RESEARCH PAPER

Comparison of the Efficacy of Traditional, Semi-Automated and Automated (AI-Based) Analysis of Cephalometric Radiographs in Diagnosis and Treatment Planning of Dental Malocclusion: A Cross-Sectional Comparative Study

Mohamed Ali Eldsoqi[1], Mostafa Mohamed Hosny[2], Layla Abdel-Fattah Amer[3], Ahmed Mohamed Bilal[4] and Ahmed Akram EL-Awady[5]*

[1]*Demonstrator, Oral Medicine and Periodontology Department Radiology Division, Faculty of Dentistry MTI University, Egypt*

[1]*Mohamed.Al-Dosouki@dnt.mti.edu.eg, 2mostafahosny.209@azhar.edu.eg, 3layla.amer@dnt.mti.edu.eg, 4ahmedbilal.209@azhar.edu.eg and 5ahmedelawady.209@azhar.edu.eg*

[2]*Professor Department of Oral Medicine, Periodontology, Diagnosis and Oral Radiology, Faculty of Dental Medicine (Boys-Cairo), Al-Azhar University, Egypt*

[3]*Lecturer of Orthodontics, Department of Orthodontics, Faculty of MTI University, Egypt*

[4]*Lecturer of Oral Medicine, Periodontology, Diagnosis and Oral Radiology, Faculty of Dental Medicine (Boys-Cairo), Al-Azhar University, Egypt*

[5]*Assistant professor of Orthodontics, Department of Orthodontics, Faculty of Dental Medicine (Boys-Cairo), Al-Azhar University, Egypt*

Corresponding author: *ahmedelawady.209@azhar.edu.eg*

Received: 01st July, 2026; Revised: 11th July 2026; Accepted: 20th July, 2026; Available Online: 28th July, 2026

ABSTRACT

Background: Cephalometric analysis is clinical application of cephalometry. It is an analysis of the dental and skeletal relationships in a human skull and it is frequently used by orthodontists, oral and maxillofacial surgeons as a treatment-planning tool. Aim: study aims to establish the accuracy and reliability of an artificial intelligence-based cephalometric analysis method in comparison to digital and standard manual methods.

Study design: retrospective comparative study.

Methodology: A total of eighty-five lateral cephalometric radiographs with complete dentition. Subjects with previous orthodontic surgeries, subjects with intraoral appliances, and any bone deformities were excluded. Each lateral cephalogram was analyzed using manual tracing method and digitally by Dolphin imaging software, and by Cephalo.AI (voxel 3di) and Cephio AI-based programs. Eighteen (linear and angular) measurements of three different analyses were calculated and recorded.

Results: The statistical evaluation revealed significant differences ($P \leq 0.05$) across all three methods concerning the 18 assessed parameters. The semi-automated Dolphin software exhibited the greatest agreement with manual tracing, especially in the evaluation of dental and soft tissue measurements. While artificial intelligence applications (Cephalo.AI and Cephio) demonstrated near-perfect reproducibility (ICC reaching 1.0) for skeletal landmarks, they displayed clinically important variations in certain linear and soft tissue measurements.

Conclusion: AI-based softwares provide rapidity and high reproducibility; however, they require precise supervision for dental and soft tissue parameters assessment. The semi-automated model (Dolphin) emerged as the highest degree of similarity, as it had no significant difference in most parameters.

Keywords: *Cephalometric analysis, Cephalo.AI, voxel 3di, Cephio, Dolphin imaging, and artificial intelligence.*

How to cite this article: Eldsoqi MA, Hosny MM, Amer LA, Bilal AM, EL-Awady AA. Comparison of the Efficacy of Traditional, Semi-Automated and Automated (AI-Based) Analysis of Cephalometric Radiographs in Diagnosis and Treatment Planning of Dental Malocclusion: A Cross-Sectional Comparative Study. Int J Drug Deliv Technol. 2026;16(76s): 896-906. DOI: 10.25258/ijddt.16.76s.87

Source of support: Nil.

Conflict of interest: None

**Author for Correspondence: ahmedelawady.209@azhar.edu.eg*

INTRODUCTION

For decades, cephalometric analysis and lateral cephalograms have been the best ways to diagnose and plan orthodontic treatment. Broadbent first used them in 1931 to study dentofacial differences as well as skeletal and dental malocclusions, which are a type of developmental malformation that can range from mild to severe.1,2 These problems with dental occlusion and the temporomandibular joints cause the jaws to be in an unnatural position, which affects the facial balance. 3 Broadbent developed the imaging technique "cephalometry" in 1931 to be used to study growth, find malocclusions, make treatment plans, and evaluate the overall results.

There are two ways to do the cephalometric analysis that have been utilized for almost 20 years. The manual approach was the most common way to do the study. The most recent advancement in dentistry is the use of digital or computerized methods to record craniofacial landmarks and find the right measurements for different types of analyses. 1, 2 The most recent advancement in AI is the ability of a technology to mimic human intelligence or make decisions that are effective and ethical based on predetermined criteria.4, 5 Machine learning is a major part of artificial intelligence. It employs the statistical patterns of data that have already been learnt to make predictions about new data and situations. This is possible because of recent advances in technology. In orthodontics, the automatic AI recognition of cephalometric landmarks represents one of numerous initiatives to implement machine learning in diverse applications. 6-8

The conventional manual approach of cephalometric analysis is susceptible to both systematic and random mistakes, in addition to being time-consuming. It has frequently been noted that there is a deficiency in consistency and reproducibility. 9, 10 But the computerized method is easier, faster, more accurate, lets you share data, and can be used for standardized testing with better reproducibility. It can get eliminates a lot of the mistakes that happen when you draw lines between landmarks or measure by hand with a protractor. It also saves time by not having to scan, prepare, and store hard copies, which means less exposure to radiation. Unfortunately, it has some problems and mistakes, such as the quality of the radiograph, magnification correction, density adjustment, contrast, and image resolution. Also, most of these programs are costly 11, 12, so the goal was to replace these old-fashioned procedures with newer computerized methods and technology, as well as the use of AI in digital radiography and other fields of dentistry.

