Global Health & Medicine. 2026; 8(3):205-214. ORIGINAL ARTICLE

DOI: 10.35772/ghm.2026.01053

Automated radiographic shoulder balance assessment in scoliosis *via* deep learning

Longhao Yang[1,2,§], Fangzheng Xu[1,2,§], Qingzhi Xiang[1,2], Jianwen Fu[3], Xiao Xia[1,2], Fuping Li[1,2], Shaobo Cheng[1,2], Yifei Qin[1,2], Yan Yu[1,2,*]

[1] Division of Spine, Department of Orthopaedics, Tongji Hospital, Tongji University School of Medicine, Tongji University, Shanghai, China;
[2] Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration (Tongji University), Ministry of Education, Shanghai, China;
[3] Spinextech Medical Technology Co., Ltd., Shanghai, China.

Abstract: The objective of this study was to develop an automated deep learning-based method for the assessment of shoulder balance in adolescent idiopathic scoliosis (AIS) patients using X-ray images in order to provide a reliable and efficient alternative to manual measurements. A total of 940 AIS radiographs were screened; 937 cases were included in the model-development cohort after quality control and were annotated for precise identification and segmentation of the T1 vertebra, both clavicles, and both coracoids. A deep learning neural network was used to segment these structures. Landmarks were extracted based on morphological image processing, and shoulder balance parameters including the clavicle angle (CA), coracoid height difference (CHD), clavicle tilt angle difference (CTAD), radiological shoulder height (RSH), and T1 tilting angle (T1TA) were calculated. The accuracy of the automated measurements was validated using an external dataset (n = 70) assessed by three senior spinal surgeons. The deep learning neural network achieved reliable segmentation performance for foreground anatomical structures, with macro-average intersection over union (IoU) values of 0.77 and 0.73 and Dice coefficients of 0.87 and 0.84 in the internal and external validation datasets, respectively. In the external dataset, the automated measurements displayed a high level of agreement with observer-averaged measurements, with intraclass correlation coefficients ranging from 0.964 to 0.994. Bland–Altman analysis revealed small mean biases across the five shoulder balance parameters, and 90.0 to 98.6% of automated measurements were within the range of interobserver variability. The proposed method provides an efficient and reproducible approach for radiographic shoulder balance assessment and may help reduce observer-dependent measurement variability.

***Keywords*:** adolescent idiopathic scoliosis, automated measurement, neural network, X-ray

1. Introduction

Shoulder balance in adolescent idiopathic scoliosis (AIS) and early-onset scoliosis (EOS) holds immense importance in both diagnostic and treatment contexts, as it influences early detection, treatment planning, prognosis evaluation, and aesthetic outcomes (*1-8*). Precise measurements of shoulder parameters, and particularly the clavicle angle (CA), coracoid height difference (CHD), clavicle tilt angle difference (CTAD), radiological shoulder height (RSH), and T1 tilting angle (T1TA), are crucial for assessing pre- and postoperative shoulder balance (*3,9-12*). However, the assessment of these parameters has been challenging due to the inherent subjectivity and variability associated with manual measurements (*12,13*). This subjectivity not only consumes considerable time but can also lead to overlooked discrepancies, especially during screening processes. A reliable, objective method for their assessment needs to be developed, as it would not only minimize the risk of measurement errors but also enhance the overall management and outcomes of AIS patients.

Achieving automated measurement of shoulder balance parameters requires the semantic segmentation of shoulder structures on X-ray images. This segmentation process should encompass the clavicles, coracoids, and T1 vertebra, as these are the primary anatomical structures commonly used in clinical practice to derive essential shoulder balance parameters. Significant progress has been made in the automated measurement of spinal balance alignment parameters in AIS, and considerable accuracy has been achieved (*14-16*), but identification and segmentation of shoulder

structures have received comparatively less attention in the research setting. Currently, there is a notable absence of established automated methods for measuring shoulder balance parameters. Automatically segmenting shoulder structures in medical images poses challenges due to anatomical overlap, low contrast, potential artifacts, and the need for high-quality annotated data. Moreover, another challenge lies in the recognition of key points within the segmented structures to facilitate parameter calculations.

