Original Article

pISSN 2234-7518 • eISSN 2005-372X
https://doi.org/10.4041/kjod.2022.52.1.3

THE KOREAN JOURNAL of ORTHODONTICS
KJO

Accuracy of one-step automated orthodontic diagnosis model using a convolutional neural network and lateral cephalogram images with different qualities obtained from nationwide multi-hospitals

Sunjin Yim[a]
Sungchul Kim[b]
Inhwan Kim[b]
Jae-Woo Park[c]
Jin-Hyoung Cho[d]
Mihee Hong[e]
Kyung-Hwa Kang[f]
Minji Kim[g]
Su-Jung Kim[h]
Yoon-Ji Kim[i]
Young Ho Kim[j]
Sung-Hoon Lim[k]
Sang Jin Sung[i]
Namkug Kim[l]
Seung-Hak Baek[m]

[a]Department of Orthodontics, School of Dentistry, Seoul National University, Seoul, Korea
[b]Department of Biomedical Engineering, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
[c]Private Practice, Incheon, Korea
[d]Department of Orthodontics, Chonnam National University School of Dentistry, Gwangju, Korea
[e]Department of Orthodontics, School of Dentistry, Kyungpook National University, Daegu, Korea
[f]Department of Orthodontics, School of Dentistry, Wonkwang University, Iksan, Korea
[g]Department of Orthodontics, College of Medicine, Ewha Womans University, Seoul, Korea
[h]Department of Orthodontics, Kyung Hee University School of Dentistry, Seoul, Korea
[i]Department of Orthodontics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
[j]Department of Orthodontics, Institute of Oral Health Science, Ajou University School of Medicine, Suwon, Korea
[k]Department of Orthodontics, College of Dentistry, Chosun University, Gwangju, Korea
[l]Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
[m]Department of Orthodontics, School of Dentistry, Dental Research Institute, Seoul National University, Seoul, Korea

Objective: The purpose of this study was to investigate the accuracy of one-step automated orthodontic diagnosis of skeletodental discrepancies using a convolutional neural network (CNN) and lateral cephalogram images with different qualities from nationwide multi-hospitals. **Methods:** Among 2,174 lateral cephalograms, 1,993 cephalograms from two hospitals were used for training and internal test sets and 181 cephalograms from eight other hospitals were used for an external test set. They were divided into three classification groups according to anteroposterior skeletal discrepancies (Class I, II, and III), vertical skeletal discrepancies (normodivergent, hypodivergent, and hyperdivergent patterns), and vertical dental discrepancies (normal overbite, deep bite, and open bite) as a gold standard. Pre-trained DenseNet-169 was used as a CNN classifier model. Diagnostic performance was evaluated by receiver operating characteristic (ROC) analysis, t-stochastic neighbor embedding (t-SNE), and gradient-weighted class activation mapping (Grad-CAM). **Results:** In the ROC analysis, the mean area under the curve and the mean accuracy of all classifications were high with both internal and external test sets (all, > 0.89 and > 0.80). In the t-SNE analysis, our model succeeded in creating good separation between three classification groups. Grad-CAM figures showed differences in the location and size of the focus areas between three classification groups in each diagnosis. **Conclusions:** Since the accuracy of our model was validated with both internal and external test sets, it shows the possible usefulness of a one-step automated orthodontic diagnosis tool using a CNN model. However, it still needs technical improvement in terms of classifying vertical dental discrepancies.

[Korean J Orthod 2022;52(1):3-19]

Key words: One-step automated orthodontic diagnosis, Convolutional neural networks, Lateral cephalogram, Multi-center study

Received March 23, 2021; Revised June 1, 2021; Accepted July 2, 2021.

Corresponding author: Seung-Hak Baek.
Professor, Department of Orthodontics, School of Dentistry, Dental Research Institute, Seoul National University, 101, Daehak-ro, Jongno-gu, Seoul 03080, Korea.
Tel +82-2-2072-3952 **e-mail** drwhite@unitel.co.kr
Corresponding author: Namkug Kim.
Professor, Department of Convergence Medicine, 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

Sunjin Yim and Sungchul Kim contributed equally to this work (as co-first authors).

How to cite this article: Yim S, Kim S, Kim I, Park JW, Cho JH, Hong M, Kang KH, Kim M, Kim SJ, Kim YJ, Kim YH, Lim SH, Sung SJ, Kim N, Baek SH. Accuracy of one-step automated orthodontic diagnosis model using a convolutional neural network and lateral cephalogram images with different qualities obtained from nationwide multi-hospitals. Korean J Orthod 2022;52:3-19.

INTRODUCTION

Accurate positioning of cephalometric landmarks is one of the most important steps in successful cephalometric analyses. Since the location and visibility of some anatomic landmarks are highly influenced by superimposition of the anatomical structures in the face between the right and left sides,[1,2] it is not easy to identify these anatomic landmarks consistently and accurately.

For the last several decades, clinicians have manually indicated the cephalometric landmarks and measured several angles and distances between these landmarks to assess dentofacial deformities.[3] Although this manual cephalometric analysis has been substituted with digital cephalometric analysis,[4,5] the process is still laborious, time-consuming, and sometimes inaccurate in detection of cephalometric landmarks.[3,6-8]

Recently, research on automatic detection of cephalometric landmarks using artificial intelligence (AI) with convolutional neural networks (CNNs) has gained popularity.[1-3,9-11] These studies have focused mainly on automatic detection of cephalometric landmarks and reported that most cephalometric landmarks were detected within a 2-mm range of accuracy.[1,10] However, these approaches still require further measurements of cephalometric parameters including distance, angle, and ratio. Although Kunz et al.[11] developed an AI algorithm to analyze 12 cephalometric parameters, they did not make a one-step automated orthodontic diagnosis tool in practice. Therefore, it is necessary to develop a one-step automated orthodontic diagnosis algorithm based on a CNN to avoid the need of additional measurements of cephalometric parameters.

In terms of a one-step CNN algorithm for classification of skeletal discrepancies, Yu et al.[8] reported > 90% accuracy, sensitivity, and specificity for diagnosis of the sagittal and vertical skeletal discrepancies in three models (Models I, II, and III). However, they intentionally excluded some portion of the data adjacent to the classification cutoff with intervals of 0.2 standard deviations (SDs) in Model II and 0.3 SDs in Model III in the test set.[8] As a result, Models II and III showed a significant increase in the values for accuracy, sensitivity, and specificity compared to Model I.[8]

The major limitations in previous studies can be summarized as follows:[1-3,8-11] (1) Most studies used lateral cephalograms from only one or two hospitals, not from nationwide several different hospitals which had different machine types, radiation exposure conditions, sensors, and image conditions; (2) No study has simultaneously reported dental and skeletal discrepancies using a one-step automated classification algorithm; and (3) If some portion of the data adjacent to the classification cutoff were excluded in the test set, there would be issues in the continuity of the test set and an exaggerated increase in accuracy. Therefore, the purpose of this study was to investigate the accuracy of a novel one-step automated orthodontic diagnosis model for determining anteroposterior skeletal discrepancies (APSDs: Class I, Class II, and Class III), vertical skeletal discrepancies (VSDs: normodivergent, hyperdivergent, and hypodivergent), and vertical dental discrepancies (VDDs: normal overbite, open bite, and deep bite) using a CNN and lateral cephalogram images with different qualities from

