Original Article

pISSN 2234-7518 • eISSN 2005-372X
https://doi.org/10.4041/kjod23.075
Korean J Orthod 2024;54(1):48-58

THE KOREAN JOURNAL of ORTHODONTICS
KJO

Accuracy of posteroanterior cephalogram landmarks and measurements identification using a cascaded convolutional neural network algorithm: A multicenter study

Sung-Hoon Han[a]
Jisup Lim[b]
Jun-Sik Kim[b]
Jin-Hyoung Cho[c]
Mihee Hong[d]
Minji Kim[e]
Su-Jung Kim[f]
Yoon-Ji Kim[g]
Young Ho Kim[h]
Sung-Hoon Lim[i]
Sang Jin Sung[g]
Kyung-Hwa Kang[j]
Seung-Hak Baek[k]
Sung-Kwon Choi[j]
Namkug Kim[b]

[a]Department of Orthodontics, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea
[b]Department of Convergence Medicine, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
[c]Department of Orthodontics, School of Dentistry, Chonnam National University, Gwangju, Korea
[d]Department of Orthodontics, School of Dentistry, Kyungpook National University, Daegu, Korea
[e]Department of Orthodontics, College of Medicine, Ewha Womans University, Seoul, Korea
[f]Department of Orthodontics, Kyung Hee University School of Dentistry, Seoul, Korea
[g]Department of Orthodontics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
[h]Department of Orthodontics, Institute of Oral Health Science, Ajou University School of Medicine, Suwon, Korea
[i]Department of Orthodontics, College of Dentistry, Chosun University, Gwangju, Korea
[j]Department of Orthodontics, School of Dentistry, Wonkwang University, Iksan, Korea
[k]Department of Orthodontics, School of Dentistry, Dental Research Institute, Seoul National University, Seoul, Korea

Objective: To quantify the effects of midline-related landmark identification on midline deviation measurements in posteroanterior (PA) cephalograms using a cascaded convolutional neural network (CNN). **Methods:** A total of 2,903 PA cephalogram images obtained from 9 university hospitals were divided into training, internal validation, and test sets (n = 2,150, 376, and 377). As the gold standard, 2 orthodontic professors marked the bilateral landmarks, including the frontozygomatic suture point and latero-orbitale (LO), and the midline landmarks, including the crista galli, anterior nasal spine (ANS), upper dental midpoint (UDM), lower dental midpoint (LDM), and menton (Me). For the test, Examiner-1 and Examiner-2 (3-year and 1-year orthodontic residents) and the Cascaded-CNN models marked the landmarks. After point-to-point errors of landmark identification, the successful detection rate (SDR) and distance and direction of the midline landmark deviation from the midsagittal line (ANS-mid, UDM-mid, LDM-mid, and Me-mid) were measured, and statistical analysis was performed. **Results:** The cascaded-CNN algorithm showed a clinically acceptable level of point-to-point error (1.26 mm vs. 1.57 mm in Examiner-1 and 1.75 mm in Examiner-2). The average SDR within the 2 mm range was 83.2%, with high accuracy at the LO (right, 96.9%; left, 97.1%), and UDM (96.9%). The absolute measurement errors were less than 1 mm for ANS-mid, UDM-mid, and LDM-mid compared with the gold standard. **Conclusions:** The cascaded-CNN model may be considered an effective tool for the auto-identification of midline landmarks and quantification of midline deviation in PA cephalograms of adult patients, regardless of variations in the image acquisition method.

Key words: Artificial intelligence, Convolutional neural network, Posteroanterior cephalograms

Received March 31, 2023; Revised September 7, 2023; Accepted October 10, 2023.

Corresponding author: Sung-Kwon Choi.
Assistant Professor, Department of Orthodontics, School of Dentistry, Wonkwang University, 460 Iksan-daero, Iksan 54538, Korea.
Tel +82-63-859-2962 **e-mail** chsk6206@wku.ac.kr
Corresponding author: Namkug Kim.
Professor, Department of Convergence Medicine, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea.
Tel +82-2-3010-6573 **e-mail** namkugkim@gmail.com

Sung-Hoon Han and Jisup Lim contributed equally to this work (as co-first authors).

How to cite this article: Han SH, Lim J, Kim JS, Cho JH, Hong M, Kim M, Kim SJ, Kim YJ, Kim YH, Lim SH, Sung SJ, Kang KH, Baek SH, Choi SK, Kim N. Accuracy of posteroanterior cephalogram landmarks and measurements identification using a cascaded convolutional neural network algorithm: A multicenter study. Korean J Orthod 2024;54(1):48-58. https://doi.org/10.4041/kjod23.075

INTRODUCTION

Artificial intelligence (AI) refers to algorithms that imitate human intelligence to recognize, solve problems, and make efficient decisions.[1,2] Artificial neural networks (ANNs) are computing systems that mimic biological neural networks found in animal brains. A convolutional neural network (CNN), a deep learning model within ANNs, extracts data characteristics and identifies their patterns, making it particularly suitable for processing visual data and addressing challenges encountered in image or video data processing using regular deep learning algorithms.[3,4]

Cephalometric analysis is an essential component of diagnostic processes but can be time-consuming and prone to analytical errors when performed by non-experts.[5-7] Therefore, there have been ongoing efforts to use the image recognition ability of CNN for automatic cephalometric landmark identification. Convolutional neural networks are designed to mimic the hierarchical organization of the human visual cortex for processing visual information and have proven successful in various image recognition domains, including cephalometric analysis.[8] Recent studies have reported high accuracy in automatically identifying cephalometric landmarks in lateral cephalograms using CNN.[9-14] Nevertheless, research on posteroanterior (PA) cephalometric analysis using cascaded CNNs, especially concerning measurement values, has been limited.

