Original Article

Yonsei Med J 2025 Aug;66(8):491-501
https://doi.org/10.3349/ymj.2024.0246

Yonsei Medical Journal
YMJ
pISSN: 0513-5796 · eISSN: 1976-2437

Deep Learning-Based Landmark Detection Model for Multiple Foot Deformity Classification: A Dual-Center Study

Su Ji Lee[1]*, Hangyul Yoon[2]*, Seongsu Bae[2], Inyoung Paik[2], Jong Hak Moon[2], Seongeun Park[1], Chan Woong Jang[3], Jung Hyun Park[4,5,6], Edward Choi[2], Eunho Yang[2], and Ji Cheol Shin[1]

[1]Department and Research Institute of Rehabilitation Medicine, Yonsei University College of Medicine, Seoul;
[2]Kim Jaechul Graduate School of AI, Korea Advanced Institute of Science and Technology, Daejeon;
[3]Department of Physical Medicine and Rehabilitation, Hallym University Sacred Heart Hospital, Hallym University College of Medicine, Anyang;
[4]Department of Rehabilitation Medicine, Gangnam Severance Hospital, Rehabilitation Institute of Neuromuscular Disease, Yonsei University College of Medicine, Seoul;
[5]Department of Medical Device Engineering and Management, The Graduate School, Yonsei University College of Medicine, Seoul;
[6]Department of Integrative Medicine, The Graduate School, Yonsei University College of Medicine, Seoul, Korea.

Purpose: To introduce heatmap-in-heatmap (HIH)-based model for automated diagnosis of foot deformities using weight-bearing foot radiographs, aiming to address the labor-intensive and variable nature of manual diagnosis.

Materials and Methods: From January 2004 to September 2022, a dual-center retrospective study was conducted. In the first center, 1561 anterior-posterior (AP) and 1536 lateral images from 806 patients were used for model training, while 374 AP and 373 lateral images from 196 patients were allocated to the validation set. For external validation at the second center, 527 AP and 529 lateral images from 270 patients were allocated. Five deformities were diagnosed using four and three angles between the predicted landmarks in the AP and lateral images, respectively. The results were compared with those of the baseline model (FlatNet).

Results: The HIH model demonstrated robust performance in diagnosing multiple foot deformities. On the test set, it outperformed FlatNet with higher accuracy (FlatNet vs. HIH: 78.9% vs. 85.1%), sensitivity (78.9% vs. 84.1%), specificity (79.0% vs. 85.9%), positive predictive value (77.3% vs. 84.4%), and negative predictive value (80.5% vs. 85.7%). Additionally, HIH exhibited significantly lower absolute pixel and angle errors, lower normalized mean errors, higher successful detection rate, faster training and inference speeds, and fewer parameters.

Conclusion: The HIH model showed robust performance in diagnosing multiple foot deformities with high efficacy in internal and external validation. Our approach is expected to be effective for various tasks using landmarks in medical imaging.

Key Words: Artificial intelligence, deep learning, foot deformities, diagnostic imaging

Received: August 21, 2024 **Revised:** February 4, 2025 **Accepted:** February 25, 2025 **Published online:** May 9, 2025

Co-corresponding authors: Eunho Yang, PhD, Kim Jaechul Graduate School of AI, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Korea.
E-mail: eunhoy@kaist.ac.kr and
Ji Cheol Shin, MD, PhD, Department and Research Institute of Rehabilitation Medicine, Severance Hospital, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea.
E-mail: jcsevrm@yuhs.ac

*Su Ji Lee and Hangyul Yoon contributed equally to this work.
•The authors have no potential conflicts of interest to disclose.

INTRODUCTION

Foot deformities include a wide spectrum of anatomical abnormalities. Although there is no gold standard for diagnosing foot deformities, utilizing weight-bearing foot radiographs is a commonly used method that measures certain angles from the given foot images and detects deformities based on predefined diagnostic criteria. However, the measurement of these angles requires the annotation of anatomical landmark points by experienced doctors, which is a labor-intensive process with high variability between clinicians.[1]

With this background, recent studies have attempted to predict foot deformities from radiograph images by finding anatomical landmarks using deep learning models.[2-6] However, most of these studies have relied solely on lateral view images and have only been able to detect pes planus. Considering that multiple foot deformities may coexist in a single patient and that a certain deformity may lead to the development of other foot deformities, this limitation hinders the applicability of prediction models in actual clinical practice.[7,8] Furthermore, most prior research has relied on computationally inefficient or outdated models,[9,10] even in recently published works.[3,5] For example, FlatNet, a recent model that has been used to research foot deformities,[3] is based on a relatively outdated model architecture, despite its recent release. Additionally, it requires a separate prediction model for each landmark, leading to a high computational cost.

In this study, our primary goal was to develop a landmark detection model for the prediction of various foot deformities while utilizing both anterior-posterior (AP) and lateral view images, which sets our work apart from previous studies in the field. While acknowledging the potential inefficiency and suboptimal performance of the existing FlatNet model, we designed landmark detection models for automated diagnosis using datasets from two distinct institutions. We adopt heatmap-in-heatmap (HIH), a relatively efficient and lightweight model for facial landmark detection,[11] to detect anatomical landmarks for multiple foot deformity diagnoses in AP and lateral view foot radiographs. The HIH model demonstrates higher landmark detection and diagnosis performance with significantly faster training and inference speeds compared to the baseline FlatNet model.

MATERIALS AND METHODS

Datasets

This study was approved by the Institutional Review Board of Severance Hospital, Yonsei University Health System, Seoul, Korea (IRB approval number: 4-2022-1124), which waived the need for informed consent due to its retrospective nature. This study was performed according to the approved protocol and the guidelines of the Declaration of Helsinki.

This retrospective study was conducted using data from patients aged ≥7 years who visited the Department of Rehabilitation Medicine of Severance Hospital and Gangnam Severance

