



Automated measurement of calcaneal inclination angle via partial semantic segmentation using U-Net

Kendal Erincik¹
 Umut Perçem Orhan Söylemez²
 Merve Gezgin²
 Baran Canku Karataş²
 Başak Atalay²

¹Çaycuma State Hospital, Department of Radiology, Zonguldak, Türkiye

²Göztepe Prof. Dr. Süleyman Yalçın City Hospital, Department of Radiology, İstanbul, Türkiye

PURPOSE

To develop and validate an interpretable deep learning pipeline for the automated measurement of the calcaneal inclination angle (CIA) on lateral weight-bearing foot radiographs using partial semantic segmentation.

METHODS

This retrospective single-center study included 884 radiographs after the removal of duplicates, partitioned into a development cohort of 779 radiographs and a held-out internal test cohort of 105 radiographs, with no patient contributing to both. Three landmark-relevant anatomical regions were manually annotated: the inferior calcaneal contour (CAL), inferior contour of the fifth metatarsal head (M5), and calcaneal aspect of the inferior calcaneocuboid joint (CC). Each region was segmented by its own single-class U-Net model, trained from scratch with an unweighted Dice + Binary Cross-Entropy loss and early stopping on the validation Dice score, and deterministic post-processing was used to derive four anatomical landmarks and to calculate the CIA. A second observer, a fifth-year radiology resident, independently re-annotated a random subset of 50 development radiographs from a blank canvas. The train/validation split was performed at the patient level, and each network was trained with five predefined seeds. Segmentation performance was evaluated on the validation set, and clinical measurement accuracy was assessed on the held-out test set against three readers who measured independently and blinded. The frozen primary model was then applied without modification to an independent external cohort of 50 radiographs from a second institution, measured by the same three readers under the identical blinded protocol.

RESULTS

Validation Dice scores across the five seeded training runs were 0.935 ± 0.010 for CAL, 0.871 ± 0.005 for M5, and 0.805 ± 0.009 for CC (mean $\pm$ standard deviation), with the pre-specified primary run reaching 0.936, 0.865, and 0.807, respectively. In the held-out test cohort, the model showed a mean absolute error (MAE) of 0.690° (95% confidence interval [CI] 0.525–0.855) against the mean of the three readers, with 92.4% of measurements within 2° and 98.1% within 3° . Bland–Altman analysis demonstrated a small negative bias of -0.425° (95% confidence interval -0.620° to -0.230°), with 95% limits of agreement from -2.427° to $+1.577^\circ$. The intraclass correlation coefficient (ICC) formula ICC(A,1) achieved 0.986 across the model and the three readers and 0.994 among the three readers alone. In the external cohort, the model achieved an MAE of 0.85° against the three-reader mean, with 92.0% of measurements within 2° and an ICC of 0.969.

CONCLUSION

Partial semantic segmentation combined with deterministic landmark extraction enables rapid and clinically accurate automated CIA measurement on lateral weight-bearing foot radiographs, with an agreement approaching, though not yet equaling, that observed among human readers.

CLINICAL SIGNIFICANCE

The proposed pipeline enabled fast, reproducible CIA measurement with high expert agreement and may support standardized foot alignment assessment after prospective validation.

KEYWORDS

Calcaneal inclination angle, deep learning, foot radiography, semantic segmentation, U-Net

Handling editor: İsmail Mese

Corresponding author: Merve Gezgin

E-mail: me.gezgin@hotmail.com

Received 05 August 2026; revision requested 21 August 2026; accepted 03 September 2026.



Epub: 06.10.2026

DOI: 10.4274/dir.2026.264385

1. Introduction

Pes planus (PP) is a postural abnormality of the foot characterized by the collapse of the arch, resulting in the sole of the foot having complete or partial contact with the ground.¹ This condition can affect the lower kinetic chain and, in severe cases, the entire body chain, impacting the biomechanics during weight-bearing activities.¹ In addition to clinical assessments, radiographic evaluation is considered an essential tool for defining the deformity and establishing treatment protocols.²

Several angular measurement techniques are employed for the diagnosis of PP; however, the calcaneal inclination angle (CIA) and talo-first metatarsal angle (Meary's angle) have been reported to be more effective for PP assessment, demonstrating superior intra- and inter-observer correlation coefficients to other radiographic markers.^{3,4} On a lateral weight-bearing radiograph (LWBR), the CIA is defined as the angle formed by the intersection of a line parallel to the floor with a line connecting the two most inferior points on the calcaneus (CAL; the inferior tip of the calcaneal tuberosity and the inferior tip of the CC).⁴

However, these measurements are time-consuming and increase the clinical workload of healthcare professionals. The utility of deep learning in various musculoskeletal radiology pathologies is well documented.⁵⁻⁹ The current literature highlights methods such as heat maps, landmark detection, and FlatNet for evaluating foot deformities.^{5,10-12} The objective of this study was to segment only the necessary bone contours, rather than the entire bone structure,

using lateral foot radiographs. By training the model on these contours, we aimed to deterministically and algorithmically identify the required landmarks for the CIA. This approach aims to develop a model that offers a faster training process while maintaining high predictive efficacy. The results indicate that our model is clinically effective, yielding results similar to those of manual expert measurements.

2. Methods

2.1. Study design

This retrospective, single-center diagnostic accuracy and algorithm development study integrated semantic segmentation and deterministic post-processing rules to enable the automated measurement of the CIA on LWBRs. The study was approved by the Göztepe Prof. Dr. Süleyman Yalçın City Hospital Clinical Research Ethics Committee (approval number: 2026/0293; date: May 14, 2026). The requirement for informed consent was waived by the Ethics Committee due to the retrospective nature of the study. The Checklist for Artificial Intelligence in Medical Imaging was used in this study.¹³ The analysis plan, including the data partition, the reference standard, the endpoints, and the sample-size justification, was written and frozen before any reader measurement was compared with any model output.

2.2. Dataset and preprocessing

For model training, Digital Imaging and Communications in Medicine (DICOM) LWBR files obtained from our hospital's Picture Ar-

chiving and Communication System (PACS) between November 2023 and April 2026 were retrospectively collected; consecutive eligible examinations over this period were retrieved from the PACS archive. Examinations were acquired at the primary institution on several digital radiography systems. Radiographs were acquired on Samsung Electronics GC80CW, Lanmage KeenRay 6.2, and Samsung DGR-C26U2B/TR units. The exclusion criteria were as follows: (1) age < 18 years and cortical bone alterations (fractures, prior surgery, or deformities) and (2) poor image quality (e.g., blurred images). To reflect real-world clinical scenarios, radiographs with slight positional variations were included in the datasets. Images containing metallic artifacts that hindered evaluation of anatomical structures, low-quality images, and radiographs with acquisition artifacts were excluded (Figure 1). Duplicate images were rejected on import using a SHA-256 hash of the normalized pixel array so that the same image re-exported in a different form can still be detected. After ensuring the anonymization of the collected DICOM files, all images were converted to PNG format for data standardization. Images were used at their full extent without cropping, and foot orientation was determined automatically from the relative horizontal positions of the forefoot and heel masks rather than being standardized by hand (Figure 2).