Posteroanterior cephalograms have been used to evaluate landmark deviation from the midsagittal reference plane in terms of angle, amount, and direction. However, studies applying AI to PA cephalograms are rare. Muraev et al.[15] reported that the accuracy of landmark identification using AI was comparable to that achieved by a human expert. Gil et al.[16] reported that the mean error of landmark identification by AI was 1.52 mm and the successful detection rate (SDR) based on errors within 2 mm was 83.3%. In contrast, validation of the reference planes is required to obtain accurate measurements of PA cephalometric variables.

Previous studies have some limitations: (1) When the gold standard for AI training is set by a single operator[11,15] or by the average coordinate value of 2 operators,[12] potential bias may be introduced. Therefore, establishing a gold standard through mutual agreement between 2 experts is essential. (2) Examining identification errors in the x- and y-coordinates and the distribution of SDR for midline landmarks is necessary. (3) For landmark identification errors and measurement accuracy, midline variables should be investigated among AI and multiple human examiners (e.g., human examiner-1 and human examiner-2) using multiple comparison test.

Therefore, this study aimed to quantify the effects of midline-related landmark identification on midline deviation measurements in PA cephalograms using a cascaded CNN algorithm.

MATERIALS AND METHODS

Subjects

A total of 2,930 PA cephalograms were obtained from 9 institutions: Seoul National University Dental Hospital (SNUDH; n = 1,591), Kyung Hee University Dental Hospital (KHUDH; n = 607), Kyungpook National University Dental Hospital (KNUDH; n = 79), Asan Medical Center (AMC; n = 205), Ajou University Dental Hospital (AUDH; n = 116), Korea University Dental Hospital (KUDH; n = 97), Chonnam National University Dental Hospital (CNUDH; n = 120), Wonkwang University Dental Hospital (WUDH; n = 67), and Ewha Womans University Medical Center (EUMC; n = 48). This study was reviewed and approved by the Institutional Review Board (IRB) of each institution (SNUDH, ERI18002; KHUDH, D19-007-003; KNUDH, KNUDH-2019-03-02-00; AMC, 2019-0927; AUDH, AJIRB-MED-MDB-19-039; KUDH, 2019AN0166; CNUDH, CNUDH-2019-004; WUDH, WKDIRB202010-06; and EUMC, EUMC 2019-04-017-009).

The inclusion criteria were as follows: 1) adult orthodontic patients with complete facial growth; 2) patients who underwent orthognathic surgery between 2013 and 2020; and 3) patients with permanent dentition. The exclusion criteria were: 1) patients who had craniofacial syndromes or systemic diseases; and 2) patients whose PA cephalogram had poor image quality, making identification of landmarks impossible.

Among the 2,930 images, 2,903 PA cephalograms were used as the final samples. All images were converted to 8-bit grayscale images (2k × 2k pixels) and saved in the DICOM file format.

Determination of PA landmarks and the gold standard

Definitions of the bilateral landmarks, including the frontozygomatic suture point (FZS) and latero-orbitale (LO), and midline landmarks, including the crista galli (Cg), anterior nasal spine (ANS), upper dental midpoint (UDM), lower dental midpoint (LDM), and menton (Me), are shown in Figure 1 and Table 1.

To set the human gold standard, 2 orthodontic professors with 12-year and 8-year clinical experience (SHH and SKC) marked the landmarks using the V-Ceph 8.0 program (Osstem, Seoul, Korea). The 2 examiners reached an agreement before marking the landmarks in 2,903 PA cephalograms. Subsequently, 2,903 images were randomly divided into training (n = 2,150), internal validation (n = 376), and test sets (n = 377) (Figure 2).

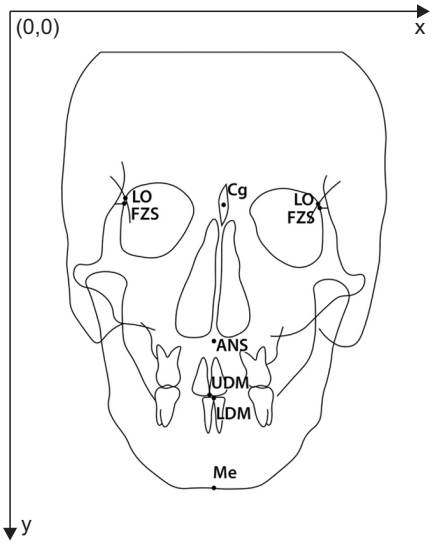

Figure 1. The posteroanterior cephalometric landmarks used in this study.
LO, latero-orbitale; FZS, frontozygomatic suture point; Cg, crista galli; ANS, anterior nasal spine; UDM, upper dental midpoint; LDM, lower dental midpoint; Me, menton.

Table 1. Definition of the posteroanterior cephalometric landmarks used in this study

Landmarks	Definition
Midline landmarks	
Cg	The middle point of the Cg
ANS	The tip of the ANS
UDM	The midpoint between the incisal margins of maxillary central incisors
LDM	The midpoint between the incisal margins of the mandibular central incisors
Me	The most inferior point of the symphysis of the mandible
Bilateral landmarks	
FZS	The intersection of the frontozygomatic suture and the inner rim of the orbit
LO	The intersection between the external orbital contour laterally and the oblique line

Cg, crista galli; ANS, anterior nasal spine; UDM, upper dental midpoint; LDM, lower dental midpoint; Me, menton; FZS, frontozygomatic suture point; LO, latero-orbitale.

Training and internal validation of the algorithm

Deep learning training using the cascaded CNN algorithm consisted of (1) determination of the region of interest (ROI) and (2) landmark prediction (Figure 3). First, RetinaNet[17] was used to extract the ROI with the x- and y-coordinates of the landmark center. The ROI was set to 2 sizes (256 × 256 and 512 × 512). Secondly, U-net[18] was used to detect the exact location of the ROI patch formed in the first step.

