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

1st author: Sung-Hoon Han,**a**

1st author: Jisup Lim, c

Jun-Sik Kim,**d**

Jin-Hyoung Cho,e

Mihee Hong, **f**

Minji Kim,**g**

Su-Jung Kim,**h**

Yoon-Ji Kim,**i**

Young Ho Kim,j

Sung-Hoon Lim,k

Sang Jin Sung,i

Kyung-Hwa Kang,**a**

Seung-Hak Baek,

corresponding author: Sung-Kwon Choi,a

corresponding author: Namkug Kim,b

Iksan-si, Seoul, Gwangju, Daegu, and Suwon-si, South Korea

aDepartment of Orthodontics, School of Dentistry, Wonkwang University, Iksan-si, South Korea.

bDepartment of Convergence Medicine, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea.

cDepartment of Convergence Medicine, University of Ulsan, College of Medicine, Asan Medical Center, Seoul, Republic of Korea

dGraduate student (MSc), Department of Biomedical Engineering, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

eDepartment of Orthodontics, School of Dentistry, Chonnam National University, Gwangju, South Korea.

fDepartment of Orthodontics, School of Dentistry, Kyungpook National University, Daegu, South Korea.

gDepartment of Orthodontics, College of Medicine, Ewha Womans University, Seoul, South Korea.

hDepartment of Orthodontics, School of Dentistry, Kyung Hee University, Seoul, South Korea.

iDepartment of Orthodontics, Asan Medical Center, College of Medicine, University of Ulsan, Seoul, South Korea.

jDepartment of Orthodontics, Institute of Oral Health Science, Ajou University School of Medicine, Suwon-si, South Korea.

kDepartment of Orthodontics, College of Dentistry, Chosun University, Gwangju, South Korea.

IDepartment of Orthodontics, School of Dentistry, Dental Research Institute, Seoul National University, Seoul, South Korea.

Sung-Hoon Han and Jisup Lim are joint first authors and contributed equally to this work.

Address correspondence to:

Sung-Kwon Choi, Department of Orthodontics, School of Dentistry, Wonkwang University, 460 Iksandae-ro, Iksan-si, Jeollabuk-do 54538, South Korea; Tel: +82-63-859-2962; e-mail, chsk6206@wku.ac.kr

or

Namkug Kim, 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, South Korea; Tel: +82-2-3010-6573; e-mail, <u>namkugkim@gmail.com</u>.

ABSTRACT

Objective: To quantify the effects of midline-related landmark identification in posteroanterior (PA) cephalogram images by cascaded convolutional neural network (CNN) algorithm on the midline deviation measurements.

Methods: A total of 2,903 PA cephalogram images obtained from nine university hospitals were divided into the training-set (n=2,150), internal validation-set (n=376), and test-set (n=377). As gold standard, two orthodontic professors marked the bilateral landmarks including frontozygomatic-suture point (FZS) and lateral-orbit point (LO), and the midline landmarks including Cg, ANS, upper dental midpoint (UDM), lower dental midpoint (LDM), and Me using V-Ceph 8.0 program. For test, Examiner-1 and Examiner-2 (3-year and 1-year orthodontic resident) and Cascaded-CNN model marked the landmarks. After point-to-point errors of landmark identification, successful detection rate (SDR, percentage within 1-, 2-, and 3-mm ranges), and distance and direction of the midline landmark deviation from the midsagittal line (ANS-mid, UDM-mid, LDM-mid, and Me-mid) were measured, statistical analysis was performed.

Results: The cascaded-CNN algorithm showed 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). Its average SDR within 2 mm range was 83.2% with high accuracy at the right LO (96.9%), left LO (97.1%), and UDM (96.9%). Its absolute measurement errors were less than 1 mm in ANS-mid, UDM-mid, and LDM-mid compared to the gold standard.

Conclusion: The cascaded-CNN model might be considered an effective tool for auto-identification of the midline landmarks and quantification of the midline deviation in PA cephalograms of adult patients, regardless of variations in image acquisition method.