To address these challenges and the need for automated shoulder balance assessment in AIS, we developed an automated pipeline for segmenting the clavicles, coracoids, and T1 vertebra and quantifying relevant shoulder balance parameters. We validated the proposed method using an external dataset by comparing automated measurements with measurements from three senior spinal surgeons. The evaluation included segmentation performance, landmark detection accuracy, agreement with observer-averaged measurements, and comparison of automated measurement error with interobserver variability.

2. Materials and Methods

2.1. Dataset collection

This retrospective study was approved by the Ethics Committee of Shanghai Tongji Hospital (Approval No. SBKT-2025-319). The Ethics Committee waived the requirement for informed consent because of the retrospective nature of the study. All sensitive patient information was anonymized before data analysis to protect patient privacy. We acquired all X-ray images following established clinical whole spine protocols, utilizing digital X-ray radiography systems including AccE GC85A vision (Samsung Electronics), CXDI-401 (Canon Medical Systems), uDR (United Imaging Healthcare), and Ti-WISH-IL (TaoImage). The images were retrieved and stored in the Digital Imaging and Communications in Medicine (DICOM) 3.0 protocol format from 2017 to 2022. A total of 940 AIS radiographs were initially collected. After quality control, three cases were excluded because of insufficient image quality. Therefore, 937 cases were included in the model-development cohort. The dataset was split at the patient level into a training set of 843 cases and an internal validation set of 94 cases. No patient appeared in more than one subset, thereby avoiding patient-level data leakage. To ensure accuracy, two attending physicians with expertise in musculoskeletal radiology independently annotated the X-ray images using the open-source annotation tool X-Anylabeling (CVHub). Annotations included the precise identification and segmentation of the T1 vertebra and both clavicles and coracoids, with disagreements resolved through consensus.

2.2. Shoulder structure segmentation

We developed a deep learning segmentation model to segment the T1 vertebra, both clavicles, and both coracoids on whole-spine radiographs. The model used RepVGG-A1 (*17*) as the backbone encoder, with the final fully connected classification layer removed. The extracted image features were passed to an upsampling segmentation head to restore the spatial resolution and generate pixel-wise masks for five foreground anatomical classes, including the left clavicle, right clavicle, left coracoid, right coracoid, and T1 vertebra. The network treated the left and right anatomical structures as independent semantic classes to facilitate subsequent landmark extraction and parameter calculation. Figure 1C and 1D show the segmentation results. All of the DICOM radiographs were converted to single-channel grayscale images before model

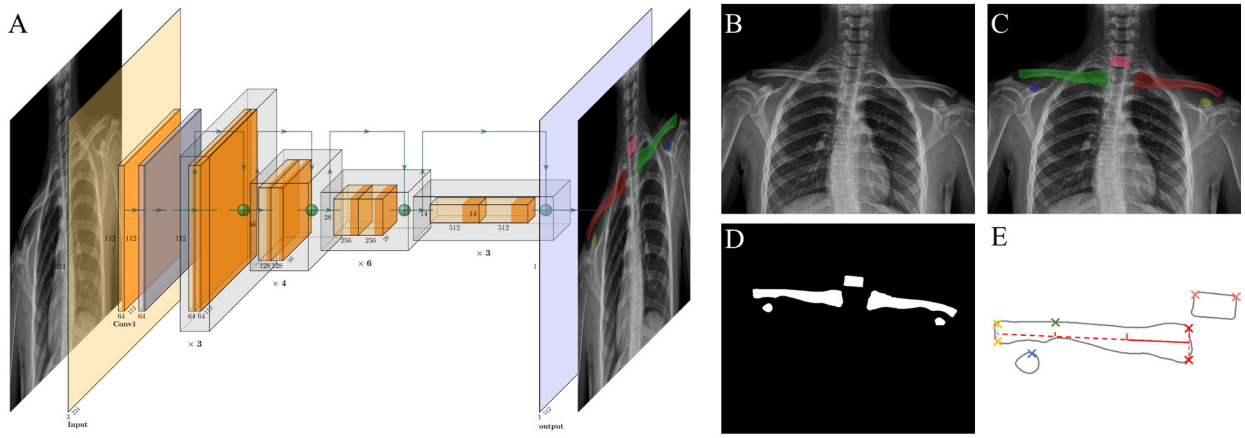

Figure 1. **(A)** Schematic of the segmentation network with a RepVGG-A1 backbone for segmenting the clavicles, coracoids, and T1 vertebra; **(B)** Original X-ray; **(C)** Segment annotations of the clavicles and coracoids; **(D)** Predicted masks of the clavicles and coracoids; **(E)** Key points of shoulder balance including CA, RSH, CHD, CTAD, and T1TA.