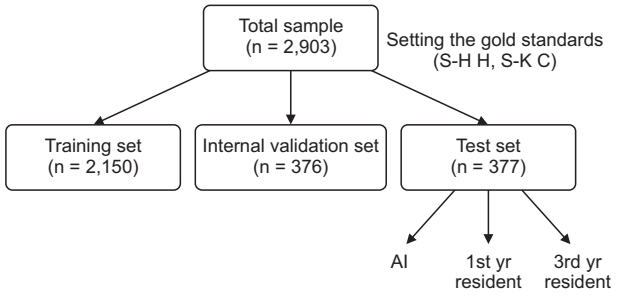

Figure 2. Flow chart showing sample allocation and study design.
AI, artificial intelligence.

RetinaNet adopts Resnet-50[19] as the backbone and uses it for learning; pre-trained weights are not used for training. The Adam optimizer combines momentum and exponentially weighted moving average gradient methods to update the network weights. The learning rate was initially set to 0.0001 and then decreased by a factor of 10 when the validation set accuracy plateaued. In total, the learning rate was decreased 3 times to complete the training.

Various augmentation methods, such as Gaussian noise, random brightness, blurring, random contraction, flipping, and random rotation, have been used in deep-learning model training. An internal validation test (n = 376) was performed to determine the optimal parameter values for machine learning.

Comparison of accuracy of landmark identification between cascaded CNN and human examiners

The cascaded CNN algorithm modeled auto-identified landmarks on the PA cephalogram images selected as the test set (n = 377). To compare the accuracy of landmark identification between the AI and orthodontic residents, 2 examiners (a third-year resident [HYS, Examiner-1] and a first-year resident [MSK, Examiner-2]) marked the landmarks on PA cephalogram images using the same conditions and methods used by the human gold standard.

Point-to-point errors in landmark identification by 2 examiners (residents) and the AI against the gold standard (2 orthodontic professors) were measured. The position of each landmark was mapped using the x- and y-coordinates to derive the mean error against the gold standard.

The inter-rater reliability test between Examiner-1 and Examiner-2 showed very high intraclass correlation coef-

https://doi.org/10.4041/kjod23.075

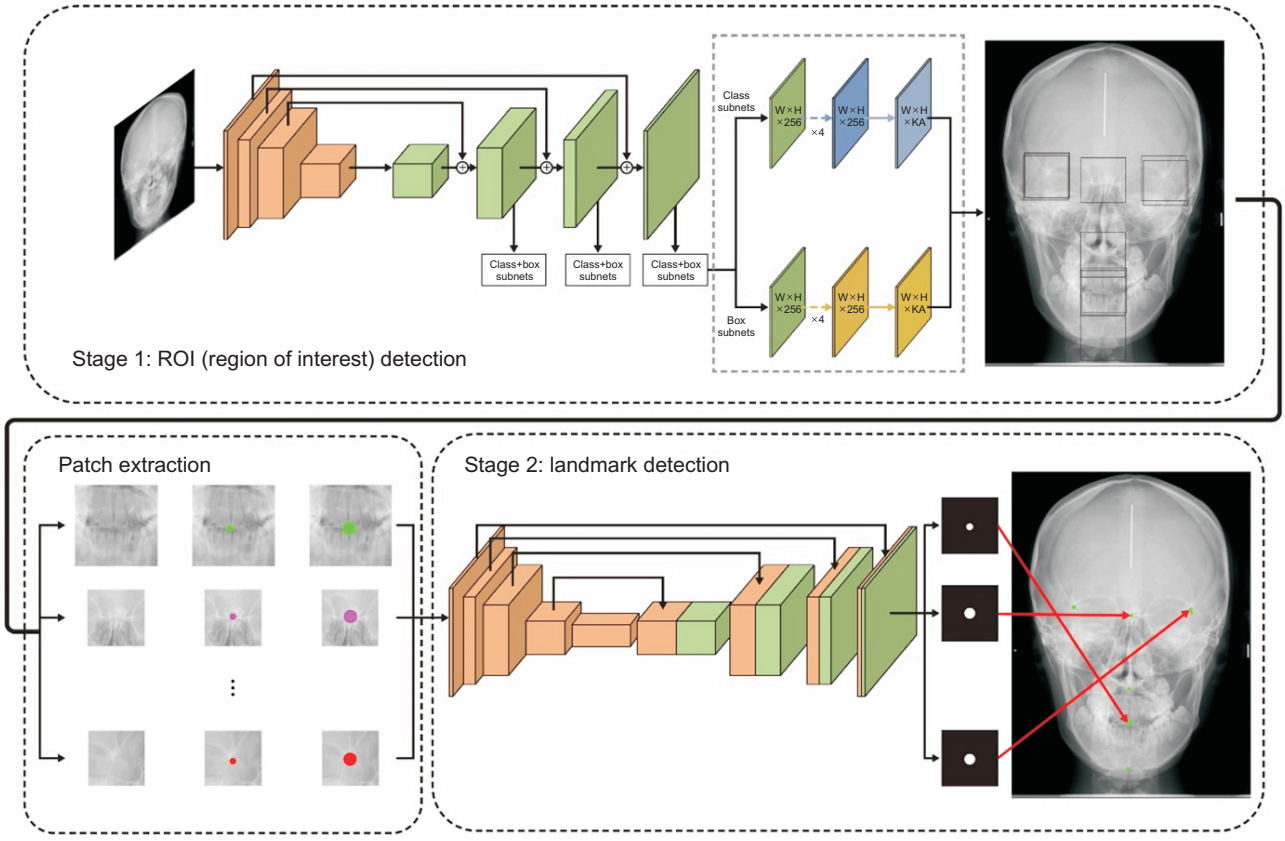

Figure 3. Cascaded convolutional neural network algorithm used in this study. Stage 1, the region of interest detection to propose the area of interest; stage 2, the landmark prediction to find the exact location of landmarks.
W, width; H, height; K, number of object classes; A, number of anchors.

ficient values (≥ 0.99) in all 9 landmarks.

The SDR was set as the percentage of landmarks within a specific range from the gold standard (< 1, < 2, and < 3 mm).