Keywords: Artificial intelligence, convolutional neural network, posteroanterior cephalograms

INTRODUCTION

Artificial intelligence (AI) refers to algorithms that imitate human intelligence for recognizing and solving problems and making efficient decisions.^{1,2} Of these algorithms, artificial neural networks (ANNs) are the computing systems that mimic the biological neural networks in the animals' brains. Convolutional neural network (CNN), one of the deep learning models that belong to ANNs, extracts the data characteristics and identifies their patterns. Since it addresses the problems that occur when processing image or video data with regular deep learning algorithms, it is suitable for recognizing or exploring visual data.^{3,4}

Cephalometric analysis is an essential part of the diagnostic process. When it is performed by a nonexpert, it takes time and effort and may produce analytical errors.⁵⁻⁷ Therefore, there have been ongoing efforts to use the image recognition ability of CNN for automatically identifying the cephalometric landmarks. CNNs are designed to mimic the hierarchical organization of the human visual cortex for processing visual information and have been successfully applied in various image recognition domains including cephalometric analysis.⁸ Recent studies have reported a high accuracy in automatically identifying the cephalometric landmarks in lateral cephalograms using CNN.⁹⁻¹⁴ Nevertheless, there has been limited research on posteroanterior (PA) cephalometric analysis using cascaded CNNs, especially concerning measurement values.

PA cephalograms have been used to evaluate the degree of angle or the amount or direction of landmark deviation from the midsagittal reference plane. However, AI studies using PA cephalograms are rare up to date. Muraev et al.¹⁵ reported that the accuracy level of landmark identification by AI was similar to that of 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%. On the contrary, validation of the reference planes is needed to obtain accurate measurements of the PA cephalometric variables.

Previous studies might have some limitations as follows: (1) When the gold standard for AI training were set by a single operator^{11,15} or by average coordinate value of two operaters,¹² it might be related with some bias. Therefore, it is necessary to establish the gold standard by mutual agreement between two experts; (2) It is necessary to examine the identification error in the x- and y-coordinates and the distribution of the SDR of the midline landmarks, respectively; and (3) Landmark Identification error and measurement accuracy the midline variables should be investigated among AI and plural human examiners (for example, human examiner-1 and human examiner-2) using multiple comparison test.

Therefore, the purpose of this study was to quantify the effects of midline-related landmark identification in PA cephalogram images by cascaded CNN algorithm on the midline deviation measurements.

MATERIALS AND METHODS

Subjects

A total of 2,930 PA cephalograms were obtained from nine 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 whose facial growth was completed; 2) patients who underwent orthognathic surgery between 2013 and 2020; and 3) patients with permanent dentition. Exclusion criteria were 1) patients who had craniofacial syndromes or systemic diseases, and 2) patients whose PA cephalogram had a poor image quality to make identification of landmarks impossible.

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

Determination of PA landmarks and the gold standard

The definitions of the bilateral landmarks including the frontozygomatic suture point (FZS) and the lateral orbit point (LO), and the midline landmarks including crista galli (Cg), anterior nasal spine (ANS), upper dental midpoint (UDM), lower dental midpoint (LDM), and menton (Me) were enumerated in Figure 1 and Table 1.

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

Training and internal validation of the algorithm

Deep learning training using the cascaded CNN algorithm consisted of (1) the determination of the region of interest (ROI) and (2) the step for landmark prediction (Figure 3). Firstly, RetinaNet¹⁷ was used to extract the ROI with the x- and y-coordinates of the center of the landmark. ROI was set to have two sizes (256 x 256 and 512 x 512). Secondly, U-net¹⁸ was used to detect the exact location of the ROI patch formed in the first step.

The RetinaNet adopted Resnet-50¹⁹ as the backbone and used it for learning, and pretrained weights were not used for training. Adam optimizer combining the momentum and exponentially weighted moving average gradients methods to update the weights of the networks. The learning late 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 end the training.

Various augmentation methods, such as gaussian noise, random brightness, blurring, random contract, flip, and random rotation, were used in the deep learning model training. Internal validation test (n=376) was performed to find the optimal parameter values for machine learning.

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

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

Point-to-point errors of landmark identification in two examiners (residents) and AI against the gold standard (two orthodontic professors) were measured. The position of each landmark was mapped by 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 a very high intraclass correlation coefficient (ICC) values (≥ 0.99) in all nine landmarks.

The SDR was set as the percentage of landmarks within a specific range from the gold standard (< 1 mm, < 2 mm, 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, respectively), reorientation of PA cephalograms was performed. To measure the PA cephalometric variables accurately, the first step is to decide which landmarks (LO vs. FZS) had higher identification accuracy. Then, the midsagittal line was set as the line passing through Cg and intersecting perpendicularly with the horizontal reference lines (LO line and FZS line).