2.3. Ground truth

To establish the reference standard (ground-truth) masks, the browser-based annotation module of RadiologyMate, a soft-

Main points

- Partial semantic segmentation enabled accurate automated calcaneal inclination angle (CIA) measurement on lateral weight-bearing foot radiographs.
- Automated CIA measurements showed excellent agreement with expert readers, with a mean absolute error below 1° in both internal and external validation cohorts.
- Deterministic landmark extraction provided an interpretable alternative to direct angle regression, demonstrating substantially lower measurement error in the held-out test cohort.
- The proposed pipeline enables rapid and reproducible CIA measurement and warrants further prospective evaluation for integrated clinical workflows.

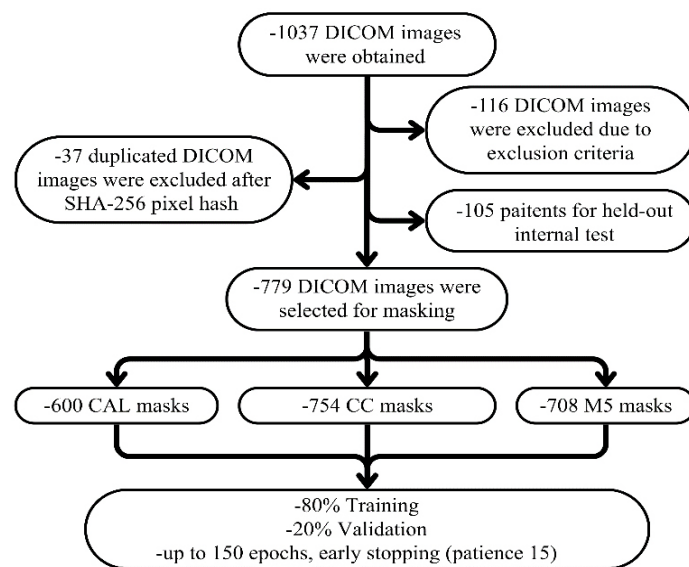


Figure 1. Flowchart. CAL, calcaneus inferior; CC, calcaneocuboid joint; DICOM, Digital Imaging and Communications in Medicine; M5, head of 5th metatarsal.

were application developed by our team, was used, in which strokes are recorded as vectors on a downscaled display image and rasterized server-side into full-resolution binary masks. Images were manually annotated by an expert radiologist. Segmentation was performed across three specific anatomical regions: the inferior contour of the fifth metatarsal head (M5), the inferior contour of the CAL, and the calcaneal aspect of the inferior CC, as shown in Figure 3. During the annotation process, the anatomical boundaries were manually delineated in images where they could be clearly identified. If a specific anatomical structure could not be definitively distinguished by the expert radiologist, its segmentation was omitted; however, the image was still included in the training set if other anatomical structures in the same radiograph could be clearly segmented. For radiographs in which a prominent calcaneal spur was observed, the inferior calcaneal boundary was not segmented. Consequently, CAL was annotated in 600 images, M5 in 708, and CC in 754. A total of 546 images had all three structures successfully annotated. Where a trained model already existed, its prediction could be offered as a pre-label for review; masks that the annotator reviewed and, where necessary, corrected before acceptance are not independent de novo annotations and are reported separately in Table 1, accounting for 44.7% of CAL masks but under 2% of the other two structures.

A second observer, a fifth-year radiology resident, independently re-annotated a random subset of 50 development radiographs, originally drawn by the first annotator.

The second annotator worked from a blank canvas: model pre-labeling is disabled in the application for every annotator other than the first. This is enforced in software rather than by instruction because a single pre-label would convert the comparison into a measure of how often the second annotator accepts the model. The masks of the second annotator were never used for training and were used only to quantify the uncertainty of the reference standard.

2.4. Data partitions

A total of 1,037 radiographs were retrieved; after applying the exclusion criteria and removing duplicates, 884 radiographs were included and partitioned into the development (779) and held-out internal test (105) cohorts. Duplicate images were rejected on import by a pixel-level hash, and a header-level review confirmed that no two records shared a Service–Object Pair In-

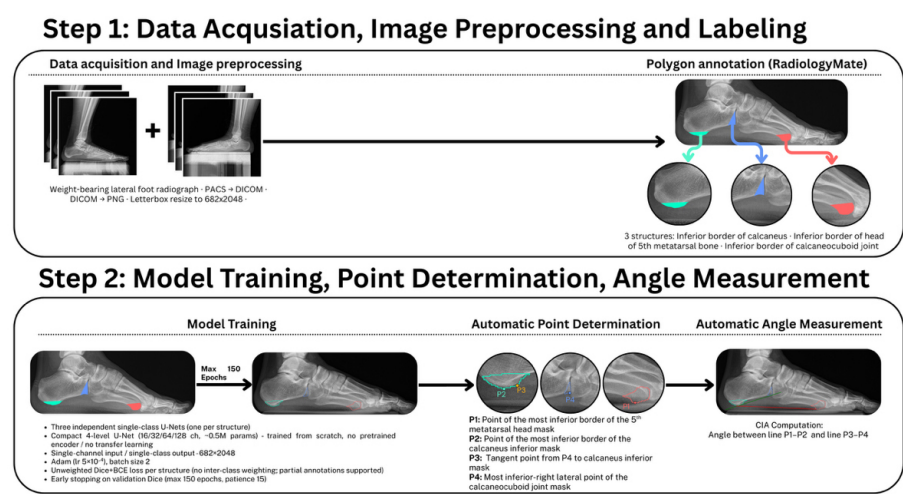


Figure 2. Flow diagram of the methodology. AI, artificial intelligence; CAL, calcaneus inferior; CC, calcaneocuboid joint; DICOM, Digital Imaging and Communications in Medicine; M5, head of 5th metatarsal; PNG, Portable Network Graphics; ROI, region of interest.