Comparison of accuracy of measurements between cascaded CNN and human examiners

After setting the horizontal reference line connecting the bilateral landmarks (right and left LO points and right and left FZS points), reorientation of the PA cephalograms was performed. To accurately measure PA cephalometric variables, the first step was to determine which landmarks (LO vs. FZS) had the highest identification accuracy. The midsagittal line was defined as the line passing through the Cg and intersecting perpendicularly with the horizontal reference lines (LO and FZS lines).

The shortest distances from midline landmarks to the midsagittal line (ANS-mid, UDM-mid, LDM-mid, and Me-mid) were measured. The deviation in the right direction was set to have a negative (-) value, whereas the deviation in the left direction was set to have a positive (+) value. The absolute values were also measured, regardless of the direction of deviation.

Statistical analysis

A one-way analysis of variance with Tukey's test was performed using SPSS software (version 12.0; SPSS Inc., Chicago, IL, USA). The statistical significance level was set at $P < 0.05$.

RESULTS

Comparison of accuracy of landmark identification between AI and human examiners

The mean point-to-point error of nine landmarks appeared to be 1.26 mm, 1.57 mm, and 1.75 mm for AI, Examiner-1, and Examiner-2, respectively. Artificial intelligence showed significantly higher accuracy than Examiner-2 for the identification of ANS, right and left FZS points, and right and left LO points ($P < 0.001$, $P < 0.01$, $P < 0.001$, $P < 0.05$, and $P < 0.001$, respectively) (Figure 4, Table 2). Although AI showed low accuracy in the identification of the right and left FZS points, it still showed higher accuracy than both Examiner-1 and Examiner-2 (1.87 mm vs. 2.26 mm and 2.33 mm, $P <$

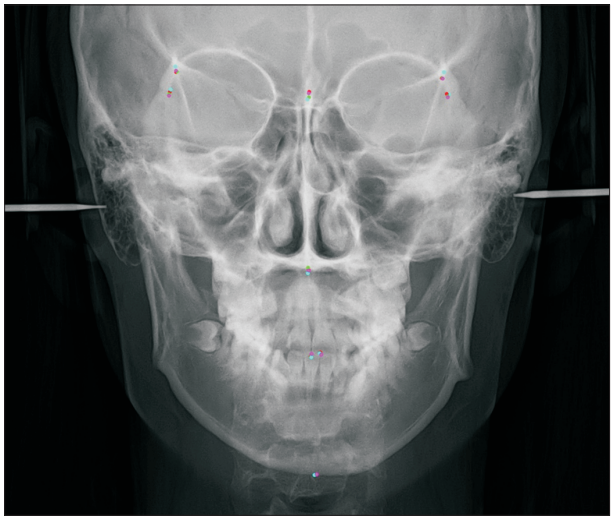

Figure 4. Examples of superimposition of the identified posteroanterior cephalometric landmarks. Red, gold standard; green, auto-identification by cascaded convolutional neural network algorithm; pink, Examiner-1; sky blue, Examiner-2.

0.01; 2.01 mm vs. 3.02 mm and 3.20 mm, $P < 0.001$). However, there was no difference in the accuracy of the identification of Cg, UDM, LDM, and Me between AI, Examiner-1, and Examiner-2.

All 3 groups showed similar patterns in the accuracy of each measurement point: high accuracy in the UDM, LDM, and right and left LO points, but low accuracy in the Cg, Me, and right and left FZS points (Figure 4, Table 2).

In terms of errors in the x-coordinate, there were no significant differences in the horizontal positioning of the Cg, ANS, UDM, LDM, and Me between the AI and human examiners. Artificial intelligence showed significantly higher identification accuracy in the horizontal positioning of the left FZS and left LO point than Examiner-2 ($P < 0.001$ and $P < 0.01$; Table 3). The error values of the horizontal positioning of all landmarks using AI were less than 1 mm, except for the UDM (Table 3).

In terms of errors in the y-coordinate, there were no significant differences in the vertical positioning of the UDM and Me between AI and human examiners. Artificial intelligence showed significantly higher identification accuracy in vertical positioning than Examiner-2 (ANS, LDM, right and left FZS points, and right and left

Table 2. The point-to-point error between artificial intelligence and human examiners

Landmarks	Point-to-point error (mm)						*P* value	Multiple comparison
	AI		Examiner-1		Examiner-2			
	Mean	SD	Mean	SD	Mean	SD		
Cg	1.76	1.98	1.97	2.51	1.73	1.52	0.215	
ANS	1.31	1.52	1.30	2.32	1.80	1.71	< 0.001***	(E1, AI) < E2
UDM	0.54	1.15	0.75	2.01	0.59	0.75	0.103	
LDM	0.97	2.27	1.08	2.15	0.94	1.27	0.594	
Me	1.61	2.59	1.53	1.66	1.34	1.54	0.264	
FZS-R	1.87	1.74	2.26	1.97	2.33	1.59	0.001**	AI < (E1, E2)
FZS-L	2.01	2.24	3.02	2.59	3.20	2.22	< 0.001***	AI < (E1, E2)
LO-R	0.58	1.15	0.71	1.62	0.82	0.64	0.022*	(AI, E1) < (E1, E2)
LO-L	0.70	1.60	0.78	2.14	1.15	1.34	0.001**	(AI, E1) < E2
P value	< 0.001***		< 0.001***		< 0.001***			
Multiple comparison	(UDM, LO R&L) < (LO R&L, LDM) < (LDM, ANS) < (ANS, Me) < (Me, Cg, FZS R & L)		(LO R&L, UDM, LDM) < (LDM, ANS, Me) < (Cg, FZS-R) < FZS-L		(UDM, LO R) < (LO-R, LDM) < (LDM, LO-L) < (LO-L, Me) < (Cg, ANS) < FZS-R < FZS-L			
Total	1.26	1.94	1.57	1.66	1.75	2.34		