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

Statistical analysis

One-way analysis of variance (ANOVA) test with Tukey's test was performed using the SPSS program (version 12.0, SPSS, Chicago, III). 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. AI showed significantly higher accuracy than Examiner-2 for 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<0.01, respectively; Figure 4, Table 2). Although AI showed low accuracy in identification of the right and left FZS points, it still showed a higher accuracy than both Examiner-1 and Examiner-2 (1.87 mm vs. 2.26 mm and 2.33 mm, P<0.01; 2.01 mm vs. 3.02 mm and 3.20 mm,

P<0.001). However, there was no difference in accuracy in identification of Cg, UDM, LDM and Me among AI, Examiner-1 and Examiner-2.

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

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

In terms of the errors in the y-coordinate, there were no significant differences in vertical positioning of UDM and Me between AI and human examiners. The AI showed significantly higher identification accuracy in vertical positioning than Examiner-2 (ANS, LDM, right and left FZS points, and right and left LO points, all P<0.001; Table 4). The error values in vertical positioning of UDM, LDM, Me, and right and left LO points by AI were less than 1 mm (Table 4).

Distribution of SDR for Al-identified landmarks

The mean SDRs of all Al-identified landmarks were 65.8% at < 1 mm, 83.2% at < 2 mm, and 89.6% at < 3 mm (Table 5). Highly accurate SDR values (\geq 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 (\leq 70% within 2 mm range) were found at 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, 3, 4, and 5), PA cephalograms were reoriented using LO line and midsagittal line. Then, the perpendicular distances between the midline landmarks and the midsagittal line (ANS-mid, UDM-mid, LDM-mid, and Me-mid) were measured (Figure 6).

When the measurements by AI and human examiners were compared to that of the gold standard, the absolute measurement errors were < 1 mm in ANS-mid, UDM-mid, and LDM-mid and was also

approximately within the clinically relevant range (< 1.5 mm) in Me (Table 6). Al did not exhibit significant differences in LDM-mid and Me-mid from human examiners. However, Examiner-2 had a higher error in LDM-mid and Me-mid than Examiner-1 (all P<0.01, Table 6).

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 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 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 Al 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, Table 2).

In the present study, when the value less than 1.5 mm is considered as clinically accurate, AI showed a high or good accuracy in identification of UDM (0.54 mm), right LO (0.58 mm) and left LO (0.70 mm), LDM (0.97 mm), and ANS (1.31 mm). These findings indicate that AI might be better for PA cephalometric landmark identification than first-year orthodontic resident. However, the accuracy in identification of Cg, Me, right and left FZS (1.76 mm, 1.61 mm, 1.87 mm, 2.01 mm) needs to be increased in future studies. It is interpreted that the low accuracy of Cg, Me, right, and left FZS is mainly due to two factors: overlapping problems that occur when converting 3D structures to 2D structures (Cg, right and left FZS), and errors that arise when identifying points on the gentle curve of the mandibular lower border (Me).

The findings that AI exhibited less than 1 mm error values in horizontal positioning of all landmarks except UDM (Table 3) and in vertical positioning of UDM, LDM, Me, and right and left LO points (Table 4) indicated that the most errors happened in the vertical positioning of the PA cephalogram landmarks due to their anatomic features.

Some of the landmarks (Cg, Me, right and left FZS) in the present study were the identical as in the

previous study of Gil et al.¹⁵ and the cascaded CNN algorithm used in this study showed higher accuracy compared to that study (1.55 mm, 0.58mm, 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 Al-identified landmarks

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

The cascaded CNN algorithm used in this study showed 83.2% of average SDR within 2 mm range (Table 5), which was almost same value (83.3%) of Gil et al.¹⁴ 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 point of maxillary incisors, 96.0% and 92.9%), and FZS points showed 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 shown depending on detailed configurations such as the composition of the sample or the annotation method and the size of the ROI.

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

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

According to the results of point-to-point error and distribution of SDR in the LO and FZS points, accuracy in the X- and Y-coordinates was much higher in LO points than FZS points (Tables 2, 3, 4 and 5). This finding was similar to Major et al's study.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, 0.50 mm in Examiner-2, Table 3), it was used as the landmark to set the midsagittal line.