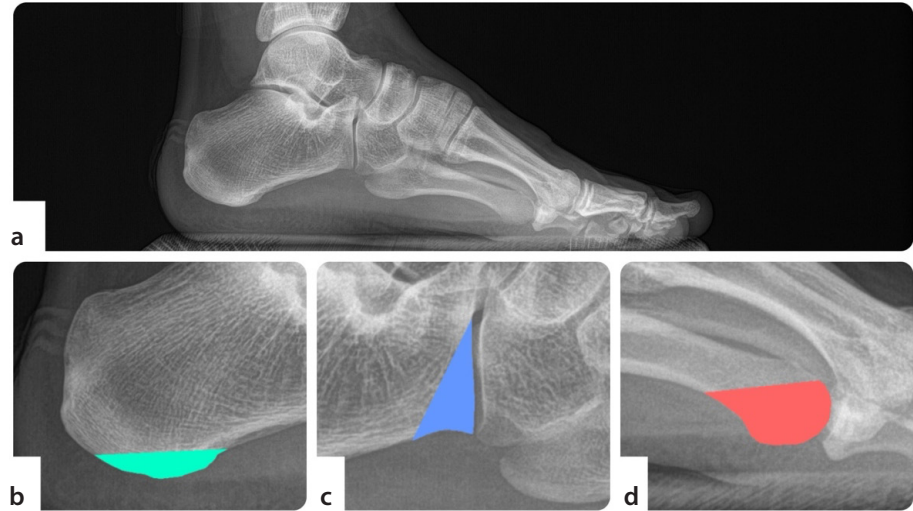


Figure 3. Partial segmentation of anatomical structures using RadiologyMate. (a) Weight-bearing lateral foot radiograph. (b) Turquoise mask: The inferior contour of the calcaneus (CAL). (c) Blue mask: The calcaneal aspect of the inferior calcaneocuboid joint (CC). (d) Red mask: The inferior contour of the fifth metatarsal head (M5).

Anatomical mask	Annotated images (n)	Dice (mean)
Calcaneus inferior (CAL)	600	0.9363
Calcaneocuboid joint (CC)	754	0.8074
Head of 5 th metatarsal (M5)	708	0.8654
All three structures annotated	546	—
Mean (overall)		0.8694

CAL, calcaneus inferior; CC, calcaneocuboid joint; M5, head of 5th metatarsal.

stance Unique Identifier. The development cohort comprises 779 radiographs from 402 patients and the held-out cohort 105 radiographs from 76 patients, with zero patient keys in common. Within the development cohort, the data were partitioned at the patient level into training (80%) and validation

(20%) sets; whole patients, grouped by the identifiers surviving anonymization (date of birth and sex), were assigned to a single split, so that no patient contributed to both training and validation. The membership of each split was recorded per training run. Because not all anatomical structures were visible or

annotatable in every radiograph, each structure was trained and evaluated only on images with available ground-truth annotations for that structure. The test cohort originates from the same institution, the same PACS, and an overlapping period as the development data; it is therefore described throughout as a held-out internal test set rather than an external validation cohort, and no claim of generalizability beyond this institution is made. The CIA measurements for these 105 cases were performed independently and in a blinded manner by the trained model and three readers.

Separately, an independent external cohort of 50 weight-bearing lateral foot radiographs from 33 patients was obtained from a second institution [external imaging equipment: SG Healthcare Jumong (2017) and Fujifilm FDR Smart X (2018)]. These radiographs were not used at any stage of development or threshold selection and were reserved solely for external validation; the frozen primary model and per-class thresholds were applied to them without modification, and all three readers measured them independently under the same blinded protocol, with the three-reader mean as the reference standard.

2.5. Model architecture and training

For the semantic segmentation of foot radiographs, a U-Net model, characterized as a fully convolutional encoder-decoder architecture, was preferred.¹⁴ Rather than a single multi-channel network, each of the three anatomical regions was segmented by its own independent single-class network so that each structure has its own loss, its own validation split, and its own best checkpoint. The architecture is a compact four-level U-Net model (16, 32, 64, and 128 channels; single-channel input; approximately 0.5 million parameters) trained from scratch; no pretrained encoder and no transfer learning were used. Images were resized to $682 \times 2,048$ pixels for model input, preserving the elongated aspect ratio of the lateral foot radiograph. The decoder network utilized skip connections to merge high-resolution feature maps from the encoder, thereby preventing the loss of spatial details and generating a binary mask.

Loss Function: Each anatomical region is segmented by its own network so that the loss is computed over a single class and requires no inter-class weighting. An unweighted hybrid Dice + Binary Cross-Entropy (BCE) loss was used for every structure.

Data Augmentation: No data augmentation was applied. Each single-class network

is trained on a single anatomical structure, and foot orientation is resolved deterministically at measurement time rather than being learned, so spatial augmentation is not required.

Hyperparameters: The network weights were optimized using the Adam algorithm, with an initial learning rate of 5×10^{-4} and a batch size of two for a maximum of 150 epochs. The mask-based Dice coefficient was calculated on the validation set at the end of each epoch, and the weights from the epoch reaching the highest validation Dice were permanently saved. Early stopping was applied to this same criterion: training was terminated when the validation Dice had not improved by more than 0.001 for 15 consecutive epochs, with retention of the best weights. To characterize run-to-run variability, each structure was trained with five predefined random seeds (11, 22, 33, 44, 55) under the patient-level split, and the mean $\pm$ standard deviation (SD) of the validation Dice scores across the five runs is reported; no run was selected on performance. A single pre-specified run (seed 11), fixed a priori, provided the weights used for all downstream testing. Per-class binarization thresholds were subsequently selected by sweeping 0.05 to 0.95 in steps of 0.05 on the recorded validation split, maximizing the validation Dice, and were then frozen and applied unchanged to the held-out cohort.

2.6. Deterministic post-processing and geometrical measurement

To avoid black-box regression models and ensure high interpretability, four critical land-

marks were automatically derived from the output segmentation masks using deterministic rules based on clinical anatomy (Figure 4):

- **P1:** The most inferior point (max Y) of the inferior M5 mask.
- **P2:** The most inferior point (max Y) of the inferior CAL mask.
- **P3:** The algorithmic tangent point on the inferior CAL mask drawn from point P4.
- **P4:** The most inferolateral point of the calcaneal aspect of the CC, determined by the maximum (X+Y) score.

The most inferolateral point of the CC mask was determined using the maximum (X+Y) score, or the maximum (Y-X) score when the foot points in the opposite direction; the orientation is resolved deterministically from the relative horizontal positions of the forefoot and heel masks, so image orientation does not need to be standardized by hand. The CIA was calculated as the angle between vector P1-P2, representing the ground plane, and vector P4-P3, representing the calcaneal slope (Figure 4d), using the following formula:

$$CIA(\theta) = \arccos [((P1 - P2) \cdot (P4 - P3)) / (\|P1 - P2\| \times \|P4 - P3\|)]$$

2.7. Evaluation metrics and statistical analysis

The model's segmentation performance at the mask level was evaluated on the validation set using the Dice coefficient, which was calculated separately for each anatomical structure (M5, CAL, and CC).