training and inference. Image intensities were clipped to reduce the influence of extreme values and then normalized to the range of 0–1. After the initial coarse localization of the T1 vertebra, a local region of interest was cropped from the center of T1. The crop width was set as the original image width, and the crop height was set as half of the crop width. Specifically, 0.2 of the crop height was retained above the center of T1 and 0.8 of the crop height was retained below the center of T1. The cropped local patch was then resized to 512 × 1,024 pixels and used as the input for the final segmentation model.

The model was trained on the 843-patient training set for 20,000 iterations with a batch size of 8. A combined cross-entropy loss and Dice loss was used as the objective function. The cross-entropy loss was used to optimize pixel-wise semantic classification, whereas the Dice loss was used to improve foreground structure overlap and reduce the influence of the class imbalance between the background and relatively small anatomical structures. The total loss was defined as the sum of cross-entropy loss and Dice loss. Common medical image augmentation strategies were applied online during training, including small-angle random rotation within ± 10°, random scaling between 0.9 and 1.1, random translation within ± 5% of the image size, random brightness and contrast adjustment within ± 20%, and mild Gaussian noise. Horizontal flipping was not used because the left and right clavicles and coracoids were defined as separate semantic classes.

The Adam optimizer was used with $\beta1 = 0.9$, $\beta2 = 0.999$, and a cosine annealing schedule was adopted to gradually decrease the learning rate from 1×10^{-3} to 1×10^{-6} over the 20,000 training iterations. Model performance on the internal validation set was monitored during training, and the checkpoint with the highest foreground macro-average Dice coefficient was selected as the final model. All experiments were implemented in Python using PyTorch and performed on an NVIDIA RTX A6000 GPU.

2.3. Shoulder balance parameters

Given the segmentation mask predicted by the network, landmarks were extracted based on morphological image processing, as shown in Figure 1E. The inner and outer upper and lower vertices of the clavicle were detected using a convex point detection algorithm. Moreover, an additional process involved a vertical search upwards from the outer upper vertices on both sides until a notable decrease in grayscale values was observed. This led to the identification of an edge point, signifying the presence of air and corresponding to the acromioclavicular joint. The vertical height difference between these two points was denoted as RSH (Figure 2D), which served as a measure of the vertical soft tissue thickness at the acromioclavicular joints.

The approximate orientation of the clavicle was determined based on the positions of these vertices. Subsequently, the inner one-third median line of the clavicle was calculated. As depicted in Figure 1E, this process entailed determining the midpoint between the inner upper and lower vertices as well as the midpoint between the outer upper and lower vertices. The angle between the inner one-third of the posture line and the horizontal line was then measured, providing the clavicle's angle of inclination. Moreover, the difference in the inclination of the two clavicles was computed and denoted as CTAD (Figure 2B).

The line connecting these two midpoints represented the posture line of the clavicle. In the outer one-third of the clavicle, we searched along the contour for the point farthest from and above the posture line in the vertical direction. This point was identified as the highest point of the clavicle. The angle between the line connecting the highest points on both sides of the clavicle and the