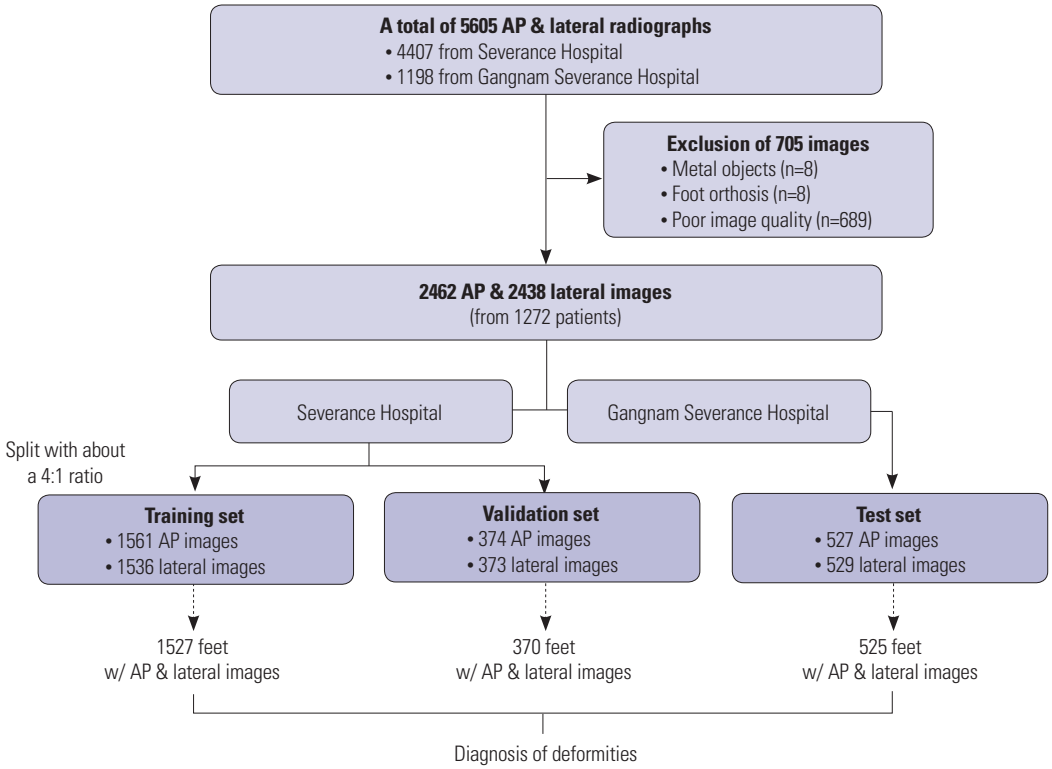


Fig. 1. Flowchart of study design and dataset construction. AP, anterior-posterior.

https://doi.org/10.3349/ymj.2024.0246

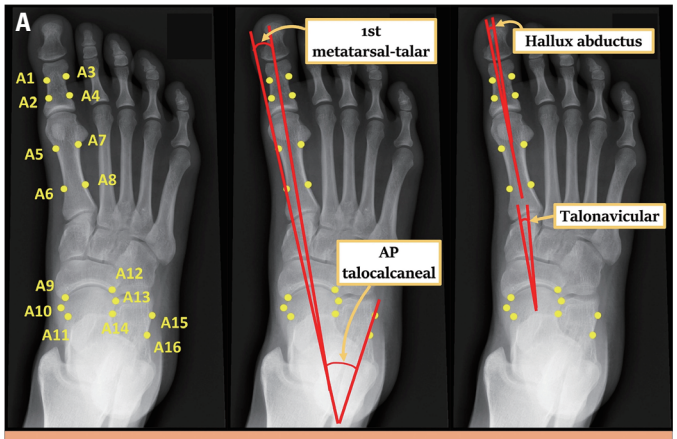
Hospital from January 2004 to September 2022 (Fig. 1). The patients included those who visited our center for musculoskeletal pain and foot deformities caused by neuromuscular diseases such as cerebral palsy and stroke. Of the initially included 5605 radiographs, 705 images were excluded due to metal objects, foot orthosis, and poor image quality. Finally, 2462 AP and 2438 lateral images were registered.

The samples from Severance Hospital were allocated to the training and validation sets according to the date of imaging, with a ratio of approximately 4:1. Consequently, 1561 AP and 1536 lateral images from 806 patients were used for model training, whereas 374 AP and 373 lateral images from 196 patients were allocated to the validation set. For external validation, 527 AP and 529 lateral images of 270 patients from Gangnam Severance Hospital were included in the test set. After training, deformities were diagnosed on the feet with both AP and lateral images available.

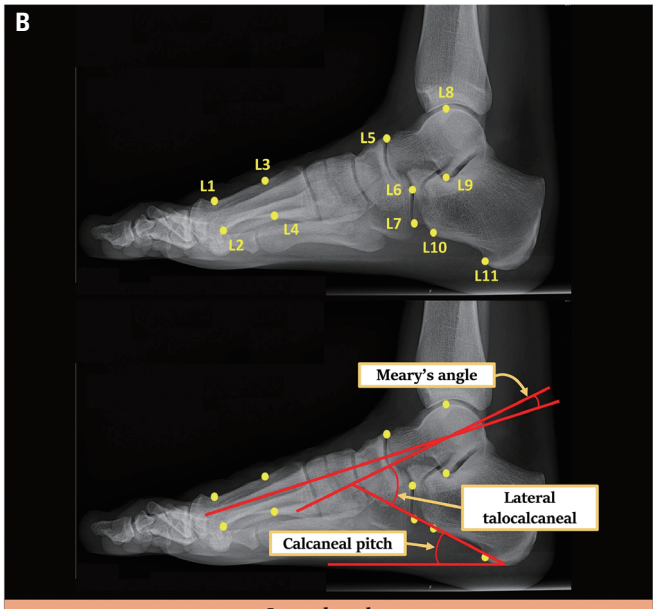
The training of deep learning models is fundamentally performed on a training set, and the trained model is then applied to a test set to evaluate its final performance. The training process on the training set involves continuously minimizing the loss function between the predicted outcomes and ground truth through iterative updates. Although the loss on the training set decreases as training progresses, there is a risk that performance on other evaluation samples may degrade, a phenomenon known as overfitting. To prevent this, at every training epoch, we evaluated the validation set and selected the model checkpoint with the best performance in the validation set as the final model. Using this final model, we conducted the evaluation on the test set. By setting aside a validation set separate from the test set, the issue of overfitting can be mitigated to some extent compared to using only a train/test split.

Anatomical landmarks and diagnosis

For diagnosis, specific anatomical criteria were used to determine 16 and 11 landmarks on the AP and lateral images,[12,13] respectively (Fig. 2 and Supplementary Fig. 1, only online). For non-adults (ages 7–18), different criteria were applied for normal angle values compared to adults, as detailed in Supplementary Table 1 (only online).[14,15] With the acquired 27 landmarks, four and three angles in the AP and lateral images were measured for diagnosis, respectively. Using these angles, five deformities were diagnosed: 1) pes planus, 2) pes cavus, 3) hindfoot valgus, 4) forefoot abduction, and 5) hallux valgus.



AP angles	
AP talocalcaneal	The angle formed by (1) the line connecting the midpoints of **A10-A13 & A11-A14** (2) And the line connecting points **A15-A16** → If (1) rotates more medially than (2), assign "+" (Otherwise "−")
Talonavicular	The angle formed by (1) the perpendicular bisector of **A9-A12** (2) and the perpendicular bisector of **A10-A13** → If (1) rotates more laterally than (2), assign "+" (Otherwise "−")
1st metatarsal-talar	The angle formed by (1) the line connecting the midpoints of **A10-A13 & A11-A14** (2) and the line connecting the midpoints of **A5-A7 & A6-A8** → If (1) rotates more laterally than (2), assign "+" (Otherwise "−")
Hallux abductus	The angle formed by (1) the line connecting the midpoints of **A1-A3 & A2-A4** (2) and the line connecting the midpoints of **A5-A7 & A6-A8** → If (1) rotates more laterally than (2), assign "+" (Otherwise "−")



Lateral angles	
Meary's angle	The angle formed by (1) the line connecting the midpoints of **L1-L2 & L3-L4** (2) and the line connecting the midpoints of **L5-L6 & L8-L9** → If (1) is gentler than (2), assign "+" (Otherwise "−")
Calcaneal pitch	The angle between (1) the horizontal plane and (2) the line formed by either **L7-L11** or **L10-L11**, whichever has the **smaller value**
Lateral talocalcaneal	The angle formed by (1) the line connecting the midpoints of **L5-L6 & L8-L9** (2) and the line from **L7** used in measuring calcaneal pitch

Fig. 2. Overview of radiographic landmarks and diagnostic angles. A total of 16 (A1–A16) and 11 (L1–L11) landmarks were annotated in the AP (A) and lateral images (B), respectively. Thereafter, four angles in the AP images and three angles in the lateral images were measured. The anatomical description of the landmarks is described in the Supplementary Fig. 1 (only online). AP, anterior-posterior.

