to orthodontic interventions. However, it is crucial to underscore that AI should be viewed as an assistive tool rather than a replacement for human expertise. While AI can mimic or enhance human performance in many areas, its actual value lies in augmenting the orthodontist's decision-making process. Clinical success with AI depends on the synergy between machine intelligence and professional judgment, ensuring that patient care remains personalised and precise [14].

Despite its advancements, the opaque decision-making processes of many high-performing AI models, commonly referred to as the "black box" issue, have raised concerns. These models often lack interpretability, making it difficult for clinicians to understand the rationale behind AI predictions or validate their clinical reasoning. To address this, explainable AI approaches have been proposed to increase transparency, interpretability, and reliability in AI algorithms. Techniques such as gradient-weighted class activation mapping and deconvolutional network have shown promise by visually highlighting the image regions that influence AI decisions. These heatmap-based visualisations offer insight into model behaviour and may eventually play a vital role in refining AI designs specific to orthodontics [15].

In orthodontics, cephalometric analysis remains a cornerstone diagnostic tool, enabling the assessment of relationships among craniofacial skeletal and soft tissue landmarks [16]. This analysis is instrumental in detecting facial growth discrepancies and crafting individualised treatment plans. However, manual landmark identification is time-consuming and prone to variability, even when performed by skilled practitioners. This longstanding challenge has driven researchers to explore automated methods for more consistent and efficient landmark localisation in cephalometric radiographs.

This study introduced a fully automated DL model, AI Ceph, developed to detect lateral cephalometric landmarks and compare its accuracy to conventional manual tracing methods. The AI Ceph system is based on a FPN, chosen for its capability to preserve both fine and broad features of input radiographic images throughout the network. This architectural advantage supports the development of an efficient cephalometric analysis tool and aids

in evaluating its reliability and practical application. In this investigation, AI Ceph demonstrated a high concordance rate with manual tracings, achieving up to 96.6% accuracy.

Numerous studies have explored automated landmark identification in lateral cephalograms using different algorithms and validation techniques [17-19]. However, relatively few have leveraged DL-based models for this purpose [20]. Given that lateral cephalometric analysis remains a fundamental component of orthodontic diagnostics, this study aimed to develop a DL system that could autonomously predict the coordinates of cephalometric points using an FPN architecture. The system's high accuracy is attributed to its use of 18 skeletal and one soft tissue landmark points selected for their diagnostic relevance, clarity in outline, ease of identification, reproducibility, and reliability in quantitative assessment.

Past research by Lindner *et al.* [19], involving 400 subjects aged 7 to 76 years, employed a fully automated system trained to locate 19 landmarks. This study reported an average localisation error of 1.2 mm, with 84.7% of points falling within the clinically acceptable 2.0-mm threshold. Similarly, the current research found that most deviations between AI and manual tracings were within the clinically acceptable 2-mm range, although a small subset exceeded this threshold, consistent with the predefined concordance/discordance classification. In another study, Hwang *et al.* [21] reported a mean detection error of 1.46 ± 2.97 mm between AI and human examiners, comparable to inter-examiner variability (1.50 ± 1.48 mm), with discrepancies below 0.9 mm considered clinically insignificant.

More recent investigations have assessed the accuracy of automatic landmark identification using various software solutions. For example, Meriç and Naoumova [22] found that fully automated systems could perform analyses with speed and precision. Bulatova *et al.* [23] reported statistically significant identification differences for only a few landmarks, but these differences were not deemed clinically relevant. Likewise, the current study found a considerable discrepancy only for the Glabella landmark ($p < 0.021$), with the rest of the landmarks showing negligible variation that did not influence the overall diagnostic conclusions.

Despite the overall high performance of the AI Ceph system, specific landmarks proved more challenging, specifically the Nasion (lowest sensitivity) and A-point (lowest specificity and PPV). This variability can probably be attributed to anatomical and radiographic factors. The Nasion, representing the junction of the frontal and nasal bones, is often a gentle curve rather than a distinct point, and its radiographic visibility can be obscured by the overlying soft tissues, making it inherently difficult for both human experts and AI to locate with high precision. Similarly, the A-point, the deepest point in the concavity of the anterior maxilla, is highly dependent on the position of the upper incisor roots and alveolar bone morphology. Its subtle curvature and anatomical variability among patients pose a significant challenge for automated detection, leading to a higher rate of false positives and lower predictive value.