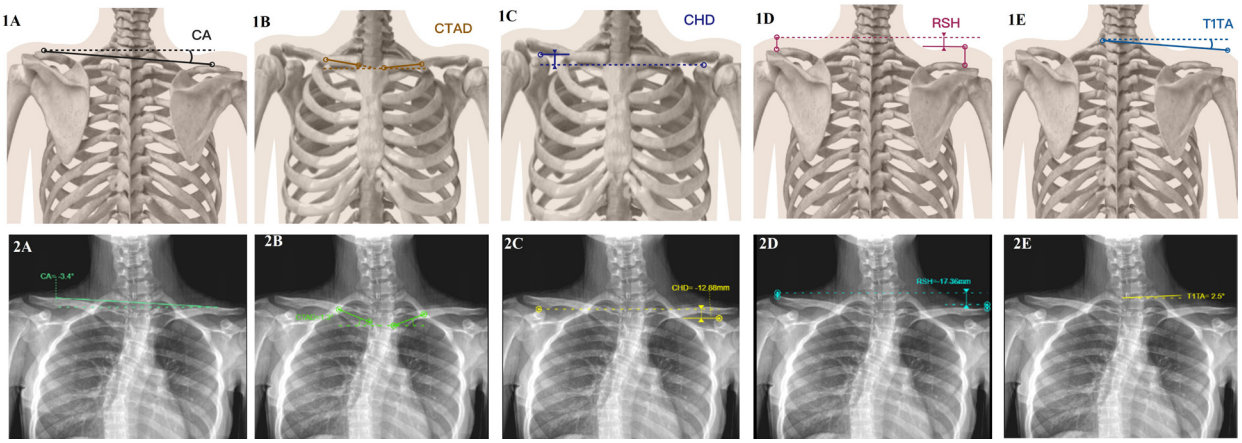

Figure 2. (1) Schematic of shoulder balance parameters; (2) Self-developed measurement tools for shoulder balance parameters. (A) Clavicular angle (CA); **(B)** Clavicle tilt angle difference (CTAD); **(C)** Coracoid height difference (CHD); **(D)** Radiological shoulder height (RSH); **(E)** T1 vertebral tilting angle (T1TA).

horizontal line was denoted as CA (Figure 2A).

Similarly, the vertical height difference between the highest points of the two sides of the masks covering the coracoid process segments was referred to as CHD (Figure 2C). In addition, the upper endplates of the T1 vertebra were detected with a convex point detection algorithm. The angle formed by the line connecting these two points and the horizontal line was defined as T1TA (Figure 2E).

For all landmark-based measurements, coordinates detected on resized local patches were first transformed back to the original DICOM image coordinate system using the inverse transformation of cropping and resizing. Linear measurements reported in millimeters, including the MRE, CHD, and RSH, were then calculated using the image-specific DICOM PixelSpacing information. For Euclidean distances such as the MRE, both row and column pixel spacing values were considered. For vertical height differences such as the CHD and RSH, the vertical coordinate difference was multiplied by the row pixel spacing. For angular measurements, landmark coordinates were first transformed into the physical coordinate system before angle calculation, thereby accounting for anisotropic pixel spacing and differences in pixel spacing across radiographic devices and institutions.

2.4. Measurement validation

To validate the accuracy of the automated measurements generated by the deep learning model, we used an external validation dataset consisting of 70 AIS X-ray images collected from Guizhou Provincial People's Hospital. For external segmentation evaluation, the T1 vertebra, both clavicles, and both coracoids in the 70 external radiographs were manually annotated using the same annotation protocol as the model-development cohort. These manual masks served as the reference standard for calculating external segmentation metrics, including recall, precision, IoU, and the Dice coefficient. In the evaluation of our method of automated measurement to detect landmarks, two crucial metrics were used: the successful detection rate (SDR) and the mean radial error (MRE). The SDR represents the percentage of successful detections at different radial error thresholds (2 mm, 3 mm, and 4 mm), while MRE quantifies the mean error in millimeters between landmark positions determined automatically and manually. To minimize the potential biases associated with manual measurements and to streamline the results, we performed a comparative analysis by comparing the positions obtained through the averaging of measurements from three human observers with the positions detected automatically *via* our method.

Three senior spinal surgeons experienced in AIS assessment independently measured the shoulder balance parameters using a custom-developed software platform equipped with specialized measurement tools. The average of the three observers' measurements was used as the observer-averaged reference for agreement analysis. In addition, observer-specific comparisons between the automated measurements and each individual observer's measurements were performed. For these observer-specific comparisons, measurement error was summarized using the mean absolute error (MAE) with standard deviation, root mean square error (RMSE), minimum and maximum absolute errors (MinAE–MaxAE), and median absolute error with interquartile range [median AE (IQR)].