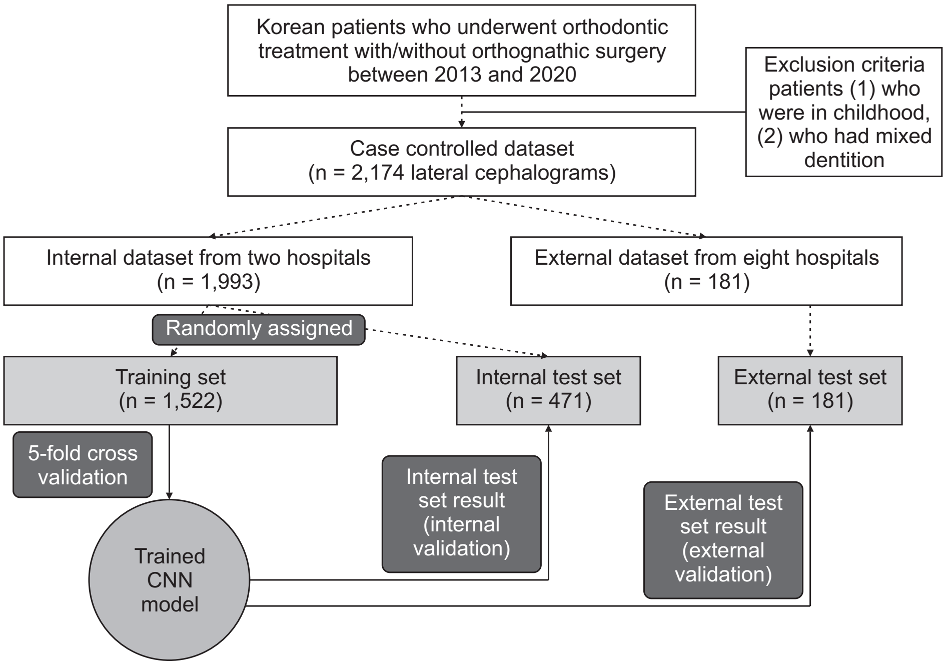
Korean patients who underwent orthodontic treatment with/without orthognathic surgery between 2013 and 2020
Exclusion criteria patients (1) who were in childhood, (2) who had mixed dentition
Case controlled dataset (n = 2,174 lateral cephalograms)
Internal dataset from two hospitals (n = 1,993)
External dataset from eight hospitals (n = 181)
Randomly assigned
Training set (n = 1,522)
Internal test set (n = 471)
External test set (n = 181)
5-fold cross validation
Internal test set result (internal validation)
External test set result (external validation)
Trained CNN model

Figure 1. Flowchart of dataset and experimental setup. CNN, convolutional neural network.

Table 1. Information on the product, radiation exposure condition, sensor, and image condition of the cephalometric radiograph system in 10 multi-centers

Cephalometric radiograph systems		**SNUDH**	**KADH**	**AJUDH**	**AMC**	**CNUDH**	**CSUDH**	**EUMC**	**KHUDH**	**KNUDH**	**WKUDH**
Product	Company	Asahi	Vatech	Planmeca	Carestream	Instrumentarium	Planmeca	Asahi	Asahi	Asahi	Planmeca
	Model	CX-90SP-II	Uni3D NC	Proline XC	CS9300	OrthoCeph OC 100	Proline XC	Ortho stage (Auto III N CM)	CX-90SP	CX-90SP-II	Promax
Radiation exposure condition	Kvp	76	85	68	80	85	80	75	70	70	Female 72, Male 74
	mA	80	10	7	12	12	12	15	15	80	10
	sec	0.32	0.9	2.3	0.63	1.6	1.8	1	0.3–0.35	0.32	1.87
Sensor	Image sensor	Cassette (CR system)	CCD sensor	CCD sensor	CCD sensor	Cassette (CR system)	Cassette (CR system)	Cassette (CR system)	Cassette (CR system)	Cassette (CR system)	CCD sensor
	Sensor size	10 × 12 (inch)	30 × 25 (cm)	10.6 × 8.85 (inch)	30 × 30 (cm)	10 × 12 (inch)	8 × 10 (inch)	8 × 12 (inch)	10 × 12 (inch)	11 × 14 (inch)	27 × 30 (cm)
Image	Image size (pixel × pixel)	2,000 × 2,510 / 2,010 × 1,670	2,360 × 1,880	1,039 × 1,200	2,045 × 2,272 / 1,012 × 2,020	2,500 × 2,048	2,392 × 1,792 / various	2,510 × 2,000	2,500 × 2,048	1,950 × 2,460 / 2,108 × 1,752	1,818 × 2,272
	Actual resolution (mm/pixel)	0.150 / 0.100	0.110	0.250	0.132 / 0.145	0.115	0.100	0.100	0.110	0.100	0.132
Lateral cephalogram images used in this study (number)		1,129	864	22	21	20	30	26	23	19	20

SNUDH, Seoul National University Dental Hospital; KADH, Kooalldam Dental Hospital; AJUDH, Ajou University Dental Hospital; AMC, Asan Medical Center; CNUDH, Chonnam National University Dental Hospital; CSUDH, Chosun University Dental Hospital; EUMC, Ewha University Medical Center; KHUDH, Kyung Hee University Dental Hospital; KNUDH, Kyungpook National University Dental Hospital; WKUDH, Wonkwang University Dental Hospital; CR, computed radiography; CCD, charge-coupled device.

https://doi.org/10.4041/kjod.2022.52.1.3

nationwide 10 unrelated dental hospitals in Korea.

MATERIALS AND METHODS

Description of the dataset

A total of 2,174 lateral cephalogram images were retrospectively obtained from the Departments of Orthodontics in nationwide 10 hospitals including Seoul National University Hospital (SNUDH), Kooalldam Dental Hospital (KADH), Ajou University Dental Hospital (AJUDH), Asan Medical Center (AMC), Chonnam National University Dental Hospital (CNUDH), Chosun University Dental Hospital (CSUDH), Ewha University Medical Center (EUMC), Kyung Hee University Dental Hospital (KHUDH), Kyungpook National University Dental Hospital (KNUDH), and Wonkwang University Dental Hospital (WKUDH) in Korea. The inclusion criteria were Korean adult patients who underwent orthodontic treatment with/without orthognathic surgery between 2013 and 2020. The exclusion criteria were (1) patients who were in childhood and adolescence and (2) patients who had mixed dentition. All datasets were strictly anonymized before use. The study protocol was reviewed and approved by the Institutional Review Board of SNUDH (ERI20022), Korean National Institute for Bioethics Policy for KADH (P01-202010-21-020), Ajou University Hospital Human Research Protection Center (AJIRB-MED-MDB-19-039), AMC (2019-0927), CNUDH (CNUDH-2019-004), CSUDH (CUDHIRB 1901 005), EUMC (EUMC 2019-04-017-003), KHUDH (D19-007-003), KNUDH (KNUDH-2019-03-02-00), and WKUDH (WKDIRB201903-01).

Lateral cephalogram images, 1,993 from two hospitals, were used for the training set (n = 1,522) and internal test set (n = 471), and 181 from eight other hospitals were used as the external test set to validate our model (Figure 1). Table 1 summarizes information on the product, radiation exposure condition, sensor, and image conditions in each hospital, which showed diverse conditions.

Setting a gold standard for the diagnosis of APSDs, VSDs, and VDDs

After detection of the cephalometric landmarks including A point, nasion, B point, orbitale, porion, gonion, menton, sella, maxilla 1 crown, maxilla 6 distal, mandible 1 crown, and mandible 6 distal by a single operator (SY), the cephalometric parameters including A point-Nasion-B point (ANB) angle, Frankfort mandibular plane angle (FMA), Jarabak's posterior/anterior facial height ratio (FHR), and overbite were calculated using V-Ceph 8.0 (Osstem, Seoul, Korea) to set a gold standard.

All cephalometric images were classified into the three classification groups by a single operator (SY) as follows. For classification of APSDs, we defined the ANB value between −1 SD and 1 SD from the ethnic norm of each sex[12] as skeletal Class I; > 1 SD as skeletal Class II; and < −1 SD as skeletal Class III. For classification of VSDs, we combined FMA and FHR values from the ethnic norm of each sex[12] for training. First, we normalized the FMA and FHR values by using the SD values. Second, the FHR values were flipped due to an opposite sign compared to the FMA values. Third, the values of FMA and flipped FHR were added because each are regarded as having equal weights. Fourth, the mean and SD values were obtained for classification into three groups. Then, we defined the values between −1 SD and 1 SD from the mean as normodivergent pattern, > 1 SD as hyperdivergent pattern, and < −1 SD as hypodivergent pattern. For classification of the VDDs, we defined the overbite value between 0 mm and 3 mm as a normal overbite, > 3 mm as a deep bite, and < 0 mm as an open bite (Tables 2 and 3).

To assess intra-examiner reliability, all classifications of APSDs, VSDs, and VDDs were performed again after one month by the same investigator (SY). Since the minimum sample size[13] was suggested as 49 from a 3 × 3 Cohen's kappa agreement test, 100 images were randomly selected to classify APSDs, VSDs, and VDDs. Cohen's kappa agreement test showed an "almost perfect" agreement (kappa value; 0.939 for APSDs, 0.984 for VSDs, and 0.907 for VDDs).[14] Therefore, the first classification results were used for further statistical analysis.

