Original Article

Evaluation of automated cephalometric analysis based on the latest deep learning method

Hye-Won Hwang[a]; Jun-Ho Moon[b]; Min-Gyu Kim[b]; Richard E. Donatelli[c]; Shin-Jae Lee[d]

ABSTRACT

Objectives: To compare an automated cephalometric analysis based on the latest deep learning method of automatically identifying cephalometric landmarks (AI) with previously published AI according to the test style of the worldwide AI challenges at the International Symposium on Biomedical Imaging conferences held by the Institute of Electrical and Electronics Engineers (IEEE ISBI).

Materials and Methods: This latest AI was developed by using a total of 1983 cephalograms as training data. In the training procedures, a modification of a contemporary deep learning method, YOLO version 3 algorithm, was applied. Test data consisted of 200 cephalograms. To follow the same test style of the AI challenges at IEEE ISBI, a human examiner manually identified the IEEE ISBI-designated 19 cephalometric landmarks, both in training and test data sets, which were used as references for comparison. Then, the latest AI and another human examiner independently detected the same landmarks in the test data set. The test results were compared by the measures that appeared at IEEE ISBI: the success detection rate (SDR) and the success classification rates (SCR).

Results: SDR of the latest AI in the 2-mm range was 75.5% and SCR was 81.5%. These were greater than any other previous AIs. Compared to the human examiners, AI showed a superior success classification rate in some cephalometric analysis measures.

Conclusions: This latest AI seems to have superior performance compared to previous AI methods. It also seems to demonstrate cephalometric analysis comparable to human examiners. (*Angle Orthod.* 2021;91:329–335.)

KEY WORDS: Artificial intelligence; Machine learning; Deep learning

INTRODUCTION

With the recent advancement in computer technology, machine learning has played a significant role in the detection and classification of certain diseases identified in medical images.[1,2] In orthodontics, there have also been efforts to utilize machine learning techniques in a variety of ways, one of which was the automated identification of cephalometric landmarks via artificial intelligence (AI).[3–10] Although research using three-dimensional images has attracted attention,[11–13] the two-dimensional cephalometric image is still important and is the most commonly utilized tool in orthodontics for diagnosis, treatment planning, and outcome prediction.[3,14–18]

As more attempts were made to incorporate AI into cephalometric analysis, worldwide AI challenges began in 2014 at the International Symposium on Biomedical Imaging conferences under the support of the Institute of Electrical and Electronics Engineers (IEEE ISBI). As accuracy measures of AI, the AI

[a] Clinical Lecturer, Department of Orthodontics, Seoul National University Dental Hospital, Seoul, Korea.

[b] Graduate Student, Department of Orthodontics, Graduate School, Seoul National University, Seoul, Korea.

[c] Assistant Professor and Program Director, Department of Orthodontics, University of Florida College of Dentistry, Gainesville, Fla., USA.

[d] Professor, Department of Orthodontics and Dental Research Institute, Seoul National University School of Dentistry, Seoul, Korea.

Corresponding author: Dr Shin-Jae Lee, Professor, Department of Orthodontics and Dental Research Institute, Seoul National University School of Dentistry, 101 Daehakro, Jongro-Gu, Seoul 03080, Korea
(e-mail: nonext@snu.ac.kr)

Accepted: November 2020. Submitted: February 2020.
Published Online: January 12, 2021

Table 1. List of Anatomical Landmarks Used in the IEEE ISBI Challenges[7,9,10]

Landmark number	Explanation
1	Sella
2	Nasion
3	Porion
4	Orbitale
5	Anterior nasal spine
6	Posterior nasal spine
7	Point A
8	Upper incisor edge
9	Lower incisor edge
10	Point B
11	Pogonion
12	Gnathion
13	Menton
14	Gonion
15	Articulare
16	Subnasale
17	Upper lip
18	Lower lip
19	Soft tissue pogonion

challenge detected 19 cephalometric landmarks (Table 1) and used them to test the success detection rate calculated from 2-mm to 4-mm error ranges for the 19 landmarks (Table 2). Since 2015, the challenge became more clinically oriented and started reporting the success classification rates for the eight cephalometric measurements shown in Tables 3, 4, and 5.[5,7–10]

In line with the trend of comparing the performance of evolving AI, the purpose of this study was to compare and evaluate an automated cephalometric analysis, based on the latest deep learning method of automatically identifying cephalometric landmarks,[4,6,19] with a number of previously published AIs.[7–9,20,21] In addition, the study assessed how accurately the latest AI performed cephalometric analysis compared to human examiners.