2.5. Statistical analysis

Agreement between automated measurements and manual measurements was evaluated by comparing automated measurements with the average measurements of the three senior spinal surgeons. Intraclass correlation coefficients (ICCs) were calculated using a two-way mixed-effects model for absolute agreement between the automated measurements and observer-averaged reference measurements. Bland–Altman analysis was performed to quantify systematic bias and 95% limits of agreement (LoA). The difference was defined as the automated measurement minus the average measurement of the three observers.

To compare automated measurement error with human observer variability, automated–observer variability was calculated as the mean absolute difference between automated measurements and individual observer measurements. Interobserver variability was calculated as the average pairwise mean absolute difference among the three observers. Interobserver variability was further used as a parameter-specific empirical acceptability threshold. A case within interobserver variability was considered where the absolute difference between the automated measurement and the observer-averaged measurement did not exceed the interobserver variability for the corresponding parameter. Pearson correlation coefficients and p-values from paired *t*-tests were calculated to assess the correlation between automatic measurements and each observer's measurements. In addition, we performed percentage cumulative error analysis. Statistical analysis was performed using the Python package SciPy (version 1.11.3).

3. Results

3.1. Deep learning neural network performance

Table 1 summarizes the segmentation performance of each foreground anatomical structure, including recall, precision, intersection over union (IoU), and the Dice coefficient, on the validation datasets. The left and right clavicles performed well, with recalls of 0.95 and 0.95,

precision scores of 0.92 and 0.94, IoU values of 0.88 and 0.89, and Dice coefficients of 0.94 for both sides. However, segmentation of the left and right coracoids yielded slightly lower scores, with a recall of 0.76 and 0.75, precision scores of 0.82 and 0.85, IoU values of 0.65 and 0.66, and Dice coefficients of 0.79 for both sides. The T1 vertebra performed extremely well, with a recall of 0.89, precision of 0.87, IoU of 0.79, and Dice coefficient of 0.88. The foreground macro-average recall, precision, IoU, and Dice coefficient in the internal validation dataset were 0.86, 0.88, 0.77, and 0.87, respectively. Because the background class was excluded from the average, these values represent the segmentation performance of clinically relevant foreground anatomical structures.

In the external validation dataset, the model displayed slightly lower but comparable segmentation performance. The left and right clavicles resulted in IoUs of 0.84 and 0.85 and Dice coefficients of 0.91 and 0.92, respectively. The left and right coracoids resulted in IoUs of 0.60 and 0.61 and Dice coefficients of 0.75 and 0.76, respectively. The T1 vertebra resulted in an IoU of 0.74 and a Dice coefficient of 0.85. The foreground macro-average recall, precision, IoU, and Dice coefficient in the external validation dataset were 0.83, 0.85, 0.73, and 0.84, respectively. The total processing time was 89.9 ± 18.3 ms per image.

The 2-mm, 3-mm, and 4-mm SDR and the MRE of the average landmarks in the external dataset are shown in Table 2. The superior coracoid points achieved an SDR of 72.62% at a 2-mm threshold, which increased to 79.10% at the 3-mm threshold and 91.65% at the 4-mm threshold. The MRE for these points was 1.83 mm. The superior clavicle points achieved SDR values of 66.70% (2-mm threshold), 83.55% (3-mm threshold), and 87.89% (4-mm threshold). The MRE for these points was 2.75 mm. The superior extraclavicular points on both sides achieved SDR values of 77.33% (2 mm), 83.10% (3 mm), and 92.65% (4 mm). The MRE for these points was 1.59 mm. The soft tissue shoulder points achieved SDR values of 69.68% (2 mm), 78.31% (3 mm), and 92.26% (4 mm), with an MRE of 1.92 mm. Finally, the superior T1 vertebral point achieved SDR values of 65.21% (2 mm), 75.24% (3 mm), and 90.54% (4 mm) along with an MRE of 2.62 mm.