Cephalometric analysis that uses AI can be done either semi-automatically or totally automatically. The entirely automatic method employs AI to find landmarks, trace them, and figure out the cephalometric measurement. The semi-automated method, on the other hand, uses both methods: manual selection of landmarks and automatic calculation. 13, 14. Most of the research that tried to find out how accurate the cephalometric methods were and compared them to each other said that AI-based automatic cephalometric analyses give clinically acceptable diagnostic performance. They have demonstrated high accuracy and precision that is comparable to that of a qualified orthodontist. They are quite helpful for orthodontists since they let them complete their jobs more quickly. 15

Previous studies showed that cephalometric readings exhibit great reproducibility when utilizing Vistadent 2.1 AT, JOE software, and manual tracing. 16 CephX needs to be better and requires manual correction to work just as well as Ceph-Ninja and Dolphin, and is quicker. But automatic approaches still need to be checked by a person to make sure they are correct. 17 Web-Ceph had problems with finding and measuring landmarks consistently, although Dolphin's semi-automatic method helped fix these problems when used carefully. 13 Overall, even if the dependability of different platforms is uneven, continuous research, including this study, is trying to find the best software for accurate, efficient, and cheap orthodontic diagnosis and planning.

Sample size calculation:
According to a previous study by Meriç and Naoumova (2020) 13, the mean values varied between 4.8 ± 1.8, 5.1 ± 1.8, 6 ± 1.4, and 6.3 ± 1.6. Using G Power statistical power Analysis program (version 3.1.9.4) for sample size determination, A total sample size (n = 85) was sufficient to detect a medium effect size (f) = 0.36, with an actual power (1-β error) of 0.88 (88%) and a significance level (α error) of 0.05 (5%) for a two-sided hypothesis test.

Study design:
The study is a retrospective and comparative study.

Inclusion criteria:
Those who matched the following criteria were included in the study: all patients are aged from 14 to 35 years old and have full permanent dentition with no extraction of any permanent first molars, and all patients are free of acute or chronic periodontitis (horizontal bone loss).

Exclusion criteria:
Any previous orthognathic surgeries, subjects with intraoral appliances 18, patients with cleft lip or palate, or any other bone deformities are excluded. 19, 20

Materials and Methods:

- For each cephalogram, cephalometric analysis was done according to three different common cephalometric analyses (skeletal Down analysis, Eastman analysis, and Tweed analysis) for comparison of three cephalometric methods: conventional, semiautomated, and automated (AI-based) method by two different softwares. for image standardization, all lateral cephalometric radiographs were acquired using standardized exposure settings to ensure image quality and reproducibility. The acquisition parameters were 90.0 kV, 14.0 mA, 3.9 seconds, with magnification 1:1 (100 %scale with a metallic reference ruler). The cephalometric images were taken with the patient in the upright standing position with the Frankfort plane parallel to the floor, keeping the teeth in centric relation and the lips relaxed. All the lateral cephalometric radiographs were taken using the same lateral cephalometric machine. After sample selection, Calibration of the actual size of each image was done based on measurement of the known 10 mm distance) between the two fixed points on the ruler present on the analog as well as digital cephalograms displayed onscreen. A single examiner of 10 years experienced radiologist, performed the cephalometric tracings for all cephalograms manually, and the landmarks were checked by two different well-experienced orthodontic observers; then the second tracing by the same examiner was done at an interval of one month to avoid fatigue of the examiner to minimize recalling bias memory effect and insure blindness of previous readings. When differences in landmark identification or measurement occurred, the observers discussed the case together and reached a consensus. If consensus was not immediately achieved, a third senior orthodontist was consulted, and the final decision was based on agreement among at least two observers. The end results were estimated by taking the average between the two readings.

Ethical approval: This study was submitted to the ethics committee in the Faculty of Dental Medicine (Boys, Cairo), Al-Azhar University, and was ethically accepted with code 928\3557. A total of 85 cephalometric radiographs of patients attending the oral medicine, periodontology, diagnosis, and oral radiology department clinic, Faculty of Dental Medicine (Boys, Cairo), Al-Azhar University.

Manual method:

Standardized digital lateral cephalometric radiographs were filmed with the same X-ray machine. The lateral cephalometric radiographs were exported in JPG format and printed using high-definition prints of digital cephalometric radiographs. The procedure involves using an overlay of an acetate tracing paper sheet of 210 mm x 160mm secured over the radiograph by Adhesive tape seated on a view-box, followed by identification of 12 landmarks. Finally, the lines are drawn with an HB 0.3 mm lead pencil and a 200 mm ruler, then linear and angular measurements (4 linear and 14 angular) of the three used analyses were measured and recorded into an Excel sheet. (Figure 1-A)