Table 2. Classification criteria for the anteroposterior skeletal discrepancies (APSDs), vertical skeletal discrepancies (VSDs), and vertical dental discrepancies (VDDs) for orthodontic analysis

Sex	APSDs		VSDs				VDDs	
	ANB		FMA		FHR		Overbite	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Female	2.4	1.8	24.2	4.6	65	9	1.5	1.5
Male	1.78	2.02	26.78	1.79	66.37	5.07		

ANB, angle among A point, nasion, and B point; FMA, Frankfort mandibular plane angle; FHR, Jarabak's posterior/anterior facial height ratio; SD, standard deviation.

Table 3. Distribution of classification groups in each diagnosis for human gold standard in the training set, internal test set, and external test set

Classifications		Training set			Internal test set			External test set									Sum	
		SNUDH	KADH	Sum	SNUDH	KADH	Sum	AJUDH	AMC	EUMC	CNUDH	CSUDH	KHUDH	KNUDH	WKUDH	Sum	Internal + external test sets	Total
APSDs	Class I	238	323	561 (36.9)	122	40	162 (34.4)	8	6	5	4	7	11	7	2	50 (27.6)	212 (32.5)	773 (35.6)
	Class II	183	263	446 (29.3)	112	44	156 (33.1)	8	8	11	8	13	4	4	6	62 (34.3)	218 (33.4)	664 (30.5)
	Class III	359	156	515 (33.8)	115	38	153 (32.5)	6	7	10	8	10	8	8	12	69 (38.1)	222 (34.0)	737 (33.9)
	Sum	780	742	1,522	349	122	471	22	21	26	20	30	23	19	20	181	652	2,174
VSDs	Normodivergent	331	389	720 (47.3)	146	50	196 (41.6)	10	6	7	9	17	10	7	7	73 (40.3)	270 (41.4)	989 (45.5)
	Hyperdivergent	314	241	555 (36.5)	135	40	175 (37.2)	5	9	12	6	3	7	8	6	56 (30.9)	231 (35.4)	786 (36.2)
	Hypodivergent	135	112	247 (16.2)	68	32	100 (21.2)	7	6	7	5	10	6	4	7	52 (28.7)	151 (23.2)	399 (18.4)
	Sum	780	742	1,522	349	122	471	22	21	26	20	30	23	19	20	181	652	2,174
VDDs	Normal overbite	440	493	933 (61.3)	196	53	249 (52.9)	11	11	10	8	9	10	10	10	79 (43.6)	328 (50.3)	1,261 (58.0)
	Open bite	209	194	403 (26.5)	99	41	140 (29.7)	4	7	9	5	9	8	4	5	51 (28.2)	191 (29.3)	594 (27.3)
	Deep bite	131	55	186 (12.2)	54	28	82 (17.4)	7	3	7	7	12	5	5	5	51 (28.2)	133 (20.4)	319 (14.7)
	Sum	780	742	1,522	349	122	471	22	21	26	20	30	23	19	20	181	652	2,174

Values are presented as number only or number (%).

APSDs, anteroposterior skeletal discrepancies; VSDs, vertical skeletal discrepancies; VDDs, vertical dental discrepancies; SNUDH, Seoul National University Dental Hospital; KADH, Kooalldam Dental Hospital; AJUDH, Ajou University Dental Hospital; AMC, Asan Medical Center; EUMC, Ewha University Medical Center; CNUDH, Chonnam National University Dental Hospital; CSUDH, Chosun University Dental Hospital; KHUDH, Kyunghee University Dental Hospital; KNUDH, Kyungpook National University Dental Hospital; WKUDH, Wonkwang University Dental Hospital.

To evaluate inter-examiner reliability, the same images used to assess intra-examiner reliability were selected. Classifications of APSDs, VSDs, and VDDs were performed by the other investigator (KL). Cohen's kappa agreement test showed an "almost perfect" agreement for APSDs and VSDs (kappa value; 0.985 for APSDs, 0.919 for VSDs) and "substantial' agreement" for VDDs (0.601).[14]

Preprocessing of the data

Augmentation techniques including cropping, padding, spatial transformations, and appearance transformation were conducted in real time.

Model architecture (Figure 2)

As the backbone of the model, DenseNet-169 pretrained with weights of the ImageNet dataset was used with group normalization (GN).[15-20] After the global average pooling (GAP) of the backbone, ArcFace was added in parallel with the softmax layer in order to overcome imbalanced data sets and obtain discriminative features during training.[21]

After training, the ArcFace head was removed, and inference was implemented using only the softmax layer as a basic CNN classifier. Because sex was included as a classification criterion of APSDs and VSDs, the one-hot vector about sex was concatenated with the feature vector after GAP.

Model training (Figures 1 and 2)

Training for APSDs, VSDs, and VDDs was performed using only a gold standard determined by a single operator (SY), not by measurement of cephalometric parameters including ANB, FMA, FHR, and overbite.

Model testing

After training was completed, one-step classification was performed with both the internal and external test sets to validate the performance of the constructed model. It took 55 seconds (sec) to diagnose the internal test set (0.1168 sec per lateral cephalogram) and 22 sec to diagnose the external test set (0.1215 sec per lateral cephalogram). The results for the internal and external test sets were compared with gold standard diagnostic data.

Analysis method

Receiver operating characteristic (ROC) analysis

The performance of our model was evaluated using accuracy, area under the curve (AUC), sensitivity, and specificity using both binary and multiple class ROC analysis.[8,22,23]

t-stochastic neighbor embedding (t-SNE)

Since this technique can visualize high-dimensional data by giving each datapoint a location in a two or three-dimensional map, it was used to check the feature distribution of the training set, internal test set, and external test set after GAP layering.[24] In each diagnosis, the labels of ground truth (GT) and prediction (PD) were set to check the distribution of each data set.

Gradient-weighted class activation mapping (Grad-CAM)[25]

As this technique can produce visual explanations of AI models, it can show the regions where the AI focuses for PD. It was used to confirm the regions where our model mainly focused on the diagnosis of APSDs, VSDs, and VDDs.

RESULTS

Metrology distribution of the APSDs, VSDs, and VDDs per dataset (Figure 3)

The continuity of the dataset between the normal groups (Class I in APSDs, normodivergent pattern in VSDs, and normal overbite in VDDs) and the other two groups (Class II and III in APSDs, hyperdivergent and hypodivergent patterns in VSDs, and open bite and deep

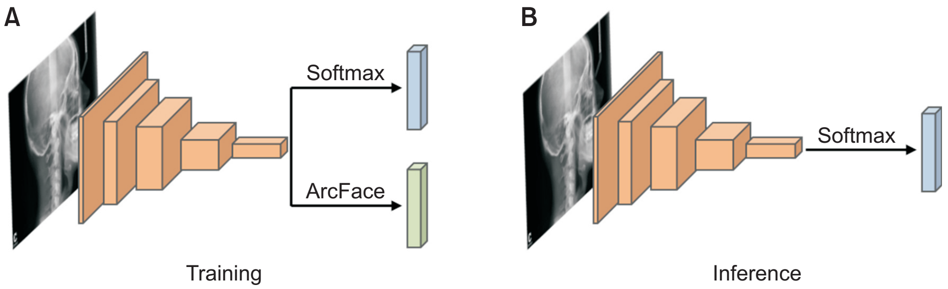

Figure 2. Diagrams of the model architecture. **A**, During training, an ArcFace head was added to the last convolutional layer of the backbone in parallel with the softmax layer. **B**, After training, the ArcFace head was removed and inference was implemented using only the softmax layer.