MATERIALS AND METHODS

The institutional review board for the protection of human subjects at Seoul National University Dental Hospital reviewed and approved the research protocol (ERI 19007).

Figure 1 summarizes the experimental design. The new AI applied in this study was based on a modification of the latest deep learning method, the YOLO version 3 algorithm.[4,6,19] The deep learning was processed by a workstation running Ubuntu 18.04.1 LTS with NVIDIA Tesla V100 GPU (NVIDIA Corporation, Santa Clara, CA, USA). During the learning process of the latest AI, a total of 1983 cephalograms were used as the training data. An examiner (examiner 1, SJL), with 30 years of clinical orthodontic practice experience, manually detected cephalometric landmarks.

Table 2. Success Detection Rate Results for the 19 Landmarks[a]

Landmark Number	Success Detection Rate (%)				MRE (mm)	SD (mm)
	2 mm	2.5 mm	3 mm	4 mm		
1	96.0	96.5	96.5	96.5	1.21	3.43
2	82.5	89.0	92.5	98.0	1.42	1.27
3	69.0	79.5	86.5	90.5	2.36	3.65
4	61.5	74.5	83.0	92.5	2.84	5.91
5	53.5	68.5	76.0	88.0	2.36	2.63
6	72.0	83.0	89.5	93.5	2.16	6.27
7	51.0	63.0	73.5	85.5	2.41	1.80
8	88.5	97.0	99.0	100.0	1.13	0.65
9	88.5	93.0	97.5	99.5	1.20	0.84
10	43.5	53.0	66.5	82.0	2.67	1.93
11	87.5	92.5	96.5	99.0	1.21	0.97
12	85.0	96.5	98.5	99.5	1.25	0.66
13	85.5	94.5	97.5	99.5	1.41	0.64
14	38.0	50.5	63.5	81.0	2.75	1.84
15	89.0	92.0	93.5	95.0	1.62	3.29
16	87.0	94.0	95.5	98.5	1.51	2.43
17	91.0	96.5	98.5	99.5	1.07	0.67
18	93.0	97.5	99.5	100.0	1.04	0.60
19	71.5	78.5	86.0	92.5	1.80	1.59
Average	75.45	83.66	88.92	94.24	1.76	2.16

[a] MRE indicates mean radial error; SD, standard deviation.

The test data consisted of 200 new images. Examiner 1 manually identified landmarks in the test data. These acted as reference points. Then, another examiner (examiner 2, HWH), a board-certified orthodontist, manually identified the same landmarks in the test data. Lastly, the trained new AI detected landmarks in the test data set. The comparison of the results from the test data set used the conventional 19 landmarks that were identical to those used at the IEEE ISBI challenges (Table 1).[7,9,10]

The performance of the cephalometric analyses were analyzed according to the same measurement values and test formats suggested by the IEEE ISBI challenges: (1) the success detection rates (SDR) of 2-, 2.5-, 3-, and 4-mm error ranges for the 19 landmarks (Table 2), and (2) the success classification rates (SCR) for eight cephalometric analysis measures (Tables 3, 4, and 5).[7,9,10]

When calculating SDR in the 2-mm error range, if the landmark detection by the AI showed an error within 2 mm from the reference, the landmark was assumed to be successfully detected. Most AI studies reported that errors within the range of 2 mm were clinically acceptable.[4,6,8–10]

For the classification of anatomical types, the same eight clinical measurements set in the IEEE ISBI 2015 challenge were analyzed. The criteria are shown in Table 3. The measurement values and clinical classification results derived by examiner 1 were set as the reference values, and the classification results by the latest AI and examiner 2 were compared through SCR. For instance, when the ANB value of a certain patient

Table 3. Classification Criteria of Anatomical Types Used in the IEEE ISBI Challenges[7,9,10,a]