3.2. Agreement and error analysis

The descriptive statistics of shoulder balance parameters in the external dataset are summarized in Table 3. On average, the CA measurement had a mean value of 0.73°, ranging from −3.9 to 12.1°. The mean absolute CA was 2.11°, ranging from 0.2 to 12.1°. The mean CHD was 3.14 mm, ranging from −16.8 to 42.3 mm. The mean

Table 1. Segmentation metrics of each class in the internal and external validation datasets

	Recall	Precision	IoU	Dice
Clavicle (left)	0.95/0.93	0.92/0.90	0.88/0.84	0.94/0.91
Clavicle (right)	0.95/0.93	0.94/0.91	0.89/0.85	0.94/0.92
Coracoid (left)	0.76/0.72	0.82/0.79	0.65/0.60	0.79/0.75
Coracoid (right)	0.75/0.71	0.85/0.81	0.66/0.61	0.79/0.76
T1 vertebrae	0.89/0.86	0.87/0.84	0.79/0.74	0.88/0.85
Average	0.86/0.83	0.88/0.85	0.77/0.73	0.87/0.84

Notes: Values are presented as internal validation/external validation.

Table 2. The 2-mm, 3-mm, 4-mm success detection rate (SDR) and the mean radial error (MRE) of the average landmarks in the external dataset

	2 mm SDR (%)	3 mm SDR (%)	4 mm SDR (%)	MRE (mm)
Superior coracoid points	72.62	79.10	91.65	1.83
Superior clavicle points	66.70	83.55	87.89	2.75
Superior extraclavicular points	77.33	83.10	92.65	1.59
Soft tissue shoulder points	69.68	78.31	92.26	1.92
Superior T1 vertebral points	65.21	75.24	90.54	2.62

Table 3. Summary of the mean and range for shoulder balance parameters in the external dataset

	Mean	Range	Mean (Absolute)	Range (Absolute)
CA (°)	0.73 ± 2.76	−3.9-12.1	2.11 ± 1.90	0.2-12.1
CHD (mm)	3.14 ± 11.30	−16.8-42.3	9.11 ± 7.32	0.0-42.3
CTAD (°)	−1.85 ± 5.90	−24.8-9.2	4.61 ± 4.10	0.1-24.8
RSH (mm)	1.47 ± 12.88	−24.2-42.7	10.12 ± 8.02	0.3-42.7
T1TA (°)	1.11 ± 4.39	−16.9-9.8	3.43 ± 2.93	0.0-16.9

absolute CHD was 9.11 mm, ranging from 0.0 to 42.3 mm. The mean CTAD was −1.85°, ranging from −24.8 to 9.2°. The mean absolute CTAD was 4.61°, ranging from 0.1 to 24.8°. The mean RSH was 1.47 mm, ranging from −24.2 to 42.7 mm. The mean absolute RSH was 10.12 mm, ranging from 0.3 to 42.7 mm. The mean T1TA was 1.11°, ranging from −16.9 to 9.8°. The mean absolute T1TA was 3.43°, ranging from 0.0° to 16.9°.

We assessed the reliability of these measurements by examining interobserver agreement among three different observers. The statistical analysis revealed strong positive relationships between measurements made by different observers, as indicated by high Pearson's correlation coefficients (*r*) ranging from 0.90 to 0.99 for all parameters. Moreover, paired *t*-tests showed no statistically significant systematic differences between observer measurements for most parameters; however, agreement was primarily evaluated using intraclass correlation coefficients and Bland–Altman analyses. The agreement between automatic measurements and each observer is shown in Table 4.

The distribution of absolute measurement errors between the automated method and each individual observer's measurements is summarized in Table 5. Across the three observers, the MAE ranged from 0.67 to 0.97° for CA, the CHD ranged from 1.00 to 2.56 mm, the CTAD ranged from 2.15 to 2.72°, the RSH ranged from 1.25 to 2.56 mm, and the T1TA ranged from 0.58 to 1.10°. The observer-specific RMSE, MinAE–MaxAE, and median AE [IQR] are also reported in Table 5.