A one-way analysis of variance followed by Tukey's test was performed.
SD, standard deviation; Cg, crista galli; ANS, anterior nasal spine; UDM, upper dental midpoint; LDM, lower dental midpoint; Me, menton; FZS, frontozygomatic suture point; R, right; L, left; LO, latero-orbitale; AI, artificial intelligence; E1, examiner-1; E2, examiner-2.
*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

https://doi.org/10.4041/kjod23.075

Table 3. The x-coordinate error (mm) between artificial intelligence and human examiners

Landmarks	AI		Examiner-1		Examiner-2		*P* value	Multiple comparison
	Mean	SD	Mean	SD	Mean	SD		
Cg	0.52	1.13	0.55	0.94	0.50	0.91	0.766	
ANS	0.46	1.07	0.42	0.74	0.56	0.72	0.057	
UDM	1.37	1.02	1.31	0.52	1.41	0.60	0.683	
LDM	0.31	1.41	0.25	1.24	0.28	1.35	0.485	
Me	0.54	1.69	0.57	1.54	0.63	1.46	0.633	
FZS-R	0.79	0.94	0.61	0.59	0.72	0.69	0.002**	(E1, E2) < (E2, AI)
FZS-L	0.86	1.43	1.20	1.05	1.26	1.26	< 0.001***	AI < (E1, E2)
LO-R	0.24	0.53	0.23	0.21	0.30	0.25	0.026*	(E1, AI) < (AI, E2)
LO-L	0.28	1.04	0.24	0.54	0.42	0.68	0.004**	(E1, AI) < E2

A one-way analysis of variance followed by Tukey's test was performed.
AI, artificial intelligence; SD, standard deviation; Cg, crista galli; ANS, anterior nasal spine; UDM, upper dental midpoint; LDM, lower dental midpoint; Me, menton; FZS, frontozygomatic suture point; R, right; L, left; LO, latero-orbitale; E1, examiner-1; E2, examiner-2.
*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Table 4. The y-coordinate error (mm) between artificial intelligence and human examiners

Landmarks	AI		Examiner-1		Examiner-2		*P* value	Multiple comparison
	Mean	SD	Mean	SD	Mean	SD		
Cg	1.55	1.76	1.93	2.03	1.80	1.60	0.012*	(AI, E2) < (E2, E1)
ANS	1.11	1.21	1.22	1.74	1.87	1.88	< 0.001***	(AI, E1) < E2
UDM	0.31	0.59	0.40	0.50	0.45	0.70	0.366	
LDM	0.35	1.87	0.64	0.58	0.53	0.67	< 0.001***	AI < E2 < E1
Me	0.58	2.08	0.73	0.81	0.60	0.94	0.187	
FZS-R	1.61	1.55	2.34	1.52	2.41	1.75	< 0.001***	AI < (E1, E2)
FZS-L	1.73	1.81	3.02	1.74	3.26	2.13	< 0.001***	AI < (E1, E2)
LO-R	0.47	1.04	0.64	0.60	0.83	0.78	< 0.001***	AI < E1 < E2
LO-L	0.58	1.25	0.68	0.90	1.17	1.35	< 0.001***	(AI, E1) < E2

A one-way analysis of variance followed by Tukey's test was performed.
AI, artificial intelligence; SD, standard deviation; Cg, crista galli; ANS, anterior nasal spine; UDM, upper dental midpoint; LDM, lower dental midpoint; Me, menton; FZS, frontozygomatic suture point; R, right; L, left; LO, latero-orbitale; E1, examiner-1; E2, examiner-2.
*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

LO points, all $P < 0.001$; Table 4). The error values in the vertical positioning of the UDM, LDM, Me, and right and left LO points by the AI were less than 1 mm (Table 4).

Distribution of successful detection rate for AI-identified landmarks

The mean SDRs of all the AI-identified landmarks were 65.8% at < 1 mm, 83.2% at < 2, and 89.6% at < 3 mm, respectively (Table 5). Highly accurate SDR values (≥ 90% within 2 mm range) were found at the right LO point (96.9%), left LO point (97.1%), and UDM (96.9%), whereas moderate SDR values (≤ 70% within 2 mm range) were found at the right FZS point (66.4%) and left FZS point (68.2%) (Figure 5, Table 5).

Comparison of measurement accuracy between AI and human examiners

Because the LO points showed higher accuracy than the FZS points (Tables 2–5), the PA cephalograms were reoriented using the LO and midsagittal lines. The perpendicular distances between the midline landmarks and the midsagittal line (ANS-mid, UDM-mid, LDM-mid, and Me-mid) were then measured (Figure 6).

When the measurements by AI and human examin-

 https://doi.org/10.4041/kjod23.075

Table 5. Distribution of the successful detection rate in artificial intelligence

Landmarks	SDR (%)		
	< 1 mm	< 2 mm	< 3 mm
Cg	50.0	71.6	79.9
ANS	52.6	82.6	93.5
UDM	93.8	96.9	97.1
LDM	79.7	89.1	91.9
Me	52.1	80.2	87.8
FZS-R	40.6	66.4	81.0
FZS-L	38.8	68.2	79.2
LO-R	94.0	96.9	97.9
LO-L	90.4	97.1	97.9
Average	65.8	83.2	89.6

SDR, successful detection rate; Cg, crista galli; ANS, anterior nasal spine; UDM, upper dental midpoint; LDM, lower dental midpoint; Me, menton; FZS, frontozygomatic suture point; R, right; L, left; LO, latero-orbitale.