Variables	Type 1	Type 2	Type 3	Type 4
ANB	3.2°–5.7°	>5.7°	<3.2°	
	Class I, normal	Class II	Class III	
SNA	79.4°–83.2°	>83.2°	<79.4°	
	Normal maxilla	Prognathic maxilla	Retrognathic maxilla	
SNB	74.6°–78.7°	<74.6°	>78.7°	
	Normal mandible	Retrognathic mandible	Prognathic mandible	
ODI	74.5° ± 6.1°	>80.5°	<68.4°	
		Deep bite tendency	Open bite tendency	
APDI	81.4° ± 3.8°	<77.6°	>85.2°	
		Class II tendency	Class III tendency	
Facial height ratio	0.65–0.75	>0.75	<0.65	
		Short face tendency	Long face tendency	
SN-GoGn angle	26.8°–31.4°	>31.4°	<26.8°	
		Mandible high angle	Mandible low angle	
Overjet	2 mm–4.5 mm	0 mm	<0 mm	>4.5 mm
		Edge-to-edge	Anterior cross bite	Large overjet

[a] IEEE indicates Institute of Electrical and Electronics Engineers; ISBI, International Symposium on Biomedical Imaging.

was identified as Class II from the analysis by examiner 1, if the same patient's ANB classifications by the AI and by examiner 2 were independently identified as Class II, the classification results by the AI and by examiner 2 were assumed to be successful.

RESULTS

SDR of the new AI for each of the 19 landmarks are shown in Table 2. Eleven of the 19 landmarks showed more than 80% SDR with an error range within 2 mm. The average SDR in the 2 mm error range was 75.45%. Figure 2 shows the detection error pattern for cephalometric landmarks.

The comparisons between the new AI with previous AIs in terms of the SCR are summarized in Table 4. The average SCR of the new AI was 81.53%, which was the highest when compared to previous AIs. In each of the eight measurements, the SCRs of the new AI were superior to previous AIs.

Comparisons of the SCR between the human examiner and the latest AI are summarized in Table 5. The latest AI showed a superior success classification rate compared to the human examiner (examiner 2) in three cephalometric measurements. In the remaining measurements, examiner 2 showed better SCR results.

DISCUSSION

The present study demonstrated how well the latest automated cephalometric analysis performed when compared to previously reported AIs. The latest AI showed better performance than previous AIs, and it demonstrated similar accuracy performance to that from human examiners. Park et al. suggested that deep learning-based techniques could provide improved accuracy compared to other popularly applied AI algorithms such as random forest.[6] Later, Hwang et al. evaluated whether the AI in the automatic identification of cephalometric landmarks could actually exceed manual markings performed by clinical experts. In general, the answer was "yes" in terms of reproducibility.[4] Although the previous studies focused on the development of AI and its reliability in terms of the Euclidean distance measures of landmark detection errors, this present study focused on its clinical relevance, ie, how accurately the new AI would perform cephalometric analyses useful for patient treatment. In addition, compared to previous AIs, the quantity of

Table 4. Comparison Between AIs in Terms of Success Classification Rate[a]

	Present Study	Wang et al. 2018[7]	Arik et al. 2017[8]	Lindner et al. 2015[20]	Ibragimov et al. 2015[21]
ANB	83.43	58.61	77.31	75.83	76.64
SNA	68.97	59.86	66.72	77.97	70.24
SNB	86.74	78.85	69.81	81.92	75.24
ODI	77.68	76.59	72.28	71.26	63.71
APDI	86.00	83.49	87.18	87.25	79.93
Facial height ratio	85.88	82.44	69.16	90.90	86.74
SN-GoGn angle	79.80	77.18	78.01	80.66	78.90
Overjet	83.72	83.20	77.45	82.11	77.53
Average	81.53	75.03	74.74	80.99	76.12

[a] AI indicates artificial intelligence.