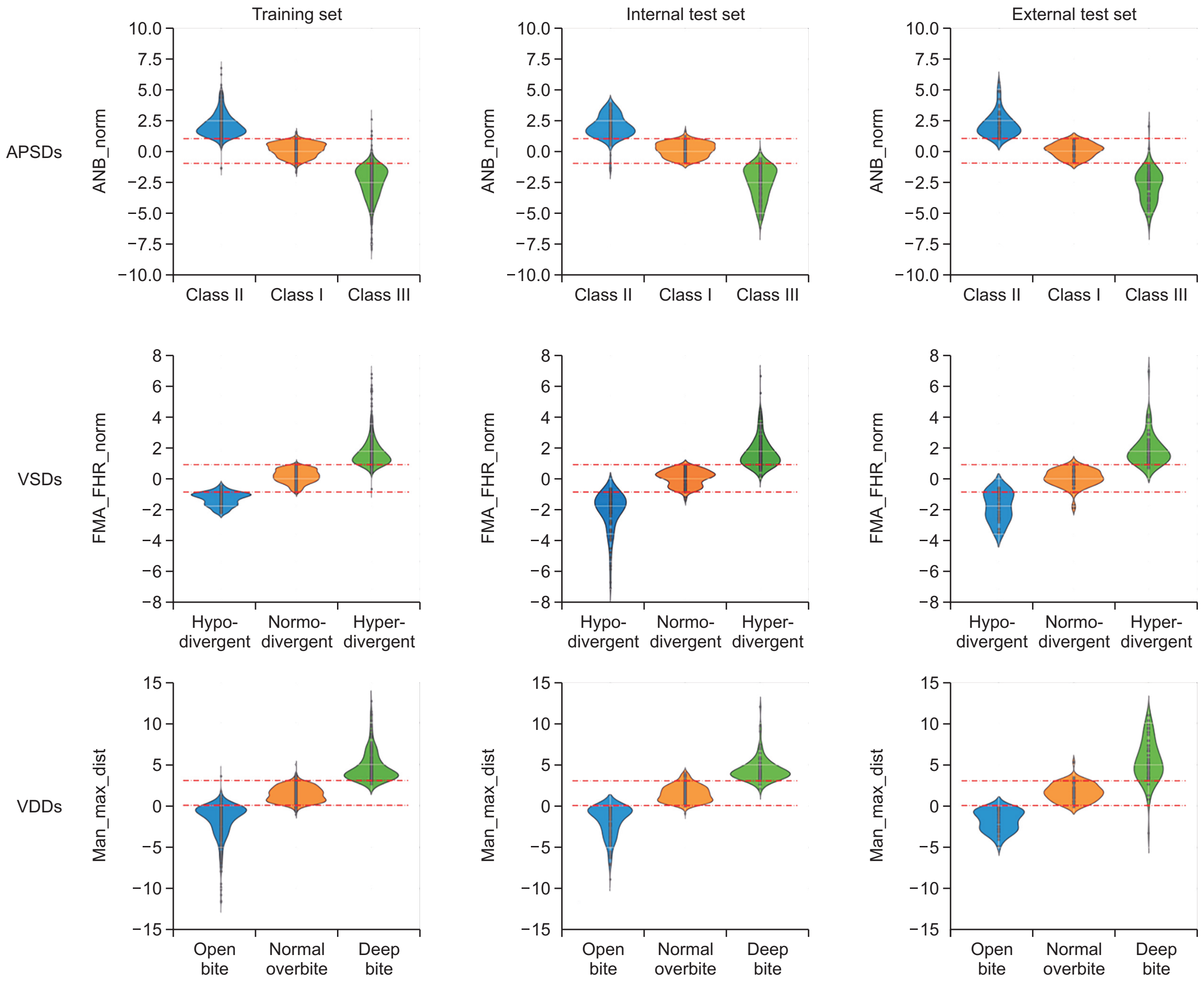

Figure 3. Metrology distribution of the anteroposterior skeletal discrepancies (APSDs: Class I, Class II, and Class III), vertical skeletal discrepancies (VSDs: normodivergent pattern, hyperdivergent pattern, and hypodivergent pattern), and vertical dental discrepancies (VDDs: normal overbite, open bite, and deep bite) per dataset. Red lines in APSDs and VSDs indicate one standard deviation of the normal classification. Red lines in VDDs indicate the boundary values, which were 0 mm and 3 mm.

ANB, angle among A point, nasion, and B point; FMA, Frankfort mandibular plane angle; FHR, Jarabak's posterior/anterior facial height ratio; norm, normalized; Man, mandible 1 crown; Max, maxilla 1 crown; dist, distance.

bite in VDDs) was confirmed.

Accuracy and AUC of the internal test set in binary ROC analysis (Table 4 and Figure 4)

In APSDs, Class III had the highest accuracy and AUC (0.9372 and 0.9807, respectively), followed by Class II (0.8972 and 0.9533, respectively) and Class I (0.8488 and 0.9212, respectively). In VSDs, hypodivergent pattern had the highest accuracy and AUC (0.9346 and 0.9824, respectively), followed by hyperdivergent pattern (0.9019 and 0.9730, respectively) and normodivergent pattern (0.8365 and 0.9186, respectively). In VDDs, open bite had the highest accuracy and AUC (0.8730 and 0.9475, respectively), followed by deep bite (0.8637 and 0.9286, respectively) and normal overbite (0.7376 and 0.8177, respectively).

In APSDs and VSDs, the total accuracy reached nearly 0.9 and the total AUC exceeded 0.95 (0.9517 and 0.9580, respectively). However, VDDs showed a relatively lower total accuracy (0.8248 vs. 0.8944 and 0.8910) and total AUC (0.8979 vs. 0.9517 and 0.9580) than APSDs and VSDs.

Table 4. Performance of our model for the diagnosis of the APSDs, VSDs, and VDDs in the internal test set and external test set using the binary ROC analysis

Classifications		Accuracy				AUC				Sensitivity				Specificity			
		Internal test set		External test set		Internal test set		External test set		Internal test set		External test set		Internal test set		External test set	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
APSDs	Class I	0.8488	0.0103	0.8320	0.0230	0.9212	0.0038	0.9042	0.0195	0.7938	0.0328	0.7840	0.0297	0.8764	0.0186	0.8504	0.0273
	Class II	0.8972	0.0057	0.8796	0.0153	0.9533	0.0026	0.9601	0.0067	0.8192	0.0334	0.7226	0.0515	0.9359	0.0161	0.9613	0.0046
	Class III	0.9372	0.0063	0.9525	0.0108	0.9807	0.0025	0.9930	0.0023	0.9111	0.0225	0.9652	0.0079	0.9497	0.0086	0.9446	0.0160
	Mean	0.8944	0.0368	0.8880	0.0518	0.9517	0.0245	0.9524	0.0382	0.8414	0.0571	0.8239	0.1076	0.9206	0.0345	0.9188	0.0516
VSDs	Normodivergent	0.8365	0.0082	0.8309	0.0267	0.9186	0.0046	0.9157	0.0151	0.8235	0.0279	0.7699	0.0416	0.8458	0.0122	0.8722	0.0178
	Hyperdivergent	0.9019	0.0035	0.9061	0.0203	0.9730	0.0047	0.9730	0.0047	0.8149	0.0273	0.9143	0.0293	0.9534	0.0190	0.9024	0.0360
	Hypodivergent	0.9346	0.0098	0.9094	0.0164	0.9824	0.0015	0.9684	0.0026	0.9000	0.0394	0.8000	0.0661	0.9445	0.0127	0.9535	0.0110
	Mean	0.8910	0.0413	0.8821	0.0410	0.9580	0.0283	0.9523	0.0273	0.8461	0.0478	0.8280	0.0757	0.9146	0.0505	0.9094	0.0398
VDDs	Normal overbite	0.7376	0.0291	0.7591	0.0230	0.8177	0.0166	0.8359	0.0152	0.6530	0.0956	0.6582	0.0664	0.8288	0.0441	0.8373	0.0557
	Open bite	0.8730	0.0130	0.8917	0.0139	0.9475	0.0053	0.9626	0.0074	0.8371	0.0366	0.8275	0.0611	0.8882	0.0304	0.9262	0.0228
	Deep bite	0.8637	0.0270	0.8586	0.0127	0.9286	0.0099	0.9238	0.0055	0.8000	0.1100	0.8196	0.0836	0.8781	0.0530	0.8723	0.0457
	Mean	0.8248	0.0654	0.8365	0.0584	0.8979	0.0582	0.9074	0.0538	0.7634	0.1111	0.7684	0.1006	0.8651	0.0468	0.8786	0.0535

APSDs, anteroposterior skeletal discrepancies; VSDs, vertical skeletal discrepancies; VDDs, vertical dental discrepancies; ROC, receiver operating characteristic; AUC, area under the curve; SD, standard deviation.