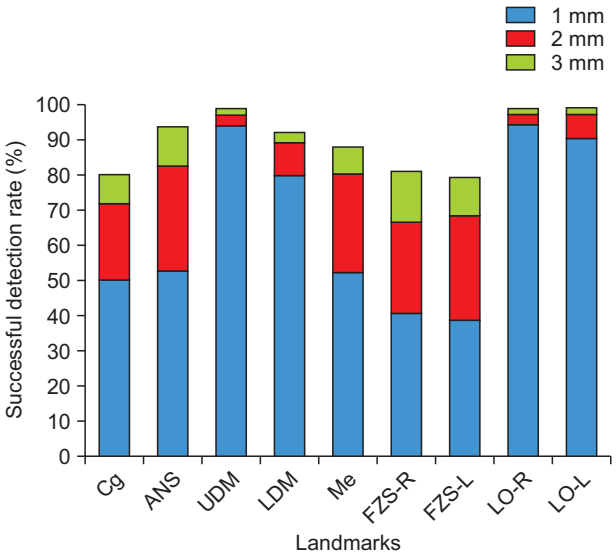

Figure 5. Comparison of the successful detection rate within the range of 1.0 mm, 2.0 mm, and 3.0 mm in each landmark.
Cg, crista galli; ANS, anterior nasal spine; UDM, upper dental midpoint; LDM, lower dental midpoint; Me, menton; FZP-R, frontozygomatic suture point right; FZP-L, frontozygomatic suture point left; LO-R, latero-orbitale right; LO-L, latero-orbitale left.

ers were compared to those of the gold standard, the absolute measurement errors were < 1 mm in ANS-mid, UDM-mid, and LDM-mid and were also within the clinically relevant range (< 2.0 mm) in the Me group (Table 6). Artificial intelligence did not exhibit significant differences between the LDM-mid and Me-mid from the human examiners. However, Examiner-2 had a higher error in LDM-mid and Me-mid than Examiner-1 (all $P < 0.01$; Table 6).

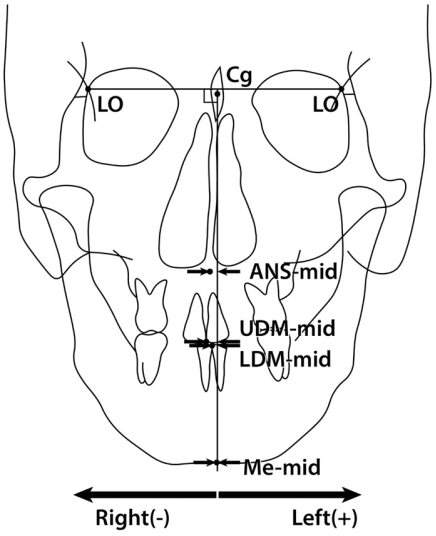

Figure 6. Landmarks and the midsagittal reference line for measurements of the distance and direction of landmark deviation from the midsagittal line (ANS-mid, UDM-mid, LDM-mid, and Me-mid) on posteroanterior cephalogram images.
LO, latero-orbitale; Cg, crista galli; ANS, anterior nasal spine; UDM, upper dental midpoint; LDM, lower dental midpoint; Me, menton; mid, midsagittal line.

In terms of the deviation direction of the midline landmarks from the midsagittal line (ANS-mid, UDM-mid, LDM-mid, and Me-mid), AI identified the midline landmarks within a range of 0.2 mm compared to the gold standard (ANS, LDM, and Me, left-sided positioning, range: 0.09–0.16; UDM to the right-sided positioning, –0.07 mm) (Table 7). However, human examiners identified all the landmarks to the right-sided positioning compared to the gold standard within a range of 0.3 mm in Examiner-1 and within a range of 0.7 mm in Examiner-2 (Table 7).

DISCUSSION

Comparison of accuracy of landmark identification between AI and human examiners

The cascaded CNN algorithm demonstrated clinically acceptable and higher accuracy in terms of PA cephalogram landmark identification error (1.26 mm vs. 1.57 mm in Examiner-1 and 1.75 mm in Examiner-2, as shown in Table 2).

In this study, AI showed high or good accuracy in the

https://doi.org/10.4041/kjod23.075

Table 6. Comparison of the absolute measurement error of posteroanterior cephalometric variables between artificial intelligence and human examiners

Measurements	Distance (mm)						*P* value	Multiple comparison
	AI-GS		E1-GS		E2-GS			
	Mean	SD	Mean	SD	Mean	SD		
ΔANS-mid	0.66	1.06	0.61	0.56	0.72	0.63	0.168	
ΔUDM-mid	0.71	1.18	0.87	0.73	0.64	0.60	0.001**	(E2-GS, AI-GS) < E1-GS
ΔLDM-mid	0.91	1.42	0.80	0.99	1.09	1.20	0.005**	(E1-GS, AI-GS) < (AI-GS, E2-GS)
ΔMe-mid	1.53	1.88	1.30	1.13	1.69	1.44	0.002**	(E1-GS, AI-GS) < (AI-GS, E2-GS)

A one-way analysis of variance followed by Tukey's test was performed.
AI, artificial intelligence; GS, gold standard; E1, examiner-1; E2, examiner-2; SD, standard deviation; ANS, anterior nasal spine; mid, midsagittal line; UDM, upper dental midpoint; LDM, lower dental midpoint; Me, menton.
**$P < 0.01$.

Table 7. Comparison of the mean measurement error of posteroanterior cephalometric variables between artificial intelligence and human examiners

Measurements	Distance (mm)						*P* value	Multiple comparison
	AI-GS		E1-GS		E2-GS			
	Mean	SD	Mean	SD	Mean	SD		
ΔANS-mid	0.09	1.24	−0.21	0.80	−0.09	0.95	< 0.001***	(E1-GS, E2-GS) < AI-GS
ΔUDM-mid	−0.07	1.38	−0.24	0.84	−0.46	1.03	< 0.001***	E2-GS < (E1-GS, AI-GS)
ΔLDM-mid	0.16	1.68	−0.19	1.26	−0.36	1.59	< 0.001***	(E2-GS, E1-GS) < AI-GS
ΔMe-mid	0.11	2.42	−0.22	1.71	−0.67	2.12	< 0.001***	E2-GS < (E1-GS, AI-GS)

A one-way analysis of variance followed by Tukey's test was performed.
A negative sign means right-side deviation.
AI, artificial intelligence; GS, gold standard; E1, examiner-1; E2, examiner-2; SD, standard deviation; ANS, anterior nasal spine; mid, midsagittal line; UDM, upper dental midpoint; LDM, lower dental midpoint; Me, menton.
***$P < 0.001$.

identification of the UDM (0.54 mm), right LO (0.58 mm), left LO (0.70 mm), LDM (0.97 mm), and ANS (1.31 mm) when clinical accuracy was defined as less than 1.5 mm. These findings indicate that AI may be better than first-year orthodontic residents for PA cephalometric landmark identification. However, the accuracy of identification of the Cg, Me, and right and left FZS (1.76 mm, 1.61 mm, 1.87 mm, and 2.01 mm) needs to be improved in future studies. The lower accuracy in identifying the Cg, Me, right, and left FZS can be attributed mainly to 2 factors: overlapping issues that occur when converting three-dimensional structures to two-dimensional structures (Cg, right, and left FZS) and errors in identifying points along the gentle curve of the mandibular lower border (Me).