Table 5. Comparison Between the Latest AI and Human Examiners in Terms of Success Classification Rates[a]

Examiner #1	Latest AI				Diagonal Average	Examiner #2				Diagonal Average
	Type1	Type2	Type3	Type 4		Type1	Type2	Type3	Type 4	
ANB					83.43					79.50
Type 1	82.76	10.34	6.90			72.41	6.90	20.69		
Type 2	23.81	76.19	0.00			28.57	71.43	0.00		
Type 3	8.67	0.00	91.33			4.67	0.67	94.67		
SNA					68.97					72.58
Type 1	67.39	15.22	17.39			57.61	7.61	34.78		
Type 2	36.59	63.41	0.00			19.51	78.05	2.44		
Type 3	16.42	7.46	76.12			17.91	0.00	82.09		
SNB					86.74					84.70
Type 1	77.78	15.56	6.67			75.56	20.00	4.44		
Type 2	7.41	92.59	0.00			7.41	92.59	0.00		
Type 3	10.16	0.00	89.84			14.06	0.00	85.94		
ODI					77.68					81.70
Type 1	78.95	10.53	10.53			81.58	2.63	15.79		
Type 2	33.33	66.67	0.00			33.33	66.67	0.00		
Type 3	12.58	0.00	87.42			3.14	0.00	96.86		
APDI					86.00					84.24
Type 1	68.42	7.89	23.68			76.32	13.16	10.53		
Type 2	5.56	94.44	0.00			16.67	83.33	0.00		
Type 3	4.17	0.69	95.14			6.94	0.00	93.06		
FHR					85.88					92.94
Type 1	83.67	0.00	16.33			91.84	0.00	8.16		
Type 2	0.00	0.00	0.00			0.00	0.00	0.00		
Type 3	9.93	1.99	88.08			5.96	0.00	94.04		
SN-GoGn					79.80					83.38
Type 1	67.19	12.50	20.31			73.44	17.19	9.38		
Type 2	14.08	84.51	1.41			14.08	85.92	0.00		
Type 3	10.77	1.54	87.69			9.23	0.00	90.77		
Overjet					83.72					93.33
Type 1	76.19	20.63	0.00	3.17		95.24	0.00	0.00	4.76	
Type 2	0.00	100.00	0.00	0.00		100.00	0.00	0.00	0.00	
Type 3	0.00	9.09	90.91	0.00		0.00	0.00	100.00	0.00	
Type 4	28.81	1.69	1.69	67.80		15.25	0.00	0.00	84.75	

[a] AI indicates artificial intelligence.

study data was almost doubled from N = 1311 to N = of 2183, which improved the power of explanation.[4,6,22]

The SDR results showed that the latest AI successfully detected most landmarks within 2 mm at close to 90%. Some landmarks showed less accurate results than others in terms of SDR. The reasons for this might be that some landmarks, such as Porion, Orbitale, and PNS are difficult to detect due to overlapping cranial base structures. Porion, Orbitale, and Gonion exist bilaterally and might cause errors in the process of determining the midpoint of those bilateral structures. Some landmarks, such as Point A, Point B, Gonion, and soft tissue Pogonion, were difficult to pinpoint exactly even in a magnified view. Orbitale and PNS showed greater standard deviation values when identified by the AI (Table 2). As Figure 2 shows, errors during the landmark identification appeared both on the x- and y-axes. Since Pogonion is defined as the anterior point of the chin region, it was easy to identify it on the x-axis. Consequently, the y-axis error was greater than the x-axis error. Similarly, error patterns could be interpreted for Pogonion, Gnathion, and Menton that represent the anterior, anteroinferior, and inferior points of the chin, respectively. These error patterns also were demonstrated between human examiners (Figure 2).

For an AI method to be useful in clinical orthodontic practice, analyses of angles or linear measurements derived from AI-detected landmarks are necessary. The results of SCRs demonstrated that the clinical classification performance of the latest AI was better than previously published AI technologies. In addition, the current study utilized approximately 2000 images in the training procedures for the new AI. This likely contributed to the improvement in the SCR over previous AIs.[22] As would be expected, measurements consisting of landmarks with more accurate SDR values resulted in more accurate SCR values. Although not all measurements showed that the new AI was performing better than humans, taking into