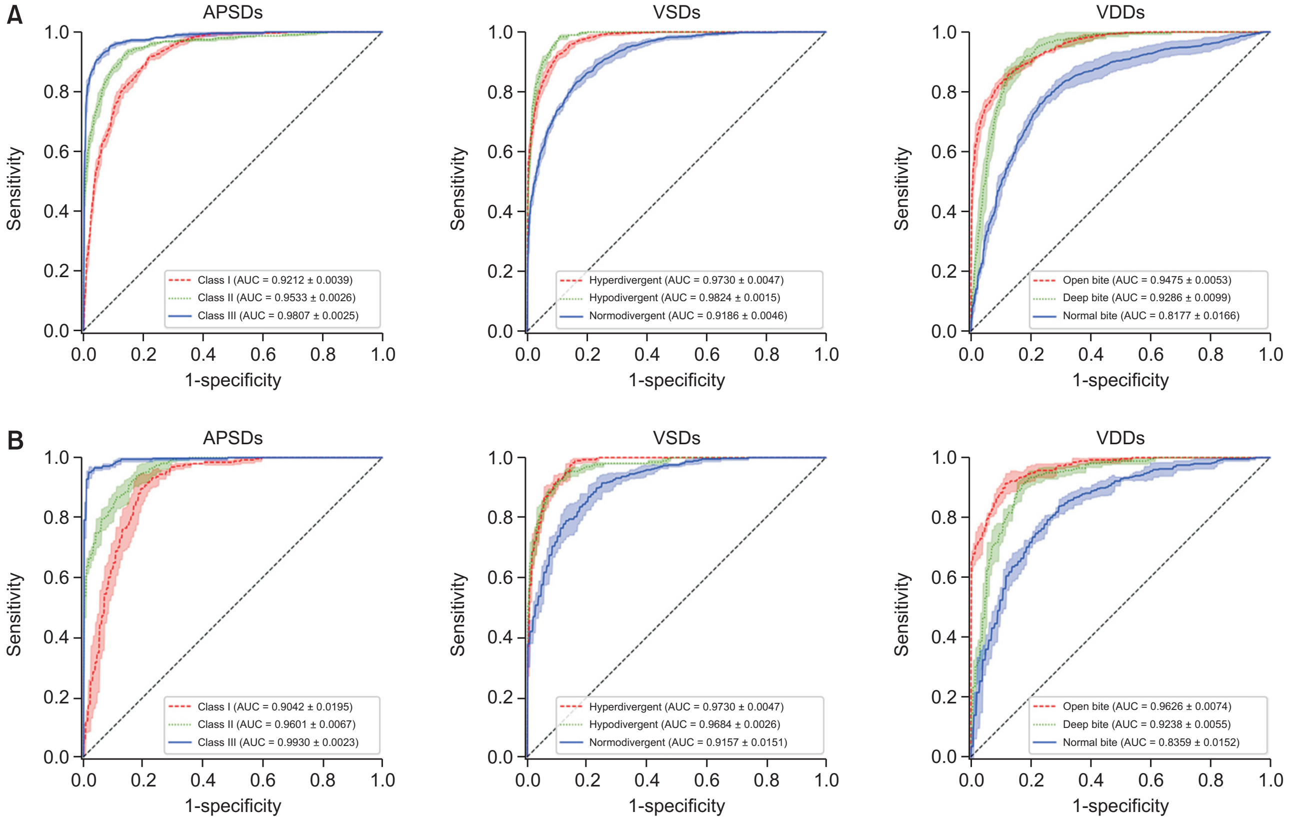

Figure 4. The results of the binary receiver operating characteristic curve analysis (**A**) in the internal test set from two hospitals and (**B**) in the external test set from other 8 hospitals for diagnosis of anteroposterior skeletal discrepancies (APSDs), vertical skeletal discrepancies (VSDs), and vertical dental discrepancies (VDDs).
AUC, area under the curve.

Accuracy and AUC of the external test set in binary ROC analysis (Table 4 and Figure 4)

In APSDs, Class III had the highest accuracy and AUC (0.9525 and 0.9930, respectively), followed by Class II (0.8796 and 0.9601, respectively) and Class I (0.8320 and 0.9042, respectively). In VDDs, open bite had the highest accuracy and AUC (0.8917 and 0.9626, respectively), followed by deep bite (0.8586 and 0.9238, respectively) and normal overbite (0.7591 and 0.8359, respectively). However, VSDs showed a different pattern between accuracy and AUC. Although the accuracy was highest for hypodivergent pattern (0.9094), followed by hyperdivergent pattern (0.9061) and normodivergent pattern (0.8309), the AUC was highest for hyperdivergent pattern (0.9730), followed by hypodivergent pattern (0.9684) and normodivergent pattern (0.9157).

In APSDs and VSDs, the total accuracy reached nearly 0.9 and the total AUC exceeded 0.95. However, VDDs showed a relatively lower total accuracy (0.8365 vs. 0.8880 and 0.8821) and total AUC (0.9074 vs. 0.9524 and 0.9523) than APSDs and VSDs.

Comparison of AUC values between internal and external test sets in binary ROC analysis (Table 4)

In APSDs and VSDs, Class III and open bite showed the highest AUC compared to other classifications (0.9807 and 0.9903 in the internal test set, 0.9475 and 0.9626 in external test set, respectively). However, VSDs showed a different pattern. The internal test set showed the highest AUC for hypodivergent pattern (0.9824), while the external test set showed the highest AUC for hyperdivergent pattern (0.9730). However, the difference in the AUC values was less than 0.01.

Comparison of AUC values between internal and external test sets in multiple ROC analysis (Table 5)

In terms of pairwise AUCs in the internal and external test sets of APSDs, VSDs, and VDDs, Class II vs. Class III ([II→III, 0.9913; II←III, 0.9920]; [II→III, 0.9992; II←III, 0.9989]; Δ value [II→III, 0.0079; II←III, 0.0069]), hyperdivergent pattern vs. hypodivergent pattern ([hyper→hypo, 0.9998; hyper←hypo, 0.9998]; [hyper→hypo, 0.9930; hyper←hypo, 0.9977]; Δ value [hyper→hypo, −0.0068; hyper←hypo,

Table 5. Performance of our model for the diagnosis of the APSDs, VSDs, and VDDs in the internal test set and external test set using the multiple ROC analysis

Classifications		Accuracy				Pairwise AUC				Pairwise sensitivity				Pairwise specificity			
		Internal test set		External test set		Internal test set		External test set		Internal test set		External test set		Internal test set		External test set	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
APSDs	Class I → Class II	0.8503	0.0086	0.8054	0.0222	0.8943	0.0106	0.8222	0.3830	0.8802	0.0283	0.9080	0.0098	0.8192	0.0299	0.7226	0.0461
	Class I ← Class II					0.9175	0.0039	0.9061	0.0136	0.8192	0.0299	0.7226	0.0461	0.8802	0.0283	0.9080	0.0098
	Class I → Class III	0.9143	0.0092	0.9277	0.0147	0.9486	0.0057	0.9780	0.0039	0.9173	0.0149	0.8760	0.0320	0.9111	0.0201	0.9652	0.0071
	Class I ← Class III					0.9698	0.0035	0.9856	0.0032	0.9111	0.0201	0.9652	0.0071	0.9173	0.0149	0.8760	0.0320
	Class II → Class III	0.9754	0.0033	0.9725	0.0142	0.9913	0.0014	0.9992	0.0009	0.9654	0.0077	0.9419	0.0299	0.9856	0.0026	1.0000	0.0000
	Class II ← Class III					0.9920	0.0013	0.9989	0.0013	0.9856	0.0026	1.0000	0.0000	0.9654	0.0077	0.9419	0.0299
VSDs	Hyper → Hypo	0.9905	0.0037	0.9778	0.0126	0.9998	0.0002	0.9930	0.0019	0.9851	0.0058	1.0000	0.0000	1.0000	0.0000	0.9538	0.0261
	Hyper ← Hypo					0.9998	0.0001	0.9977	0.0003	1.0000	0.0000	0.9538	0.0261	0.9851	0.0058	1.0000	0.0000
	Hyper → Normo	0.8755	0.0040	0.8791	0.0223	0.9593	0.0063	0.9587	0.0068	0.8149	0.0244	0.9143	0.0262	0.9296	0.0257	0.8521	0.0485
	Hyper ← Normo					0.9034	0.0088	0.9329	0.0119	0.9296	0.0257	0.8521	0.0485	0.8149	0.0244	0.9143	0.0262
	Hypo → Normo	0.8959	0.0139	0.8688	0.0212	0.9669	0.0024	0.9459	0.0042	0.9000	0.0352	0.8000	0.0591	0.8939	0.0231	0.9178	0.0173
	Hypo ← Normo					0.9451	0.0153	0.8972	0.0316	0.8939	0.0231	0.9178	0.0173	0.9000	0.0352	0.8000	0.0591
VDDs	Open → Deep	0.9766	0.0112	0.9706	0.0186	0.9982	0.0012	0.9924	0.0044	0.9814	0.0116	0.9922	0.0096	0.9683	0.0412	0.9490	0.0319
	Open ← Deep					0.9951	0.0066	0.9956	0.0042	0.9683	0.0412	0.949	0.0319	0.9814	0.0116	0.9922	0.0096
	Open → Normal	0.8463	0.0141	0.8538	0.0201	0.9308	0.0063	0.9434	0.0084	0.8414	0.0318	0.8314	0.0520	0.8490	0.0363	0.8684	0.0284
	Open ← Normal					0.8190	0.0452	0.8373	0.0341	0.8490	0.0363	0.8684	0.0284	0.8414	0.0318	0.8314	0.0520
	Deep → Normal	0.8066	0.0338	0.8062	0.0132	0.8911	0.0130	0.8775	0.0089	0.8000	0.0984	0.8275	0.0788	0.8088	0.0741	0.7924	0.0682
	Deep ← Normal					0.8156	0.0388	0.8345	0.0277	0.8088	0.0741	0.7924	0.0682	0.8000	0.0984	0.8275	0.0788

ROC curve analysis with multiple classification tasks was performed.