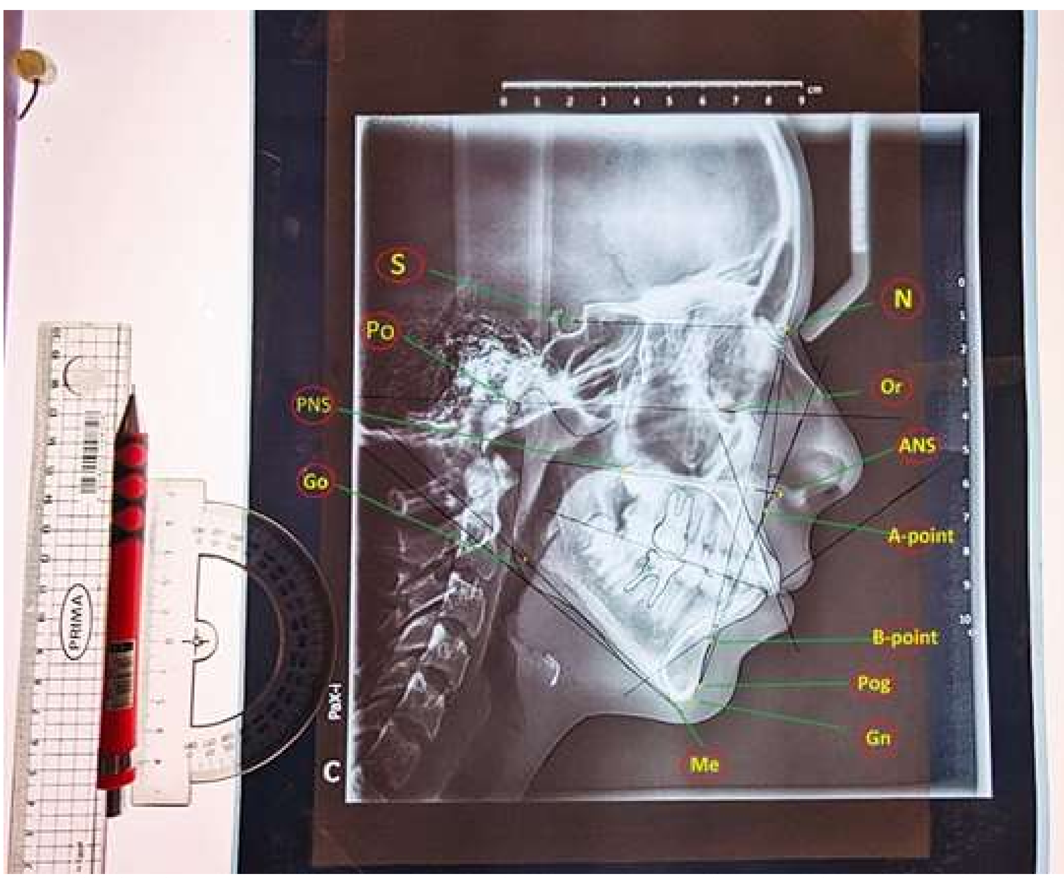


Fig. (1-A): 12 landmarks are plotted, lines are drawn with an HB 0.3 mm lead pencil and a 200 mm ruler, then linear and angular measurements (4 linear and 14 angular) of the three used analyses were measured.

Semiautomated method:

combination of manual and automatic methods. In this procedure after importing the personal data of the patient, the digital radiograph JPG image was imported into computerized software Dolphin imaging-2D module Dolphin imaging system, Canada U.S.A (version 11. 5. 04. 36), next step was Correction of any magnification using a 10mm distance between two points on the cephalostat arm in the cephalogram, then the landmarks including bony and soft tissue structures were identified by the arrow manually by the operator (figure 1-B). Whereas the linear and angular measurements of cephalometric parameters used in the study were calculated automatically by software, then recorded into the Excel sheet.

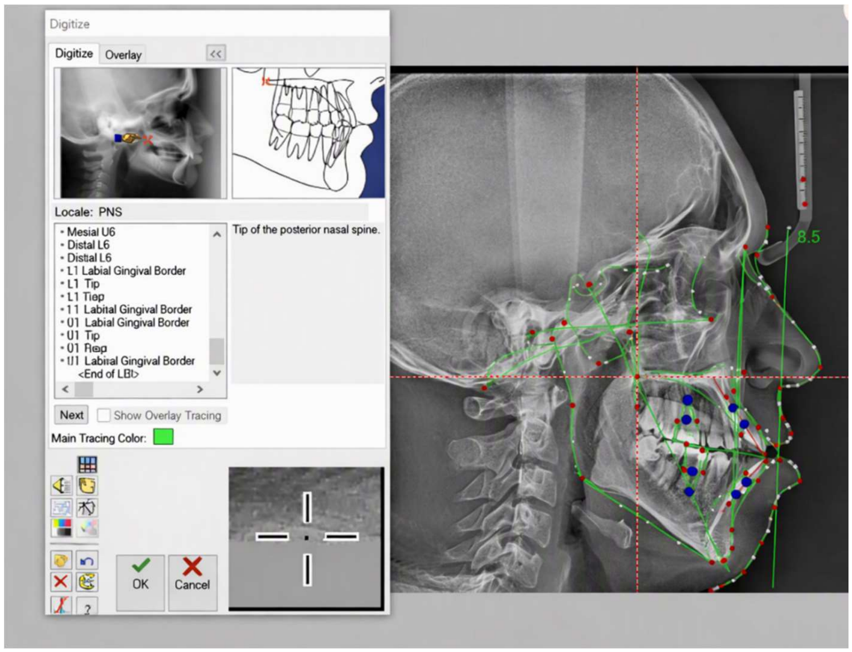
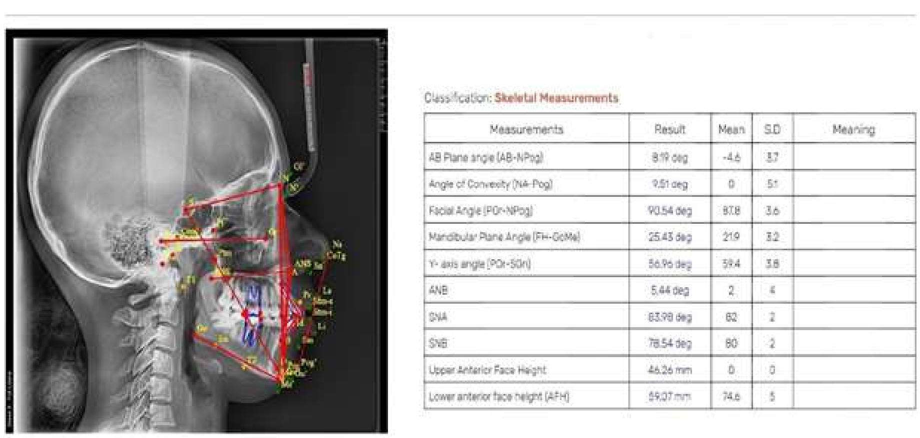


Fig. (1-B): Calibration of the image on ruler scale, the landmarks including bony and soft tissue structures were identified by the arrow manually by the operator using Dolphin software.

AI-based method

AI-driven softwares: (Cephalo.AI.voxel.3di) and Cephio programs were used for cephalometric analysis. After importing the digital radiograph into the computer, a new patient file was generated, and a lateral cephalogram JPG image was imported into the program. Correction of any magnification was done using a 10mm distance between two points on the cephalostat arm in the cephalogram. After calibration of the images, all landmarks needed for three analyses (skeletal down analysis, Eastman analysis, and Tweed analysis) were detected using AI option in each software automatically. After this step, the results were obtained in pdf format and recorded in the Excel sheet. (Figure 1-C).