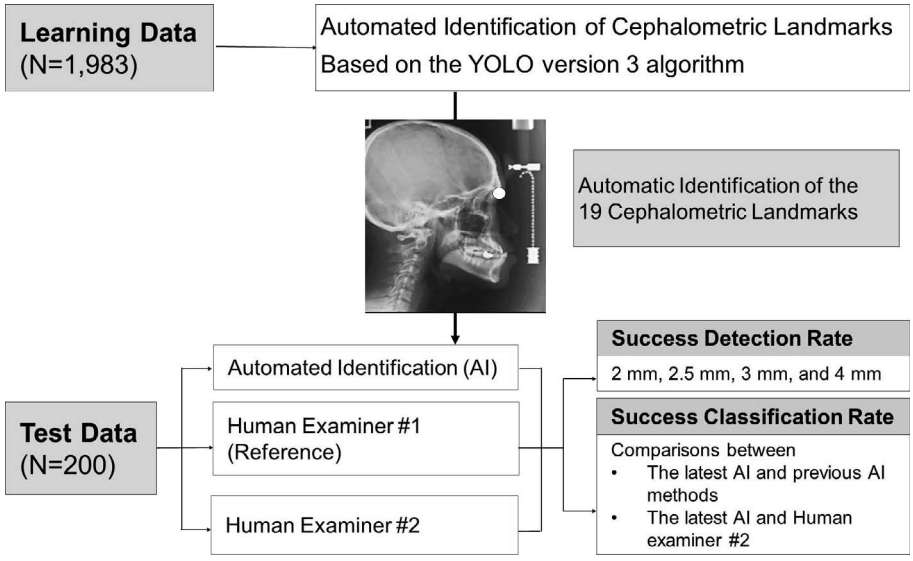

Figure 1. The experimental design of the present study.

account the overall results, the AI performance could be considered comparable to a human's analysis. In clinical practice, identifying anatomical landmarks, tracing anatomical structures, and diagnosing patients' problems used to be a significant, time-consuming task. Consequently, saving labor, time, and effort with the help of this improved AI would be advantageous to clinicians.

Previously published algorithms can be divided into two categories: random forest[23] and convolutional neural network.[1,24] Random forest is a machine learning algorithm developed in 2001 and has a relatively long history. Until fairly recently, it seemed the mainstream in the development of automated cephalometric analysis systems. For example, the majority of teams participating in the IEEE ISBI challenges in 2014 and 2015 applied the random forest algorithm. A major drawback of random forest is its complexity. The decision tree that makes up the random forest is a logical structure. However, with many decision tree collections, it is difficult to intuitively identify relationships that may exist in input data. In addition, the prediction process using random forest may take more time and require more computational resources than other algorithms. In comparison, convolutional neural network is a type of deep learning that has come into the spotlight relatively recently.[25] It was inspired by the natural visual recognition mechanism of living things and is known to be more suitable for image processing. The YOLO version 3 algorithm that was modified and applied in this present study was based on the convolutional neural network.[19]

Medical image analysis using deep learning technology could help clinicians make their diagnosis and treatment planning more efficient. Despite the ability to deliver better performance through deep learning methods, some limitations require careful attention when applying deep learning for use in clinical practice. First, deep learning structures would require a huge amount of training data and computing power. In some cases of deep learning procedures, even though the input and output values are known, their internal structure cannot be well explained. Finally, deep learning might be affected by the noise issues inherent in medical imaging. Deep learning algorithms can misread objects if even only a little noise, which would be invisible to humans, is added to the image.[2]

Despite the limitations, it seems that the performance of the latest deep learning algorithms continues to improve over the years. Deep learning techniques have the potential to be a great help in the development of medical image analysis, including orthodontic cephalometric analysis. In addition, if a clinician applies the deep learning AI cephalometric analysis with some manual modification, the results could be even better.

CONCLUSIONS

- The latest AI seems to have superior performance than previous AIs. It demonstrated comparable cephalometric analysis to human examiners.
- It is envisioned that AI will maintain, and even improve, its effectiveness under supervision by orthodontists.