APSDs, anteroposterior skeletal discrepancies; VSDs, vertical skeletal discrepancies; VDDs, vertical dental discrepancies; ROC, receiver operating characteristic; AUC, area under the curve; SD, standard deviation; Hyper, hyperdivergent; Hypo, hypodivergent; Normo, normodivergent; Open, open bite; Deep, deep bite; Normal, normal overbite.

–0.0021]), and open bite vs. deep bite ([open→deep, 0.9982; open←deep, 0.9951]; [open→deep, 0.9924; open←deep, 0.9956]; Δ value [open→deep, –0.0058; open←deep, 0.0005]) showed the highest values in both the internal and external test sets and the smallest differences compared to other pairwise classifications.

t-SNE of APSDs, VSDs, and VDDs per dataset (Figure 5)

The GT in the training set, internal test set, and external test set showed that dots with different colors were mixed irregularly in the classification cutoff areas (dotted circle in Figure 5, GT) between the normal group (Class I in APSDs, normodivergent pattern in VSDs, and normal overbite in VDDs) and the other two groups (Class II and III for APSDs, hyperdivergent and hypodivergent patterns for VSDs, and open bite and deep bite for VDDs).

However, in the AI PD, the areas with irregular mixing had almost disappeared enough to indicate a cutoff line between the normal group and the other two groups in the training set, internal test set, and external test set (Figure 5, PD). This indicated that our model succeeded in creating good separation between the three classification groups in each diagnosis, resulting in consistent classification within each group.

Grad-CAM for each diagnosis (Figure 6)

Heat maps show differences in the location and size of the focus areas between three classification groups in

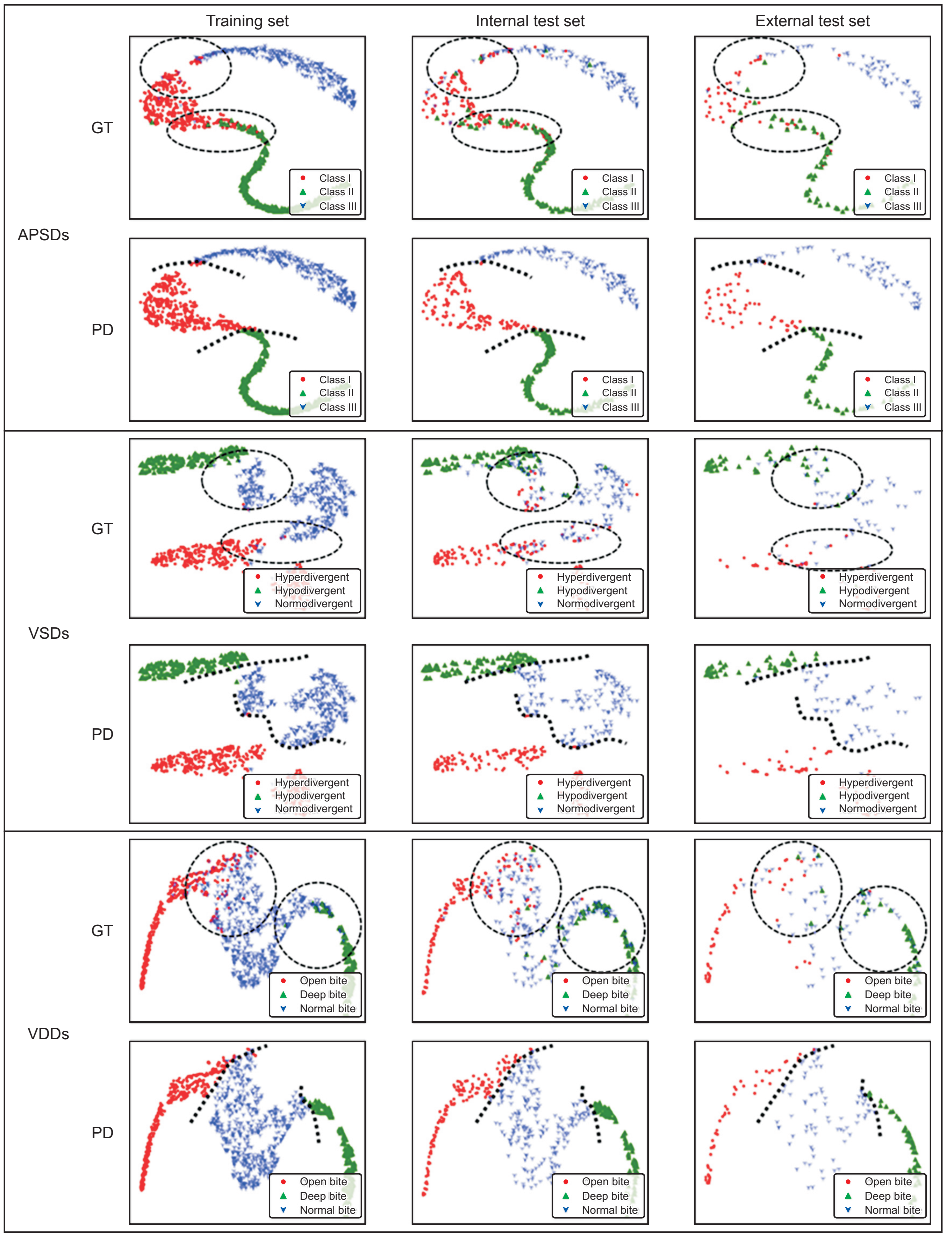

Figure 5. The results of t-stochastic neighbor embedding in anteroposterior skeletal discrepancies (APSDs), vertical skeletal discrepancies (VSDs), and vertical dental discrepancies (VDDs) per dataset. The labels of ground truth (GT) and prediction (PD) were set to check their distribution. Dotted circles indicate areas with irregular mixing. Dotted lines indicate cutoff lines.

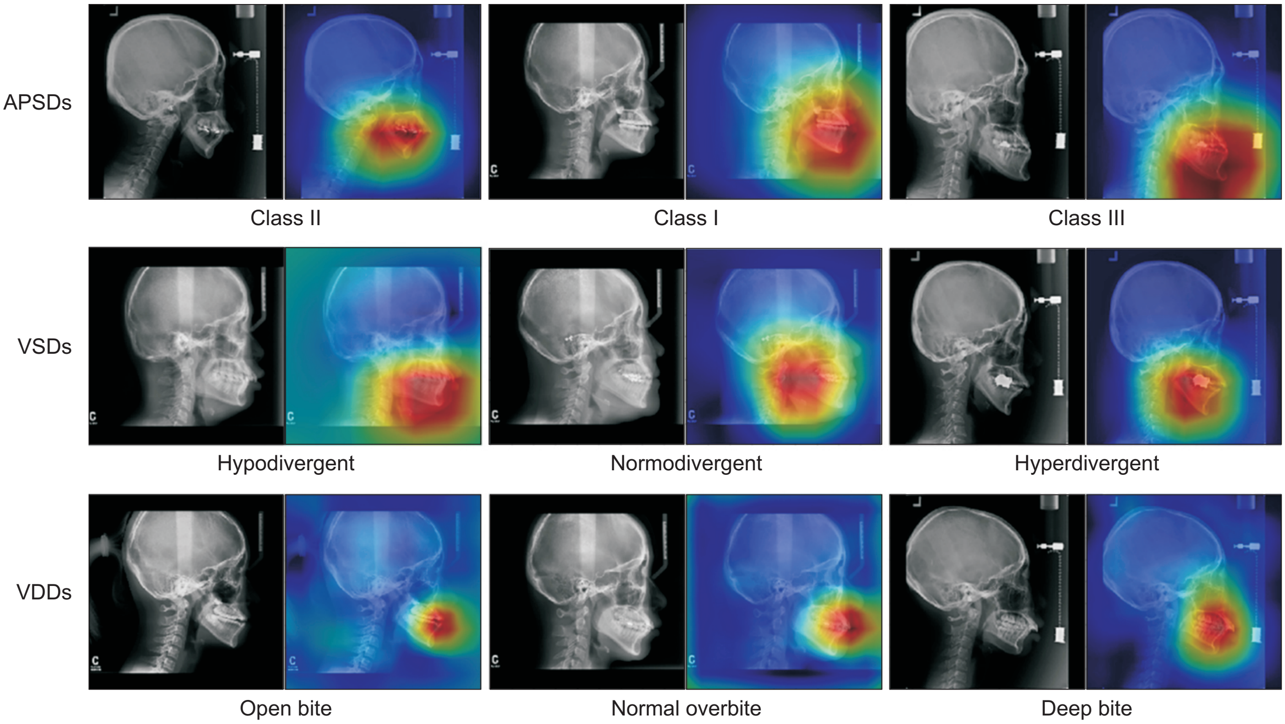

Figure 6. Gradient-weighted class activation mapping plots for anteroposterior skeletal discrepancies (APSDs), vertical skeletal discrepancies (VSDs), and vertical dental discrepancies (VDDs).