Data preprocessing

All images from Severance Hospital and Gangnam Severance Hospital were taken using devices from DK (Seongnam, Korea) and Philips (Amsterdam, Netherlands) Medical Systems. The pixel spacing of the images was adjusted to 0.15×0.15 mm using the ImageIO 2.15.0 and Scipy 1.5.4 Python libraries. If a single image contained both feet, it was cropped and separated so that each image contains only one foot. De-identification was performed using Deid 0.3.22 and Pydicom 2.3.0 Python libraries.

The landmark annotation processes were performed using Slicer 3D (https://www.slicer.org/) by three doctors: A (board-certified physician of rehabilitation medicine with 7 years of experience), B (in-training doctor of rehabilitation medicine with 4 years of experience), and C (board-certified physician of radiation oncology with 5 years of clinical experience and deep learning research). Doctors A, B, and C independently annotated 3228, 905, and 767 radiographs, respectively, without discussion. Subsequently, the three doctors gathered and collectively reviewed the annotation results.

Models

In this study, our proposed model (HIH) was compared to the baseline model (FlatNet), which is based on commonly used methods in several previous studies.[9,16,17] For each architecture, the models for the AP and lateral images were created and trained separately (Fig. 3A). The models with the lowest landmark prediction losses in validation set were selected as the final prediction models.

The baseline, FlatNet, is a two-stage landmark detection model. The first stage uses a patch detection model, whereas the second stage predicts the location of each landmark within the previously suggested region.[5] A separate architecture is required for every landmark in the second stage (i.e., a total of 27 models), which significantly increases the time and memory complexity.

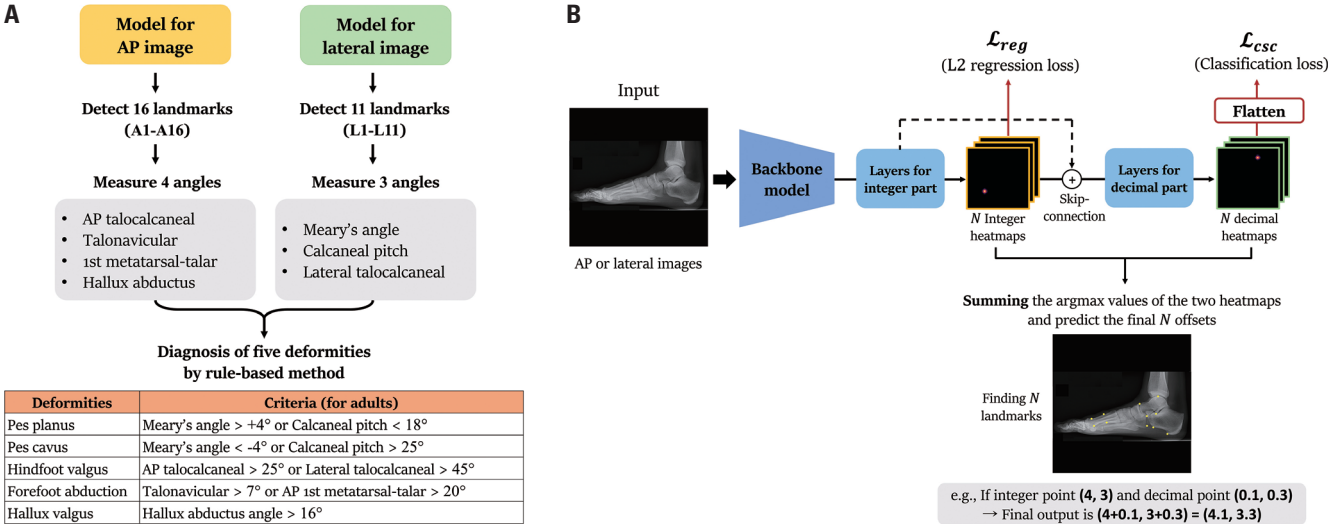
On the other hand, HIH predicts all landmarks with a single model (Fig. 3B). The model generates two types of heatmaps (integer and decimal) to predict landmark offsets through the application of the HourGlass architecture.[11] The model incorporates decimal heatmaps to represent subpixel coordinates, which can mitigate quantization errors in pixel-level tasks. Unlike the original study, we applied two 7×7 convolutional layers to the output integer heatmaps and added a residual connection from the original heatmaps. This modification was made because we wanted to make the final predictions by considering the correlation between landmarks by aggregating the independently predicted heatmap results for each landmark and refining them once again. The results of the ablation study on this matter are described in the Supplementary Table 2 (only online). All the inputs were padded to be square, and the input and output sizes were set as 512×512 and 128×128, respectively.

All the codes were implemented using Python 3.6.9 and PyTorch 1.10.2, and the details for model implementation are described in the Supplementary Material (only online). The code for the proposed deep learning model is publicly available at: https://github.com/hangyulyoon/foot-deformity.

Evaluation of model performance and statistical analysis

We measured the evaluation metrics for 1) diagnosis, 2) landmark detection, and 3) computational efficiency. To assess the diagnostic abilities, metrics including accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were used.

The mean absolute errors of pixel distances and diagnostic



Deformities	Criteria (for adults)
Pes planus	Meary's angle > +4° or Calcaneal pitch < 18°
Pes cavus	Meary's angle < -4° or Calcaneal pitch > 25°
Hindfoot valgus	AP talocalcaneal > 25° or Lateral talocalcaneal > 45°
Forefoot abduction	Talonavicular > 7° or AP 1st metatarsal-talar > 20°
Hallux valgus	Hallux abductus angle > 16°

Fig. 3. Overall frameworks of our proposed approach. (A) Two models are trained to predict AP and lateral landmarks, and diagnoses for the five-foot deformities are performed using the angles between the landmarks. Criteria for non-adults are described in the Supplementary Table 1 (only online). (B) Overall framework of our HIH model. The model outputs two types of heatmaps (integer and decimal heatmaps) for N landmarks. During training, two losses (regression and classification loss terms) are used. AP, anterior-posterior; HIH, heatmap-in-heatmap.

https://doi.org/10.3349/ymj.2024.0246

angles were calculated to evaluate landmark detection performance, and then they were compared using paired t-tests. In addition, normalized mean error (NME) and successful detection rate (SDR) were used. The NME is calculated by dividing the distance between the ground truth and predicted points into a normalization factor. SDR of χ (mm) is the proportion of the predicted points within χ (mm) of the ground truth points.

