Clinical validity and precision of deep learning-based cone-beam computed tomography automatic landmarking algorithm

Jung-Eun Park

The Graduate School

Yonsei University

Department of Dentistry

Clinical validity and precision of deep learning-based cone-beam computed tomography automatic landmarking algorithm

A Dissertation Submitted

to the Department of Dentistry

at the Graduate School of Yonsei University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Dental Science

Jung-Eun Park

December 2024

This certifies that the Doctoral Dissertation of
Jung-Eun Park is approved.

_____

Thesis Supervisor: Hyung-Seog Yu

_____

Jung-Yul Cha

_____

Yoon-Jeong Choi

_____

Sung-Hwan Choi

_____

Ui-Lyong Lee

The Graduate School

Yonsei University

December 2024

감사의 글

저에게 아낌없는 지도와 가르침을 주시고, 항상 따뜻하게 격려해주신 유형석 교수님께 진심으로 깊은 존경과 감사의 말씀을 드립니다. 교수님의 조언과 배려 덕분에 박사 학위 과정과 더불어 논문을 잘 마무리할 수 있었습니다. 또한 이 논문을 위해 바쁘신 와중에도 귀한 시간을 내주시어 미흡한 부분을 보완해 나갈 수 있도록 지도해주신 차정열 교수님, 부족한 논문을 꼼꼼하게 검토해주신 최윤정 교수님, 논문을 위해 함께 고민해주시고 세심한 지도를 해주신 최성환 교수님, 부족한 저를 믿고 이끌어 주신 이의룡 교수님께도 깊이 감사드립니다. 제가 연세대학교 교정과에서 공부할 수 있는 기회를 주신 황충주 교수님, 김경호 교수님, 이기준 교수님, 정주령 교수님, 박선형 교수님, 이지현 교수님, 김하림 교수님께도 감사드립니다.

논문 작업의 기반을 마련해주시고, 아이디어와 큰 도움을 주신 김영준 박사님, 매번 아낌없는 도움을 주신 김한나 박사님과 윤성원 선생님께도 큰 감사의 마음을 전합니다. 악교정 수술과 교정학 공부를 할 수 있도록 이끌어주신 최영준 교수님과 최원철 교수님께도 감사드립니다. 논문 작업에 도움을 준 연세대학교 교정과 의국 선생님들, 그리고 논문 심사 준비과정에 있어 많은 조언과 도움을 준 대학원 동기들에게 감사의 마음을 전합니다.

언제나 저를 믿어 주시며 응원해주시는 사랑하고 존경하는 부모님과 언니에게도 고마운 마음을 전합니다. 마지막으로 늘 아낌없는 응원을 보내주는 정지원에게 이 지면을 빌어 고마운 마음을 전합니다.

2024 년 12 월

저자 씀

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Clinical validity and precision of deep learning-based cone-beam computed tomography automatic landmarking algorithm

Jung-Eun Park

Department of Dentistry
The Graduate School, Yonsei University
(Directed by Prof. Hyung-Seog Yu, D.D.S., M.S., Ph.D)

This study was performed to assess the clinical validity and accuracy of a deep learning-based automatic landmarking algorithm for cone-beam computed tomography (CBCT). Three-dimensional (3D) CBCT head measurements obtained through manual and automatic landmarking were compared.

A total of 80 CBCT scans were divided into 3 groups: non-surgical (39 cases); surgical without hardware, namely surgical plates and mini-screws (9 cases); and surgical with hardware (32 cases). Each CBCT scan was analyzed to obtain 53 measurements, comprising 27 lengths, 21 angles, and 5 ratios, which were determined based on 65 landmarks identified using either a manual or a 3D automatic landmark detection method.

In comparing measurement values derived from manual and artificial intelligence landmarking, 6 items displayed significant differences: R U6CP-L U6CP, R L3CP-L L3CP, S-N, Or_R-R U3CP, L1L to Me-GoL, and GoR-Gn/S-N (P<0.05). Of the 3 groups, the surgical scans without hardware exhibited the lowest error, reflecting the smallest difference in measurements between human- and artificial intelligence-based landmarking. The time required to identify 65 landmarks was approximately 40-60 minutes per CBCT volume when done manually, compared to 10.9 seconds for the artificial intelligence method (PC specifications: GeForce 2080Ti, 64GB RAM, and an Intel i7 CPU at 3.6 GHz).

There were 4 differences in length, 1 difference in angle, and 1 difference in ratio, and in the surgical group, there was a significant difference in the measurement errors for length and ratio depending on the presence or absence of the hardware. It is necessary to re-evaluate whether the clinical application of the measurement values is meaningful by comparing the differences between human and between AI and human in the measurement values (length, angle, ratio). In conclusion, although there are still limitations, there is a clear advantage in terms of efficiency, and it can be used in the clinical range.

Key Words: Cone-Beam Computed Tomography; Anatomic Landmarks; Cephalometry; Deep Learning; Orthognathic Surgery

Clinical validity and precision of deep learning-based cone-beam computed tomography automatic landmarking algorithm

Jung-Eun Park

Department of Dentistry
The Graduate School, Yonsei University
(Directed by Prof. Hyung-Seog Yu, D.D.S., M.S., Ph.D)

I. Introduction

Cephalometric analysis is essential for diagnosis and treatment planning in orthodontic and orthognathic surgery. Traditionally, 2-dimensional (2D) cephalometric radiography has been used to evaluate the craniomaxillofacial (CMF) region. However, this imaging modality projects a 3-dimensional (3D) CMF structure onto a 2D plane, leading to image distortion. This distortion can manifest as the overlapping of anatomical structures and the enlargement or reduction of specific areas (Gribel BF et al., 2011).

Several experts have proposed the use of 3D cephalometric analysis with computed tomography or cone-beam computed tomography (CBCT) images (Lee SH et al., 2014; Olszewski

R et al., 2006). CBCT enables clinicians to visualize anatomy in 3 dimensions without overlap, providing comprehensive information on anatomical spatial relationships (Mah JK et al., 2010).

Accurate landmarking is essential for proper diagnosis. However, manual landmarking can be repetitive and laborious, often yielding inconsistencies between and within practitioners. To alleviate the challenges associated with manual landmarking, numerous studies have explored the use of automatic landmark detection systems (Lindner C et al., 2016; Vandaele R et al., 2018).

A wealth of information can be obtained from 3D cephalometry; however, 3D landmarking is difficult, labor-intensive, time-consuming, and heavily dependent on expertise and experience (Lagravere MO et al., 2010; Hassan B et al., 2013). The complexity of processing 3D data contributes to these challenges, as does the substantial effort required to create a 3D labeled dataset. Furthermore, since no open dataset is available for 3D CMF landmarks, many studies have had to rely on small datasets and have limited their measurements to landmarks on the bone surface.

As indicated above, research on 2D and 3D automatic landmarking is ongoing. However, few studies have examined the utility and clinical applicability of cephalometric analysis based on landmarks identified through automatic processes. The clinically acceptable margin of error for landmark placement depends on the error value when implemented in a clinical setting (Ghowsi A et al., 2022). When conducting cephalometric analyses, it is essential to ascertain the impact of errors in linear, angular, and ratio measurements across all 3 dimensions (x, y, and z coordinates). Accuracy results reported for most AI models should be re-evaluated to reflect clinical reproducibility. Because a 1–2 mm deviation in other two measurement points can mostly affect linear and angular values, and consequently orthodontic diagnosis and treatment. The impact of cephalometric analysis method on the accuracy of AI landmarking has not yet been studied sufficiently. When selecting linear and angular measurements of landmarks, it may be useful to evaluate whether AI can achieve

clinically acceptable results in terms of accuracy and whether it can obtain a sufficiently complete cephalometric analysis for an accurate orthodontic diagnosis (Polizzi A and Leonardi R, 2024).

Jeon and Lee (Jeon S and Lee KC, 2021) compared 26 measurements obtained from a convolutional neural network (CNN)-based 2D automatic head landmarking system with those derived from conventional landmarking across 35 lateral head radiographs. Gupta et al. (Gupta A et al., 2016) compared 51 measurements between a knowledge-based 3D automatic head landmarking system and manual landmarking for 30 CBCT scans. Both studies concluded that automated cephalometric analysis is comparable in accuracy to manual calculations (Jeon S and Lee KC, 2021; Gupta A et al., 2016).

In dentistry, metal artifacts are commonly observed on CBCT images due to materials used in orthodontics, surgical applications, and dental restorations. The presence of metal artifacts along the path of the radiation beam causes photon depletion and scattering, resulting in characteristic light and dark banding artifacts on the CBCT image (Barrett JF and Keat N; 2004). These artifacts obscure the adjacent anatomy and impede diagnosis; furthermore, they can interfere with the image segmentation of maxillary and mandibular teeth and bone structures for computer-guided therapy (Hung K et al., 2020).

As mentioned above, there are papers on automatic cephalometric analysis, but there are few studies on automatic cephalometric analysis targeting orthognathic surgery patient data. It is necessary to find out whether AI automatic cephalomeric analysis can analyze orthognathic surgery patient image data well and whether it is affected by the presence or absence of hardware including the plate used in the surgery.

This study aimed to evaluate the clinical validity and accuracy of a deep learning-based CBCT automatic landmarking algorithm in 3D automatic cephalometry and analysis. The authors posited

that the values produced by the algorithm would be comparable to those obtained by humans and would promote efficiency by reducing time. Additionally, the authors hypothesized that no errors would be observed in the measured values attributable to hardware and screws after surgery.

II. Materials and Methods

1. 3D manual landmarking

Under the supervision of a clinical physician, two biomedical experts identified 3D landmarks in 821 CBCT images acquired with an i-CAT 17-19 device (Imaging Sciences International, Hatfield, PA, USA) and 148 CBCT images acquired with different NewTom models (5G, VGi EVO, VGi Mark 3, VGi Mark 4; NewTom, Imola, Italy).