each diagnosis. These indicate that our model can effectively use the information in the lateral cephalogram images.

DISCUSSION

The present study has some meaningful outcomes as follows: (1) Despite the different quality of lateral cephalogram images from diverse conditions of cephalometric radiograph systems in nationwide 10 hospitals (Table 1), a clinically acceptable accuracy of diagnosis was obtained for APSDs, VSDs, and VDDs; and (2) since it was possible to give a proper diagnosis for APSDs, VSDs, and VDDs with input of lateral cephalograms only, our model showed the possibility of general-purpose one-step orthodontic diagnosis tool.

Clinical meaning of the comparison results between internal and external test sets in binary and multiple ROC analysis

Since the differences in AUC values for APSDs, VSDs, and VDDs in both binary and multiple ROC analyses were almost insignificant (Tables 4 and 5), it could be regarded that our model was well-validated in the external test set.

Comparison of accuracy with a previous study using binary ROC analysis results

Compared to model 1 of Yu et al.,[8] our model showed slightly lower scores for total accuracy (< 0.011) and slightly higher scores for total AUC (< 0.020) (Table 6). Although our dataset had some disadvantages including a relatively smaller number of images in the dataset and an imbalanced data set compared to Yu et al.'s study[8] (n = 5,890 lateral cephalogram images, and even distribution of data set after under-sampling), our model exhibited nearly the same performance as model 1 by Yu et al.[8] To overcome this disadvantageous environment, we elaborated on constructing the proper architecture of our model using GN, ArcFace, and a softmax layer (Figure 2).

Excluding specific data, especially in the test set, may increase the risk of sample selection bias and lead to inaccurate validation of the model. Therefore, in the present study, all datasets with a whole distribution were included to properly validate the model (Figure 3).

Difference in the AUC values of in Class II and Class III groups in APSDs and hyperdivergent and hypodivergent groups in VSDs in binary and multiple ROC analysis

The hypodivergent group showed a higher AUC score than the hyperdivergent group in the internal test set, while the hyperdivergent group showed a higher AUC

Table 6. Comparison of the binary ROC analysis results between multi-models in a previous study and a single model in this study

Models	APSDs								VSDs							
	Sensitivity		Specificity		Accuracy		AUC		Sensitivity		Specificity		Accuracy		AUC	
	Yu et al's study[8]	This study	Yu et al's study[8]	This study	Yu et al's study[8]	This study	Yu et al's study[8]	This study	Yu et al's study[8]	This study	Yu et al's study[8]	This study	Yu et al's study[8]	This study	Yu et al's study[8]	This study
Model I (no exclusion of data set)	0.8575	0.8414	0.9288	0.9206	0.9050	0.8944	0.938	0.9517	0.8427	0.8461	0.9213	0.9146	0.8951	0.8910	0.937	0.9580
Model II (exclusion of data set within interval of 0.2 SD)	0.9079	NA	0.9539	NA	0.9386	NA	0.970	NA	0.9222	NA	0.9611	NA	0.9481	NA	0.985	NA
Model III (exclusion of data set within interval of 0.3 SD)	0.9355	NA	0.9677	NA	0.9570	NA	0.978	NA	0.9459	NA	0.9729	NA	0.9640	NA	0.984	NA

ROC, receiver operating characteristic; APSDs, anteroposterior skeletal discrepancies; VSDs, vertical skeletal discrepancies; AUC, area under the curve; SD, standard deviation; NA, not applicable.

than the hypodivergent group in the external test set (0.9824 vs. 0.9730 in the internal test set, respectively; 0.9684 vs. 0.9730 in the external test set, respectively; Table 4).

The Class III group showed higher AUC values than the Class II group, which was in accordance with the results of Yu et al.[8] for both internal and external test sets (0.9807 vs. 0.9533 in the internal test set, respectively; 0.9930 vs. 0.9601 in the external test set, respectively; Table 4). The reason might be a difference in the location and size of the focus areas in the diagnosis of VSDs and APSDs (i.e., relatively larger difference between Class II and Class III groups compared to between the hyperdivergent and hypodivergent groups; Figure 6). Further studies are necessary to investigate the reason why the Class III group showed a higher AUC than the Class II group.

Lower AUC values in VDDs compared to APSDs and VSDs in binary ROC analysis

The lower AUCs in VDDs in both internal and external test sets (Table 4) and relatively unclear separation of the normal overbite group from the deep bite and open bite groups in the GT of the t-SNE result (Figure 5) might be due to two reasons: (1) the imbalanced data composition in the training set, internal test set, and external test set (normal overbite, 61.3%, 52.9% and 43.6%; open bite, 26.5%, 29.7% and 28.2%; deep bite, 12.2%, 17.4% and 28.2%, respectively; Table 3) or (2) an inherent problem in the superimposed image between the anterior teeth.

Current status of CNN-based orthodontic diagnosis

Most previous CNN studies have focused on detecting cephalometric landmarks and/or calculating cephalometric variables for a two-step automated diagnosis.[1-3,8-11] The study design, methods, and results of previous CNN studies are summarized in Table 7. In the present study, we proposed a one-step orthodontic diagnosis model, which only needs input of lateral cephalograms. The degree of performance of the AI model used in this study was comparable to the human gold standard (Tables 4 and 5). Automated AI-assisted procedures might save clinicians valuable time and labor in classification of skeletodental characteristics in a large sample size. However, it still needs an ultimate decision from a human expert, especially in borderline cases.

Limitations of this study and suggestions for future studies

The present study has some limitations. First, this study had a relative imbalance in the data sets of some centers. Second, more demographic, clinical, and cephalometric parameters should be included in setting the

https://doi.org/10.4041/kjod.2022.52.1.3

Table 7. Summary of the study design, methods and results in the orthodontic diagnosis of previous CNN studies and this study

Author (year)	Samples	Model and its application	Data set	Results
Arık et al. (2017)[1]	• 400 publicly available cephalograms • 19 landmarks • 8 cephalometric parameters • 2 human examiners	• Deep learning with CNN and shape-based model • Landmark detection • Cephalometric analysis	• Training set: 150 • Test set: 250	• High anatomical landmark detection accuracy (~1% to 2% higher success detection rate for a 2-mm range compared with the top benchmarks in the literature) • High anatomical type classification accuracy (~76% average classification accuracy for test set)
Park et al. (2019)[9]	• 1,028 lateral cephalograms • 80 landmarks • 1 human examiner	• Deep learning with YOLOv3 and SSD • Landmark detection	• Training set: 1,028 • Test set: 283	• The YOLOv3 algorithm outperformed SSD in accuracy for 38 of 80 landmarks • The other 42 of 80 landmarks did not show a statistically significant difference between YOLOv3 and SSD • Error plots of YOLOv3 showed not only a smaller error range but also a more isotropic tendency • The mean computational time spent per image was 0.05 seconds and 2.89 seconds for YOLOv3 and SSD, respectively • YOLOv3 showed approximately 5% higher accuracy compared with the top benchmarks in the literature
Nishimoto et al. (2019)[3]	• 219 lateral cephalograms from internet • 10 skeletal landmarks • 12 cephalometric parameters • Human examiners – not mentioned	• Personal desktop computer • CNN • Landmark detection • Cephalometric analysis	• Training set: 153 (expanded 51 folds) • Test set: 66	• Average and median prediction errors were 17.02 and 16.22 pixels • *No difference in Angles and lengths* between CNN and manually plotted points • Despite the variety of image quality, using cephalogram images on the internet is a feasible approach for landmark prediction
Hwang et al. (2020)[10]	• 1,028 lateral cephalograms • 80 landmarks • 2 human examiners	• Deep learning with YOLOv3 • Landmark detection	• Training set: 1,028 • Test set: 283	• Upon repeated trials, AI always detected identical positions on each landmark • Human intra-examiner variability of repeated manual detections demonstrated a detection error of 0.97 ± 1.03 mm • The mean detection error between AI and human: 1.46 ± 2.97 mm • The mean difference between human examiners: 1.50 ± 1.48 mm • Comparisons in the detection errors between AI and human examiners: less than 0.9 mm, which did not seem to be clinically significant
Kunz et al. (2020)[11]	• 1,792 cephalograms • 18 landmarks • 12 orthodontic parameters • 12 human examiners	• CNN deep learning algorithm • Landmark detection • Cephalometric analysis • Humans' gold standard: median values of the 12 examiners	• Training set: 1,731 • Validation set: 61 • Test set: 50	• No clinically significant differences between humans' gold standard and the AI's predictions