For computational efficiency, three metrics were measured: 1) number of parameters, 2) training time (per epoch), and 3) inference time (per sample). All tests were performed in the same environment and with the same graphics processing unit (GPU).

All statistical analyses were conducted using the Scipy 1.5.4 Python library. To compare the characteristics between all datasets, chi-square and one-way analysis of variance tests were employed for categorical and continuous variables, respectively. A paired t-test was used to examine whether the prediction results from the two given models were significantly different.

RESULTS

Patient characteristics

The characteristics of the datasets are summarized in Table 1. The training set included a relatively higher proportion of non-adults (54.8%) compared to the validation (29.6%) and test (14.8%) sets (p<0.001). In addition, there was a significant difference in the distribution of sex (p=0.003). The table also shows the number of deformities and the mean angles for diagnosis of each dataset with feet that have both the AP and lateral images.

The ratios of deformities, including hindfoot valgus (p<0.001), forefoot abduction (p<0.001), and hallux valgus (p=0.007), differed significantly across the datasets. Additionally, significant differences were observed in the mean angles: talonavicular (p<0.001), 1st metatarsal-talar (p=0.015), Meary's angle (p<0.001), calcaneal pitch (p<0.001), and lateral talocalcaneal angle (p< 0.001).

Table 1. Patient Demographics and Dataset Characteristics

	Training set	Validation set	Test set	*p* value*
Number of subjects	(n=806)	(n=196)	(n=270)	-
Age (yr)				<0.001
≥18	364 (45.2)	138 (70.4)	230 (85.2)	
<18	442 (54.8)	58 (29.6)	40 (14.8)	
Sex				0.003
Male	381 (47.3)	103 (52.6)	1010 (37.4)	
Female	425 (52.7)	93 (47.4)	169 (62.6)	
Number of images				-
AP view	1561	374	527	
Lateral view	1536	373	529	
Number of feet				-
Total	1583	376	531	
w/ both the AP & lateral images	1527	370	525	
Deformity	(n=1527)	(n=370)	(n=525)	
Pes planus	1079 (70.7)	247 (66.8)	344 (65.5)	0.055
Pes cavus	198 (13.0)	45 (12.2)	56 (10.7)	0.382
Hindfoot valgus	554 (36.3)	197 (53.2)	310 (59.0)	<0.001
Forefoot abduction	976 (63.9)	270 (73.0)	382 (72.8)	<0.001
Hallux valgus	365 (23.9)	109 (29.5)	157 (29.9)	0.007
Angles in AP image (°)				
AP talocalcaneal	27.25±11.29	28.24±10.70	26.49±10.78	0.066
Talonavicular	22.03±17.77	19.08±15.26	15.41±9.77	<0.001
1st metatarsal-talar	-11.57±14.91	-9.70±13.50	-9.99±11.06	0.015
Hallux abductus	13.08±10.21	13.79±8.79	13.80±8.47	0.211
Angles in lateral image (°)				
Meary's angle	8.54±13.24	5.56±10.11	4.92±7.86	<0.001
Calcaneal pitch	13.96±7.41	16.12±6.93	17.90±5.40	<0.001
Lateral talocalcaneal	33.28±11.93	36.74±10.53	39.71±8.16	<0.001

AP, anterior-posterior.
The deformities and diagnostic angles were evaluated in feet with both the AP and lateral view images.
*Distributions of the three datasets were compared using the chi-square and one-way ANOVA tests.

Deformity diagnosis

Table 2 shows the diagnostic ability for multiple deformities and the overall statistics using the micro-average. In the validation set, our HIH model outperformed the FlatNet model in terms of overall accuracy (FlatNet vs. HIH: 85.2% vs. 88.1%), sensitivity (82.8% vs. 89.2%), PPV (85.2% vs. 85.9%), and NPV (85.2% vs. 90.1%). In terms of each individual deformity, our model showed higher accuracy in diagnosing pes planus (85.7% vs. 89.7%), hindfoot valgus (77.6% vs. 83.5%), and forefoot abduction (83.2% vs. 90.0%) compared to the baseline. In addition, although the accuracies of the other deformities were similar between the two models, our model showed a higher sensitivity for hallux valgus (66.1% vs. 81.7%).

In the test set, our model demonstrated higher performance in all overall evaluation metrics. Our model showed better accuracy (78.9% vs. 85.1%), sensitivity (78.9% vs. 84.1%), specificity (79.0% vs. 85.9%), PPV (77.3% vs. 84.4%), and NPV (80.5% vs. 85.7%) compared to the baseline. The diagnostic accuracy of our model was also high for pes planus (82.7% vs. 85.5%), hindfoot valgus (68.2% vs. 78.9%), forefoot abduction (74.9% vs. 85.1%), and hallux valgus (78.9% vs. 85.9%). Furthermore, for a relatively minor deformity such as pes cavus (n=56), our model showed relatively good sensitivity (67.9% vs. 80.4%), while the baseline model did not.

Landmark prediction performance

The average absolute differences in pixel distance and angles for diagnosis between the ground truth and predicted results are shown in Table 3. Our model demonstrated low pixel error in both the validation (AP 7.02±5.84 vs. 4.61±4.46 pixels, $p<0.001$; lateral 10.60±16.26 vs. 7.38±6.13 pixels, $p<0.001$) and test (AP 4.70±4.07 vs. 2.73±1.93 pixels, $p<0.001$; lateral 5.50±8.05 vs. 3.81±4.16 pixels, $p<0.001$) sets compared to the baseline. Prediction examples of the baseline and our models are shown in Fig. 4.

Our model also exhibited lower absolute angle prediction errors in most cases. In the validation set, our absolute angle differences were lower in AP talocalcaneal (12.35°±10.07° vs. 8.66°±7.88°, $p<0.001$), talonavicular (7.80°±8.28° vs. 4.72°±4.09°, $p<0.001$), and 1st metatarsal-talar (10.33°±9.01° vs. 6.26°±6.91°, $p<0.001$) angles than those of the baseline. In the test set, our model also demonstrated lower absolute angle errors com-