Classification: Skeletal Measurements

Measurements	Result	Mean	S.D	Meaning
AB Plane angle (AB-NPog)	8.19 deg	-4.6	3.7	
Angle of Convexity (NA-Pog)	9.51 deg	0	5.1	
Facial Angle (POr-NPog)	90.54 deg	87.8	3.6	
Mandibular Plane Angle (FH-GoMe)	25.43 deg	21.9	3.2	
Y- axis angle (POr-SGn)	56.96 deg	59.4	3.8	
ANB	5.44 deg	2	4	
SNA	83.98 deg	82	2	
SNB	78.54 deg	80	2	
Upper Anterior Face Height	46.26 mm	0	0	
Lower anterior face height (AFH)	59.07 mm	74.6	5	

Fig. (1-C): the end result of chosen analyses or specific parameters that cephalo.AI (voxel) shows in PDF format

DESCRIPTIVE STATISTICS

The study sample included 85 patients with an average age of 23.29 years in the range 15-35 years. Of these, 28 (32.94%) were males, and 57 (67.05%) were females. We measured 18 different Parameters as follows: 5 parameters according to Down analysis (skeletal), 3 parameters according to Tweed analysis, and 10 parameters according to Eastman analysis to determine the presence of statistically significant differences between the four methods.

Numerical data were explored for normality by checking the distribution of data and using tests of normality (Kolmogorov-Smirnov and Shapiro-Wilk tests). Data were presented as mean, standard deviation (SD), median, and range values. For parametric data, a repeated measures ANOVA test was used to compare methods. Bonferroni's post-hoc test was used for pairwise comparisons when the ANOVA test was significant. For non-parametric data, Friedman's test was used to compare between methods. Dunn's test was used for pair-wise comparisons when Friedman's test is significant. the effect sizes in the Statistical Analysis section Specifically, Partial Eta Squared for parametric ANOVA tests was classified as small (0.01), medium (0.06), and large (0.14) according to Cohen (1988), while Kendall's W for non-parametric tests was interpreted based on Tomczak & Tomczak (2014) benchmarks. Agreement between the methods, as well as intra-observer agreement, was assessed using the Intra-Class Correlation Coefficient (ICC). Values of ICC above 0.8 indicate very good agreement. The significance level was set at $P \leq 0.05$. Statistical analysis was performed with IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.

RESULTS

The statistical results comparing 18 orthodontic measurements across four methods; Manual, Dolphin, Cephalo.AI (Voxel), and Cephio, indicate significant differences between the methods ($P \leq 0.05$) for all parameters. All data showed normal (parametric) distribution except for angle of convexity, AB to facial plane, ANB, Lower incisor -APog, and Lower lip to E-plane measurements, which showed non-normal (non-parametric) distribution. The reported effect sizes demonstrate that the statistically significant differences observed between platforms carry meaningful effect magnitudes (Table 1). However, the proximity to the "Manual" gold standard varied by measurement, as shown in Table 2.

Dolphin software had no significant difference with the manual method in Facial angle, Frankfort Mandibular Angle (FMA), Y-Axis Angle, SNB, Inter-incisal angle, Lower incisor -Mandibular Plane (Li-MP), upper incisor - Maxillary plane (I-Max plane), Li-APog, Upper facial height, Lower lip to E-plane, FMA (of Tweed), Incisor Mandibular Plane Angle (IMPA), and Frankfort Mandibular Incisor Angle (FMIA).

While Cephalo.AI (Voxel) agreed with the manual method in the following angles of convexity, AB to facial plane, ANB, Li-MP, and IMPA, finally, Cephio software had no significant difference with the manual method in FMA, SNA, SNB, I-Max plane, Li-APog, Upper facial height, Lower facial height, and FMA (of tweed).

According to the Intra-Class Correlation Coefficient (ICC), ICC values above 0.8 indicate very good agreement. there was very good intra-observer agreement with all measurements indicating reliable reproducibility except for the following: angle of convexity in manual method, SNA in manual method, Interincisal angle in manual method, lower incisor – APog in manual method, Lower lip -E plane in manual method, Incisor mandibular plane angle in manual method, Interincisal angle in Dolphin software, lower incisor -mandibular plane in Dolphin software, Lower lip- E plane in Voxel software, Lower lip-E plane in Cephio software. (Table 3).

Table 1: Descriptive statistics and results of repeated measures ANOVA test and Friedman's test for comparison between measurements by the four methods