2. 3D automatic landmarking

Using an automatic 3D landmarking algorithm utilizing a 2-stage coarse-to-fine approach, 65 landmarks were measured. All CBCT datasets were acquired for diagnostic purposes and exported in Digital Imaging and Communications in Medicine format. The personal data of all patients, including names and registration numbers, were anonymized. The dataset was compiled without regard to sex, age, or race and included perioperative information, with a focus on cases of orthognathic surgery. Consequently, approximately 40% of the CBCT scans contained surgical hardware, such as surgical plates and mini-screws.

Sixty-five 3D landmarks in 12 anatomical groups, including the bone, skin, dental crown, tooth root, neural canal (center or opening), and sella, were cataloged (Table 1). The time required to measure these 65 landmarks was 10.9 seconds per volume, with point-to-point errors of 1.7 ± 0.1 mm (99% confidence interval). The threshold for a clinically acceptable successful detection rate (SDR) was set at 3 mm, with SDRs of 88.16% at this level and 94.35% at 4 mm (Torosdagli N et al., 2014). All 65 landmarks were detected near their true positions, even in the presence of various types of orthognathic surgical hardware.

Table 1. Anatomical groups and included landmarks.

Group	Landmarks
Mid-sagittal Mx. (bone)	A (A-point), ANS (anterior nasal spine), PNS (posterior nasal spine), N (nasion)
Mid-sagittal Mn. (bone)	B (B-point), Pog (pogonion), Gn (gnathion), Me (menton)
Mid-sagittal Mx. (soft tissue)	Sls (soft tissue A-point), Pn (pronasale), Soft N (soft tissue nasion), Sts (stomion superius), Soft Gabella (soft tissue gabella), Ala R (right alar base), Ala L (left alar base)
Mid-sagittal Mn. (soft tissue)	Soft Pog (soft tissue pogonion), Si (mentolabial sulcus), Sti (stomion inferius)
Skull	G (crista galli), Ba (basion), S (sella), Po_R (right porion), Po_L (left porion)
Lateral Mn.	Go_R (right gonion), Go_L (left gonion), M_R (right mental foramen), M_L (left mental foramen), MF_R (right mandibular foramen), MF_L (left mandibular foramen)
Tooth crown (Mx.)	R U1CP (center of right maxillary incisor crown), L U1CP (center of left maxillary incisor crown), R U3CP (tip of right maxillary canine crown), L U3CP (tip of left maxillary canine crown), R U6CP (tip of mesiobuccal cusp of right maxillary first molar crown), L U6CP (tip of mesiobuccal cusp of left maxillary first molar crown)
Tooth roots (Mx.)	R U1RP (root of right maxillary incisor), L U1RP (root of left maxillary incisor), R U3RP (root of right maxillary canine), L U3RP (root of left maxillary canine), R U6RP (mesiobuccal root of right maxillary first molar), L U6RP (mesiobuccal root of left maxillary first molar)
Tooth crown (Mn.)	R L1CP (center of right mandibular incisor crown), L L1CP (center of left mandibular incisor crown), R L3CP (tip of right mandibular canine crown), L L3CP (tip of left mandibular canine crown), R L6CP (tip of mesiobuccal cusp of right mandibular first molar crown), L L6CP (tip of mesiobuccal cusp of left mandibular first molar crown)
Tooth roots (Mn.)	R L1RP (root of right mandibular incisor), L L1RP (root of left mandibular incisor), R L3RP (root of right mandibular canine), L L3RP (root of left mandibular canine), R L6RP (mesiobuccal root of right mandibular first molar), L L6RP (mesiobuccal root of left mandibular first molar)
Bone around orbit	ZyFr_R (right zygomaticofrontal suture), ZyFr_L (left zygomaticofrontal suture), RO_R (right roof of orbit), Or_R (right orbitale), RO_L (left roof of orbit), Or_L (left orbitale)
Condyle	Cl_R (right condylus lateralis), Cm_R (right condylus medialis), Co_R (right condylion), Cl_L (left condylus lateralis), Cm_L (left condylus medialis), Co_L (left condylion)

Mx: maxilla, Mn: mandible.

3. 3D landmark measurement

Among the 821 sets of CBCT images from Chung-Ang University Hospital in Seoul, Korea(Institutional Review Board number: 1922-007-362), which were used in a study on 3D automatic landmarking using deep learning, 80 sets of CBCT images that were not used for training were selected as the study subjects. A total of 80 CBCT were categorized into a non-surgical group (39 cases) and a surgical group (41 cases). The dataset was constructed independent of gender, age, and race and included orthognathic surgery cases. All CBCT datasets were collected for diagnostic purposes through anonymization and encryption before generating ground truth data.

The latter was further subdivided into 9 cases without hardware and 32 cases with hardware. All CBCT scans were anonymized and assigned new serial numbers for the study.

Based on the 65 landmarks, 53 measurements (27 lengths, 21 angles, and 5 ratios) were taken. The classification and measurement values were based on the methodology outlined by Gupta et al (Gupta A et al., 2016).

The length measurements were categorized into 3 groups: 1) bilateral, obtained from 2 symmetrical landmarks in the parasagittal plane (Fig. 1); 2) midsagittal, derived from 2 landmarks in the midsagittal plane (Fig. 2); and 3) midsagittal to bilateral, acquired using 3 landmarks—1 in the central sagittal plane and 2 symmetrically located in the parasagittal plane (Fig. 3).

The angles were classified into 3 types: midsagittal (calculated using 3 landmarks within the midsagittal plane); midsagittal to bilateral (determined by 4 landmarks, 2 in the midsagittal plane and 2 symmetrically positioned in the parasagittal plane) (Fig. 4).; and planar (either four landmarks or two landmarks and one horizontal plane (Fig. 5).

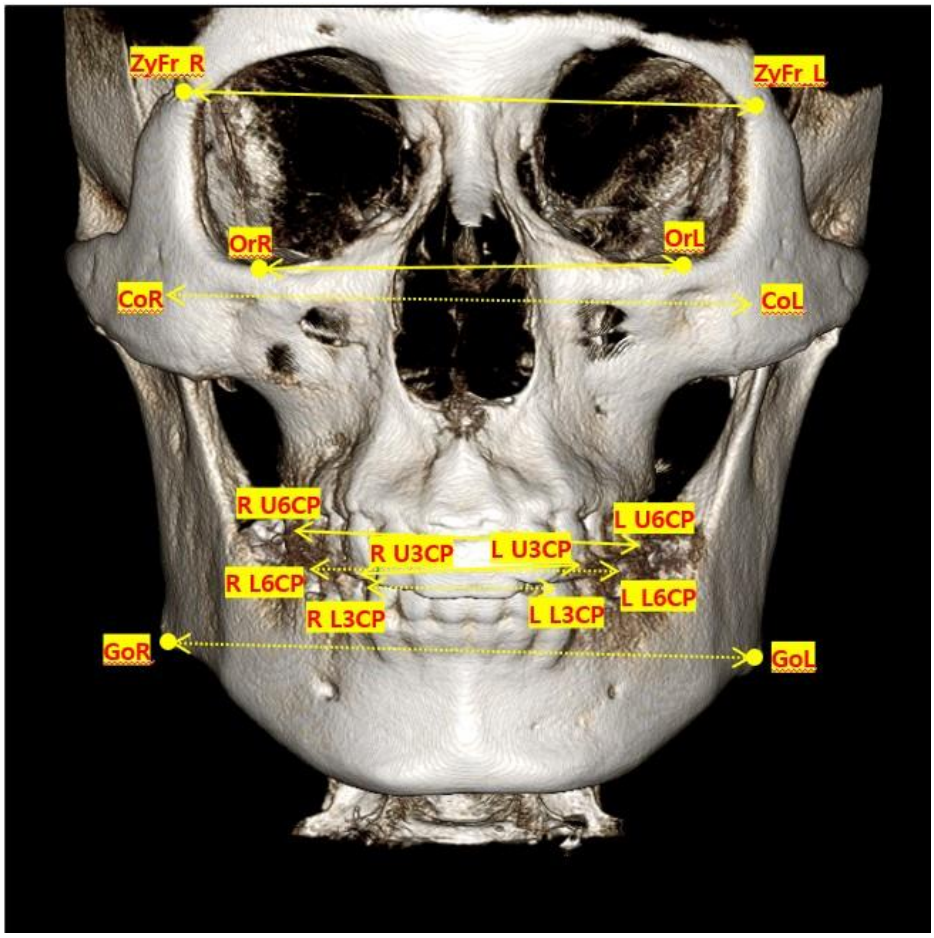

Figure 1. Bilateral linear 3D cephalometric measurements obtained through two symmetrically present landmarks, one in right parasagittal and one in left parasagittal plane.

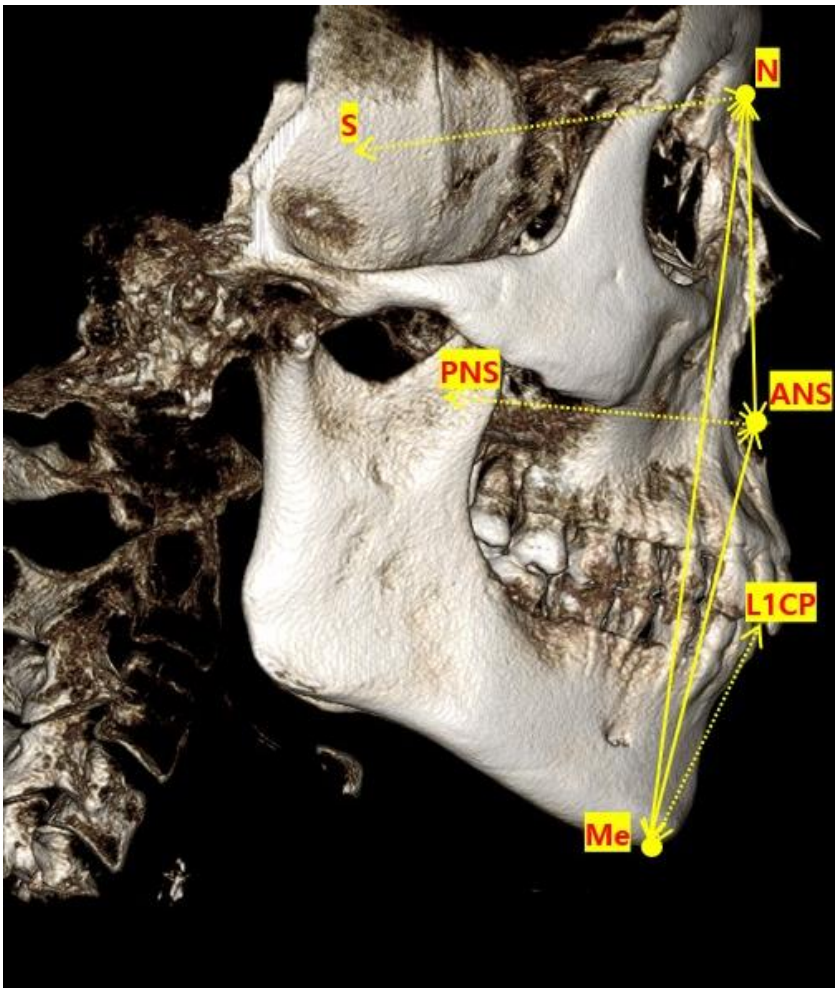

Figure 2. Midsagittal linear 3D cephalometric measurements obtained through two landmarks in midsagittal plane.