Table 2. Diagnostic Abilities for Multiple Foot Deformities

	Models	Accuracy	Sensitivity	Specificity	PPV	NPV
Validation set (n=370)						
Pes planus (n=247)	FlatNet	317/370 (85.7)	219/247 (88.7)	98/123 (79.7)	219/244 (89.8)	98/126 (77.8)
	HIH	332/370 (89.7)*	227/247 (91.9)*	105/123 (85.4)*	227/245 (92.7)*	105/125 (84.0)*
Pes cavus (n=45)	FlatNet	343/370 (92.7)*	34/45 (75.6)*	309/325 (95.1)*	34/50 (68.0)*	309/320 (96.6)*
	HIH	337/370 (91.1)	33/45 (73.3)	304/325 (93.5)	33/54 (61.1)	304/316 (96.2)
Hindfoot valgus (n=197)	FlatNet	287/370 (77.6)	155/197 (78.7)	132/173 (76.3)	155/196 (79.1)	132/174 (75.9)
	HIH	309/370 (83.5)*	173/197 (87.8)*	136/173 (78.6)*	173/210 (82.4)*	136/160 (85.0)*
Forefoot abduction (n=270)	FlatNet	308/370 (83.2)	239/270 (88.5)	69/100 (69.0)	239/270 (88.5)	69/100 (69.0)
	HIH	333/370 (90.0)*	252/270 (93.3)*	81/100 (81.0)*	252/271 (93.0)*	81/99 (81.8)*
Hallux valgus (n=109)	FlatNet	321/370 (86.8)*	72/109 (66.1)	249/261 (95.4)*	72/84 (85.7)*	249/286 (87.1)
	HIH	318/370 (85.9)	89/109 (81.7)*	229/261 (87.7)	89/121 (73.6)	229/249 (92.0)*
Overall	FlatNet	1576/1850 (85.2)	719/868 (82.8)	857/982 (87.3)*	719/844 (85.2)	857/1006 (85.2)
	HIH	1629/1850 (88.1)*	774/868 (89.2)*	855/982 (87.1)	774/901 (85.9)*	855/949 (90.1)*
Test set (n=525)						
Pes planus (n=344)	FlatNet	434/525 (82.7)	299/344 (86.9)	135/181 (74.6)	299/345 (86.7)	135/180 (75.0)
	HIH	449/525 (85.5)*	305/344 (88.7)*	144/181 (79.6)*	305/342 (89.2)*	144/183 (78.7)*
Pes cavus (n=56)	FlatNet	473/525 (90.1)*	38/56 (67.9)	435/469 (92.8)*	38/72 (52.8)*	435/453 (96.0)
	HIH	472/525 (89.9)	45/56 (80.4)*	427/469 (91.0)	45/87 (51.7)	427/438 (97.5)*
Hindfoot valgus (n=310)	FlatNet	358/525 (68.2)	212/310 (68.4)	146/215 (67.9)	212/281 (75.4)	146/244 (59.8)
	HIH	414/525 (78.9)*	242/310 (78.1)*	172/215 (80.0)*	242/285 (84.9)*	172/240 (71.7)*
Forefoot abduction (n=382)	FlatNet	393/525 (74.9)	310/382 (81.2)	83/143 (58.0)	310/370 (83.8)	83/155 (53.5)
	HIH	447/525 (85.1)*	335/382 (87.7)*	112/143 (78.3)*	335/366 (91.5)*	112/159 (70.4)*
Hallux valgus (n=157)	FlatNet	414/525 (78.9)	126/157 (80.3)*	288/368 (78.3)	126/206 (61.2)	288/319 (90.3)
	HIH	451/525 (85.9)*	124/157 (79.0)	327/368 (88.9)*	124/165 (75.2)*	327/360 (90.8)*
Overall	FlatNet	2072/2625 (78.9)	985/1249 (78.9)	1087/1376 (79.0)	985/1274 (77.3)	1087/1351 (80.5)
	HIH	2233/2625 (85.1)*	1051/1249 (84.1)*	1182/1376 (85.9)*	1051/1245 (84.4)*	1182/1380 (85.7)*

HIH, heatmap-in-heatmap; PPV, positive predictive value; NPV, negative predictive value.
All statistics are presented as predicted number/total number (%). The overall statistics were calculated by micro-averaging the values for each deformity.
*Higher values.

https://doi.org/10.3349/ymj.2024.0246

Table 3. Absolute Errors of Pixels and Diagnostic Angles from the Predicted Landmarks

	Validation set (n=370)			Test set (n=525)		
	FlatNet	HIH	*p* value	FlatNet	HIH	*p* value
Pixel errors (pixels)						
AP images	7.02±5.84	4.61±4.46	<0.001*	4.70±4.07	2.73±1.93	<0.001*
Lateral images	10.60±16.26	7.38±6.13	<0.001*	5.50±8.05	3.81±4.16	<0.001*
Angle errors (°)						
AP talocalcaneal	12.35±10.07	8.66±7.88	<0.001*	16.17±15.93	7.94±6.12	<0.001*
Talonavicular	7.80±8.28	4.72±4.09	<0.001*	8.63±7.44	5.03±4.20	<0.001*
Metatarsal-talar	10.33±9.01	6.26±6.91	<0.001*	13.19±13.30	6.26±4.79	<0.001*
Hallux abductus	3.67±3.12	3.67±2.78	0.997	4.41±4.70	3.53±2.58	<0.001*
Meary's angle	3.75±4.36	3.53±3.16	0.398	3.49±3.43	3.48±2.60	0.927
Calcaneal pitch	1.31±1.19	1.40±1.12	0.215	2.05±2.61	1.53±1.33	<0.001*
Lateral talocalcaneal	2.63±2.65	2.42±1.99	0.206	3.51±4.63	2.56±1.98	<0.001*

AP, anterior-posterior; HIH, heatmap-in-heatmap.
*p<0.05.

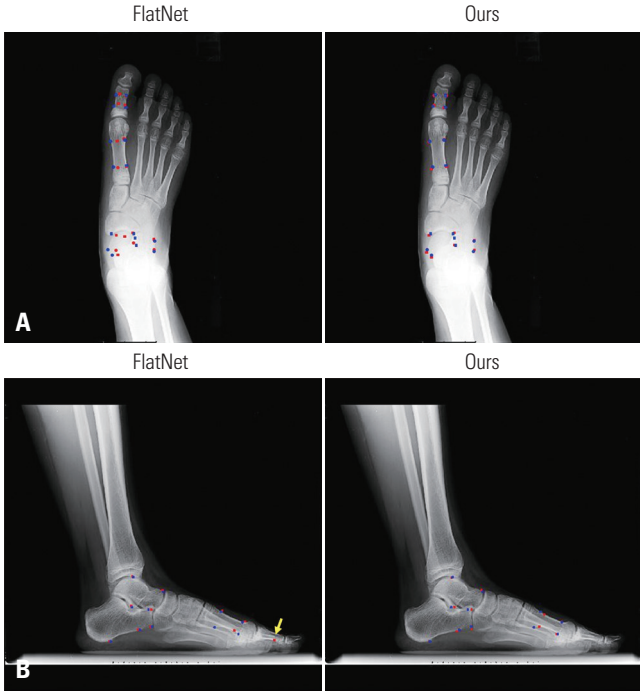


Fig. 4. Prediction results between the two models in AP and lateral images from the same patient. The ground truth and predicted points are colored blue and red, respectively. (A) In the AP image, our model predicts all landmarks better overall. (B) In the lateral image, the predictions of the two models are similar in most cases. However, the outputs of FlatNet have an outlier with a large pixel error (yellow arrow). AP, anterior-posterior.

pared to the baseline: AP talocalcaneal (16.17°±15.93° vs. 7.94°±6.12°, p<0.001), talonavicular (8.63°±7.44° vs. 5.03°±4.20°, p<0.001), 1st metatarsal-talar (13.19°±13.30° vs. 6.26°±4.79°, p<0.001), hallux abductus (4.41°±4.70° vs. 3.53°±2.58°, p<0.001), calcaneal pitch (2.05°±2.61° vs. 1.53°±1.33°, p<0.001), and lateral talocalcaneal (3.51°±4.63° vs. 2.56°±1.98°, p<0.001) angles.