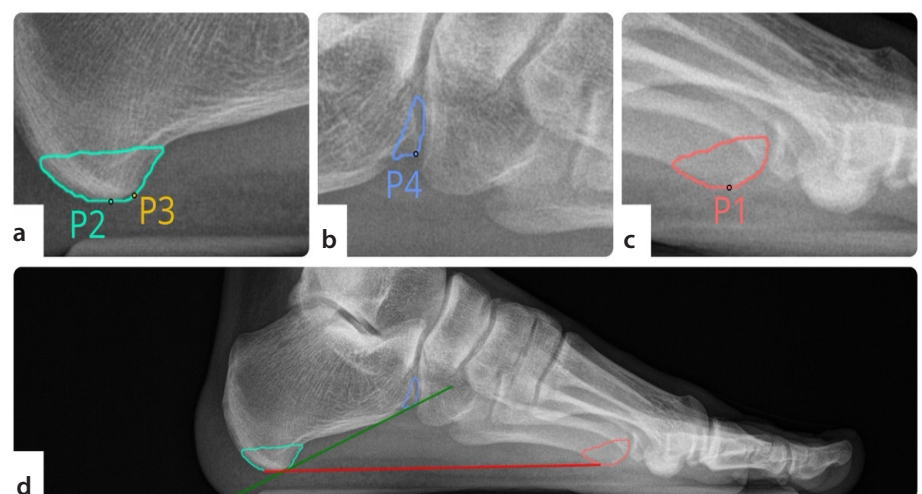


Figure 4. The masks predicted by the model and the points and lines obtained through deterministic landmark detection: (a) P2: The most inferior point (max Y) of the inferior calcaneus mask, and P3: The algorithmic tangent point on the inferior calcaneus mask drawn from point P4. (b) P4: The most inferolateral point of the calcaneal aspect of the calcaneocuboid joint, determined by the maximum (X + Y) score. (c) P1: The most inferior point (max Y) of the inferior fifth metatarsal head mask. (d) Green line: Vector from P4 to P3. Red line: Vector from P1 to P2.

The clinical accuracy of the geometric measurements was assessed on the held-out internal test set of 105 cases that was entirely independent of the training process. Three readers measured every case independently and blinded, each placing P1, P2, and P4 on the radiograph and then fixing a line pivoting on P4 to obtain P3, using the same angle formula as the model so that manual and automated values lie on the same scale. Reader 1, an attending (expert) radiologist, also produced the training masks; Readers 2 and 3 were radiology residents in their fifth and fourth years of training, respectively, and had no involvement in the training data. All three readers measured within the application under the blind-reading protocol described above. Blinding was enforced by the software, which withheld the model prediction, the other readers' values, every difference column, and all summary statistics while blind reading was active; landmark coordinates and the elapsed time were recorded automatically for every case. The primary reference is the mean of the three readers. The following metrics were used: mean absolute error (MAE) for the CIA, proportion of cases falling below clinical acceptability thresholds (1°, 2°, 3°, and 5°), Bland-Altman analysis (mean bias and 95% limits of agreement), and intraclass correlation coefficient (ICC). The same analyses were repeated between readers to quantify inter-observer agreement.

Limits of agreement are reported with 95% confidence intervals (CIs) computed as $SE = \sqrt{(3s^2/n)}$, and agreement was analyzed within the Bland-Altman framework. Because 29 studies contributed two radiographs each, all MAE CIs were repeated with a cluster bootstrap over patients (5,000 resamples). The ICC values are ICC(A,1)—two-way random effects, absolute agreement, and single measurement—with the form stated explicitly as recommended. Inter-annotator segmentation agreement between the two observers is reported with the Dice and Jaccard coefficients, the average symmetric surface distance, and the 95th-percentile Hausdorff distance, since overlap alone is a poor summary for elongated partial structures, together with ICC(A,1) on the mask area. The propagation of landmark error into angular error was assessed both by random isotropic perturbation of P1–P4 and by decomposing the observed reader-to-reader landmark disagreement into components parallel and perpendicular to the line each landmark defines.

A classic power calculation does not apply since no two groups are compared; the test

cohort size is justified by the precision of the agreement estimates. For an anticipated SD of the differences of 1.0°, the 95% confidence half-width of each limit of agreement is 0.48° at $n = 50$, 0.39° at $n = 75$, and 0.33° at $n = 105$. Because some studies contributed two radiographs, the effective sample size is below 105, and these figures are therefore reported as an upper bound.

2.8. Software and hardware environment

Dataset preprocessing, model training, and all deterministic algorithms were implemented using Python (v3.12). The architecture and optimization processes were built upon the PyTorch framework. OpenCV and NumPy were used for contour determination and tangent algorithms. Annotation, training, and measurement were performed in RadiologyMate, a purpose-built local web application developed by our team, running on a workstation with a consumer graphics processing unit; no cloud service was used for any step involving patient images. The Python source code for the deterministic landmark detection and angle computation has been deposited in a public repository. The repository URL and DOI have been omitted from the blinded manuscript and are provided to the editor separately. Competing interests: RadiologyMate is a non-commercial application developed by the authors and made freely available for download at www.radiologymate.com; there is no associated patent, license, or commercial interest, and the authors declare no financial conflicts of interest.

3. Results

3.1. Dataset characteristics

A total of 1,037 LWBRs were retrospectively retrieved from the PACS from between November 2023 and April 2026; after exclusions and duplicate removal, 884 were included in this study [mean age 45.4 ± 14.0 years, range 19–82; 660 women (75%), 224 men (25%)] (Table 2). The anatomical structures were

annotated as follows: CAL in 600 images, M5 in 708 images, and CC in 754 images. The number of images in which all three structures were successfully annotated was 546 (Table 1). The development cohort of 779 radiographs was partitioned into training (80%) and validation (20%) sets. A held-out internal test set of 105 radiographs from 76 imaging studies, entirely separate from the training process and sharing no patient with it, was established for comparison between the model and the three readers.

3.2. Segmentation performance

Each structure was trained under the same protocol with early stopping on the validation Dice. For the primary run, of the planned 150 epochs, training halted at epoch 47 for CAL, epoch 62 for M5, and epoch 58 for CC, with the best weights obtained at epochs 32, 47, and 43, respectively. Among the three target structures, the highest segmentation accuracy was observed for the CAL mask (Dice = 0.9363), reflecting the abundance of training annotations and the relatively large pixel area of the structure (Table 1). The M5 mask achieved a Dice of 0.8645 and the CC mask 0.8074, the latter being the smallest structure with the least distinct cortical boundary. Across the five seeded runs, the validation Dice was stable, with a mean $\pm$ SD of 0.935 ± 0.010 (CAL), 0.871 ± 0.005 (M5), and 0.805 ± 0.009 (CC); no run was selected on performance, and the pre-specified primary run (seed 11) used for all downstream analyses fell within this range for every structure. The threshold sweep changed the operating point only marginally: for all three structures, every threshold between 0.05 and 0.80 lay within 2% of the best pooled validation Dice, and the selected values were 0.50 (CAL), 0.30 (M5), and 0.40 (CC).