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

In the present study, there was significant differences between Examiner-1 and Examiner-2 in 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 PA cephalometric variables depends on the horizontal position of each landmark. The mean measurement errors did not show 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 clinical experience between Examiner-1 and Examiner-2 (Tables 2, 3 and 4). However, since different results could be produced by examiner' skill level, it would be needed to investigate difference in measurement errors using examiners with different skill level.^{15,22}

Limitations of this study and suggestions for future study

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

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 auto-identification of the midline landmarks and quantification of the midline deviation in PA cephalograms of adult patients, regardless of variations in image acquisition method.

Acknowledgements

This research was supported by grants from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute and 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.

Conflict of interest statement

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

REFERENCES

- 1. McCarthy J. Artificial intelligence, logic and formalizing common sense. In: Thomason RH, eds. Philosophical logic and artificial intelligence. Dordrecht: Springer; 1989.161-90.
- 2. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017;695:536-40.
- 3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
- 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.
- Broadbent BH. A new x-ray technique and its application of orthodontia. Angle Orthod 1931;1:45-66.
- Kazandjian S, Kiliaridis S, Mavropoulos A. Validity and reliability or a new edge-based computerized method for identification of cephalometric landmarks. Angle Orthod 2006;76:619-24.
- 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.
- 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.
- Wang CW, Huang CT, Hsieh MC, Li CH, Chang SW, Li WC, et al. Evaluation and comparison of analandmarklandmark detection methods for cephalometric x-ray images: a grand challenge. IEEE Trans Med Imaging 2015;34:1890-900.
- 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.
- 11. Hwang HW, Park JH, Moon JH, Yu Y, Kim H, Her SB, et al. Automated identifaction of cephalometric landmarks: Part 2-Might it be better than human? Angle Orthod 2020;90:69-76.
- 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.
- 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.
- 14. Hong MH, Kim IH, Cho JH, Kang KH, Kim MJ, Kim SJ, et al. Accuracy of artificial intelligenceassisted landmark identification in serial lateral cephalograms of Class III patients who underwent orthodontic treatment and two-jaw orthognathic surgery. Korean J Orthod 2022;52:287-297.
- 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.
- 16. Gil SM, Kim IH, Cho JH, Hong MH, Kim MJ, Kim SJ, et al. Accuracy of auto-identification of the posteroanterior cephalometric landmarks using cascade convolutional neural network algorithm and cephalometric images of different quality from nationwide multiple centers. Am J Orthod Dentofacial Orthop 2022;161:e361-71.
- 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.

- Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention. Springer 2015;234-41.
- 19. K He, X Zhang, S Ren, J Sun. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- 20. Lee HJ, Lee SG, 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.
- 21. Major PW, Johnson DE, Hesse KL, Glover KE. Landmark identification error in posterior anterior cephalometrics. Angle Orthod 1994;64:447-54.
- 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.

FIGURES & FIGURE LEGENDS



Figure 1. The posteroanterior (PA) cephalometric landmarks used in this study.



Figure 2. Flow chart showing sample allocation and study design.



Figure 3. Cascaded CNN algorithm used in this study. Stage 1, the ROI detection to propose the area of interest; stage 2, the landmark prediction to find the exact location of landmarks.



Figure 4. Examples of superimposition of the identified PA cephalometric landmarks. Red, gold standard; Green, auto-identification by **cascaded** convolutional neural network (CNN) algorithm; Pink, Examiner-1; Sky blue, Examiner-2.



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.



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 PA cephalogram images.

TABLES

Table 1. Definitions 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 landn	narks					
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

	Point-to	-point err		Multiple				
Landmarks	AI		Examine	er-1	Examiner-2		<i>p</i> -value	comparison
	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	0.000***	(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 Right	1.87	1.74	2.26	1.97	2.33	1.59	0.001**	AI < (E1, E2)
FZS Left	2.01	2.24	3.02	2.59	3.20	2.22	0.000***	AI < (E1, E2)
LO Right	0.58	1.15	0.71	1.62	0.82	0.64	0.022*	(AI, E1) < (E1, E2)
LO Left	0.70	1.60	0.78	2.14	1.15	1.34	0.001**	(AI, E1) < E2
<i>p</i> -value	0.000***	-	0.000***		0.000***			
Multiple comparison	(UDM, LO R&L) < (LO R&L, LDM) < (LDM, ANS) < (ANS, Me) < (Me, Cg, FZS R&L)		(LO R8 LDM) ANS, M FZS R) <	kL, UDM, < (LDM, e) < (Cg, < FZS L	(UDM, LC R, LDM) LO L) < (< (Cg, AI R < FZS L) R) < (LO < (LDM, (LO L, Me) NS) < FZS -		
Total	1.26	1.94	1.57	1.66	1.75	2.34		

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

A one-way ANOVA followed by Tukey's test was performed.