Table 4 lists the NME and SDR values obtained from the prediction results. In all datasets and image views, our model

Table 4. Normalized Mean Errors and Successful Detection Rates of the Baseline and Our Models

	Model	NME ↓ ($\times 10^{-3}$)	SDR ↑ (%)			
			<1 mm	<2 mm	<3 mm	<4 mm
Validation set (n=370)						
AP images	FlatNet	8.341	59.4	89.1	98.2	99.4
	HIH	5.467	82.6	97.2	99.2	99.5
Lateral images	FlatNet	7.976	46.4	81.8	92.4	95.4
	HIH	5.506	58.4	87.9	95.5	98.1
Test set (n=525)						
AP images	FlatNet	8.522	82.6	98.1	99.3	99.8
	HIH	4.974	95.7	99.8	100.0	100.0
Lateral images	FlatNet	7.833	79.6	96.8	98.4	98.9
	HIH	5.408	89.2	99.2	99.8	99.9

↓, the lower, the better; ↑, the higher, the better; AP, anterior-posterior; HIH, heatmap-in-heatmap; NME, normalized mean error; SDR, successful detection rate.

showed lower NME and higher SDR at 1, 2, 3, and 4 mm compared to the baseline. The gap in SDR between the two models increased as the cut-off value decreased in both datasets (e.g., SDR 4 mm 99.4% vs. 99.5%, 3 mm 98.2% vs. 99.2%, 2 mm 89.1% vs. 97.2%, 1 mm 59.4% vs. 82.6% in AP images of the validation set). The NMEs of the individual landmarks are shown in Fig. 5. Using our model, landmarks A16 and A15 (proximal and distal lateral border of the calcaneus) of AP images showed the highest and second highest NME in the validation set, and landmarks A5 and A7 (medial and lateral inflections between the head and neck of 1st metatarsal bone) of AP images showed the highest and second highest NME in both test sets, respectively. For lateral images, landmarks L8 (cephalad margin of the talar body) and L6 (inferior border of the talar head) in the validation set and landmarks L8 and L9 (cephalad and caudal margins of the talar body) in the test set showed the highest and second highest NME, respectively. Our model demonstrated a lower NME compared to the base-

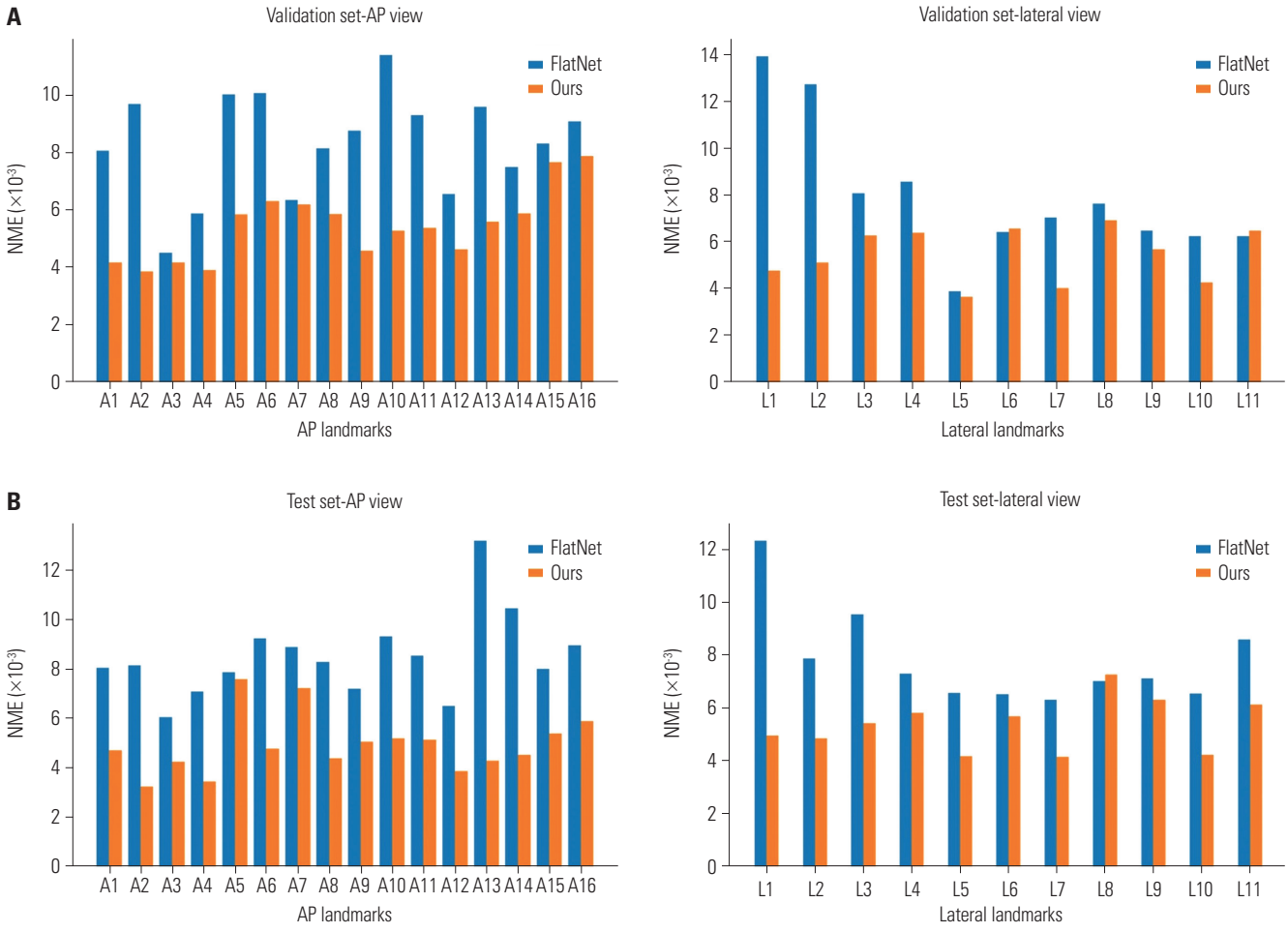


Fig. 5. (A) Normalized mean error for the landmarks in AP and lateral view images of the validation set. (B) Normalized mean error for the landmarks in AP and lateral view images of the test set. A lower NME value indicates a lesser pixel distance error between the ground truth and predicted coordinates. NME, normalized mean error; AP, anterior-posterior.