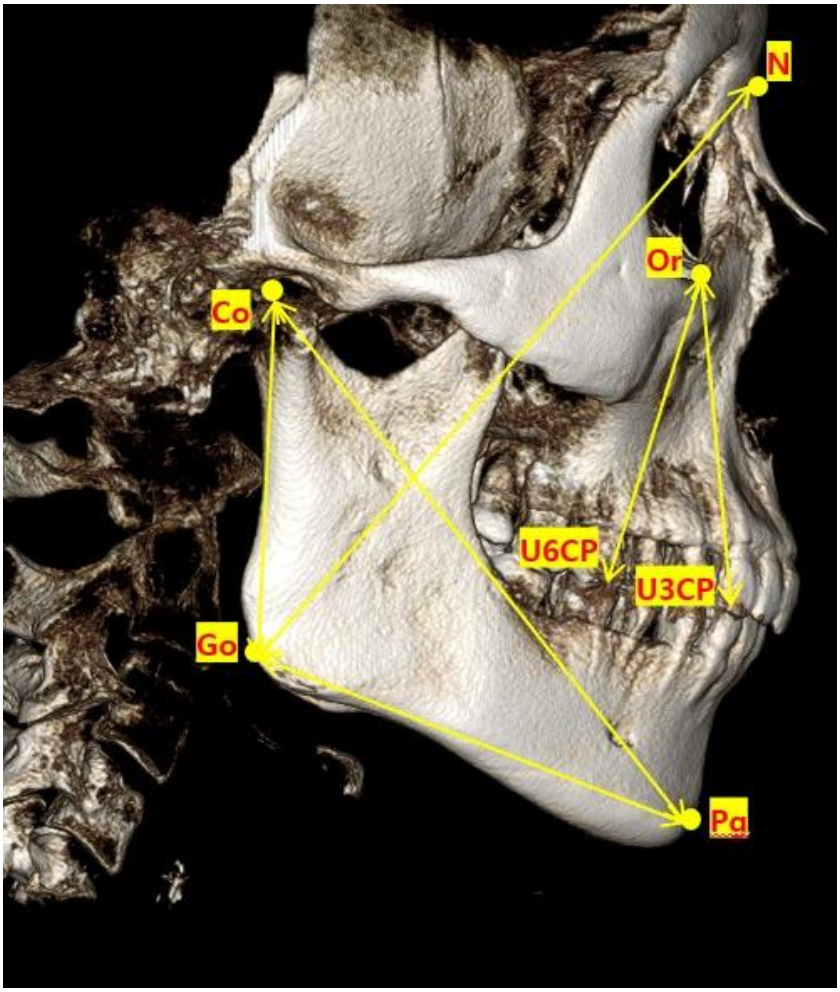

Figure 3. Midsagittal to bilateral linear 3D cephalometric measurements obtained through three landmarks, one landmark in midsagittal plane and other two landmarks symmetrically located in parasagittal planes on the right and left side of skull.

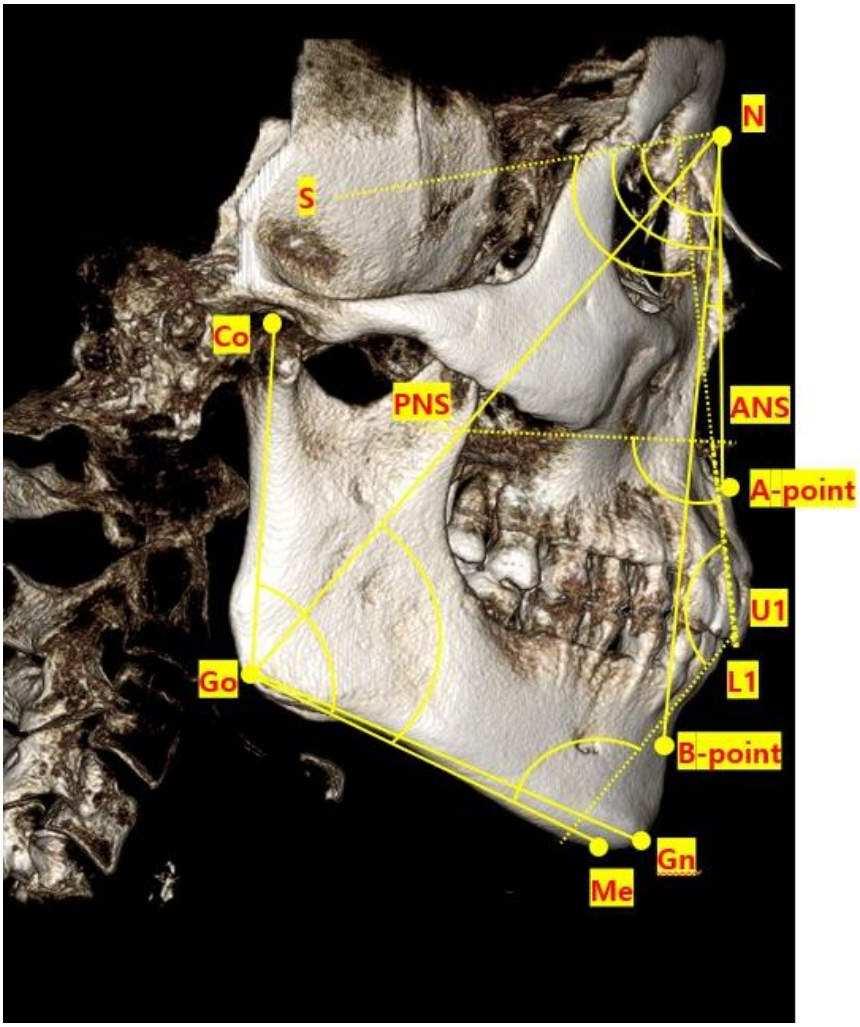

Figure 4. Angular 3D cephalometric measurements (midsagittal and midsagittal to bilateral) based on three points.

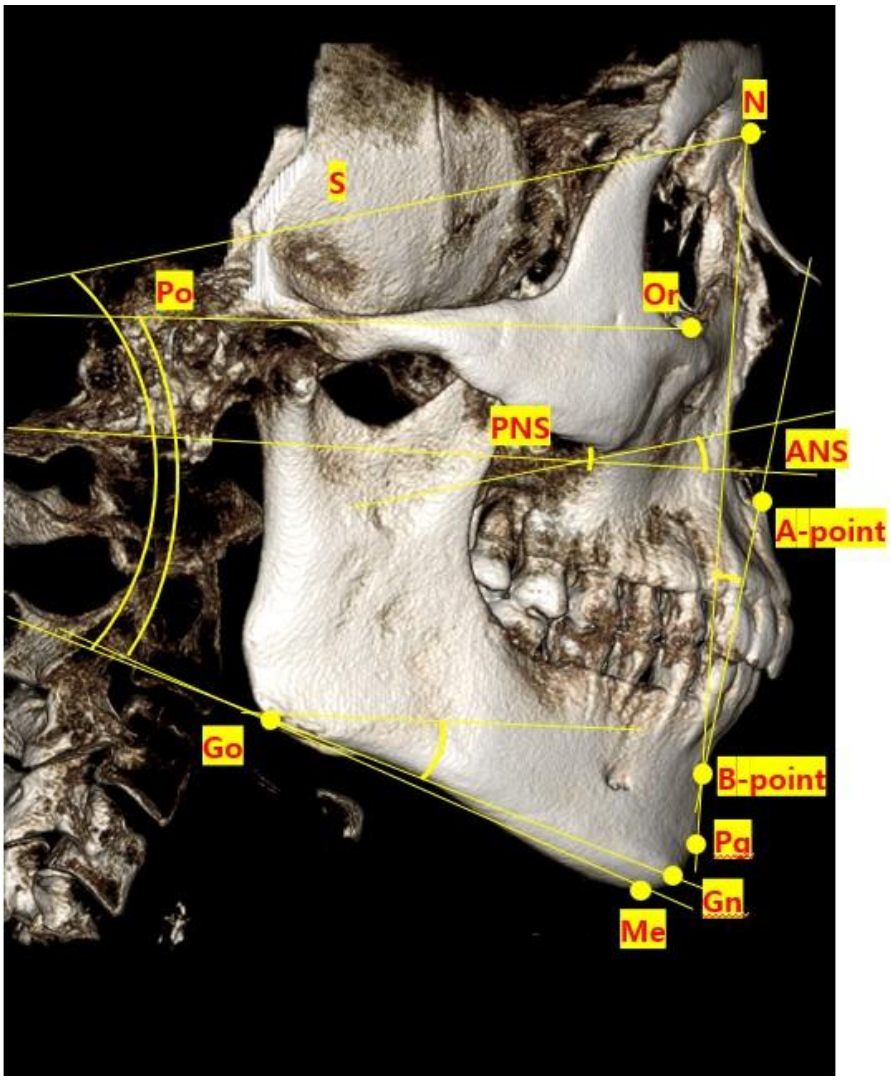

Figure 5. Angular 3D cephalometric measurements obtained through either four landmarks or two landmarks and one horizontal plane.