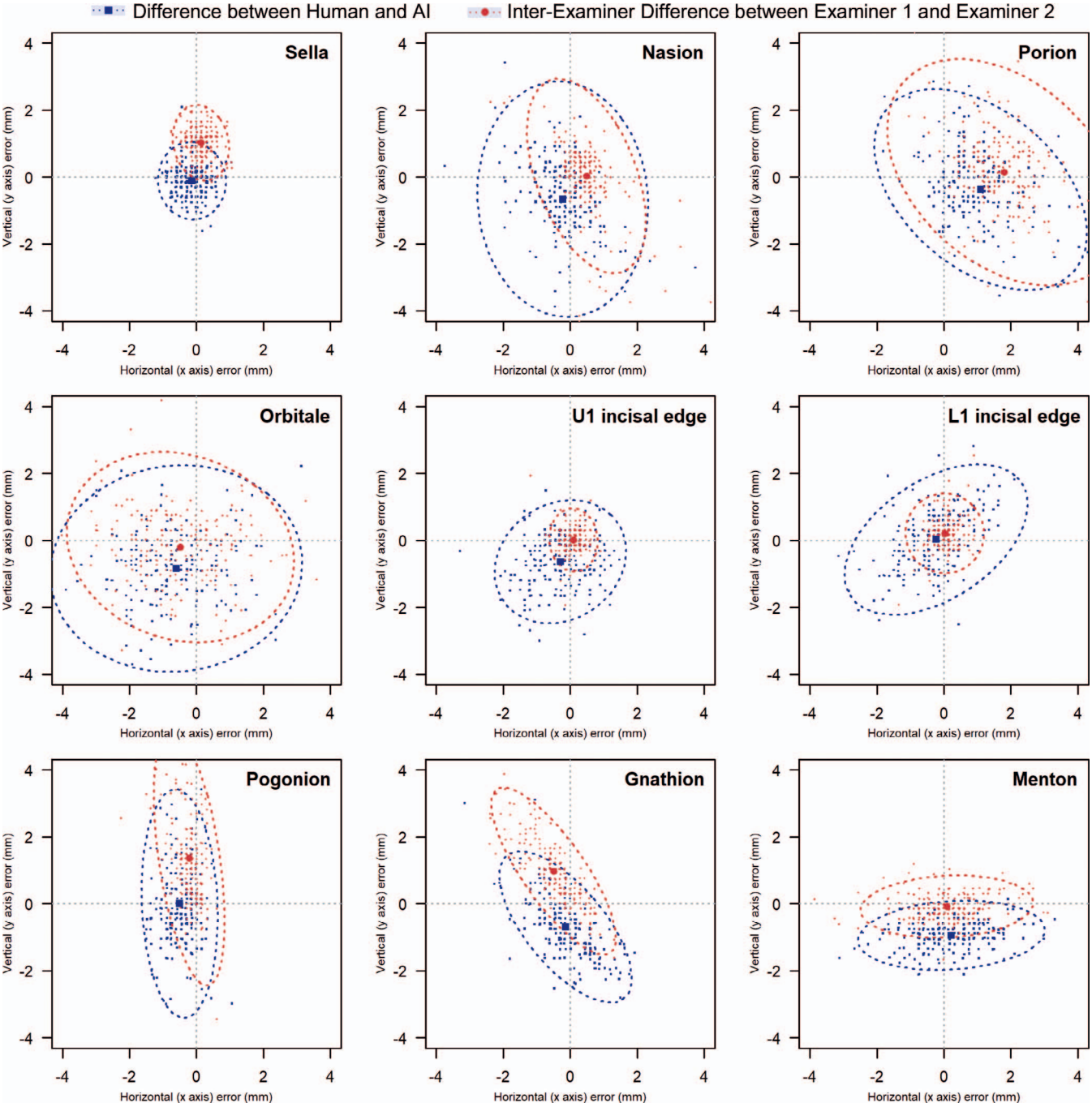

Figure 2. Scatter plots with 95% confidence ellipses for the landmark detection errors.

ACKNOWLEDGMENT

The data presented in the present study were part of a doctoral dissertation (HWH). This study was partly supported by grant 05-2020-0021 from the SNUDH Research Fund.