The training and validation curves for each structure are shown in Figure 5. The validation Dice rose steeply during the first 20–30 epochs and then plateaued, and in each case, early stopping terminated train-

Table 2. Demographic and imaging characteristics of the study population		
Characteristic	Development cohort	Held-out test cohort
Number of patients, n	402	76
Number of images, n	779	105
Age, years		
Mean $\pm$ SD	45.3 $\pm$ 14.0	46.7 $\pm$ 14.0
Median (IQR)	46 (33–57)	51 (34–55)
Range	19–82	22–77
The held-out test set was not used for polygon annotation; CIA measurements in this set were performed independently by an expert radiologist (Reader 1) and two radiology residents in their fifth (Reader 2) and fourth (Reader 3) years of training. CIA, calcaneal inclination angle; IQR, interquartile range; SD, standard deviation.		

ing 15 epochs after the last improvement; therefore, the reported weights are those of the best epoch rather than the last. A gap between training and validation loss is expected with the combined Dice + BCE objective and does not by itself indicate overfitting; equally, early stopping does not exclude it. The checkpoint was selected on the highest validation Dice, and training was stopped once that metric plateaued, which limits overfitting to the training data, and the small training-to-validation Dice gap is consistent with only limited overfitting.

3.3. Inter-annotator agreement

Between the two observers, Dice was 0.848 ± 0.080 for CAL, 0.912 ± 0.044 for M5, and 0.819 ± 0.115 for CC, with a pooled value of 0.860 ± 0.093 over 150 mask pairs. The average symmetric surface distance was 0.30, 0.10, and 0.18 mm, respectively, and the 95th-percentile Hausdorff distance was 1.97, 0.74, and 1.02 mm, respectively (Table 3). The intraclass correlation on mask area was markedly lower (0.456, 0.809, and 0.652) because the second annotator drew the calcaneal band systematically 17% wider and

the other two structures approximately 3% narrower; absolute-agreement ICC penalizes such a systematic width offset even when the boundary position agrees, which the sub-millimeter surface distances indicate it does.

Passing the masks of the second annotator through the identical deterministic geometry changed the CIA by $0.389^\circ \pm 0.294^\circ$ (median 0.372° , maximum 1.157°), with a bias of $+0.210^\circ$, limits of agreement of -0.658° to $+1.078^\circ$, and ICC(A,1) of 0.996; 94% of cases agreed within 1° and all were within 2° .

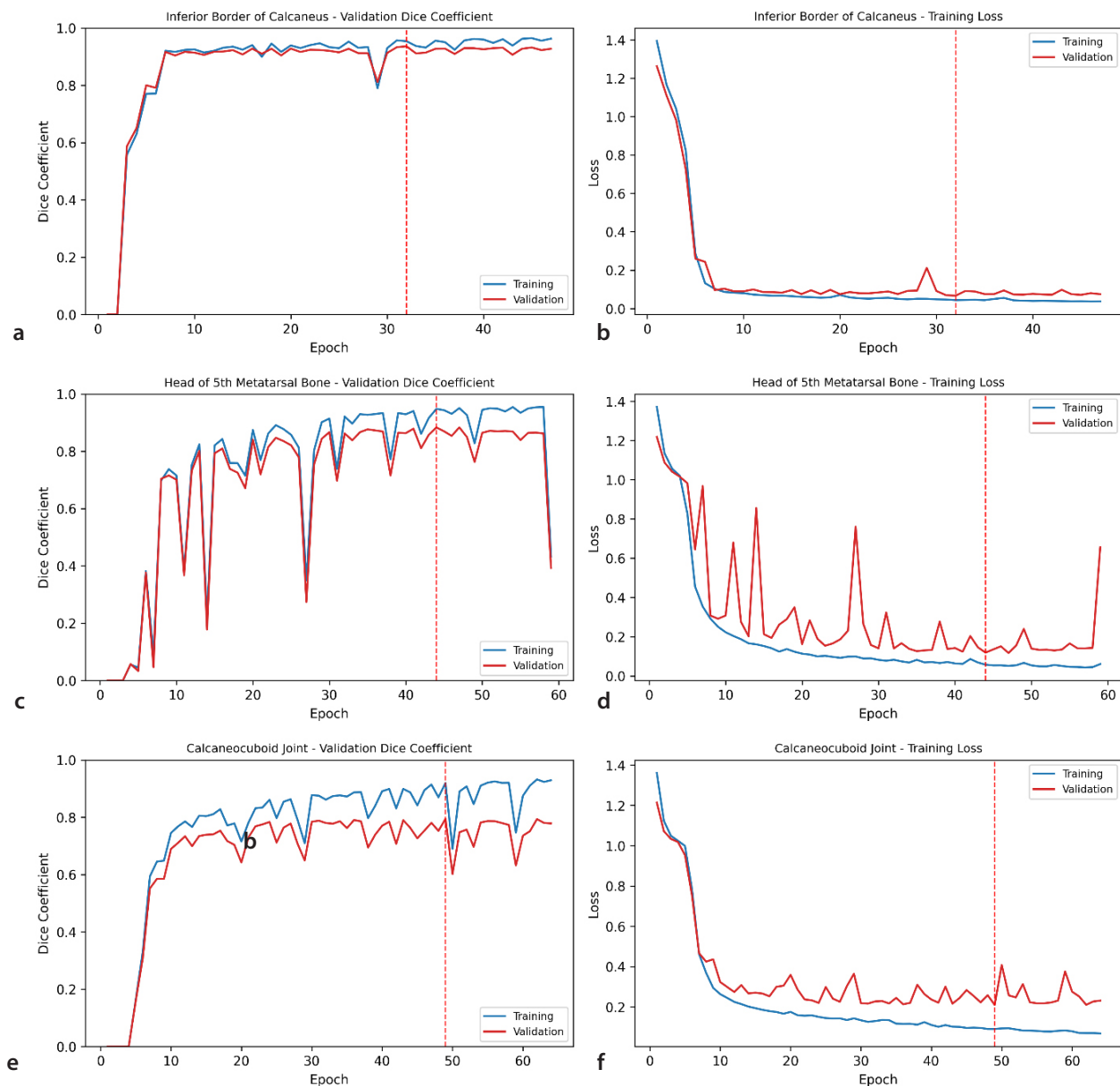


Figure 5. Training and validation curves for the three per-structure segmentation models (inferior border of the calcaneus, CAL; head of the fifth metatarsal, M5; and the calcaneocuboid region, CC). For each structure the left panel shows the training and validation Dice coefficient and the right panel the training and validation loss. The dashed red line marks the best-validation-Dice epoch at which early stopping (patience 15) restored the best weights. The curves are from a representative seed-11 training run under the patient-level split; the deployed primary-run validation Dice values are those reported in Table 1 (CAL 0.936, M5 0.865, CC 0.807), and run-to-run variability across the five predefined seeds is reported as mean $\pm$ standard deviation in Section 3.2.