Measurements were obtained through 3D vector calculation. Length was determined by multiplying the 3D voxel index of a landmark by the spacing values along the x, y, and z axes, followed by application of the 3D Euclidean distance formula. For angles, the angle between the vectors $\bar{u}$ and $\bar{v}$, represented by the pair of landmarks, was calculated. This angle was derived using the second law of cosines. When determining the smaller of the angles formed by the 2 vectors, the direction of the vectors was verified and factored into the calculation. In mathematics, the angle between 2 vectors is a value between 0 and 180 degrees. The angle between 2 vectors $\bar{u}$ and $\bar{v}$ can be calculated using the dot product and inverse trigonometric functions.

The following formulas were used to calculate the distances and angles in P(x_1, y_1, z_1), Q(x_2, y_2, z_2), R(x_3, y_3, z_3), and S(x_4, y_4, z_4):

$$D_{PQ} = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2}$$

$where\ X_i = x_i \times (spacing\ of\ x\text{-}\!\!-axis), Y_i = y_i \times (spacing\ of\ y\text{-}axis),\ and\ Z_i = z_i \times (spacing\ of\ z\text{-}axis)$

$$\theta_{PQRS} = cos^{-1}\frac{\bar{u}\bar{v}}{|\vec{u}||\vec{v}|}$$

$where\ \vec{u} = (x_2 - x_1)\hat{\imath} + (y_2 - y_1)\hat{\jmath} + (z_2 - z_1)\hat{k}$ *and* $\vec{v} = (x_4 - x_3)\hat{\imath} + (y_4 - y_3)\hat{\jmath} + (z_4 - z_3)\hat{k}$

4. Statistical analysis

The means and standard deviations (SDs) of the measurements, obtained using landmarks identified by humans or artificial intelligence (AI), were determined. The means, medians, and SDs of the errors between the 2 measured values were also calculated. For the 80 CBCT scans, the Shapiro-Wilk test was applied to assess the normality of the measurements based on human- and AI-identified landmarks. Subsequently, an unpaired *t*-test was used to evaluate whether a significant difference existed between the 2 groups. Additionally, a Bland-Altman plot was employed to visually represent the differences between groups.

An unpaired *t*-test was used to compare the nonsurgical, surgical, hardware present, and hardware absent groups. Differences between the groups were visually expressed using violin plots. Statistical significance was established at $P<0.05$. Shapiro-Wilk and unpaired *t*-tests were conducted using SPSS GradPacks Statistics 28 (IBM Corp., Armonk, NY, USA). Bland-Altman and violin plots were created using Microsoft Excel (Version 2208 Build 16.0.15601.20148, 64-bit; Microsoft, Redmond, WA, USA).

III. Results

1. Comparison of measurement values obtained by human and AI methods

The means and SDs of the measurements, as well as the means, medians, and SDs of the differences between the 2 methods, were calculated (Tables 2-4). The unpaired *t*-test revealed statistically significant differences for 6 of the 53 measured values when comparing landmarks detected by humans and AI (R U6CP-L U6CP, R L3CP-L L3CP, S-N, Or_R-R U3CP, L1L to Me-Go, and GoR-Gn/S-N; $P<0.05$). The measurements with the highest mean error values were CoL-CoR (3.700 mm) for length, U1R to L1R (3.587°) and U1L to L1L (3.169°) for angle, and N-Me/N-ANS (0.043) for ratio.

The difference between values measured using human-identified and AI-detected landmarks was illustrated with a Bland-Altman plot. The limits of agreement and the width obtained in the Bland-Altman analysis are presented in Table 5. The measurements with the largest widths were CoR-CoL (17.49 mm) for length, U1R to L1R (20.06°) for angle, and N-Me/N-ANS (0.22) for ratio.

Table 2. Comparison of linear cephalometric measurements between manual and artificial intelligence methods. (unit: mm)

Linear measurement parameters	Manual	Artificial intelligence	Error	p value
Bilateral measurement				
ZyFr_R-ZyFr_L	102.09 ± 6.08	101.83 ± 5.15	1.13 ± 1.15	0.155
GoL-GoR	92.21 ± 7.73	92.38 ± 7.38	0.94 ± 0.70	0.201
CoL-CoR	103.68 ± 6.29	103.24 ± 5.01	3.70 ± 2.50	0.377
OrL-OrR	59.63 ± 3.39	59.33 ± 3.12	1.93 ± 1.54	0.281
R U3CP-L U3CP	35.09 ± 2.64	34.94 ± 2.08	1.35 ± 1.63	0.535
R U6CP-L U6CP	52.66 ± 3.47	52.14 ± 2.79	1.70 ± 1.34*	0.030 *
R L3CP-L L3CP	27.09 ± 1.87	27.78 ± 1.66	1.35 ± 1.01*	0.000 *
R L6CP-L L6CP	47.38 ± 2.99	47.64 ± 2.43	1.63 ± 1.10	0.245
Midsagittal measurement				
N-Me	123.15 ± 7.69	122.97 ± 7.36	1.25 ± 1.13	0.341
N-ANS	54.39 ± 3.29	54.29 ± 3.00	1.33 ± 1.01	0.572
ANS-Me	69.73 ± 6.78	69.62 ± 6.52	1.02 ± 0.94	0.476
ANS-PNS	46.92 ± 4.18	47.24 ± 3.44	1.38 ± 1.30	0.139
S-N	63.98 ± 3.73	63.60 ± 3.68	0.85 ± 0.74*	0.002 *
Me-L L1CP	42.16 ± 3.53	42.25 ± 3.46	0.68 ± 0.88	0.466
Me-R L1CP	42.14 ± 3.50	42.18 ± 3.44	0.62 ± 0.71	0.656
Midsagittal to bilateral measurement				
GoL-Pg	88.07 ± 5.73	87.84 ± 5.37	1.69 ± 1.21	0.317
GoR-Pg	87.61 ± 5.86	87.61 ± 5.74	1.79 ± 1.38	0.980
GoL-N	119.18 ± 7.59	118.81 ± 7.22	1.29 ± 1.16	0.057
GoR-N	119.45 ± 7.43	119.01 ± 7.08	1.49 ± 1.40	0.051
CoL-GoL	57.1 ± 6.14	56.97 ± 5.86	1.98 ± 1.57	0.629
CoR-GoR	57.58 ± 6.25	57.48 ± 5.46	2.12 ± 1.85	0.745
CoL-Pg	125.29 ± 7.32	125.28 ± 7.12	1.01 ± 0.80	0.969
CoR-Pg	125.43 ± 7.66	125.49 ± 7.67	1.14 ± 1.03	0.709
Or_L-L U3CP	55.93 ± 4.28	55.70 ± 4.05	0.76 ± 0.77	0.060
Or_R-R U3CP	55.55 ± 4.35	55.33 ± 4.11	0.73 ± 0.62*	0.041 *
Or_L-L U6CP	51.33 ± 4.24	51.38 ± 4.02	0.72 ± 0.57	0.660
Or_R-R U6CP	51.51 ± 4.20	51.52 ± 3.96	0.72 ± 0.60	0.882

*P<0.05. Refer to Table 1 for abbreviations.

Table 3. Comparison of angular cephalometric measurements between manual and artificial intelligence methods. (unit: mm)

Angular measurement parameters	Manual	Artificial intelligence	Error	p value
Midsagittal measurement				
S-N-A	82.16 ± 3.92	82.27 ± 3.33	1.41 ± 1.22	0.587
S-N-B	80.18 ± 4.77	80.24 ± 4.39	1.25 ± 1.15	0.752
A-N-B	3.74 ± 2.15	3.64 ± 2.15	0.72 ± 0.56	0.366
U1L to ANS-PNS	64.65 ± 7.7	65.13 ± 7.1	2.65 ± 2.1	0.206
U1R to ANS-PNS	65.19 ± 7.92	65.29 ± 7.02	2.57 ± 2.4	0.797
U1L-SN	104.46 ± 8.15	103.89 ± 7.1	2.37 ± 1.91	0.092
U1R-SN	103.94 ± 8.21	103.78 ± 7.03	2.57 ± 1.97	0.669
Midsagittal to bilateral measurement				
N-GoL-Me	71.74 ± 4.5	71.8 ± 4.42	0.64 ± 0.61	0.493
N-GoR-Me	71.71 ± 4.5	71.8 ± 4.36	0.64 ± 0.69	0.380
CoL-GoL-Me	121.28 ± 6.65	121.38 ± 6.41	1.45 ± 1.06	0.609
CoR-GoR-Me	121.46 ± 6.16	121.49 ± 6.05	1.38 ± 1.26	0.873
U1L to L1L	130.88 ± 10.37	130.62 ± 8.9	3.17 ± 2.91	0.590
U1R to L1R	131.18 ± 10.87	131.16 ± 9.37	3.59 ± 3.63	0.981
L1L to Me-GoL	96.44 ± 7.77	95.64 ± 6.96	2.74 ± 2.27*	0.045 *
L1R to Me-GoR	94.8 ± 7.4	95.22 ± 7.22	2.72 ± 2.53	0.321
Planar measurement				
A-B X N-Pog	5.14 ± 2.77	5.21 ± 3.18	1.16 ± 0.92	0.678
S-N X GoL-Gn	47.59 ± 5.23	47.25 ± 5.02	1.55 ± 1.49	0.152
S-N X GoR-Gn	46.83 ± 5.25	47.06 ± 4.79	1.5 ± 1.07	0.274
ANS-PNS X S-N	168.69 ± 4.57	168.88 ± 4.02	1.61 ± 1.46	0.440
PoL-OrLxGoL-Me	30.91 ± 5.78	30.85 ± 5.61	0.94 ± 0.78	0.663
PoR-OrRxGoR-Me	30.32 ± 5.35	30.35 ± 5.23	1.12 ± 0.89	0.865

*$P<0.05$. Refer to Table 1 for abbreviations.