Agreement analysis was performed between the automated measurements and the average measurements

Table 4. Pearson correlation coefficients and paired *t*-tests between automated measurements and each observer

	Observer 1		Observer 2		Observer 3	
	Pearson's *r*	Paired *t*-test (*p* value)	Pearson's *r*	Paired *t*-test (*p* value)	Pearson's *r*	Paired *t*-test (*p* value)
CA (°)	0.92	0.35	0.91	0.38	0.96	0.74
CHD (mm)	0.96	0.74	0.98	0.08	0.99	0.37
CTAD (°)	0.90	0.91	0.90	0.40	0.92	0.28
RSH (mm)	0.99	0.40	0.97	0.26	0.97	0.96
T1TA (°)	0.97	0.23	0.99	0.18	0.95	0.12

Table 5. Distribution of measurement errors between automated measurements and each observer

	Observer 1			
	MAE ± SD	RMSE	MinAE–MaxAE	Median AE [IQR]
CA (°)	0.75 ± 0.68	1.01	0.00–4.10	0.60 [0.30–1.00]
CHD (mm)	2.56 ± 2.03	3.26	0.02–9.00	2.09 [1.35–3.37]
CTAD (°)	2.15 ± 1.78	2.78	0.00–8.70	1.65 [1.00–2.67]
RSH (mm)	1.25 ± 1.01	1.60	0.05–3.88	1.09 [0.45–1.71]
T1TA (°)	0.86 ± 0.66	1.08	0.00–2.80	0.70 [0.40–1.20]
	Observer 2			
	MAE ± SD	RMSE	MinAE–MaxAE	Median AE [IQR]
CA (°)	0.97 ± 0.74	1.22	0.00–3.20	0.80 [0.40–1.40]
CHD (mm)	1.47 ± 0.92	1.73	0.04–3.76	1.57 [0.59–2.11]
CTAD (°)	2.72 ± 1.97	3.36	0.10–8.30	2.20 [1.20–3.98]
RSH (mm)	2.34 ± 1.95	3.04	0.01–9.23	2.05 [0.73–3.13]
T1TA (°)	0.58 ± 0.37	0.68	0.00–1.40	0.50 [0.30–0.87]
	Observer 3			
	MAE ± SD	RMSE	MinAE–MaxAE	Median AE [IQR]
CA (°)	0.67 ± 0.48	0.82	0.00–2.40	0.60 [0.30–0.97]
CHD (mm)	1.00 ± 0.76	1.25	0.00–2.90	0.88 [0.43–1.38]
CTAD (°)	2.15 ± 1.45	2.58	0.10–5.60	1.90 [0.93–3.15]
RSH (mm)	2.56 ± 1.72	3.08	0.01–6.73	2.35 [1.27–3.66]
T1TA (°)	1.10 ± 0.86	1.39	0.00–3.60	0.90 [0.50–1.48]

Notes: Absolute errors were calculated between automated measurements and each individual observer. Median AE and the IQR were calculated from the absolute errors. *Abbreviations*: AE, absolute error; MAE, mean absolute error; SD, standard deviation; IQR, interquartile range; RMSE, root mean square error.

Table 6. Agreement between automated measurements and observer-averaged measurements

	ICC	Mean bias	95% LoA	Automated–observer variability	Within interobserver variability rate
CA (°)	0.974	0.09	−1.11-1.30	0.80	63/70 (90.0%)
CHD (mm)	0.994	0.12	−2.34-2.58	1.67	69/70 (98.6%)
CTAD (°)	0.964	0.21	−3.06-3.48	2.34	67/70 (95.7%)
RSH (mm)	0.993	-0.18	−3.29-2.92	2.05	66/70 (94.3%)
T1TA (°)	0.991	-0.07	−1.14-1.00	0.85	68/70 (97.1%)

Notes: Mean bias and 95% LoA were calculated as automated measurements minus the average measurements made by the three observers. Automated–observer variability was expressed as the mean absolute difference between automated measurements and individual observer measurements. Interobserver variability was expressed as the average pairwise mean absolute difference among the three observers. Within interobserver variability was defined as the proportion of cases in which the absolute difference between the automated measurement and the observer-averaged measurement did not exceed the interobserver variability for the corresponding parameter. *Abbreviations*: ICC, intraclass correlation coefficient; LoA, limits of agreement.