The results reveal that AI exhibited less than 1 mm error in the horizontal positioning of all landmarks except the UDM (Table 3) and in the vertical positioning of the UDM, LDM, Me, and right and left LO points (Table 4). This suggests that most errors occurred in the vertical positioning of the PA cephalogram landmarks due to their anatomical features.

Several landmarks (Cg, Me, right and left FZS) in this study were identical to those in a previous study by Gil et al.[15] The cascaded CNN algorithm employed in this study showed higher accuracy compared to that study (1.55 mm, 0.58 mm, 1.61 mm, and 1.73 mm vs. 1.89 mm, 1.99 mm, 1.83 mm, and 1.96 mm, respectively).

Distribution of SDR for AI-identified landmarks

The results showed a relatively low SDR for the Cg and right and left FZS points (71.6%, 66.4%, and 66.2%, respectively), and a high SDR for the UDM, LDM, and right and left LO points (96.9%, 89.1%, 96.9%, and 97.1%, respectively) (Table 5). Therefore, the right and left LO points could be used as the horizontal reference line in the PA cephalometric analysis rather than the right and left FZS points (Table 5).

The cascaded CNN algorithm used in this study showed 83.2% of average SDR within the 2 mm range

(Table 5), which was almost the same value (83.3%) as Gil et al.[16] Comparing each landmark with this study, Me point showed higher SDR (80.2 % vs. 72.7%), Cg and dental landmarks showed similar SDR (Cg, 71.6%; UDM, 96.9%; LDM, 89.1% vs. Cg, 74.7%; right and left crown points of maxillary incisors, 96.0% and 92.9%), and FZS points showed a lower SDR value (FZS-R, 66.4%; FZS-L, 68.2% vs. FZS-R, 77.8%; FZS-L, 70.7%).

This indicates that even if an identical AI algorithm is used, various results can be obtained depending on the detailed configuration, such as the composition of the sample, the annotation method, and the size of the ROI.

Comparison of measurement accuracy of PA cephalometric variables between AI and human examiners

When selecting a horizontal reference line, it is necessary to use bilateral landmarks in the upper facial structures that do not change significantly with growth or treatment. Depending on which landmarks and horizontal reference lines are used, the measurement values of the lower facial structures may be completely changed.[20] In a previous study by Gil et al.,[16] the FZS exhibited an average error of approximately 2 mm. This deviation could lead to an error of 5 mm at the Me point. Consequently, the calibration of the reference plane was deemed necessary.

According to the results of the point-to-point error and distribution of the SDR in the LO and FZS points, the accuracy in the x- and y-coordinates was much higher in the LO points than in the FZS points (Tables 2–5). This finding is similar to that reported by Major et al.[21] Therefore, the horizontal reference line was set as the line connecting the left and right LO points. Since the identification accuracy of Cg in the x-coordinate seemed to be very high in both AI and examiners (0.52 mm in AI, 0.55 mm in Examiner-1, and 0.50 mm in Examiner-2) (Table 3), it was used as the landmark to set the midsagittal line.

The AI showed that the absolute measurement error values were within the clinically relevant range in the ANS-mid, UDM-mid, LDM-mid (< 1.0 mm), and Me-mid (1.53 mm) (Table 6). These variables were affected by the horizontal position of the Cg, ANS, UDM, LDM, Me, and midsagittal lines and not by the vertical position of each landmark. Therefore, the horizontal measurement errors in the lateral direction were regarded as negligible.

In the present study, there were significant differences between Examiner-1 and Examiner-2 in the accuracy of landmark identification for the right and left LO points in the x-coordinate (0.23 mm vs. 0.30 mm; 0.24 mm vs. 0.42 mm) (Table 3) and ANS and right and left LO points in the y-coordinate (1.22 mm vs. 1.87 mm; 0.64 mm vs. 0.83 mm; 0.68 mm vs. 1.17 mm) (Table 4). However, the measurement errors for the PA cephalometric variables depend on the horizontal position of each landmark. The mean measurement errors did not show a clinically significant difference (all < 0.67 mm), despite statistical differences (all $P < 0.001$; Table 7). Therefore, measurement errors in human examiners might be different from landmark identification errors, despite the clinical experience of Examiner-1 and Examiner-2 (Tables 2–4). However, because different results could be produced by the examiner's skill level, it is necessary to investigate the differences in measurement errors using examiners with different skill levels.[15,22]

Limitations of this study and suggestions for future study

In the present study, the duration of clinical experience of human examiners was relatively short (3 years for Examiner-1 and 1 year for Examiner-2), and the difference in the duration of clinical experience between Examiner-1 and Examiner-2 was small (2 years). Therefore, further studies are required to compare the accuracy of examiners with varying durations of clinical experience.

CONCLUSIONS

According to the results of point-to-point error, SDR, and distance and direction of midline landmark deviation from the midsagittal plane, the cascaded CNN model used in this study might be considered an effective tool for the auto-identification of midline landmarks and quantification of midline deviation in PA cephalograms of adult patients, regardless of variations in the image acquisition method.