Table 4. Comparison of ratios between manual and artificial intelligence methods.

Ratio parameters	Manual	Artificial intelligence	Error	p value
N-Me/N-ANS	2.27 ± 0.13	2.27 ± 0.13	0.04 ± 0.04	0.952
S-GoL/N-Me	0.73 ± 0.05	0.72 ± 0.05	0.01 ± 0.01	0.528
S-GoR/N-Me	0.73 ± 0.05	0.73 ± 0.05	0.02 ± 0.02	0.738
GoL-Gn/S-N	1.38 ± 0.09	1.39 ± 0.08	0.04 ± 0.03	0.119
GoR-Gn/S-N	1.38 ± 0.09	1.39 ± 0.08	0.03 ± 0.03*	0.025 *

*P<0.05. Refer to Table 1 for abbreviations.

Table 5. Bland-Altman analysis of the difference between manual and artificial intelligence-based cephalometric measurements.

	95% limit of agreement		Width		95% limit of agreement		Width
	Upper limit	Lower limit			Upper limit	Lower limit	
Linear measurement				**Angular measurement**			
Bilateral measurement				*Midsagittal measurement*			
ZyFr_R-ZyFr_L	3.38	−2.87	6.24	S-N-A	3.54	−3.77	7.31
GoR-GoL	2.11	−2.44	4.55	S-N-B	3.28	−3.41	6.69
CoR-CoL	9.19	−8.3	17.49	A-N-B	1.88	−1.69	3.57
OrR-OrL	5.12	−4.52	9.64	U1L to ANS-PNS	6.1	−7.06	13.17
R U3CP-L U3CP	4.29	−3.99	8.28	U1R to ANS-PNS	6.8	−7.01	13.81
R U6CP-L U6CP	4.66	−3.61	8.27	U1L-SN	6.45	−5.31	11.75
R L3CP-L L3CP	2.33	−3.72	6.05	U1R-SN	6.53	−6.22	12.74
R L6CP-L L6CP	3.59	−4.1	7.69	*Midsagittal to bilateral measurement*			
Midsagittal measurement				N-GoL-Me	1.66	−1.8	3.46
N-Me	3.48	−3.12	6.59	N-GoR-Me	1.74	−1.92	3.66
N-ANS	3.38	−3.17	6.55	CoL-GoL-Me	3.43	−3.63	7.06
ANS-Me	2.84	−2.61	5.45	CoR-GoR-Me	3.64	−3.7	7.34
ANS-PNS	3.37	−3.99	7.36	U1L to L1L	8.71	−8.19	16.9
S-N	2.47	−1.71	4.18	U1R to L1R	10.04	−10.01	20.06
Me-L L1CP	2.08	−2.27	4.35	L1L to Me-GoL	7.62	−6.03	13.64
Me-R L1CP	1.8	−1.9	3.7	L1R to Me-GoR	6.84	−7.66	14.5
Midsagittal to bilateral measurement				*Midsagittal to bilateral measurement*			
GoL-Pg	4.29	−3.82	8.11	A-B X N-Pog	2.84	−2.98	5.82
GoR-Pg	4.45	−4.44	8.9	S-N X GoL-Gn	4.51	−3.82	8.33
GoL-N	3.69	−2.96	6.65	S-N X GoR-Gn	3.37	−3.83	7.2
GoR-N	4.36	−3.47	7.83	ANS-PNS X S-N	4.06	−4.43	8.49
CoL-GoL	5.1	−4.83	9.93	PoL-OrLxGoL-Me	2.46	−2.34	4.81
CoR-GoR	5.63	−5.42	11.05	PoR-OrRxGoR-Me	2.78	−2.84	5.62
CoL-Me	2.53	−2.52	5.05	**Ratio parameters**			
CoR-Me	2.96	−3.09	6.05	N-Me/N-ANS	0.11	−0.11	0.22
Or_L-L U3CP	2.31	−1.85	4.16	S-GoL/N-Me	0.04	−0.04	0.07
Or_R-R U3CP	2.05	−1.61	3.66	S-GoR/N-Me	0.04	−0.04	0.09
Or_L-L U6CP	1.76	−1.85	3.61	GoL-Gn/S-N	0.08	−0.1	0.18
Or_R-R U6CP	1.81	−1.85	3.66	GoR-Gn/S-N	0.07	−0.09	0.09

Refer to Table 1 for abbreviations.

2. Comparison of measured values by surgical history and hardware presence

The 80 CBCT scans included a non-surgical group (39 cases) and a surgical group (41 cases), with the latter including 9 cases without hardware and 32 cases with hardware. For each group, measurements derived from human-identified and AI-detected landmarks were compared for the 27 linear parameters using an unpaired *t*-test (Fig. 6). In the linear cephalometric measurements, the non-surgical group displayed a mean error of 1.32 mm and an SD of 1.30 mm. The surgical group without hardware had a mean of 1.10 mm (SD, 1.08 mm), while the group with hardware had a mean of 1.45 mm (SD, 1.49 mm). The lowest error was observed in the surgical group without hardware. No significant difference in measurement agreement was observed between the non-surgical and surgical groups ($P<0.05$). However, a significant difference was noted between surgical subgroups based on the presence or absence of hardware ($P<0.05$). These findings were graphically represented using a violin plot (Fig. 7).

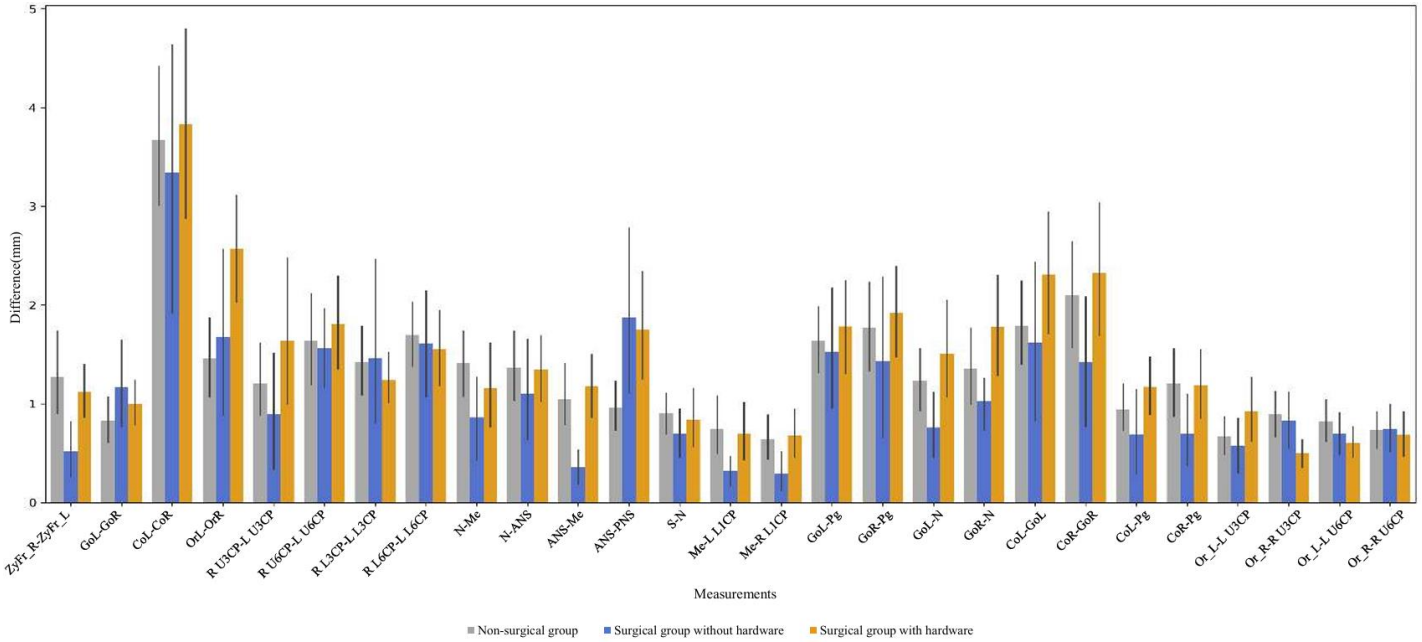

Figure 6. Comparison of linear cephalometric measurements between manual and artificial intelligence methods by patient group. Black error bars represent the 95% confidence standard deviation range for each item value.

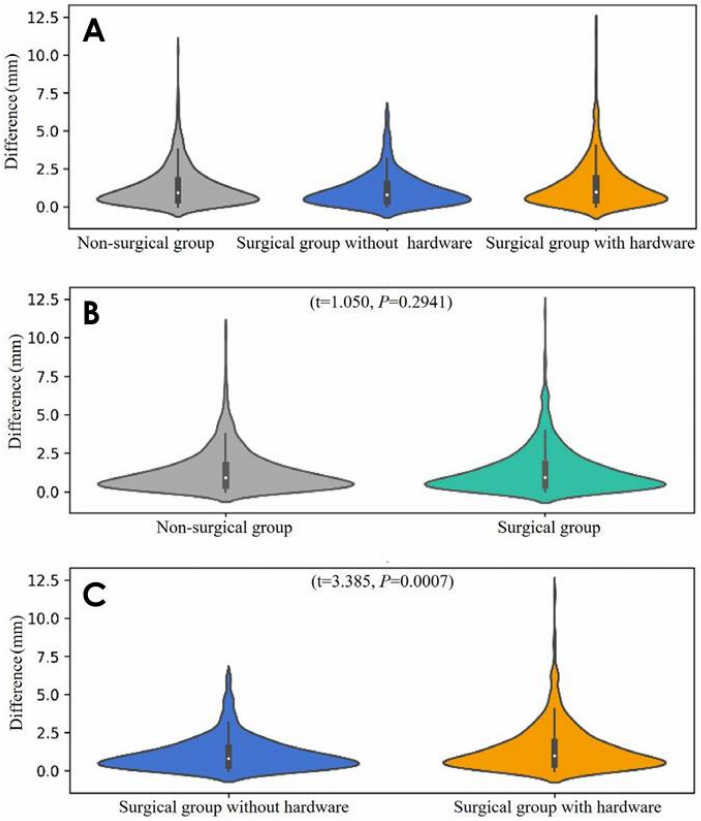

Figure 7. Violin plots representing the difference in linear measurements between the manual and artificial intelligence methods. The thick vertical bar in the violin plot represents the interquartile range, while the thin vertical line indicates the 95% confidence interval; the extremes of this thin line denote the maximum and minimum values. The central white dot signifies the median. The width of a violin plot reflects the density of the data, with wider sections indicating a higher frequency of values and narrower sections representing a lower frequency. The difference values are distributed around the median for all 3 groups: non-surgical, surgical without hardware, and surgical with hardware. A. Violin plots of linear measurements for the 3 groups. B. Violin plots of linear measurements for the non-surgical and surgical groups. C. Violin plots of linear measurements for the surgical group, comparing cases with and without hardware.