Table 7. Continued

Author (year)	Samples	Model and its application	Data set	Results
Yu et al. (2020)[8]	• 5,890 lateral cephalograms and demographic data from one institute • 4 cephalometric parameters • 2 human examiners	• One-step diagnostic system for skeletal classification • Multimodal CNN model	<Model I> Sagittal • Training set: n = 1,644 • Validation set: n = 351 • Test set: n = 351 Vertical • Training set: n = 1,912 • Validation set: n = 375 • Test set: n = 375	• Vertical and sagittal skeletal diagnosis: > 90% sensitivity, specificity, and accuracy • Vertical classification: highest accuracy at 96.40 (95% CI, 93.06 to 98.39; model III) • Binary ROC analysis: excellent performance (mean area under the curve > 95%) • Heat maps of cephalograms: visually representing the region of the cephalogram
Kim et al. (2020)[2]	• 2,075 lateral cephalograms from two institutes • 400 open dataset • 23 landmarks • 8 cephalometric parameters • 2 human examiners	• Stacked hourglass deep learning • Two-stage automated algorithm • Web-based application • Landmark detection • Cephalometric analysis	Evaluation group 1: • Training set: n = 1,675 • Validation set: n = 200 • Test set: n = 200 Evaluation group 2: • Training set: n = 1,675 • Validation set: n = 175 • Test set: n = 225 Evaluation group 3: • ISBI 2015 test set: n = 400	• Landmark detection error: 1.37 ± 1.79 mm • Successful classification rate: 88.43%
This study (2020)	• 2,174 lateral cephalograms from ten institutes • 4 cephalometric parameters • 1 human examiners	• One-step diagnostic system for skeletal and dental discrepancy • CNN including Densenet-169, Arcface, Softmax • External validation	• Training set: n = 1,522 from 2 institutes • Internal test set: n = 471 from 2 institutes • External test set: n = 181 from the other 8 institutes	• Binary ROC analysis: Accuracy and area under the curve were high in both internal and external test set (range: 0.8248–0.8944 and 0.8979–0.9580 in internal test set; 0.8821–0.8880 and 0.9074–0.9524 in external test set) in diagnosis of the skeletal and dental discrepancies • Multiple ROC analysis: Accuracy and area under the curve were high in both internal and external test set (range:0.8066–0.9905 and 0.8156–0.9998 in internal test set; 0.8054–0.9725 and 0.8222–0.9992 in external test set) in diagnosis of the skeletal and dental discrepancies • t-SNE analysis succeeded in creating the well-separated boundaries between the three classification groups in each diagnosis • Grad-CAM showed different patterns and sizes of the focus areas according to three classification groups in each diagnosis

CNN, convolutional neural network; YOLO, "you only look once" real-time object detection; SSD, single shot detector; ISBI, International Symposium on Biomedical Imaging; AI, artificial intelligence; CI, confidence interval; ROC, receiver operating characteristic; t-SNE, t-stochastic neighbor embedding; Grad-CAM, gradient-weighted class activation mapping.

gold standard and training AI models in future studies.

As suggestions for future studies, it is necessary to develop a one-step automated classification algorithm for diagnosis of transverse and asymmetry problems. Prospective studies with larger diagnostic cohort data sets will allow more robust validation of the model.

CONCLUSION

• The accuracy of our model was well-validated with internal test sets from two hospitals as well as external test sets from eight other hospitals without issues regarding the continuity of the data sets or exaggerated accuracy.

• Our model shows the possible usefulness of a one-step automated orthodontic diagnosis tool for classifying skeletal and dental discrepancies with input of lateral cephalograms only in an end-to-end manner. However, it still needs technical improvement in terms of classifying VDDs.

CONFLICTS OF INTEREST

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

ACKNOWLEDGEMENTS

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C1638). This article was based on the study of Dr. Yim's PhD dissertation. We thank Professor Won-hee Lim and Dr. Keunoh Lim for their contribution in performing the inter-examiner reliability test.

REFERENCES

1. Arık SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging (Bellingham) 2017;4:014501.
2. Kim H, Shim E, Park J, Kim YJ, Lee U, Kim Y. Web-based fully automated cephalometric analysis by deep learning. Comput Methods Programs Biomed 2020;194:105513.
3. Nishimoto S, Sotsuka Y, Kawai K, Ishise H, Kakibuchi M. Personal computer-based cephalometric landmark detection with deep learning, using cephalograms on the internet. J Craniofac Surg 2019;30:91-5.
4. Erkan M, Gurel HG, Nur M, Demirel B. Reliability of four different computerized cephalometric analysis programs. Eur J Orthod 2012;34:318-21.
5. Wen J, Liu S, Ye X, Xie X, Li J, Li H, et al. Comparative study of cephalometric measurements using 3 imaging modalities. J Am Dent Assoc 2017;148:913-21.
6. Rudolph DJ, Sinclair PM, Coggins JM. Automatic computerized radiographic identification of cephalometric landmarks. Am J Orthod Dentofacial Orthop 1998;113:173-9.
7. Mosleh MA, Baba MS, Malek S, Almaktari RA. Ceph-X: development and evaluation of 2D cephalometric system. BMC Bioinformatics 2016;17(Suppl 19):499.
8. 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.
9. 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.
10. 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.
11. 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.
12. Korean Association of Orthodontics Malocclusion White Paper Publication Committee. Cephalometric analysis of normal occlusion in Korean adults. Seoul: Korean Association of Orthodontists; 1997.
13. Bujang MA, Baharum N. Guidelines of the minimum sample size requirements for Cohen's Kappa. Epidemiol Biostat Public Health 2017;14:e12267.
14. McHugh ML. Interrater reliability: the Kappa statistic. Biochem Med (Zagreb) 2012;22:276-82.
15. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211-52.
16. Huang G, Liu Z, van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Paper presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21-26; Honolulu, USA. Piscataway: IEEE, 2017. p. 2261-9.
17. Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, et al. Accurate, large minibatch SGD: training ImageNet in 1 hour [Internet]. arxiv; 2017 Jun 8 [updated 2018 Apr 30; cited 2020 Aug 7]. Available from: https://arxiv.org/abs/1706.02677.
18. Jia X, Song S, He W, Wang Y, Rong H, Zhou F, et al.

Highly scalable deep learning training system with mixed-precision: training ImageNet in four minutes [Internet]. arxiv; 2018 Jul 30 [cited 2020 Aug 7]. Available from: https://arxiv.org/abs/1807.11205.

19. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [Internet]. arxiv; 2015 Feb 11 [updated 2015 Mar 2; cited 2020 Sep 8]. Available from: https://arxiv.org/abs/1502.03167.
20. Wu Y, He K. Group normalization. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. ECCV 2018: Computer vision – ECCV 2018. Cham: Springer; 2018. p. 3-19.
21. Deng J, Guo J, Xue N, Zafeiriou S. ArcFace: additive angular margin loss for deep face recognition. Paper presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15-20; Long Beach, USA. Piscataway: IEEE, 2019. p. 4690-9.
22. Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. Med Decis Making 2000;20:323-31.
23. Li J, Fine JP. ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. Biostatistics 2008;9:566-76.
24. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579-605.
25. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Paper presented at: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22-29; Venice, Italy. Piscataway: IEEE, 2017. p. 618-26.