Table 3. Inter-annotator agreement between the two observers on 50 development radiographs re-annotated from a blank canvas

Structure	Dice (mean $\pm$ SD)	Jaccard	ASSD (mm)	HD95 (mm)	ICC(A,1) mask area (95% CI)
Calcaneus inferior (CAL)	0.848 $\pm$ 0.080	0.744 $\pm$ 0.116	0.295 $\pm$ 0.304	1.971 $\pm$ 1.778	0.456 (−0.04 to 0.72)
Head of 5 th metatarsal (M5)	0.912 $\pm$ 0.044	0.841 $\pm$ 0.072	0.097 $\pm$ 0.107	0.743 $\pm$ 0.930	0.809 (0.20 to 0.93)
Calcaneocuboid joint (CC)	0.819 $\pm$ 0.115	0.708 $\pm$ 0.154	0.180 $\pm$ 0.237	1.020 $\pm$ 1.020	0.652 (0.40 to 0.80)
Pooled (150 mask pairs)	0.860 $\pm$ 0.093	—	—	—	—

ASSD, average symmetric surface distance; CAL, calcaneus inferior; CC, calcaneocuboid joint; CI, confidence interval; HD95, 95th percentile Hausdorff distance; ICC, intraclass correlation coefficient; M5, head of 5th metatarsal; SD, standard deviation.

Substituting one structure at a time, the CC mask contributed most (0.338°) despite being the smallest structure because it defines the pivot of the tangent line, followed by CAL (0.212°) and M5 (0.126°).

3.4. Held-out clinical validation

The CIA measurements from the model and the three readers in the held-out test cohort ($n = 105$) are presented in Supplementary Table 1. The mean CIA values were comparable across all four raters: model = $21.12^\circ \pm 5.52^\circ$, Reader 1 = $21.56^\circ \pm 5.79^\circ$, Reader 2 = $21.49^\circ \pm 5.77^\circ$, and Reader 3 = $21.57^\circ \pm 5.72^\circ$, spanning a range of 4.72° – 34.80° (Supplementary Table 1).

Inter-reader agreement was excellent, with pairwise MAEs ranging from 0.362° (Reader 1 vs. Reader 3) to 0.591° (Reader 1 vs. Reader 2), small systematic biases (-0.083° to $+0.072^\circ$), and narrow 95% limits of agreement (Supplementary Table 2).

3.4.1. Agreement with the reference standard

The MAE of the model against the mean of the three readers was 0.690° (95% CI 0.525 – 0.855 ; cluster-bootstrap CI 0.555 – 0.859), with a median of 0.458° and a maximum of 6.11° . Against the individual readers, the MAE was 0.675° for Reader 1, 0.876° for Reader 2, and 0.743° for Reader 3 (Supplementary Table 1). The readers agreed with one another more closely than the model agreed with any of them: 0.591° between Readers 1 and 2, 0.362° between Readers 1 and 3, and 0.486° between Readers 2 and 3. Reader 1 also produced the training masks, so the comparison that bears on circularity is that the model agreed with Reader 1 only 0.082° more closely than with the mean of the two independent readers—a difference smaller than the spread among the readers themselves (0.114°).

The model exhibited a small negative bias of -0.425° (95% CI: -0.620° to -0.230° ; patient-cluster CI -0.623° to -0.250°) relative to the three-reader mean, indicating that on av-

erage the model slightly underestimated the CIA. The limits of agreement were -2.427° (95% CI -2.748° to -2.082°) to $+1.557^\circ$ (95% CI $+1.191^\circ$ to $+1.857^\circ$). Between readers, the biases were $+0.072^\circ$ (Readers 1 and 2; limits of agreement -1.444° to $+1.589^\circ$), -0.011° (Readers 1 and 3; -0.949° to $+0.927^\circ$), and -0.083° (Readers 2 and 3; -1.392° to $+1.225^\circ$).

3.4.2. Threshold analysis

Against the three-reader mean, 80.9% of model predictions fell within 1° , 92.4% within 2° , 98.1% within 3° , and 99.0% within 5° . Between readers, agreement within 2° was reached in 100% of cases for Readers 1 and 2, 100% for Readers 1 and 3, and 99.0% for Readers 2 and 3; therefore, the human raters essentially always remained within the 2° tolerance, whereas the model exceeded it in eight cases.

3.4.3. Intraclass correlation

The ICC(A,1) across all four raters (model plus three readers) was 0.9863 (95% CI: 0.9808–0.9904), indicating excellent absolute agreement. When restricted to the three readers, the ICC was 0.9936 (95% CI: 0.9912–0.9955). The pairwise ICC values ranged from 0.9756 (model–Reader 2) to 0.9966 (Readers 1 and 3), with the model–Reader 1 pair yielding 0.9813 (Supplementary Table 1). A Friedman test indicated a small but statistically significant difference across the four raters ($\chi^2 = 17.4$, $P < 0.001$), consistent with the model's slight negative bias relative to the readers.

3.5. Error propagation

Random isotropic displacement of the landmarks produced the angular errors listed in Supplementary Table 3. Displacing P3 or P4 is approximately 2.6 times as damaging as displacing P1 or P2, which follows from the geometry: the tangent line P4–P3 averaged 50.6 mm and the ground line P1–P2 averaged 129.2 mm, giving sensitivities of 1.13° and 0.44° per millimeter of perpendicular displacement, respectively.

The stored reader landmarks allowed this to be checked against real observer behavior. Decomposing every reader-to-reader landmark disagreement into the component parallel to the line that landmark defines and the component perpendicular to it, the perpendicular component—and the only angle-relevant component—was 0.92 mm for P1, 0.36 mm for P2, 0.43 mm for P3, and 0.51 mm for P4. The parallel components were far larger: a mean of 41.7 pixels for P1, and in one radiograph, two readers placed P1 133 mm apart, whereas their angles differed by only 0.04° . Readers frequently disagree about the point and agree about the line. A first-order model driven only by the perpendicular components reproduced the observed inter-reader differences (predicted 0.571° vs. observed 0.591° for Readers 1 and 2; 0.353° vs. 0.362° for Readers 1 and 3).

3.6. Failure modes

Eight of the 105 radiographs exceeded 2° and two exceeded 3° , the largest error being -6.11° . In all eight, the readers agreed closely with one another (inter-reader SD 0.05° – 1.02° , mean 0.30° against 0.38° across the cohort); these are therefore model failures rather than radiographs on which the measurement is intrinsically ambiguous. These cases had a reference CIA of 19.1° – 34.5° (mean 25.8° , vs. a cohort mean of 21.5°), and the two errors exceeding 3° occurred in the feet with the highest inclination (reference CIA 29.8° and 34.5°), so the residual errors were concentrated toward higher-arch configurations.

The CC mask, which defines the pivot P4 and is both the smallest of the three structures and the one with the lowest validation Dice, is the most likely origin of these residual errors, consistent with the error-propagation analysis. No failure carried a poor-image-quality flag, a calcaneal spur was present in only one of the eight cases (in which the inferior calcaneal border was, by protocol, not segmented), and faint cortical borders did not otherwise predict error.