Measurement	Manual		Dolphin		Voxel		Cephio		*P*-value	*Effect size (Partial Eta Squared)*
	Mean	SD	Mean	SD	Mean	SD	Mean	SD		
Facial angle (°)	87.05 ^{A}	3.64	87.02 ^{A}	3.47	87.61 ^{B}	3.42	88.07 ^{C}	3.42	<0.001*	0.155
FMA (°)	28.15 ^{A}	5.54	28.62 ^{A}	5.1	25.63 ^{B}	5.89	28.34 ^{A}	5.87	<0.001*	0.37
S-GN (°)	60.56 ^{A}	4.19	60.93 ^{A}	4.21	59.76 ^{B}	4.14	70.28 ^{C}	5.01	<0.001*	0.87
SNA (°)	81.15 ^{A}	4.53	82.38 ^{C}	4.89	80.64 ^{B}	4.27	81.27 ^{A}	4.45	<0.001*	0.224
SNB (°)	77.77 ^{A}	4.37	78.13 ^{A}	4.65	76.94 ^{B}	3.98	77.38 ^{A}	4.21	<0.001*	0.152
Inter-incisal angle (°)	115.65 ^{A}	10.81	115.9 ^{A}	11.93	111.1 ^{B}	16.29	119.37 ^{C}	10.97	<0.001*	0.177
Li-MP (mm)	97.01 ^{A}	6.98	96.76 ^{A}	7.38	97.68 ^{A}	7.27	93.79 ^{B}	12.04	<0.001*	0.085
I-Max Plane (°)	117.19 ^{A}	6.48	117.75 ^{A}	6.46	119.75 ^{B}	9.15	116.64 ^{A}	6.83	<0.001*	0.088
Upper facial height (mm)	49.52 ^{A}	3.74	49.44 ^{A}	4.21	50.13 ^{B}	4.04	50.05 ^{A}	3.74	0.009*	0.045
Lower facial height (mm)	65.61 ^{A}	6.57	66.97 ^{B}	6.98	62.87 ^{C}	7.38	65.71 ^{A}	6.51	<0.001*	0.478
FMA (°)	29.26 ^{A}	5.72	28.63 ^{A}	5.09	25.32 ^{B}	6.51	28.28 ^{A}	5.77	<0.001*	0.403
IMPA (°)	96.05 ^{A}	7.05	96.7 ^{A}	7.34	97.22 ^{A}	8.64	94.85 ^{B}	7.62	<0.001*	0.083
FMIA (°)	54.56 ^{A}	8.12	54.67 ^{A}	8.67	56.01 ^{B}	10.18	56.75 ^{B}	8.39	<0.001*	0.070
	Median (Range)	Mean (SD)	Median (Range)	Mean (SD)	Median (Range)	Mean (SD)	Median (Range)	Mean (SD)	*P*-value	*Effect size (w)*
Angle of convexity (°)	6.5 (-17, 23) ^{A}	6.16 (7.35)	9.1 (-16.7, 22.2) ^{B}	7.64 (7.86)	7.11 (-14.98, 22.99) ^{A}	6.71 (7.27)	8.32 (-12.7, 24.68) ^{B}	7.4 (7.34)	<0.001*	0.149
AB to facial plane (mm)	-6 (-15, 8) ^{A}	-4.96 (4.83)	-6.7 (-18.1, 9.3) ^{B}	-5.52 (5.2)	-6.15 (-14.64, 9.1) ^{A}	-4.89 (4.95)	-6.53 (-16.3, 6.5) ^{B}	-5.92 (4.5)	0.002*	0.057
ANB (°)	4 (-7, 10) ^{A}	3.36 (3.26)	4.7 (-7.2, 10.1) ^{B}	4.19 (3.59)	3.9 (-6.41, 10.05) ^{A}	3.61 (3.18)	4.06 (-4.8, 10.42) ^{B}	3.88 (3.12)	<0.001*	0.126
Li-A Pog (mm)	5 (0, 16) ^{A}	5 (3.19)	4.9 (-2.4, 15.8) ^{A}	4.88 (3.56)	5.32 (0.19, 16.32) ^{B}	5.48 (3.13)	4.8 (-1.67, 16.06) ^{A}	4.77 (3.26)	<0.001*	0.093
Lower lip to E-plane (mm)	2 (-3, 15) ^{A}	2.61 (2.98)	1.9 (-5.5, 15.8) ^{A}	2.28 (3.69)	1.15 (-5.52, 10.65) ^{B}	1.69 (3.15)	1.6 (-5.21, 8.74) ^{B}	1.65 (2.88)	<0.001*	0.166

**: Significant at $P \leq 0.05$, Different superscripts in the same row indicate statistically significant difference between methods*

Table 2: The level of agreement and proximity to the "Manual" gold standard varied by measurement based on the Intra-Class Correlation Coefficient (ICC) values. ICC above 0.8 indicates very good agreement

Measurement	Accurate Method(s) (No Sig. Difference from Manual)	Less Accurate Method(s) (Significant Difference from Manual)	Agreement Strength (ICC) and Notes
1. Facial angle (º)	Dolphin	Voxel & Cephio (Higher)	Very Good (>0.8) for all.
2. Angle of convexity (º)	Voxel	Dolphin & Cephio (Higher)	Very Good (>0.8) for all.
3. AB to facial plane (º)	Voxel	Dolphin & Cephio (Higher)	Very Good (>0.8) for all.
4. FMA (º)	Dolphin & Cephio	Voxel (Lower)	Very Good (>0.8) for all.
5.Y-Axis(º) (S-Gn)	Dolphin	Voxel (Lower) & Cephio (Higher)	Very Good for Dolphin/Voxel; Acceptable for Cephio (0.738).
6. SNA (º)	Cephio	Dolphin (Higher) & Voxel (Lower)	Very Good (>0.8) for all.
7. SNB (º)	Dolphin & Cephio	Voxel (Lower)	Very Good (>0.8) for all.
8. ANB (º)	Voxel	Dolphin & Cephio (Higher)	Very Good (>0.8) for all.
9. Inter-incisal angle (º)	Dolphin	Voxel (Lower) & Cephio (Higher)	Very Good (Dolphin/Cephio); Moderate (Voxel 0.543).
10. Li-MP (mm)	Dolphin & Voxel	Cephio (Lower)	Very Good (Dolphin/Voxel); Below Average (Cephio 0.411).
11. I-Max plane (º)	Dolphin & Cephio	Voxel (Higher)	Very Good (Dolphin); Moderate (Voxel/Cephio).
12. Li-A Pog (mm)	Dolphin & Cephio	Voxel (Higher)	Very Good (>0.8) for all.
13. Upper facial height (mm)	Dolphin & Cephio	Voxel (Higher)	Very Good (>0.8) for all.