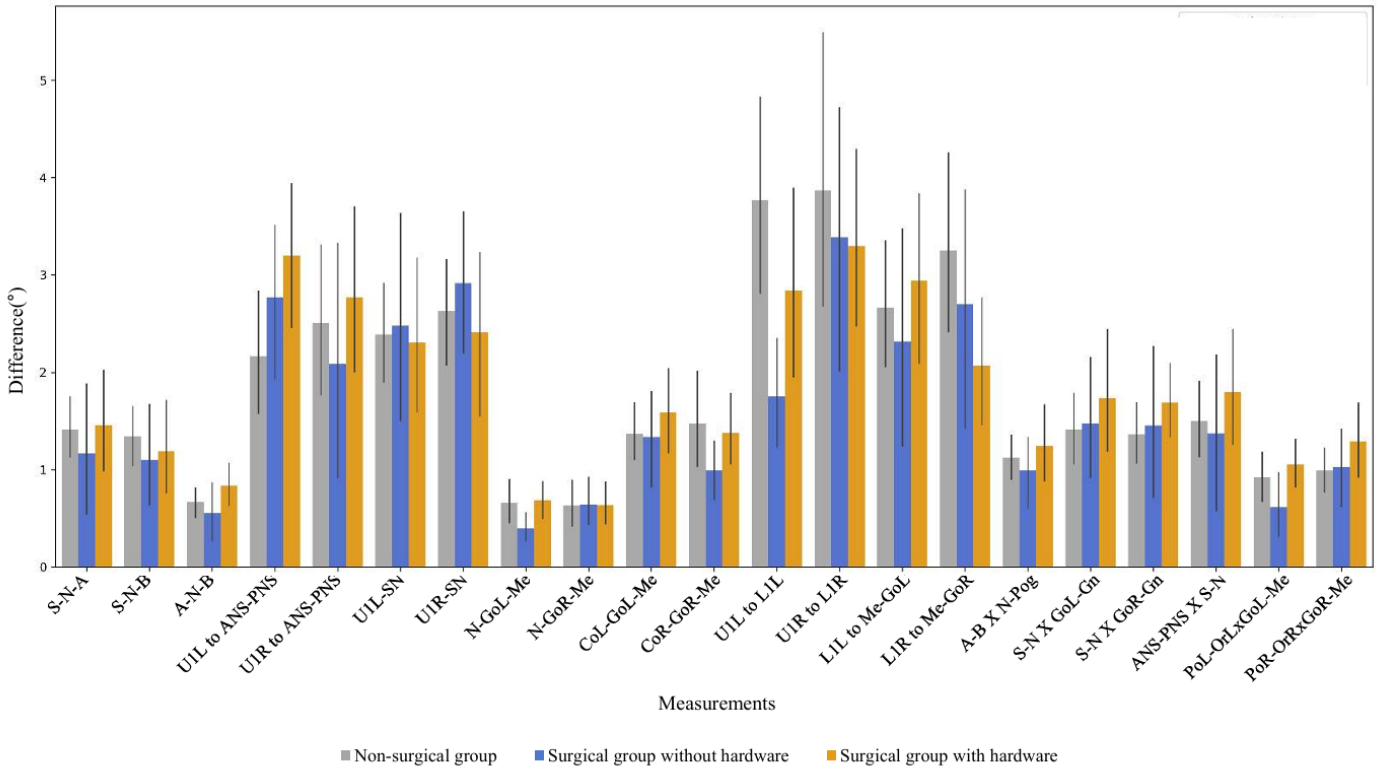

Figure 8. Comparison of angular cephalometric measurements between manual and artificial intelligence methods by patient group.

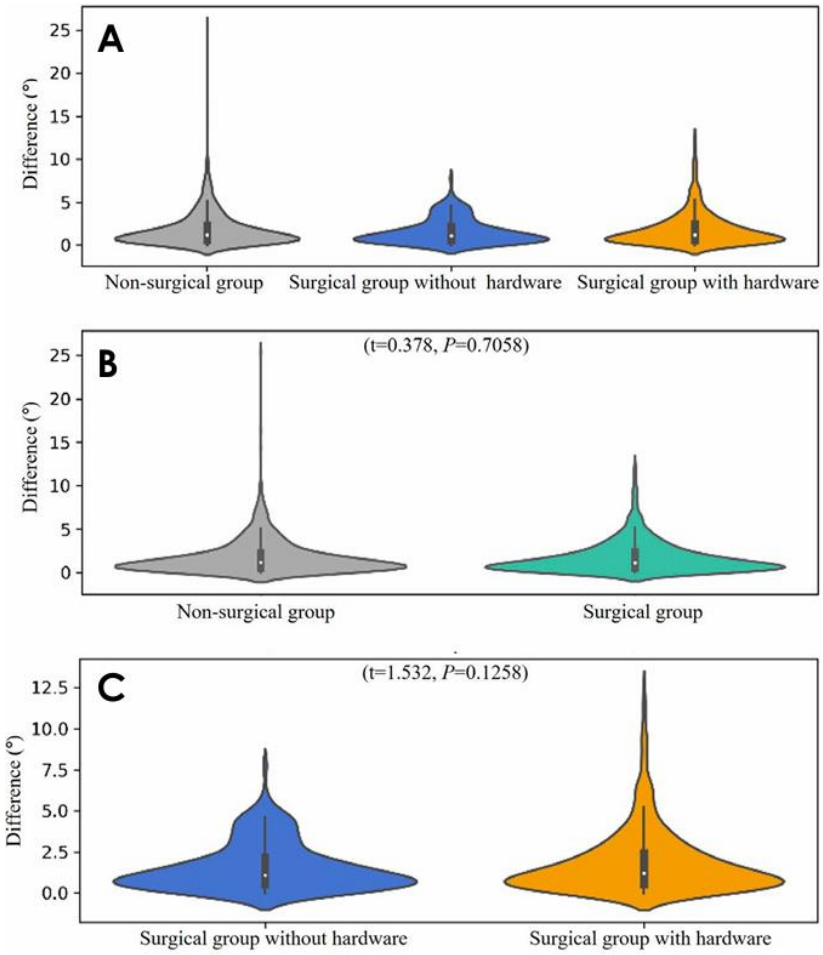

Figure 9. Violin plots representing the difference in angular measurements between the manual and artificial intelligence methods. The difference values are distributed around the median for all 3 groups: non-surgical, surgical without hardware, and surgical with hardware. A. Violin plots of angular measurements for the 3 groups. B. Violin plots of angular measurements for the non-surgical and surgical groups. C. Violin plots of angular measurements for the surgical group, comparing cases with and without hardware.

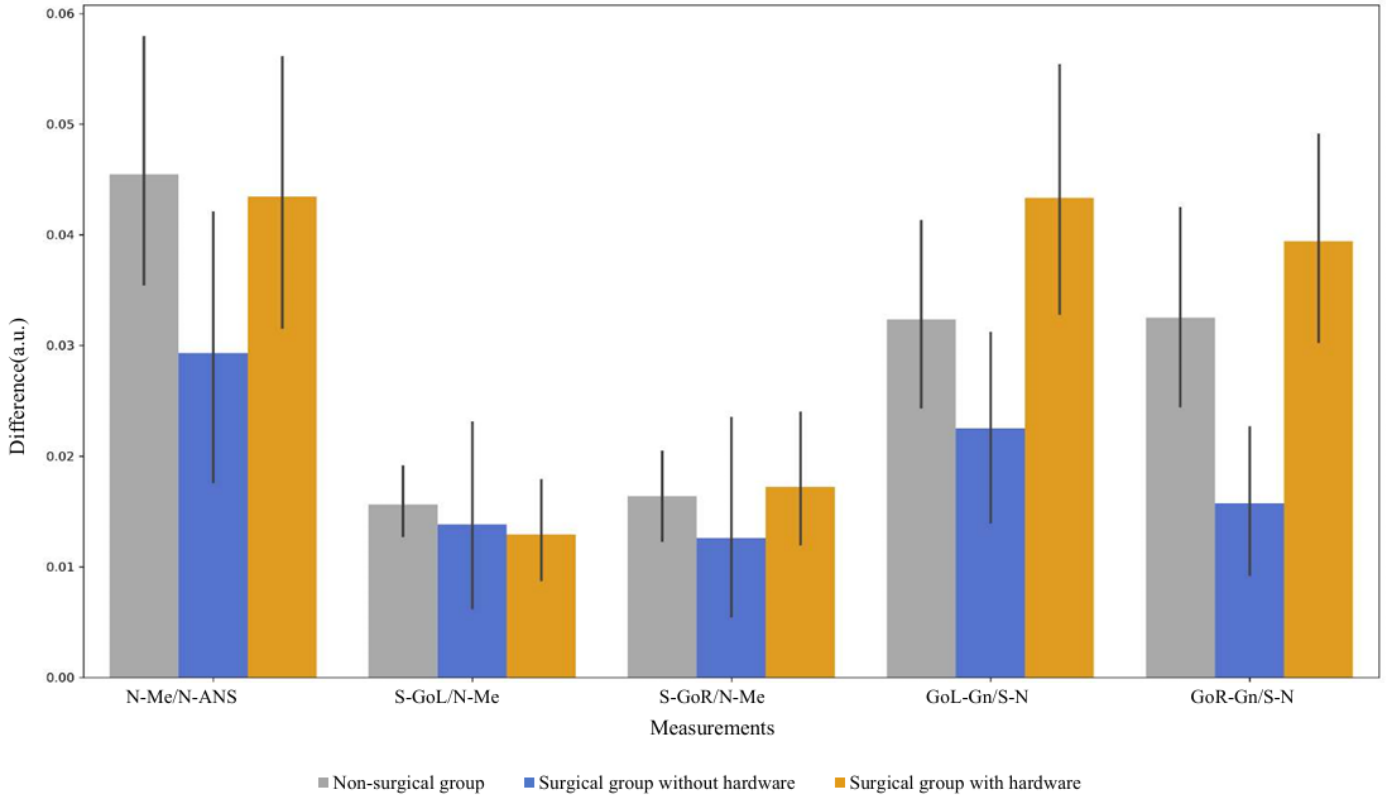

Figure 10. Comparison of cephalometric measurement ratios between manual and artificial intelligence methods by patient group.