3.7. External validation

An external cohort of 50 weight-bearing lateral foot radiographs from 33 patients (37 women, 13 men; mean age 47.3 ± 12.4 years, range 20–73), acquired at a second institution and not used at any stage of model development or threshold selection, was measured by all three readers under the identical blinded protocol. Applying the frozen primary model and per-class thresholds without modification, the MAE against the three-reader mean was 0.85° (95% CI 0.63–1.08; patient-cluster CI 0.64–1.10), with a median of 0.68° and a maximum of 4.01°. Agreement fell within 1° in 66.0%, within 2° in 92.0%, and within 3° in 98.0% of cases. The model showed a negative bias of –0.74° (95% limits of agreement –2.54° to +1.05°) and an ICC(A,1) of 0.969 against the reader mean, whereas the three readers again agreed very closely with one another (inter-reader ICC 0.996–0.998; MAE 0.25°–0.30°). External performance was therefore modestly lower than internal performance, with a somewhat larger underestimation bias, but remained within a clinically acceptable range (Table 4).

3.8. Comparison of design choices

Two of the design choices originally presented as advantages were examined in controlled comparisons, each reusing the same deterministic geometry pipeline and a matched compact U-Net model so that only the variable under study differed. The three independent single-class networks were first compared against a single multi-class network with one shared encoder and three output channels, trained under the identical patient-level split and seed. The multi-class network reached a validation Dice of 0.932 (CAL), 0.874 (M5), and 0.802 (CC), comparable to the per-class values (0.936 / 0.865 / 0.807), and—after its per-channel decision thresholds were swept independently—a held-out MAE of 0.68° with 96.0% of measurements within 2°, comparable to the deployed per-class pipeline (0.690°, 92.4%). Accuracy was therefore equivalent between the two configurations. The per-class design was retained not because it is more accu-

rate but because it removes inter-class loss weighting entirely—the balancing of a single combined loss across structures of markedly different size and difficulty that had been questioned—and because it lets each structure retain its own best checkpoint. Consistent with this, the shared network required strongly asymmetric per-channel thresholds (0.90/0.05/0.50) to balance the three structures, whereas the per-class networks used moderate, similar thresholds (0.50/0.30/0.40); the single network thus reproduced at the thresholding stage the balancing problem that separate networks avoid by construction.

The deterministic landmark-and-geometry stage was then compared against direct end-to-end regression of the angle from the image, using the same backbone and training regimen. The regression model reached a held-out MAE of 5.00°, with only 20.0% of measurements within 2° and a maximum error of 16.36°, against 0.690°, 92.4%, and 6.11° for the deterministic pipeline—roughly an order of magnitude less accurate and without the intermediate landmarks that let a clinician inspect or correct an individual measurement. Both comparisons are summarized in Supplementary Table 4.

3.9. Subgroup analysis

To examine whether accuracy was uniform across the held-out cohort, per-radiograph absolute errors against the three-reader mean were stratified post-hoc (n = 105; Supplementary Table 5). Stratified by the reference angle, agreement was closest in feet with lower arches (MAE 0.59°, 95.5% within 2°, below 20°; 0.57°, 93.5%, between 20° and 25°) and fell in high-arch feet (0.97°, 86.7% within 2°, maximum 6.11°, at or above 25°), consistent with the concentration of failures toward higher inclinations reported in Section 3.6. Nineteen radiographs carried a calcaneal-spur quality flag, on which the CAL is by protocol not segmented; accuracy on these was not inferior (0.57° vs. 0.72°; 94.7% vs. 91.9% within 2°), and none produced an error exceeding 3°. Performance was similar between the sexes (0.69° in women, 0.67° in

men). Feet with fractures, deformities, previous surgery, or metallic hardware were excluded at enrollment and are therefore not represented in these subgroups.

4. Discussion

In this study, we developed an automated pipeline for CIA measurement based on region-specific semantic segmentation of the critical anatomical boundaries. Our approach differs from those of previous studies in several key methodological aspects.

First, although traditional models often segment entire osseous structures,^{15,16} our pipeline targets only the specific anatomical regions required for CIA computation: the CAL, M5 region, and CC junction. This task-specific framing reduced the time spent on segmentation relative to full-bone contours. Because our ground-truth annotations strictly prioritized specific functional boundaries and other borders were delineated with less emphasis on precise cortical margins, the resulting Dice scores for these small regions did not reach the high values typical of whole-bone segmentation.^{15,17} The consequence of this for the measurement itself was quantified directly (Sections 3.3 and 3.5): although the two observers’ masks overlapped only moderately, the angles derived from them agreed closely. The explanation is geometric: the CIA is determined by the direction of two lines, so only displacement perpendicular to a line can alter it, and the observed perpendicular components of reader disagreement were sub-millimeter, whereas the parallel components were an order of magnitude larger. Dice, an area-overlap statistic, is sensitive to losing extent at the ends of an elongated structure, which is precisely the error that leaves a line’s direction unchanged. This argument is specific to angular measurements defined by line directions and does not generalize to tasks in which regional extent is itself the quantity of interest.

Second, the architecture is intentionally modest: a compact four-level U-Net model of approximately 0.5 million parameters was trained from scratch, without a pretrained

Table 4. Agreement in the external validation cohort (50 radiographs, 33 patients)

Comparison	MAE (95% CI)	Bias	95% limits of agreement	Within 2°	ICC(A,1)
Model vs. 3-reader mean	0.85° (0.63–1.08)	–0.74°	–2.54° to +1.05°	92.0%	0.969
Reader 1 vs. Reader 2	0.30° (0.25–0.37)	–0.02°	–0.77° to +0.74°	100%	0.997
Reader 1 vs. Reader 3	0.30° (0.23–0.38)	–0.10°	–0.87° to +0.67°	100%	0.996
Reader 2 vs. Reader 3	0.25° (0.20–0.31)	–0.09°	–0.71° to +0.53°	100%	0.998

CI, confidence interval; ICC, intraclass correlation coefficient; MAE, mean absolute error.

encoder or transfer learning. We make no claim of architectural novelty. The design choice we would defend is the separation of the three structures into independent single-class networks, which lets each structure be optimized on its own loss and retain its own best checkpoint instead of competing with others for a single combined score. A direct comparison supported this choice: a matched single multi-class network measured the angle with equivalent accuracy (Section 3.8, Supplementary Table 4) but required strongly asymmetric per-channel thresholds to balance the three structures, the very balancing that separate networks avoid by construction.