14. Lower facial height (mm)	Cephio	Dolphin (Higher) & Voxel (Lower)	Very Good (>0.8) for all.
15. Lower lip to E-plane (mm)	Dolphin	Voxel & Cephio (Lower)	Very Good (Dolphin/Cephio); Good/Borderline (Voxel 0.797).
16. FMA (º)	Dolphin & Cephio	Voxel (Lower)	Very Good (>0.8) for all.
17. IMPA (º)	Dolphin & Voxel	Cephio (Lower)	Very Good (Dolphin/Cephio); Good (Voxel 0.763).
18. FMIA (º)	Dolphin	Voxel & Cephio (Higher)	Very Good (Dolphin/Cephio); Good (Voxel 0.748

Table 3: Results of Intra-Class Correlation Coefficient (ICC) and its 95% Confidence Interval (95% CI) for intra-observer agreement regarding all measurements

Measurement	ICC	95% CI Lower	95% CI Upper	Measurement	ICC	95% CI Lower	95% CI Upper
Facial Angle (FA) in Manual	0.921	0.716	0.98	Lower Incisor - Mandibular Plane (LI-MP) in Manual	0.85	0.507	0.96
FA in Dolphin	0.851	0.51	0.961	LIMP in Dolphin	0.783	0.341	0.941
FA in Voxel	0.997	0.968	0.999	LIMP in Voxel	0.999	0.994	1
FA in Cephio	0.999	0.997	1	LIMP in Cephio	0.999	0.998	1
Angle of Convexity (AC) in Manual	0.514	-0.127	0.852	Incisor -Maxillary plane (I-MAX) in Manual	0.843	0.488	0.958
AC in Dolphin	0.962	0.856	0.99	I-MAX in Dolphin	0.878	0.585	0.968
AC in Voxel	1	0.998	1	I-MAX in Voxel	0.997	0.988	0.999
AC in Cephio	0.999	0.998	1	I-MAX in Cephio	0.999	0.998	1
AB -Facial Plane in Manual	0.917	0.702	0.979	Lower Incisor - Apog (LI-Apog) in Manual	0.698	0.166	0.916
AB -Facial Plane in Dolphin	0.931	0.749	0.982	LI-Apog in Dolphin	0.722	0.213	0.923
AB -Facial Plane in Voxel	1	0.999	1	LI-Apog in Voxel	0.972	0.891	0.993
AB -Facial Plane in Cephio	0.999	0.996	1	LI-Apog in Cephio	0.994	0.976	0.998
FMA in Manual	0.92	0.713	0.98	Upper facial height in Manual	0.945	0.797	0.986
FMA in Dolphin	0.948	0.807	0.987	Upper facial height in Dolphin	0.938	0.771	0.984
FMA in Voxel	0.998	0.994	1	Upper facial height in Voxel	0.995	0.979	0.999
FMA in Cephio	0.999	0.996	1	Upper facial height in Cephio	0.995	0.981	0.999
Y axis angle in Manual	0.883	0.601	0.97	Lower facial height in Manual	0.966	0.869	0.991
Y axis angle in Dolphin	0.865	0.55	0.965	Lower facial height in Dolphin	0.951	0.816	0.988
Y axis angle in Voxel	0.998	0.99	0.999	Lower facial height in Voxel	0.999	0.997	1
Y axis angle in Cephio	0.997	0.99	0.999	Lower facial height in Cephio	0.999	0.997	1
SNA in Manual	0.793	0.367	0.944	Lower lip-E Plane in Manual	0.749	0.267	0.931

SNA in Dolphin	0.854	0.518	0.962	Lower lip-E Plane in Dolphin	0.802	0.387	0.947
SNA in Voxel	0.998	0.992	0.999	Lower lip-E Plane in Voxel	0.447	-0.212	0.827
SNA in Cephio	0.995	0.981	0.999	Lower lip-E Plane in Cephio	0.628	0.042	0.893
SNB in Manual	0.959	0.843	0.99	FMA (tweed) in Manual	0.804	0.392	0.948
SNB in Dolphin	0.984	0.939	0.996	FMA in Dolphin	0.942	0.784	0.985
SNB in Voxel	0.996	0.985	0.999	FMA in Voxel	0.998	0.994	1
SNB in Cephio	0.997	0.989	0.999	FMA in Cephio	1	1	1
ANB in Manual	0.869	0.56	0.966	Incisor mandibular plane angle (IMPA) in Manual	0.743	0.256	0.929
ANB in Dolphin	0.924	0.724	0.98	IMPA in Dolphin	0.845	0.493	0.959
ANB in Voxel	0.998	0.992	1	IMPA in Voxel	0.999	0.995	1
ANB in Cephio	0.999	0.996	1	IMPA in Cephio	1	0.998	1
Interincisal angle (IN) in Manual	0.714	0.196	0.92	Frankfort mandibular plane angle (FMIA) in Manual	0.912	0.688	0.977
IN in Dolphin	0.678	0.128	0.909	FMIA in Dolphin	0.9	0.652	0.974
IN in Voxel	1	0.999	1	FMIA in Voxel	1	1	1
IN in Cephio	0.999	0.996	1	FMIA in Cephio	0.999	0.997	1

DISCUSSION

Digital methods for lateral cephalometric analysis are widely used. Before using any digital tools on the market, the most important thing is that tracing must be accurate, no matter if the software is computer-based or web-based. Consequently, this study evaluated the precision of two commercially available AI-driven algorithms for lateral cephalometric analysis against the traditional cephalometric method. [21] the authors state that manual method is widely utilized in orthodontics, despite the technique is time-consuming and has the disadvantage of being subject to significant potential liability of errors, remains the predominant technique for tracing, identifying landmarks, and measuring distances and angles between landmark locations [22], statistical analysis showed that there were significant differences ($P \leq 0.05$) between all 18 parameters when we compared manual, semi-automated (Dolphin), and fully automated (Cephio and Voxel) methods. For instance, Cephio showed that it was in agreement with manual tracing for the SNA angle, which was (81.27 ± 4.45) and SNB angle (77.38 ± 4.21) mean and standard deviation, which means it is very reliable for detecting cranial base landmarks.