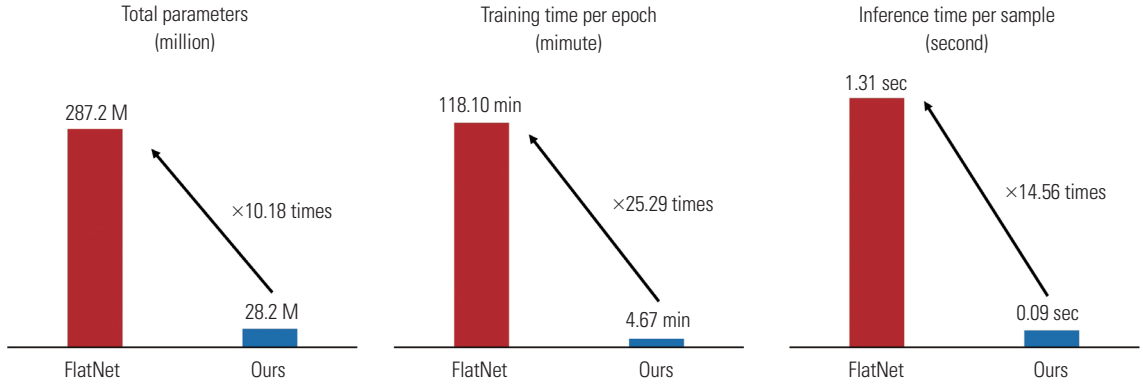


Fig. 6. Comparison of the computational efficiency between the two models. All the measurements were conducted using a single NVIDIA TITAN Xp 12GB GPU, in lateral images.

line in all datasets and image views except L6 and L11 (posterior tuberosity of the calcaneus) of the validation set and L8 of the test set, suggesting that our model predicted the outcomes with lower pixel errors in most landmark positions.

Computational efficiency

The computational efficiencies of the baseline and our models for lateral images are summarized in Fig. 6. Compared to our model, the baseline has approximately 11 times more parameters (287.2 M vs. 28.2 M) and approximately 25 and 15 times longer training time per epoch (118.10 min vs. 4.67 min) and inference time per sample (1.31 sec vs. 0.09 sec) with the given single GPU. In other words, the whole training time of the baseline model goes up to approximately 98.42 hours for

https://doi.org/10.3349/ymj.2024.0246

50 epochs, while the whole training time of our model is approximately 3.89 hours for 50 epochs. In addition, the total inference times to predict the outcomes of 500 samples are approximately 10.92 and 0.75 minutes if we use the baseline and our models, respectively.

DISCUSSION

This study discovered that multiple foot deformities can be diagnosed by finding landmarks in foot radiographs based on deep learning models, particularly the HIH model. HIH detects not only the presence of pes planus but also other frequently observed foot deformities associated with its pathophysiology, including hindfoot valgus and forefoot abduction. Compared with the baseline FlatNet model, our model showed better deformity diagnosis and landmark detection abilities, with significantly faster training and inference speeds.

To the best of our knowledge, this is the first study to use a deep learning model to detect various foot deformities other than pes planus while utilizing both AP and lateral view foot radiographs. Previous studies have primarily focused on diagnosing pes planus using lateral foot radiographs. For instance, Ryu, et al.[5] used FlatNet to predict pes planus in 100 and 17 lateral foot radiographs for internal and external validation, respectively. In external validation, the absolute average errors were 0.61°±0.45° for calcaneal pitch, 2.59°±2.40° for Meary's angle, and 2.26°±2.18° for talocalcaneal angle. However, this study did not report diagnostic metrics such as accuracy, sensitivity, and specificity. Similarly, Koo, et al.[4] designed a single-center study using a segmentation model to predict pes planus with 300 and 95 training and validation samples, respectively. In the validation set, the average accuracy, sensitivity, and specificity for Meary's angle were 90.18%, 90.48%, and 89.94%, respectively, while for the calcaneal pitch, they were 96.84%, 96.88%, and 96.77%, respectively. However, Koo's study was a single-center study and had the limitation of not performing external validation.

The foot's three-dimensional structure makes it challenging to diagnose using only lateral views, necessitating the inclusion of AP views for a comprehensive assessment. Unlike previous studies that focused solely on lateral views, this study incorporated both AP and lateral views in the diagnostic process. As a result, our model demonstrated superior performance across various diagnostic metrics, including accuracy, compared to the baseline model. Furthermore, the HIH model showed statistically significant lower errors in pixel distances and diagnostic angles from the predicted landmarks than FlatNet, confirming its enhanced performance. Additionally, in the analysis of NME and SDR, which are metrics for evaluating model performance, HIH was found to be superior to the baseline. We also observed improved computational efficiency in the same GPU environment.

Early screening and intervention are crucial factors that positively impact the prognosis of pes planus.[18,19] Untreated pediatric pes planus that is not appropriately managed can progress and worsen until adulthood.[20] Additionally, even asymptomatic pes planus should be observed carefully, as it can potentially develop into conditions such as metatarsal stress fractures.[21] Therefore, it is important to regularly monitor patients and establish a proper therapeutic plan at an appropriate time. During the decision process, it is essential to assess the status of pes planus and the presence of associated deformities, including hindfoot valgus and forefoot abduction.[7,8] However, diagnosing multiple types of foot deformities using plain radiographs is labor-intensive, time-consuming, and often subject to inter-rater variability.[1,22] Considering these aspects, our proposed approach can improve the efficiency and accuracy of physicians' clinical decisions, which can also lead to a better patient prognosis of foot deformities, including pes planus.

Our baseline model, FlatNet, uses a two-stage approach to detect landmarks. This method has been used in several studies to find anatomical landmarks in medical images.[9,16,17] However, this method requires two-stage training, and a U-Net is required for landmark detection in the second stage. As a result, the computational complexity and training time of the model increase as the number of landmarks increases. These disadvantages were evident in the comparison of the training and inference speeds with our model. Furthermore, our model considers the coordinates of other landmarks when predicting the offset of a single landmark. These strengths indicate the potential of the HIH model for anatomical landmark detection in the medical domain.

This study had several limitations. First, the datasets were constructed using retrospectively collected data. Second, the datasets used in this study consisted of images from people who visited hospitals, resulting in a relatively high proportion of abnormal cases. Therefore, it is necessary to validate whether our model performs well on class-imbalanced datasets with a higher proportion of normal subjects. Third, there were differences in the demographic and imaging characteristics among the datasets in this study. However, this may indicate that our model can show robust performance, even when the data distribution differs from the training set. Fourth, inter-observer variability was not measured as the ground truth landmark offsets were defined as the coordinates where all three doctors completely agreed. In addition, we did not examine the degree to which the diagnostic abilities of each clinician could improve when using the model. Therefore, additional research is needed to determine whether the use of the deep learning model we have proposed can reduce diagnosis time or improve clinical diagnostic accuracy when used by actual healthcare professionals. Fifth, in the case of landmark-based prediction, the variability of diagnostic ability may vary depending on the robustness of landmark prediction. Sixth, since diagnostic criteria can change over time, there is a possibility of differences in

outcomes as a result. Seventh, since HIH can only diagnose the five deformities mentioned, including pes planus, it is necessary to develop a model capable of more diverse diagnoses in future studies. Additionally, this study analyzed images obtained from multiple X-ray machines, which may have caused differences in landmark settings due to issues such as resolution. Further studies are needed to investigate these aspects and enhance landmark detection models in future research.

In conclusion, our HIH model based on deep learning was able to diagnose various foot deformities simultaneously and outperformed the baseline model. Furthermore, our model demonstrated robust performance, even on datasets collected from another institution that were not included during the training set construction. Our results indicate the potential of the HIH model for radiological diagnosis using anatomical landmarks. Further studies are required to determine whether our proposed approach can be effectively applied to various tasks and datasets in medical imaging that utilize landmarks.