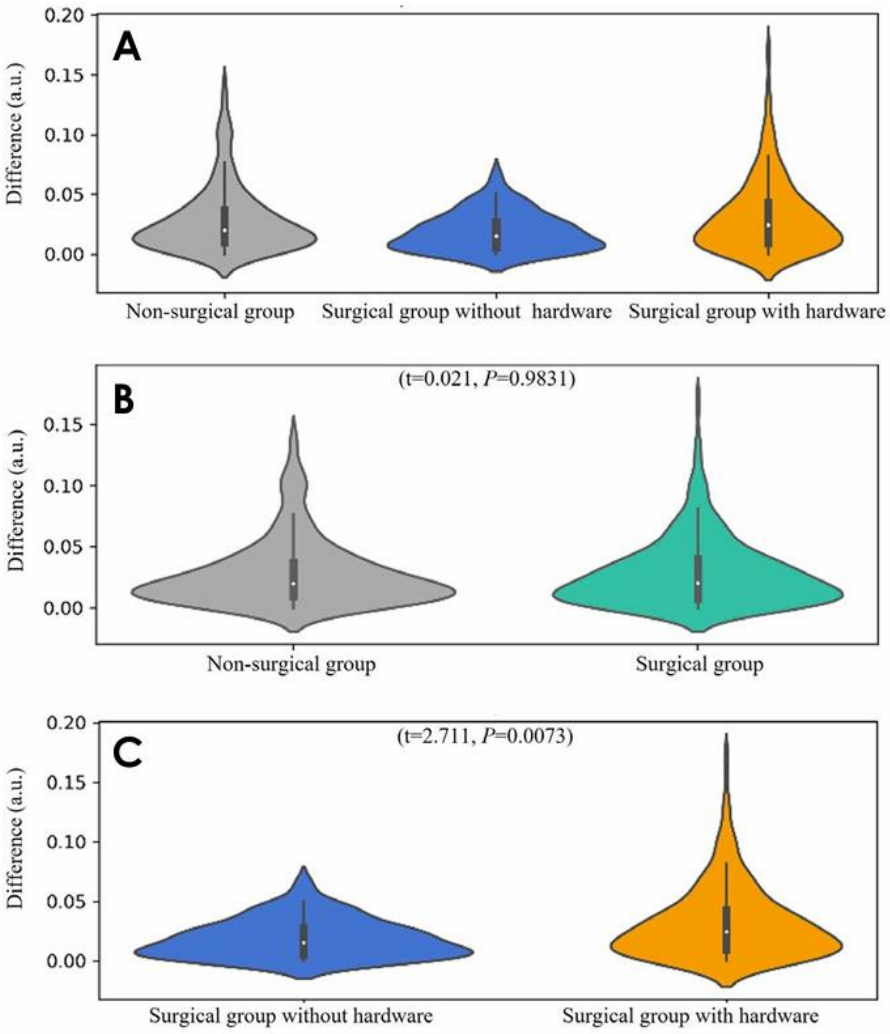

Figure 11. Violin plots representing the difference in ratio measurements between the manual and artificial intelligence methods. Difference values are distributed around the median for all 3 groups: non-surgical, surgical without hardware, and surgical with hardware. A. Violin plots of the difference in ratios for the 3 groups. B. Violin plots of the difference in ratios for the non-surgical and surgical groups. C. Violin plots of the difference in ratios for the surgical group, comparing cases with and without hardware.

Finally, measurement values for the 5 ratio items were compared in a similar fashion using an unpaired *t*-test (Fig. 10). For the non-surgical group, the mean error between human- and AI-landmarked measurements was 0.0285 (SD, 0.0275); for the surgical group without hardware, the mean was 0.0188 (SD, 0.0158); and for the surgical group with hardware, the mean was 0.0313 (SD, 0.0297). The postoperative group without hardware exhibited the lowest error. No significant difference was noted between the non-surgical and surgical groups ($P>0.05$). However, a significant difference was noted between surgical subgroups based on the presence or absence of hardware ($P<0.05$). These findings were graphically represented using a violin plot (Fig. 11).

Similarly, for the 21 angle items, the measurements derived from human- and AI-identified landmarks were compared using an unpaired *t*-test (Fig. 8). In the angular cephalometric measurements, the mean error for the non-surgical group was 1.81° (SD, 2.05°). The surgical group without hardware had a mean error of 1.60° (SD, 1.46°), while the surgical group with hardware had a mean error of 1.83° (SD, 1.94°). The postoperative group without hardware exhibited the lowest error. No significant difference was noted between the non-surgical and surgical groups ($P>0.05$), nor was a significant difference present between surgical subgroups based on the presence or absence of hardware ($P<0.05$). These findings were graphically represented using a violin plot (Fig. 9).

IV. Discussion

In recent years, AI technology has made remarkable advances in medical imaging, especially in dentistry and maxillofacial surgery. AI-automated landmarking is becoming increasingly important in orthodontics and maxillofacial surgery by providing important anatomical landmarks for treatment planning. This paper aims to present a comprehensive exploration of whether 3D AI-based automatic craniometric landmark identification is clinically useful.

Among the 6 items displaying significant differences, 2 measurements included the left gonion (GoL) or the right gonion (GoR): L1L to Me-GoL and GoR-Gn/S-N (P<0.05). In a previous study by these authors, the SDRs of these landmarks were particularly low. All detection methods from the "2014 Automatic Cephalometric X-Ray Landmark Detection: a grand challenge", conducted by the Institute of Electrical and Electronics Engineers International Symposium on Biomedical Imaging, misrepresented the gonion landmark. This resulted in a minimum error greater than 4 mm from the ground truth point (Lee JH et al., 2020). This discrepancy suggests that either the dataset failed to capture the high variability around these landmarks, or errors were present during manual annotation.

Furthermore, of the 6 significant items, 2 measurements involved the sella: S-N and GoR-Gn/S-N (P<0.05). The sella is a fiducial point located at the center of a cavity that, by definition, is a cephalometric landmark easily detectable on 2D head radiographs. However, it is difficult to identify on 3D CBCT because the skull structure does not create 3D contours. Makram et al. (Makram M and Kamel H, 2014) proposed a system that automatically localizes 20 three-dimensional hard tissue cephalometric landmarks using Reeb graphs. In their study, the mean error of the sella was notably high, at 2.6 mm. Given the challenges associated with the sella, various methods have been

attempted for landmark detection. Montúfar et al. (Montúfar J et al., 2018) employed a technique involving the circle adjustment of the sub-volume slice of the sella using Hough transformation to generate an anatomical geometric contour of the sella.

Four of the 6 items - R U6CP-L U6CP, R L3CP-L L3CP, Or_R-R U3CP, and L1L to Me-GoL (P<0.05) - included landmarks related to the teeth. The identification of landmarks associated with teeth can be affected by the surrounding anatomical structures, leading to potential errors even for clinicians. This is particularly true for the mandibular incisors, which are often difficult to discern due to their typical overlap with the maxillary incisors (Jeon S and Lee KC, 2021).

Since manual landmarking has been considered as a criterion for evaluating the accuracy of AI-based fully automated cephalometric analysis, inaccuracy related to inter- and intra-operator variability in manual landmarking identification can be seen as one of the major limitations.

In this study, 80 CBCT scans were analyzed. Using unpaired *t*-tests, comparisons were made between measurement values based on manual and AI landmarking. Of 53 measurements, statistically significant differences were observed for 4 lengths, 1 angle, and 1 ratio. When assessing the errors in assessments of length, angle, and ratio based on the designated measurement points on a 3-dimensional structure, the greatest error was found for length. In contrast, even errors at the measurement points had minimal impact on angles and ratios. Given that measured angles and ratios, more so than lengths, are valuable in planning orthodontic treatment or orthognathic surgery, the findings of this study are promising for clinical application.

The 80 CBCT scans were categorized into a non-surgical group (39 cases) and a surgical group (41 cases), with the latter including 9 cases without hardware and 32 cases with hardware. When comparing the manual and AI-based measurements for the 5 ratio items in each group, the cohort in which hardware was removed postoperatively exhibited the lowest measurement error across length,

angle, and ratio values. No significant differences were detected in any of the measurement groups when comparing preoperative and postoperative data. Postoperatively, however, significant differences in measurement errors for length and ratio were observed depending on whether hardware was present (Figs. 6-11).

Noise in CBCT images, along with metal artifacts from dental prostheses and implants, complicates the accurate delineation of teeth and bones. Hardware and screws were expected to introduce errors; however, the AI method performed well, regardless of hardware and screw presence. Consequently, this algorithm may serve as a valuable tool for assessing the extent of preoperative to postoperative change.

Minnema et al. (Minnema J et al., 2019) developed a deep learning algorithm based on mixed-scale density CNNs for the segmentation of teeth and bones on CBCT images containing metal structures. The algorithm appeared capable of excluding metal artifacts and accurately segmenting teeth and bone structures. These findings indicate that CNNs can identify voxel-level features in CBCT images that humans cannot distinguish.