AUTHOR CONTRIBUTIONS

Conceptualization: SHH, SKC, NK. Data curation: JHC, MH, MK, SJK, YJK, YHK, SHL, SJS, KHK, SHB. Formal analysis: SHH, SKC, NK. Funding acquisition: SHB. Investigation: SHH, SKC, NK. Methodology: SHH, SKC, NK. Project administration: SKC, NK. Resources: JHC, MH, MK, SJK, YJK, YHK, SHL, SJS, KHK, SHB. Software: JL, JSK, NK. Supervision: SHB, SKC, NK. Validation: JL, NK. Visualization: JL. Writing–original draft: SHH. Writing–review & editing: SKC, NK, SHB.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

FUNDING

This research was supported by grants from the Ko-

rea Health Technology R&D Project through the Korea Health Industry Development Institute funded by the Ministry of Health & Welfare (HI18C1638) and the Technology Innovation Program (20006105) funded by the Ministry of Trade, Industry & Energy, Republic of Korea.

REFERENCES

1. McCarthy J. Artificial intelligence, logic and formalizing common sense. In: Thomason RH, ed. Philosophical logic and artificial intelligence. Dordrecht: Springer; 1989. p. 161-90. https://doi.org/10.1007/978-94-009-2448-2_6
2. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017;69S:S36-40. https://doi.org/10.1016/j.metabol.2017.01.011
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44. https://doi.org/10.1038/nature14539
4. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. Jpn J Radiol 2018;36:257-72. https://doi.org/10.1007/s11604-018-0726-3
5. Broadbent BH. A new x-ray technique and its application to orthodontia. Angle Orthod 1931;1:45-66. https://meridian.allenpress.com/angle-orthodontist/article/1/2/45/55162/A-NEW-X-RAY-TECHNIQUE-and-ITS-APPLICATION-TO
6. Kazandjian S, Kiliaridis S, Mavropoulos A. Validity and reliability of a new edge-based computerized method for identification of cephalometric landmarks. Angle Orthod 2006;76:619-24. https://pubmed.ncbi.nlm.nih.gov/16808568/
7. Yu HJ, Cho SR, Kim MJ, Kim WH, Kim JW, Choi J. Automated skeletal classification with lateral cephalometry based on artificial intelligence. J Dent Res 2020;99:249-56. https://doi.org/10.1177/0022034520901715
8. Kim IH, Kim YG, Kim S, Park JW, Kim N. Comparing intra-observer variation and external variations of a fully automated cephalometric analysis with a cascade convolutional neural net. Sci Rep 2021;11:7925. https://doi.org/10.1038/s41598-021-87261-4
9. Wang CW, Huang CT, Hsieh MC, Li CH, Chang SW, Li WC, et al. Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge. IEEE Trans Med Imaging 2015;34:1890-900. https://doi.org/10.1109/TMI.2015.2412951
10. Park JH, Hwang HW, Moon JH, Yu Y, Kim H, Her SB, et al. Automated identification of cephalometric landmarks: Part 1-Comparisons between the latest deep-learning methods YOLOV3 and SSD. Angle Orthod 2019;89:903-9. https://doi.org/10.2319/022019-127.1
11. 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. https://doi.org/10.2319/022019-129.1
12. Song Y, Qiao X, Iwamoto Y, Chen YW. Automatic cephalometric landmark detection on x-ray images using a deep-learning method. Appl Sci 2020;10:2547. https://doi.org/10.3390/app10072547
13. Kunz F, Stellzig-Eisenhauer A, Zeman F, Boldt J. Artificial intelligence in orthodontics: evaluation of a fully automated cephalometric analysis using a customized convolutional neural network. J Orofac Orthop 2020;81:52-68. https://doi.org/10.1007/s00056-019-00203-8
14. Hong M, Kim I, Cho JH, Kang KH, Kim M, Kim SJ, et al. Accuracy of artificial intelligence-assisted landmark identification in serial lateral cephalograms of Class III patients who underwent orthodontic treatment and two-jaw orthognathic surgery. Korean J Orthod 2022;52:287-97. https://doi.org/10.4041/kjod21.248
15. Muraev AA, Tsai P, Kibardin I, Oborotistov N, Shirayeva T, Ivanov S, et al. Frontal cephalometric landmarking: humans vs artificial neural networks. Int J Comput Dent 2020;23:139-48. https://pubmed.ncbi.nlm.nih.gov/32555767/
16. Gil SM, Kim I, Cho JH, Hong M, Kim M, Kim SJ, et al. Accuracy of auto-identification of the postero-anterior cephalometric landmarks using cascade convolution neural network algorithm and cephalometric images of different quality from nationwide multiple centers. Am J Orthod Dentofacial Orthop 2022;161:e361-71. https://doi.org/10.1016/j.ajodo.2021.11.011
17. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42:318-27. https://doi.org/10.1109/TPAMI.2018.2858826
18. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. Medical image computing and computer-assisted intervention - MICCAI 2015. Cham: Springer; 2015. p. 234-41. https://doi.org/10.1007/978-3-319-24574-4_28
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, USA. Piscataway: Institute of Electrical and

Electronics Engineers (IEEE), 2016. https://doi.org/10.1109/CVPR.2016.90
20. Lee HJ, Lee S, Lee EJ, Song IJ, Kang BC, Lee JS, et al. A comparative study of the deviation of the menton on posteroanterior cephalograms and three-dimensional computed tomography. Imaging Sci Dent 2016;46:33-8. https://doi.org/10.5624/isd.2016.46.1.33
21. Major PW, Johnson DE, Hesse KL, Glover KE. Landmark identification error in posterior anterior cephalometrics. Angle Orthod 1994;64:447-54. https://pubmed.ncbi.nlm.nih.gov/7864466/
22. Na ER, Aljawad H, Lee KM, Hwang HS. A comparative study of the reproducibility of landmark identification on posteroanterior and anteroposterior cephalograms generated from cone-beam computed tomography scans. Korean J Orthod 2019;49:41-8. https://doi.org/10.4041/kjod.2019.49.1.41