In contrast, Voxel matched the manual method (3.36 ± 3.26) only for the ANB angle (3.61 ± 3.18). While Dolphin, on the other hand, recorded higher SNA values (82.38 ± 4.89) than the manual measurements (81.15 ± 4.53), these discrepancies, are supported by Saifeldin's (2023) observations [21], who suggests that it is frequently arising from the anatomical difficulty in accurately identifying A-Points and B- Points, which are low-density landmarks that lack clarity on radiographic images, as well as midfacial concavity overlap at Points A and B, this challenge also observed by Oltramari-Navarro et al. (2013).[23]

In terms of vertical skeletal relationships, Dolphin and Cephio both had agreement with manual tracing for the Frankfort Mandibular Angle (FMA). Dolphin also showed a high level of alignment on Y-axis (S-GN) measurements, with an ICC greater than 0.8. The most surprising finding was that Cephio's S-GN mean value of 70.28° was much higher than the manual mean value of 60.56°. This 10-degree difference suggests that there may be a systematic error in finding the Gnathion or Sella, unlike its acceptable results given in FMA angle, Tsolakis et al. (2022) and Bao et al. (2023) propose that these discrepancies arise from the challenges AI systems face in recognizing landmarks on curved anatomical boundaries or due to smooth contour of the mandibular symphysis at Gnathion which directly accounts for the shift observed in the Y-axis. [4, 18] As for Linear facial height measurements, they revealed additional discrepancies: Cephio agreed with the manual tracing for lower facial height (65.71 ± 6.51), but Dolphin recorded higher values (66.97 ± 6.98), and Voxel recorded much lower values (62.87 ± 7.38). Voxel also had statistically higher mean values for upper facial height than any other method. Davidovitch et al. (2022) recorded similar discrepancies, observing that digital software frequently has difficulty in accurately identifying the Anterior Nasal Spine (ANS). [24] As for the dental parameters, AI had the most trouble because of overlapping shadows and low radiopacity. Dolphin aligned with manual tracing for the inter-incisal angle (115.65 ± 10.81), whereas Voxel offered lower values (111.1 ± 16.29) and Cephio produced higher values (119.37 ± 10.97). Farooq et al. (2016) and Lalakiya et al. (2016) observed that manual tracing frequently produces more logical mean values, as human operators can utilize anatomical knowledge to reduce image noise [9, 25], this is why Dolphin, which allows operator to make adjustments, is more accurate for dental parameters than AI systems that are fully automated. Dolphin and Voxel both agreed with manual tracing (96.05 ± 7.05) for the Incisor Mandibular Plane Angle (IMPA).

However, Cephio recorded much lower values (94.85 ± 7.62), which suggests that its algorithm has difficulty with the long axis of lower incisors when there is crowding and superimposition. the analysis of soft tissue further showed the limitation of fully automated systems; the measurement from the lower lip to the E-plane indicated that Dolphin (2.28 ± 3.69) was the only method showing no significant difference from manual tracing (2.61 ± 2.98), while Voxel (1.69 ± 3.15) and Cephio (1.65 ± 2.88) did not aligned with the manual gold standard, these findings were strongly agreed with Zughair et al(2025).[26]As for the assessment of the reliability of the results, a difference is evident with regard to the comparison between the readings of the human and AI-algorithms.

Notably, AI technologies such as Voxel and Cephio came close to attaining high efficiency with their reproducibility results. The ICCs of the skeletal parameters, such as Frankfort Mandible Angle (FMA), were 0.999 or even 1.0. Previous studies by Bao et al. (2023) and Jeon & Lee (2021) showed that AI-based analyses provide a precision that mimics or even exceeds that of a professional orthodontist, and could standardize criteria that are historically prone to human error.[4, 11] also, the study showed variations in the manual method results, especially with the Angle of Convexity (ICC = 0.514) and the inter-incisal angle (0.714). According to Table 3, however, the efficiency of AI wasn't enough for soft tissue characteristics. These findings highlight that AI precision is strongly dictated by image contrast and structural complexity: high-contrast cortical boundaries (such as Sella and Nasion in FMA) yield near-perfect reliability (ICC = 0.999-1.0), whereas dental root overlap and gradual soft-tissue density gradients introduce algorithm drift unless refined by clinician oversight.

The results of the study also revealed a decline in the intra-observer agreement to 0.447 using Voxel and 0.628 using Cephio, with the objective of measuring the lower lip with reference to the E-plane. While the manual method showed moderate agreement at 0.749, it means that AI is good at high-contrast bony boundaries but not so good at soft tissue gradients and linear measurements that need to be calibrated by hand. Meriç & Naoumova (2020) noted that web-based AI tends towards decreased efficiency concerning the detection of soft tissue landmarks compared to the manual method using Dolphin. [13]. It must be mentioned that factors influencing AI performance across different measurements, such as anatomical landmark clarity and contrast, structural overlap, low radiopacity and tissue density and radiographic quality variations, should be taken into consideration.

In interpreting cephalometric research, a clear distinction must be made between statistical significance and clinical relevance. Although statistical analysis revealed significant differences (P value ≤ 0.05) across the evaluated parameters, many of these mean discrepancies are clinically negligible. In orthodontic literature, variations of less than 2.0 degrees for angular parameters and less than 1.5mm for linear measurements are generally considered clinically acceptable.

The "gold standard" is no longer the human but a hybrid of the two, with the AI system providing reliability of reproducibility in the results and the human that providing the precise supervision and adjustment required for complex diagnostic parameter, particularly in soft-tissue measurements. Gürler (2026).[27] The present study had limitations of its retrospective single-center design which limits generalization, the sample size was statistically adequate, but broader inclusion of malocclusion types would strengthen results. The evaluation was restricted to a specific set of AI platforms, while newer algorithms may perform differently. Automated accuracy also depends on radiographic image quality, which can vary with patient positioning and contrast. Finally, the study evaluated landmark accuracy, leaving future work to assess processing speed and diagnostic agreement among clinicians.

Future scope: AI algorithms need better accuracy for identifying dental and soft-tissue contours (lips/nose). Confidence Scoring may be used; Software should flag low-confidence landmarks so clinicians can perform quick manual corrections (*Human-in-the-Loop*). Also, Diversify Training Datasets, as AI models must train on varied datasets including Class I, II, and III skeletal cases to eliminate algorithmic bias. And finally, expand to 3D CBCT imaging as AI should transition from 2D cephalograms to 3D CBCT analysis to overcome structural superposition errors.