The algorithm employed in the present study was based on deep learning techniques. Dot et al. (Dot G et al. 2020) compared the results of 11 studies to assess the accuracy and reliability of automatic CBCT cephalometric landmarking; the 2 algorithms that demonstrated the best performance employed deep learning methods. In this study, the deep learning-based algorithm reported an average error of less than 2 mm for all landmarks, comparable to the inter-operator variability observed in manual landmarking.

In this study, landmarks were not manually adjusted after 3D automatic landmarking. If manual adjustments were made to landmarks with a high likelihood of error (such as teeth, sellae, and gonions) following automatic landmarking, the accuracy of the measured values could be further

improved. Alternatively, hybrid analysis methods that determine specific landmarks through various approaches, such as the Montúfar sella measurement, can be employed (Montúfar J et al, 2018).

The time required to manually measure 65 landmarks was approximately 40-60 minutes per CBCT volume, although this time was impossible to fully capture because the workers took intermittent breaks. In contrast, the AI algorithm completed the task in 10.9 seconds, with the following PC specifications: GeForce 2080Ti, 64 GB RAM, and an Intel i7 CPU at 3.6 GHz. Since the landmark-based calculation of measurements is identical in the manual and AI methods, the AI method markedly reduced the time needed to identify a landmark and determine the measurement value.

The accuracy of measurements obtained with the deep learning-based CBCT automatic landmarking algorithm was comparable to that based on human-identified landmarks. By decreasing the time needed to calculate these measurements, the use of such an algorithm can improve the efficiency of diagnosis and treatment.

In this study, measurements of length demonstrated the lowest accuracy. However, as angles and ratios are more commonly utilized than length in patient diagnosis, the findings confirm that employing measurements derived from AI-based landmarks is suitable for diagnostic purposes.

In the comparison between surgical and non-surgical groups, no significant differences were found in linear measurements, angular measurements, or ratio parameters. Similarly, no significant differences were observed between preoperative and postoperative data for any patient group. Additionally, the presence of skeletal deformity did not impact the accuracy of automatic landmark identification.

The technology discussed in this report is anticipated to increase clinician efficiency and minimize diagnostic errors. This will facilitate the use of 3D cephalometric analyses for clinicians

of all experience levels. Furthermore, the availability of a user-friendly, web-based application for 3D automatic landmarking will broaden access for clinicians. At present, no clear standard exists for 3D cephalometric analysis, largely due to the time and effort involved as well as constraints related to its application in corrective procedures, surgical diagnosis, and treatment planning. The findings of this study may assist clinicians in incorporating 3D cephalometric analysis into their practice, irrespective of their level of experience.

V. Conclusion

There were 4 differences in length, 1 difference in angle, and 1 difference in ratio, and in the surgical group, there was a significant difference in the measurement errors for length and ratio depending on the presence or absence of the hardware. It is necessary to re-evaluate whether the clinical application of the measurement values is meaningful by comparing the differences between human and between AI and human in the measurement values (length, angle, ratio). In conclusion, although there are still limitations, there is a clear advantage in terms of efficiency, and it can be used in the clinical range.

VI. References

Gribel BF, Gribel MN, Frazäo DC, McNamara JA Jr, Manzi FR (2011). Accuracy and reliability of craniometric measurements on lateral cephalometry and 3D measurements on CBCT scans. *Angle Orthod* 81: 26-35.

Lee SH, Kil TJ, Park KR, Kim BC, Kim JG, Piao Z, et al (2014). Three-dimensional architectural and structural analysis - a transition in concept and design from Delaire's cephalometric analysis. *Int J Oral Maxillofac Surg* 43: 1154-60.

Olszewski R, Cosnard G, Macq B, Mahy P, Reychler H (2006). 3D CT-based cephalometric analysis: 3D cephalometric theoretical concept and software. *Neuroradiology* 48: 853-62.

Mah JK, Huang JC, Choo H (2010). Practical applications of cone-beam computed tomography in orthodontics. *J Am Dent Assoc* 141 Suppl 3: 7S-13S.

Lindner C, Wang CW, Huang CT, Li CH, Chang SW, Cootes TF (2016). Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Sci Rep* 6: 33581.

Vandaele R, Aceto J, Muller M, Peronnet F, Debat V, Wang CW, et al (2018). Landmark detection in 2D bioimages for geometric morphometrics: a multi-resolution tree-based approach. *Sci Rep* 8: 538.

Lagravere MO, Low C, Flores-Mir C, Chung R, Carey JP, Heo G, et al (2010). Intraexaminer and interexaminer reliabilities of landmark identification on digitized lateral cephalograms and formatted 3-dimensional cone-beam computerized tomography images. *Am J Orthod Dentofacial Orthop* 137: 598-604.

Hassan B, Nijkamp P, Verheij H, Tairie J, Vink C, van der Stelt P, et al (2013). Precision of identifying cephalometric landmarks with cone beam computed tomography in vivo. *Eur J Orthod* 35: 38-44.

Ghowsi A, Hatcher D, Suh H, Wile D, Castro W, Krueger J, et al (2022). Automated landmark identification on cone-beam computed tomography: accuracy and reliability. *Angle Orthod* 92: 642-54.

Polizzi A, Leonardi R. Automatic cephalometric landmark identification with artificial intelligence : An umbrella review of systematic reviews (2024). *J Dent*. 146:105056.

Jeon S, Lee KC (2021). Comparison of cephalometric measurements between conventional and automatic cephalometric analysis using convolutional neural network. *Prog Orthod* 22: 14.

Gupta A, Kharbanda OP, Sardana V, Balachandran R, Sardana HK (2016). Accuracy of 3D cephalometric measurements based on an automatic knowledge-based landmark detection algorithm. *Int J Comput Assist Radiol Surg* 11: 1297-309.

Barrett JF, Keat N (2004). Artifacts in CT: recognition and avoidance. *Radiographics* 24: 1679-91.

Hung K, Yeung AW, Tanaka R, Bornstein MM (2020). Current applications, opportunities, and limitations of AI for 3D imaging in dental research and practice. *Int J Environ Res Public Health* 17: 4424.

Torosdagli N, Liberton DK, Verma P, Sincan M, Lee JS, Bagci U (2019). Deep geodesic learning for segmentation and anatomical landmarking. *IEEE Trans Med Imaging* 38: 919-31.

Lee JH, Yu HJ, Kim MJ, Kim JW, Choi J (2020). Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. *BMC Oral Health* 20: 270.

Makram M, Kamel H (2014). Reeb graph for automatic 3D cephalometry. *Int J Image Process* 8: 17-29.

Montúfar J, Romero M, Scougall-Vilchis RJ (2018). Hybrid approach for automatic cephalometric landmark annotation on cone-beam computed tomography volumes. *Am J Orthod Dentofacial Orthop* 154: 140-50.

Minnema J, van Eijnatten M, Hendriksen AA, Liberton N, Pelt DM, Batenburg KJ, et al (2019). Segmentation of dental cone-beam CT scans affected by metal artifacts using a mixed-scale dense convolutional neural network. *Med Phys* 46: 5027-35.

Dot G, Rafflenbeul F, Arbotto M, Gajny L, Rouch P, Schouman T (2020). Accuracy and reliability of automatic three-dimensional cephalometric landmarking. *Int J Oral Maxillofac Surg* 49: 1367-78.

국문 요약

3 차원컴퓨터단층촬영(CBCT)을 이용한 자동 랜드마킹 알고리즘의 임상적 타당성 및 정확성

(지도교수 유 형 석)

연세대학교 대학원 치의학과

박 정 은

본 연구의 목적은 수동 랜드마킹과 딥러닝으로 학습된 랜드마킹을 기반으로 하는 3 차원 CBCT 두부 계측값을 비교하여, 딥러닝 기반 CBCT 자동랜드마킹 알고리즘의 임상적 유효성 및 정확성 평가하는 것이다.

80 개의 CBCT 를 비수술 그룹(39 case), 하드웨어가 있는 수술그룹(32 case), 하드웨어가 없는 수술 그룹(9 case) 로 구분하였으며, 총 53 개의 계측값(measurements)(27 개의 길이(Length), 21 개의 각도(angle), 5 개의 비율(ratio))을 측정하여, 다음과 같은 결론을 얻었다.

1. 사람과 인공지능이 찍은 landmarking 을 기반으로 하는 계측값을 비교했을 때, 6 개 항목 R U6CP-L U6CP (P=0.030), R L3CP-L L3CP (P=0.000), S-N

(P=0.002), Or_R-R U3CP (P=0.041), L1L to Me-GoL (P=0.045), GoR-Gn/S-N (P=0.025))에서 통계적으로 유의미한 차이가 있었다.

2. 비수술 그룹과 수술 그룹에 대하여 사람과 인공지능이 landmarking 을 기반으로 하는 계측값을 비교하였을 때, 하드웨어가 없는 수술 그룹이 가장 낮은 계측치 오차를 가졌다.

3. 65 개의 landmarking 에 소요되는 시간은, 수동 방식의 경우, CT 볼륨 하나 당 대략 40-60 분인 반면, 인공지능 방식의 경우, 10.9 초 (PC 사양: a GeForce 2080Ti, 64G RAM, and Intel i7 CPU 3.6 GHz)였다.

길이에서 4 개, 각도, 비율에서 각 1 개씩 차이가 있었고, 수술 그룹에서는 하드웨어 유무에 따라 길이와 비율에 대한 계측치 오차가 유의미한 차이를 보였다. 계측값(길이, 각도, 비율)에서 사람 간의 차이와 AI 와 사람간 차이를 비교하여, 측정값의 임상적 적용 의미가 있는지 다시 평가할 필요가 있다. 결론적으로 아직까지 한계는 존재하지만, 효율성 측면에서는 확실한 이점이 존재하며, 임상적 범위에서 사용 가능하다.

핵심되는 말 : 3 차원컴퓨터단층촬영(CBCT), 해부학적 계측점, 자동 두부 방사선 계측, 딥러닝, 악교정 수술