Lee et al.¹⁸ utilized a heatmap-in-heatmap regression approach and reported enhanced performance over the segmentation-based FlatNet method implemented by Ryu et al.¹² in the context of foot deformity assessment. Ryu et al.¹² reported a calcaneal pitch external MAE of 1.17° – 1.23° , whereas Lee et al.¹⁸ reported $1.53^{\circ} \pm 1.33^{\circ}$; our held-out MAE was 0.690° (Supplementary Table 1). These figures are not directly comparable: the aforementioned studies measured a different angle, with different anatomical reference points, on different cohorts, and against externally acquired data rather than a held-out internal set. They are cited to indicate the order of magnitude of the error achievable with this class of method. Replacing the deterministic landmark-and-geometry stage with direct angle regression on the same backbone raised the held-out error to 5.00° with only 20.0% of cases within 2° (Section 3.8), underscoring the value of the interpretable geometric stage over end-to-end regression.

Despite these promising findings, our study has certain methodological limitations that must be acknowledged. First, the principal test cohort is internal: it originates from one institution, one PACS, and an overlapping time period. To probe generalizability, we additionally evaluated an external cohort of 50 radiographs from a second institution, on which the model retained clinically acceptable agreement (Section 3.7) although with a slightly larger negative bias; nonetheless a single external set does not establish generalizability across the full range of centers, protocols, and populations, and broader multicenter validation remains necessary. Second, Reader 1 also produced the training masks, which introduces a potential dependence between the reference standard and the training data; the model agreed only marginally more closely with that reader than with the two independent

readers (Section 3.4.1), so any bias toward one clinician's interpretation is small, but it cannot be excluded at this sample size. Third, nearly half the calcaneal masks were model pre-labels that the annotator reviewed and, where necessary, corrected rather than de novo drawings, which introduces a degree of self-confirmation into that structure's training data. Fourth, the second annotator re-annotated 50 radiographs, which is sufficient to estimate inter-annotator agreement but not to support a consensus reference standard for the training masks. Fifth, the pipeline provides no per-prediction uncertainty estimate, so a case in which the CC mask has failed is not distinguishable from the output alone. Finally, the retrospective design precluded evaluation of the model's direct integration into a real-time clinical workflow. Because radiographs with metallic artifacts, fractures, deformities, previous surgery, or poor image quality were excluded, some of the more difficult cases encountered in routine practice were not represented; this narrowing of the clinical spectrum is a potential source of spectrum bias, and the reported accuracy should be interpreted with this in mind. These limitations highlight the necessity for future prospective, multicenter studies involving multiple readers to further validate the clinical robustness and generalizability of the algorithm.

In routine clinical practice, the evaluation of weight-bearing foot radiographs using conventional PACS software is time-consuming and prone to error, particularly for non-expert radiologists and residents. Our study demonstrated that partial semantic segmentation combined with deterministic landmark extraction measures the CIA with an MAE of 0.690° against three blinded readers, with 92.4% of measurements within 2° , approaching but not yet equaling the agreement the readers achieved with one another; on an external cohort from a second institution, the MAE was 0.85° (92.0% within 2°). The uncertainty of the reference standard was quantified directly: two observers annotating independently produced angles agreeing to 0.39° . Because the measurement depends on the direction of two lines rather than on regional overlap, moderate Dice coefficients did not translate into proportionate angular error, and the residual failures concentrated in a single identifiable structure in feet with higher arches. Manual measurement required a mean of 11–32 seconds per image across the three readers, whereas the automated pipeline completed all geometric extractions and angle calculations

in less than a second; this is an observation on the measurement step alone and does not by itself imply improved departmental throughput.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Tables 1-5 Link: <https://d2v96fxpocvxx.cloudfront.net/beb8919b-f013-4ea1-b1c8-40332e840fe1/content-images/5634168a-3a8f-4a1a-ae53-e43b42e-f436a.pdf>

References

1. 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(1):80-87. [\[Crossref\]](#)
2. Canella C. Imaging evaluation of hallux valgus. *Radiol Bras.* 2020;53(3):VII. [\[Crossref\]](#)
3. Younger AS, Sawatzky B, Dryden P. Radiographic assessment of adult flatfoot. *Foot Ankle Int.* 2005;26(10):820-5. [\[Crossref\]](#)
4. Sensiba PR, Coffey MJ, Williams NE, Mariscalco M, Laughlin RT. Inter- and intraobserver reliability in the radiographic evaluation of adult flatfoot deformity. *Foot Ankle Int.* 2010;31(2):141-145. [\[Crossref\]](#)
5. Wang Q, Wang JH, Ji DD, Zhang Y, Zhao HM. Automated measurement of hallux valgus angles using a deep learning model: validation and comparison with surgeons of varying expertise. *Sci Rep.* 2026;16(1):12280. [\[Crossref\]](#)
6. von Schacky CE, Wilhelm NJ, Schäfer VS, et al. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology.* 2021;301(2):398-406. [\[Crossref\]](#)
7. Li J, Li S, Li X, et al. Primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model. *Eur Radiol.* 2023;33(6):4237-4248. [\[Crossref\]](#)
8. Wei W, Huang Y, Zheng J, et al. YOLOv11-based multi-task learning for enhanced bone fracture detection and classification in X-ray images. *J Radiat Res Appl Sci.* 2025;18(1):101309. [\[Crossref\]](#)
9. Kwolek K, Liszka H, Kwolek B, Gądek A. Measuring the angle of hallux valgus using segmentation of bones on X-ray images. In: *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions.* Cham: Springer; 2019. p.313-325. [\[Crossref\]](#)
10. Wang Q, Ji D, Wang J, et al. HVAngleEst: a dataset for end-to-end automated hallux

- valgus angle measurement from X-ray images. *Sci Data*. 2025;12:915. [\[Crossref\]](#)
11. Noh WJ, Lee MS, Lee BD. Deep learning-based automated angle measurement for flatfoot diagnosis in weight-bearing lateral radiographs. *Sci Rep*. 2024;14(1):18411. [\[Crossref\]](#)
 12. Ryu SM, Shin K, Shin SW, 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. [\[Crossref\]](#)
 13. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6(4):e240300. [\[Crossref\]](#)
 14. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Lecture Notes in Computer Science. Vol. 9351. Cham: Springer; 2015. p. 234-241. [\[Crossref\]](#)
 15. Ning X, Ru T, Zhu J, et al. Deep learning-assisted segmentation of X-ray images for rapid and accurate assessment of foot arch morphology and plantar soft tissue thickness. *Sci Rep*. 2024;14(1):20028. [\[Crossref\]](#)
 16. 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. [\[Crossref\]](#)
 17. Koo J, Hwang S, Han SH, et al. Deep learning-based tool affects reproducibility of pes planus radiographic assessment. *Sci Rep*. 2022;12(1):12891. [\[Crossref\]](#) Erratum in: *Sci Rep*. 2022;12(1):15169. [\[Crossref\]](#)
 18. Lee SJ, Yoon H, Bae S, et al. Deep learning-based landmark detection model for multiple foot deformity classification: a dual-center study. *Yonsei Med J*. 2025;66(8):491-501. [\[Crossref\]](#)