*p < 0.05; **p < 0.01; ***p < 0.001.

Landmarks	AI		Examiner-1		Examiner-2		<i>p</i> -value	Multiple
	Mean	SD	Mean	SD	Mean	SD	<i>p</i>	comparison
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	0.6	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 Right	0.79	0.94	0.61	0.59	0.72	0.69	0.002**	(E1, E2) < (E2, AI)
FZS Left	0.86	1.43	1.2	1.05	1.26	1.26	0.000***	AI < (E1, E2)
LO Right	0.24	0.53	0.23	0.21	0.3	0.25	0.026*	(E1, AI) < (AI, E2)
LO Left	0.28	1.04	0.24	0.54	0.42	0.68	0.004**	(E1, AI) < E2

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

A one-way ANOVA followed by Tukey's test was performed.

*p < 0.05; **p < 0.01; ***p < 0.001.

Landmarks	AI		Examiner-1		Examiner-2		<i>p</i> -value	Multiple	
	Mean	SD	Mean	SD	Mean	SD		comparison	
Cg	1.55	1.76	1.93	2.03	1.80	1.6	0.012*	(AI, E2) < (E2, E1)	
ANS	1.11	1.21	1.22	1.74	1.87	1.88	0.000***	(AI, E1) < E2	
UDM	0.31	0.59	0.4	0.5	0.45	0.7	0.366		
LDM	0.35	1.87	0.64	0.58	0.53	0.67	0.000***	AI < E2 < E1	
Me	0.58	2.08	0.73	0.81	0.60	0.94	0.187		
FZS Right	1.61	1.55	2.34	1.52	2.41	1.75	0.000***	AI < (E1, E2)	
FZS Left	1.73	1.81	3.02	1.74	3.26	2.13	0.000***	AI < (E1, E2)	
LO Right	0.47	1.04	0.64	0.6	0.83	0.78	0.000***	AI < E1 < E2	
LO Left	0.58	1.25	0.68	0.9	1.17	1.35	0.000***	(Al, E1) < E2	

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

A one-way ANOVA followed by Tukey's test was performed.

*p < 0.05; **p < 0.01; ***p < 0.001.

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							

Table 5. Distribution of the successful detection rate (SDR) in AI

Table 6. Comparison of the absolute measurement error of PA cephalometric variables between AI

 and human examiners

	Distanc	e (mm)							
Measurements	AI-GS		E1-GS E2		E2-GS		p-value	comparison	
	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)	

One-way ANOVA test and Tukey's test were performed.

***p* < 0.01.

E1, examiner-1; E2, examiner-2; GS, gold standard.

	Distance	e (mm)			A d. 16:01-				
Measurements	AI-GS		E1-GS		E2-GS		<i>p</i> -value	comparison	
	mean	SD	mean	SD	mean	SD			
ΔANS-mid	0.09	1.24	-0.21	0.80	-0.09	0.95	0.000***	(E1-GS, E2-GS) < AI-GS	
ΔUDM-mid	-0.07	1.38	-0.24	0.84	-0.46	1.03	0.000***	E2-GS < (E1- GS, AI-GS)	
ΔLDM-mid	0.16	1.68	-0.19	1.26	-0.36	1.59	0.000***	(E2-GS, E1-GS) < AI-GS	
ΔMe-mid	0.11	2.42	-0.22	1.71	-0.67	2.12	0.000***	E2-GS < (E1- GS, AI-GS)	

Table 7. Comparison of the mean measurement error of PA cephalometric variables between AI and human examiners

A one-way ANOVA followed by Tukey's test was performed.

A negative sign means right-side deviation.

****p* < 0.001.

E1, examiner-1; E2, examiner-2; GS, gold standard.