REFERENCES

1. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. *J Med Syst*. 2018;42:226.
2. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging*. 2018;9:611–629.
3. Moon JH, Hwang HW, Lee SJ. Evaluation of an automated superimposition method for computer-aided cephalometrics. *Angle Orthod*. 2020;90:390–396.
4. Hwang HW, Park JH, Moon JH, et al. Automated identification of cephalometric landmarks: part 2- Might it be better than human? *Angle Orthod*. 2020;90:69–76.
5. Qian J, Cheng M, Tao Y, Lin J, Lin H. CephaNet: An improved faster R-CNN for cephalometric landmark detec-

tion. Paper presented at: Institute of Electrical and Electronics Engineers' International Symposium on Biomedical Imaging; April 2019; Venice, Italy.
6. Park JH, Hwang HW, Moon JH, et al. Automated identification of cephalometric landmarks: part 1-Comparisons between the latest deep-learning methods YOLOV3 and SSD. *Angle Orthod*. 2019;89:903–909.
7. Wang S, Li H, Li J, Zhang Y, Zou B. Automatic analysis of lateral cephalograms based on multiresolution decision tree regression voting. *J Healthc Eng*. 2018;2018:1797502.
8. Arik SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. *J Med Imaging*. 2017;4:014501.
9. Wang CW, Huang CT, Lee JH, et al. A benchmark for comparison of dental radiography analysis algorithms. *Med Image Anal*. 2016;31:63–76.
10. Wang CW, Huang CT, Hsieh MC, et al. Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge. *IEEE Trans Med Imaging*. 2015;34:1890–1900.
11. Kang TJ, Eo SH, Cho H, Donatelli RE, Lee SJ. A sparse principal component analysis of Class III malocclusions. *Angle Orthod*. 2019;89:768–774.
12. Chen S, Wang L, Li G, et al. Machine learning in orthodontics: introducing a 3D auto-segmentation and auto-landmark finder of CBCT images to assess maxillary constriction in unilateral impacted canine patients. *Angle Orthod*. 2020;90:77–84.
13. Au J, Mei L, Bennani F, Kang A, Farella M. Three-dimensional analysis of lip changes in response to simulated maxillary incisor advancement. *Angle Orthod*. 2020;90:118–124.
14. Suh HY, Lee HJ, Lee YS, Eo SH, Donatelli RE, Lee SJ. Predicting soft tissue changes after orthognathic surgery: the sparse partial least squares method. *Angle Orthod*. 2019;89:910–916.
15. Yoon KS, Lee HJ, Lee SJ, Donatelli RE. Testing a better method of predicting postsurgery soft tissue response in Class II patients: a prospective study and validity assessment. *Angle Orthod*. 2015;85:597–603.
16. Lee YS, Suh HY, Lee SJ, Donatelli RE. A more accurate soft-tissue prediction model for Class III 2-jaw surgeries. *Am J Orthod Dentofacial Orthop*. 2014;146:724–733.
17. Suh HY, Lee SJ, Lee YS, et al. A more accurate method of predicting soft tissue changes after mandibular setback surgery. *J Oral Maxillofac Surg*. 2012;70:e553–562.
18. Lee HJ, Suh HY, Lee YS, et al. A better statistical method of predicting postsurgery soft tissue response in Class II patients. *Angle Orthod*. 2014;84:322–328.
19. Redmon J, Farhadi A. Yolov3: an incremental improvement. *arXiv preprint arXiv:1804.02767*. 2018.
20. Lindner C, Cootes TF. Fully automatic cephalometric evaluation using random forest regression-voting. Paper presented at: Institute of Electrical and Electronics Engineers' International Symposium on Biomedical Imaging; April 2015; Brooklyn, NY.
21. Ibragimov B, Likar B, Pernus F, Vrtovec T. Computerized cephalometry by game theory with shape-and appearance-based landmark refinement. Paper presented at: Institute of Electrical and Electronics Engineers' International Symposium on Biomedical Imaging; April 2015; Brooklyn, NY.
22. Moon JH, Hwang HW, Yu Y, Kim MG, Donatelli RE, Lee SJ. How much deep learning is enough for automatic identification to be reliable? A cephalometric example. *Angle Orthod*. 2020;90:823–830.
23. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
24. Gu JX, Wang ZH, Kuen J, et al. Recent advances in convolutional neural networks. *Pattern Recognit*. 2018;77:354–377.
25. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.