To further evaluate AI Ceph's performance, a validation approach was employed using a dataset of cephalometric radiographs containing both correctly and incorrectly labelled landmarks. Of the 66 landmarks assessed, 33 were accurately labelled, and 33 were deliberately mislabelled. The majority of landmarks demonstrated sensitivity levels exceeding 80%; however, two landmarks fell below this threshold: Nasion (72.7%) and A-point (78.8%). Gonion achieved perfect sensitivity (100%), indicating flawless identification across all cases. When comparing concordance percentages before and after additional training with mislabelled data, improvements were observed across several landmarks, although the increase was statistically significant only for the Glabella point. This suggests that AI performance in landmark detection can be enhanced through continued training, reflecting the adaptive nature of DL systems.

The statistically significant improvement for the Glabella landmark after training with mislabelled data is a particularly noteworthy finding. This suggests that the AI model possesses a degree of robustness, allowing it to learn the general location and features of a well-defined anatomical structure even from imperfect or "noisy" datasets. The clinical implication of this is significant; it suggests that future AI training protocols may not require perfectly annotated datasets, which are time-consuming and expensive to create. Instead, models could potentially be trained effectively on larger, real-world clinical archives that contain minor inconsistencies, thereby accelerating the development and deployment of AI tools in orthodontics. This adaptability is crucial for creating algorithms that can perform reliably in diverse clinical settings.

The mislabelled landmark experiment should be interpreted as a controlled robustness analysis rather than evidence that inaccurate annotations are desirable. While the observed improvement for the Glabella landmark suggests a degree of resilience to minor annotation noise, this finding does not negate the importance of high-quality expert labelling. Instead, it highlights the potential for DL systems to tolerate limited real-world variability in annotation quality when trained on sufficiently large and diverse datasets.

While several studies report mean radial errors and SDRs at multiple thresholds, the present investigation prioritised a clinically interpretable binary tolerance (≤ 2.0 mm) supported by CI estimation and image-level resampling. This choice reflects common orthodontic decision-making thresholds and reduces overinterpretation of small numerical differences that may not translate into clinical relevance. Future work with larger external

test cohorts will enable more granular error distribution modelling and landmark-specific threshold optimisation.

This AI study has several limitations. Technically, the use of lateral cephalograms introduces projection and identification errors due to geometric distortion and challenges in locating certain landmarks. Anatomical variability further complicates accurate landmark detection. Additionally, the "black-box" nature of DL models makes their decision-making processes difficult to interpret, limiting transparency and trust in AI outputs. Another limitation is that all images were acquired using a single imaging device, which may limit the applicability of the findings to other equipment or clinical settings. Moreover, the use of JPG files rather than higher-fidelity digital imaging and communications in medicine images may have affected landmark detection accuracy. Finally, evolving regulatory and legal frameworks pose challenges, as clear guidelines are still needed to ensure the responsible and ethical use of AI in clinical settings. Furthermore, the study lacked external validation, and the independent test set consisted of only 33 radiographs, so the findings should be interpreted cautiously. Future studies should assess the model on larger and more diverse external datasets to determine broader generalisability.

Conclusions

AI is transforming orthodontics, offering promising tools to enhance diagnostic precision and clinical efficiency. In this study, the AI-based system demonstrated a high level of concordance with manual cephalometric landmark identification, reflecting its potential utility in routine orthodontic workflows. Notably, the model demonstrated the ability to learn from imperfect data, with a statistically significant improvement in identifying the Glabella landmark after retraining. While most discrepancies in landmark placement were within the clinically acceptable threshold of 2 mm, a small subset exceeded this limit and were explicitly categorised as discordant (> 2 mm) according to the predefined criterion, ensuring consistent and transparent reporting. This distinction clarifies the application of the binary concordance criterion and underscores the model's overall reliability. However, these findings should be interpreted as preliminary and exploratory, given the study's inherent limitations, including the use of a single imaging device, JPG image format, and a relatively small independent test set without external validation. Therefore, the generalisability of the model to other clinical settings, imaging systems, and populations remains limited. Despite these encouraging results, AI-assisted cephalometric analysis should not yet be considered a full replacement for expert human tracing. The involvement of a trained clinician remains essential to validate and, when necessary, adjust landmark positions to ensure diagnostic accuracy. Further validation using larger, externally sourced datasets is required before routine clinical implementation can be recommended. As AI technologies continue to evolve, ongoing validation and human oversight will be critical for their safe and effective integration into orthodontic practice.

Disclosures

1. Institutional review board statement: This retrospective diagnostic accuracy study was approved by the Institutional Review Board of Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU IRB-2023-46), and it adhered to the principles outlined in the Declaration of Helsinki.
2. Assistance with the article: None.
3. Financial support and sponsorship: None.
4. Conflicts of interest: None.