Conclusion: There were significant differences across all 18 parameters when comparing the three methods. While AI-based programs achieved perfect reproducibility with ICC values frequently reaching 1.0 or 0.999 for skeletal parameters, they do not work efficiently with soft tissue landmarks and linear measurements, such as Lower lip to E-Plane measurement, that require improvement especially Voxel which deviated in both dental and soft tissue measurements and requires manual adjustment for vertical and dental precision. Cephio proved reliable for cranial base landmarks. Dolphin emerged as the method with the highest degree of agreement with the manual method; it had no significant difference in most parameters except for angle of convexity, AB-facial plane, SNA, ANB, and lower facial height. Overall, while several AI and digital tracing outputs exhibited statistically significant differences compared to manual tracing, the majority of these variations remain within clinically acceptable limits. Therefore, statistical significance should not be equated with clinical error, though manual verification remains recommended.

REFERENCES

1. Aksakallı S, Yılancı H, Görükmez E, Ramoğlu Sİ. Reliability assessment of orthodontic apps for cephalometrics. Turkish journal of orthodontics. 2016;29(4):98.
2. Albarakati S, Kula K, Ghoneima A. The reliability and reproducibility of cephalometric measurements: a comparison of conventional and digital methods. Dentomaxillofacial Radiology. 2012;41(1):11-7.
3. Arık SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. Journal of Medical Imaging. 2017;4(1):014501.
4. 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.
5. Büyük SK, Hatal S. Artificial intelligence and machine learning in orthodontics. Ortadoğu Tıp Dergisi. 2019;11(4):517-23.
6. Celik E, Polat-Ozsoy O, Toygar Memikoglu TU. Comparison of cephalometric measurements with digital versus conventional cephalometric analysis. The European Journal of Orthodontics. 2009;31(3):241-6.
7. Çoban G, Öztürk T, Hashimli N, Yağci A. Comparison between cephalometric measurements using digital manual and web-based artificial intelligence cephalometric tracing software. Dental Press Journal of Orthodontics. 2022;27(4):e222112.
8. Erkan M, Gurel HG, Nur M, Demirel B. Reliability of four different computerized cephalometric analysis programs. The European Journal of Orthodontics. 2012;34(3):318-21.
9. Farooq MU, Khan MA, Imran S, Sameera A, Qureshi A, Ahmed SA, et al. Assessing the reliability of digitalized cephalometric analysis in comparison with manual cephalometric analysis. Journal of clinical and diagnostic research: JCDR. 2016;10(10):ZC20.
10. Greenberg BL. Etiology of skeletal malocclusion. Craniomaxillofacial reconstructive and corrective bone surgery: Springer; 2019. p. 39-43.
11. Jeon S, Lee KC. Comparison of cephalometric measurements between conventional and automatic cephalometric analysis using convolutional neural network. Progress in Orthodontics. 2021;22(1):14.
12. Junaid N, Khan N, Ahmed N, Abbasi MS, Das G, Maqsood A, et al., editors. Development, application, and performance of artificial intelligence in cephalometric landmark identification and diagnosis: a systematic review. Healthcare; 2022: MDPI.
13. Meriç P, Naoumova J. Web-based fully automated cephalometric analysis: comparisons between app-aided, computerized, and manual tracings. Turkish journal of orthodontics. 2020;33(3):142.
14. Mohammad-Rahimi H, Nadimi M, Rohban MH, Shamsoddin E, Lee VY, Motamedian SR. Machine learning and orthodontics, current trends and the future opportunities: A scoping review. American Journal of Orthodontics and Dentofacial Orthopedics. 2021;160(2):170-92. e4.
15. Naoumova J, Lindman R. A comparison of manual traced images and corresponding scanned radiographs digitally traced. The European Journal of Orthodontics. 2009;31(3):247-53.
16. Richardson A. An investigation into the reproducibility of some points, planes, and lines used in cephalometric analysis. American journal of orthodontics. 1966;52(9):637-51.
17. Tikku T, Khanna R, Maurya R, Srivastava K, Bhushan R. Comparative evaluation of cephalometric measurements of monitor-displayed images by Nemoceph software and its hard copy by manual tracing. Journal of oral biology and craniofacial research. 2014;4(1):35-41.
18. Tsolakis IA, Tsolakis AI, Elshebiny T, Matthaios S, Palomo JM. Comparing a fully automated cephalometric tracing method to a manual tracing method for orthodontic diagnosis. Journal of clinical medicine. 2022;11(22):6854.
19. Yassir YA, Salman AR, Nabbat SA. The accuracy and reliability of WebCeph for cephalometric analysis. Journal of Taibah University Medical Sciences. 2022;17(1):57-66.
20. Zaheer R, Shafique HZ, Khalid Z, Shahid R, Jan A, Zahoor T, et al. Comparison of semi and fully automated artificial intelligence driven softwares and manual system for cephalometric analysis. BMC Medical Informatics and Decision Making. 2024;24(1):271.
21. Saifeldin H. Comparison between manual lateral cephalometric analysis and Artificial Intelligence Driven platforms. Ain Shams Dental Journal. 2023;29(1):1-5.
22. Rudolph D, Sinclair P, Coggins J. Automatic computerized radiographic identification of cephalometric landmarks. American Journal of Orthodontics and Dentofacial Orthopedics. 1998;113(2):173-9.
23. Navarro RdL, Oltramari-Navarro PVP, Fernandes TMF, Oliveira GFd, Conti ACdCF, Almeida MRd, et al. Comparison of manual, digital and lateral CBCT cephalometric analyses. Journal of Applied Oral Science. 2013;21:167-76.

24. Davidovitch M, Sella-Tunis T, Abramovicz L, Reiter S, Matalon S, Shpack N. Verification of convolutional neural network cephalometric landmark identification. Applied Sciences. 2022;12(24):12784.

25. Patel D, Chaudhary S. The comparison of computer aided digital cephalometric radiograph with manual tracing. Int J Adv Res. 2016;4:621-6.

26. Zughair, Saif Aldeen Kareem, et al. "Evaluation of artificial intelligence-based cephalometric tracing versus semi-automatic and manual tracing." *BMC Oral Health* 25.1 (2025): 1429.

27. Gürler, Güray, Mustafa Serdar Toroglu, and Oruc Yener Cam. "Agreement of WebCeph-Based Automated and Expert-Adjusted Cephalometric Analyses with Manual and Dolphin Tracings." *Diagnostics* 16.12 (2026): 1836.