ACKNOWLEDGEMENTS

This study is supported by a research grant of Research Institute of Rehabilitation Medicine, Yonsei University College of Medicine for 2024, and the Medical Scientist Training Program from the Ministry of Science & ICT of Korea.

AUTHOR CONTRIBUTIONS

Conceptualization: Su Ji Lee, Hangyul Yoon, Eunho Yang, and Ji Cheol Shin. Data curation: Su Ji Lee, Hangyul Yoon, Seongeun Park, Chan Woong Jang, Jung Hyun Park, and Ji Cheol Shin. Formal analysis: Su Ji Lee and Hangyul Yoon. Funding acquisition: Eunho Yang and Ji Cheol Shin. Investigation: Su Ji Lee, Hangyul Yoon, Seongsu Bae, Inyoung Paik, Jong Hak Moon, Seongeun Park, and Chan Woong Jang. Methodology: Su Ji Lee, Hangyul Yoon, Seongsu Bae, Inyoung Paik, and Jong Hak Moon. Project administration: Jung Hyun Park, Eunho Yang, and Ji Cheol Shin. Resources: Su Ji Lee, Jung Hyun Park, and Ji Cheol Shin. Software: Hangyul Yoon, Seongsu Bae, Inyoung Paik, and Jong Hak Moon. Supervision: Edward Choi, Eunho Yang, and Ji Cheol Shin. Validation: Su Ji Lee, Hangyul Yoon, and Edward Choi. Visualization: Su Ji Lee and Hangyul Yoon. Writing—original draft: Su Ji Lee and Hangyul Yoon. Writing—review & editing: Su Ji Lee, Hangyul Yoon, Edward Choi, Eunho Yang, and Ji Cheol Shin. Approval of final manuscript: all authors.

ORCID iDs

Su Ji Lee https://orcid.org/0000-0002-6376-0125
Hangyul Yoon https://orcid.org/0000-0002-9515-1623
Seongsu Bae https://orcid.org/0000-0003-4482-8953
Inyoung Paik https://orcid.org/0009-0002-8082-022X
Jong Hak Moon https://orcid.org/0000-0002-6708-3918
Seongeun Park https://orcid.org/0000-0002-0249-3617
Chan Woong Jang https://orcid.org/0000-0002-5037-0080
Jung Hyun Park https://orcid.org/0000-0003-3262-7476
Edward Choi https://orcid.org/0000-0002-5958-3509
Eunho Yang https://orcid.org/0000-0003-2188-0169
Ji Cheol Shin https://orcid.org/0000-0002-1133-1361

REFERENCES

1. Yildiz K, Cetin T. Interobserver reliability in the radiological evaluation of flatfoot (pes planus) deformity: a cross-sectional study. J Foot Ankle Surg 2022;61:1065-70.
2. Ryu SM, Shin K, Shin SW, Lee S, Kim N. Enhancement of evaluating flatfoot on a weight-bearing lateral radiograph of the foot with U-Net based semantic segmentation on the long axis of tarsal and metatarsal bones in an active learning manner. Comput Biol Med 2022;145:105400.
3. Ryu SM, Shin K, Shin SW, Lee SH, Seo SM, Cheon SU, et al. Automated landmark identification for diagnosis of the deformity using a cascade convolutional neural network (FlatNet) on weight-bearing lateral radiographs of the foot. Comput Biol Med 2022;148:105914.
4. Koo J, Hwang S, Han SH, Lee J, Lee HS, Park G, et al. Deep learning-based tool affects reproducibility of pes planus radiographic assessment. Sci Rep 2022;12:12891.
5. Ryu SM, Shin K, Shin SW, Lee SH, Seo SM, Cheon SU, et al. Automated diagnosis of flatfoot using cascaded convolutional neural network for angle measurements in weight-bearing lateral radiographs. Eur Radiol 2023;33:4822-32.
6. Gül Y, Yaman S, Avcı D, Çilengir AH, Balaban M, Güler H. A novel deep transfer learning-based approach for automated pes planus diagnosis using X-ray image. Diagnostics (Basel) 2023;13:1662.
7. Deland JT. Adult-acquired flatfoot deformity. J Am Acad Orthop Surg 2008;16:399-406.
8. Raikin SM, Winters BS, Daniel JN. The RAM classification: a novel, systematic approach to the adult-acquired flatfoot. Foot Ankle Clin 2012;17:169-81.
9. 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.
10. Yang F, Choi W, Lin Y. Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers [accessed on 2023 July 10]. Available at: http://doi.org/10.1109/CVPR.2016.234.
11. Lan X, Hu Q, Chen Q, Xue J, Cheng J. HIH: towards more accurate face alignment via heatmap in heatmap. arXiv [Preprint]. 2021 [accessed on 2023 July 10]. Available at: https://doi.org/10.48550/arXiv.2104.03100.
12. Lamm BM, Stasko PA, Gesheff MG, Bhave A. Normal foot and ankle radiographic angles, measurements, and reference points. J Foot Ankle Surg 2016;55:991-8.
13. Lau BC, Allahabadi S, Palanca A, Oji DE. Understanding radiographic measurements used in foot and ankle surgery. J Am Acad Orthop Surg 2022;30:e139-54.
14. Bourdet C, Seringe R, Adamsbaum C, Glorion C, Wicart P. Flatfoot in children and adolescents. Analysis of imaging findings and therapeutic implications. Orthop Traumatol Surg Res 2013;99:80-7.
15. Hamel J, Hörterer H, Harrasser N. Is it possible to define reference values for radiographic parameters evaluating juvenile flatfoot deformity? A case-control study. BMC Musculoskelet Disord 2020;21:838.
16. Kim IH, Kang J, Jeong J, Kim JS, Nam Y, Ha Y, et al. A fully automated landmark detection for spine surgery planning with a cascaded convolutional neural net. Inform Med Unlocked 2022;32:101045.
17. Jiang F, Guo Y, Zhou Y, Yang C, Xing K, Zhou J, et al. Automated calibration system for length measurement of lateral cephalometry based on deep learning. Phys Med Biol 2022;67:225016.
18. Azhagiri R, Malar A, Hemapriya J, Sumathi G. The cause and fre-

quency of PES planus (flat foot) problems among young adults. Asian J Med Sci 2021;12:107-11.
19. Chibuzom CN, Egele CS, Ndukwu CU. The prevalence of flat foot among school-aged children in a Nigerian population: prevalence of flat foot among school-aged children in a Nigerian. Trop J Med Res 2022;21:113-20.
20. Mosca VS. Flexible flatfoot in children and adolescents. J Child Orthop 2010;4:107-21.
21. Simkin A, Leichter I, Giladi M, Stein M, Milgrom C. Combined effect of foot arch structure and an orthotic device on stress fractures. Foot Ankle 1989;10:25-9.
22. de Cesar Netto C, Kunas GC, Soukup D, Marinescu A, Ellis SJ. Correlation of clinical evaluation and radiographic hindfoot alignment in stage II adult-acquired flatfoot deformity. Foot Ankle Int 2018;39:771-9.