References

1. Allareddy V, Rengasamy Venugopalan S, Nalliah RP, Caplin JL, Lee MK, Allareddy V. Orthodontics in the era of big data analytics. Orthod Craniofac Res 2019; 22: 8-13.
2. Ahmad Z, Rahim S, Zubair M, Abdul-Ghafar J. Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review. Diagn Pathol 2021; 16: 24. DOI: 10.1186/s13000-021-01085-4.
3. Albarakati S, Kula K, Ghoneima A. The reliability and reproducibility of cephalometric measurements: a comparison of conventional and digital methods. Dentomaxillofac Radiol 2012; 41: 11-17.
4. Alessandri-Bonetti A, Sangalli L, Salerno M, Gallenzi P. Reliability of artificial intelligence-assisted cephalometric analysis. A pilot study. BioMedInformatics 2023; 3: 44-53.
5. Ali A, Chandna AK, Munjal A. Accuracy and reliability of soft tissue landmarks using three-dimensional imaging in comparison with two-dimensional cephalometrics: a systematic review. J Indian Orthod Soc 2020; 54: 289-296.
6. Alqahtani J, Alhemaid G, Alqahtani H, Abughandar A, AlSaadi R, Algarni I, et al. Digital diagnostics and orthodontic practice. J Healthc Sci 2022; 2: 112-117.
7. Alsubai S. A critical review on the 3D cephalometric analysis using machine learning. Computers 2022; 11: 154. DOI: 10.3390/computers11110154.
8. Bahaa K, Noor G, Yousif Y. The artificial intelligence approach for diagnosis, treatment and modelling in orthodontic. In: Naretto S, editor. Principles in contemporary orthodontics. London: InTech; 2011, p. 451-492.
9. Cunningham C, Scheuer L, Black S. Developmental juvenile osteology. 2nd ed. London: Academic Press; 2016.

10. Mohan A, Sivakumar A, Nalabothu P. Evaluation of accuracy and reliability of OneCeph digital cephalometric analysis in comparison with manual cephalometric analysis – a cross-sectional study. BDJ Open 2021; 7: 22. DOI: 10.1038/s41405-021-00077-2.
11. Jayachandran S. Digital imaging in dentistry: a review. Contemp Clin Dent 2017; 8: 193-194.
12. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep 2014; 16: 441. DOI: 10.1007/s11886-013-0441-8.
13. Ahmed N, Abbasi MS, Zuberi F, Qamar W, Halim MSB, Maqsood A, Alam MK. Artificial intelligence techniques: analysis, application, and outcome in dentistry – a systematic review. Biomed Res Int 2021; 2021: 9751564. DOI: 10.1155/2021/9751564.
14. Gill SS, Xu M, Ottaviani C, Patros P, Bahsoon R, Shaghaghi A, et al. AI for next generation computing: emerging trends and future directions. Internet Things 2022; 19: 100514. DOI: 10.1016/j.iot.2022.100514.
15. Liu J, Zhang C, Shan Z. Application of artificial intelligence in orthodontics: current state and future perspectives. Healthcare (Basel) 2023; 11: 2760. DOI: 10.3390/healthcare11202760.
16. Kamoen A, Dermaut L, Verbeeck R. The clinical significance of error measurement in the interpretation of treatment results. Eur J Orthod 2001; 23: 569-578.
17. Chen SK, Chen YJ, Yao CCJ, Chang HF. Enhanced speed and precision of measurement in a computer-assisted digital cephalometric analysis system. Angle Orthod 2004; 74: 501-507.
18. Erkan M, Gurel HG, Nur M, Demirel B. Reliability of four different computerized cephalometric analysis programs. EurJ Orthod 2012; 34: 318-321.
19. Lindner C, Wang CW, Huang CT, Li CH, Chang SW, Cootes TF. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. Sci Rep 2016; 6: 33581. DOI: 10.1038/srep33581.
20. Uysal T, Baysal A, Yagci A. Evaluation of speed, repeatability, and reproducibility of digital radiography with manual versus computer-assisted cephalometric analyses. European J Orthod 2009; 31: 523-528.
21. Hwang HW, Park JH, Moon JH, Yu Y, Kim H, Her SB, et al. Automated identification of cephalometric landmarks: Part 2 – Might it be better than human? Angle Orthod 2020; 90: 69-76.
22. Meriç P, Naoumova J. Web-based fully automated cephalometric analysis: comparisons between app-aided, computerized, and manual tracings. Turk J Orthod 2020; 33: 142-149.
23. Bulatova G, Kusnoto B, Grace V, Tsay TP, Avenetti DM, Sanchez FJC. Assessment of automatic cephalometric landmark identification using artificial intelligence. Orthod Craniofac Res 2021; 24: 37-42.