of the three observers. The ICCs for absolute agreement ranged from 0.964 to 0.994, indicating high agreement across all shoulder balance parameters, as shown in Table 6. Bland–Altman analysis revealed small mean biases, including 0.09° for the CA, 0.12 mm for the CHD, 0.21° for the CTAD, −0.18 mm for the RSH, and −0.07° for the T1TA. The corresponding 95% LoA was −1.11 to 1.30° for the CA, −2.34 to 2.58 mm for the CHD, −3.06 to 3.48° for the CTAD, −3.29 to 2.92 mm for the RSH, and −1.14 to 1.00° for the T1TA. The corresponding correlation scatter plots, Bland–Altman plots, and cumulative error curves are shown in Figure 3.

The automated–observer variability was smaller than the corresponding interobserver variability for all five parameters. Specifically, the automated–observer variability values were 0.80° for the CA, 1.67 mm for the CHD, 2.34° for the CTAD, 2.05 mm for the RSH, and 0.85° for the T1TA, whereas the corresponding interobserver variability values were 1.11°, 2.44 mm, 3.25°, 2.93 mm, and 1.31°, respectively. When interobserver variability was used as a parameter-specific empirical threshold, 90.0% of CA, 98.6% of CHD, 95.7% of CTAD, 94.3% of RSH, and 97.1% of T1TA measurements were within the range of interobserver variability.

4. Discussion

In this study, we developed an automated deep learning-based method for shoulder balance assessment in AIS radiographs and evaluated its performance at the segmentation, landmark detection, and measurement levels. For anatomical structure segmentation, the model achieved foreground macro-average IoU values of 0.77 and 0.73 and Dice coefficients of 0.87 and 0.84 in the internal and external validation datasets, respectively, indicating stable segmentation performance across the two validation cohorts. Among the segmented structures, the clavicles performed best, with IoUs of 0.88–0.89 and Dice coefficients of 0.94 in the internal validation dataset and IoUs of 0.84–0.85 and Dice coefficients of 0.91–0.92 in the external validation dataset. The T1 vertebra also displayed acceptable segmentation performance, with IoUs of 0.79 and 0.74 and Dice coefficients of 0.88 and 0.85 in the internal and external datasets, respectively. In contrast, the coracoids displayed relatively lower segmentation performance, with IoUs of 0.65–0.66 internally and 0.60–0.61 externally, reflecting the difficulty of segmenting small anatomical structures with overlapping radiographic projections. Previous studies have shown that small structures and class imbalance can disproportionately affect overlap-based metrics such as the IoU and Dice coefficient (*18-20*). Therefore, the relatively lower coracoid IoU and Dice values should be interpreted together with landmark-level and measurement-level performance. Importantly, the CHD depends primarily on the localization of the superior coracoid landmarks rather than complete mask overlap. In this study, the superior coracoid landmarks achieved an MRE of 1.83 mm and an SDR of 91.65% at the 4-mm threshold. The CHD also showed a small mean bias of 0.12 mm, a 95% LoA of −2.34 to 2.58 mm, and 98.6% of cases were within interobserver variability. These findings suggest that although coracoid segmentation remained challenging, its influence on CHD measurement was limited in the present validation cohort. At the measurement level, the automated method demonstrated a high level of agreement with observer-averaged measurements, with ICCs ranging from 0.964 to 0.994 across all five shoulder balance parameters. Bland–Altman analysis revealed small mean biases, including 0.09° for the CA, 0.12 mm for the CHD, 0.21° for the CTAD, −0.18 mm for the RSH, and −0.07° for the T1TA. Moreover, the automated–observer variability was lower than the corresponding interobserver variability for all parameters, and 90.0% to 98.6% of automated measurements were within the range of interobserver variability. These findings suggest that the proposed method achieved not only reliable anatomical segmentation but also measurement accuracy comparable to manual assessment by experienced observers.

The precise segmentation of shoulder structures has the potential for valuable applications in clinical tasks such as fracture diagnosis, joint assessment, and surgical planning. The Japanese Society of Radiological

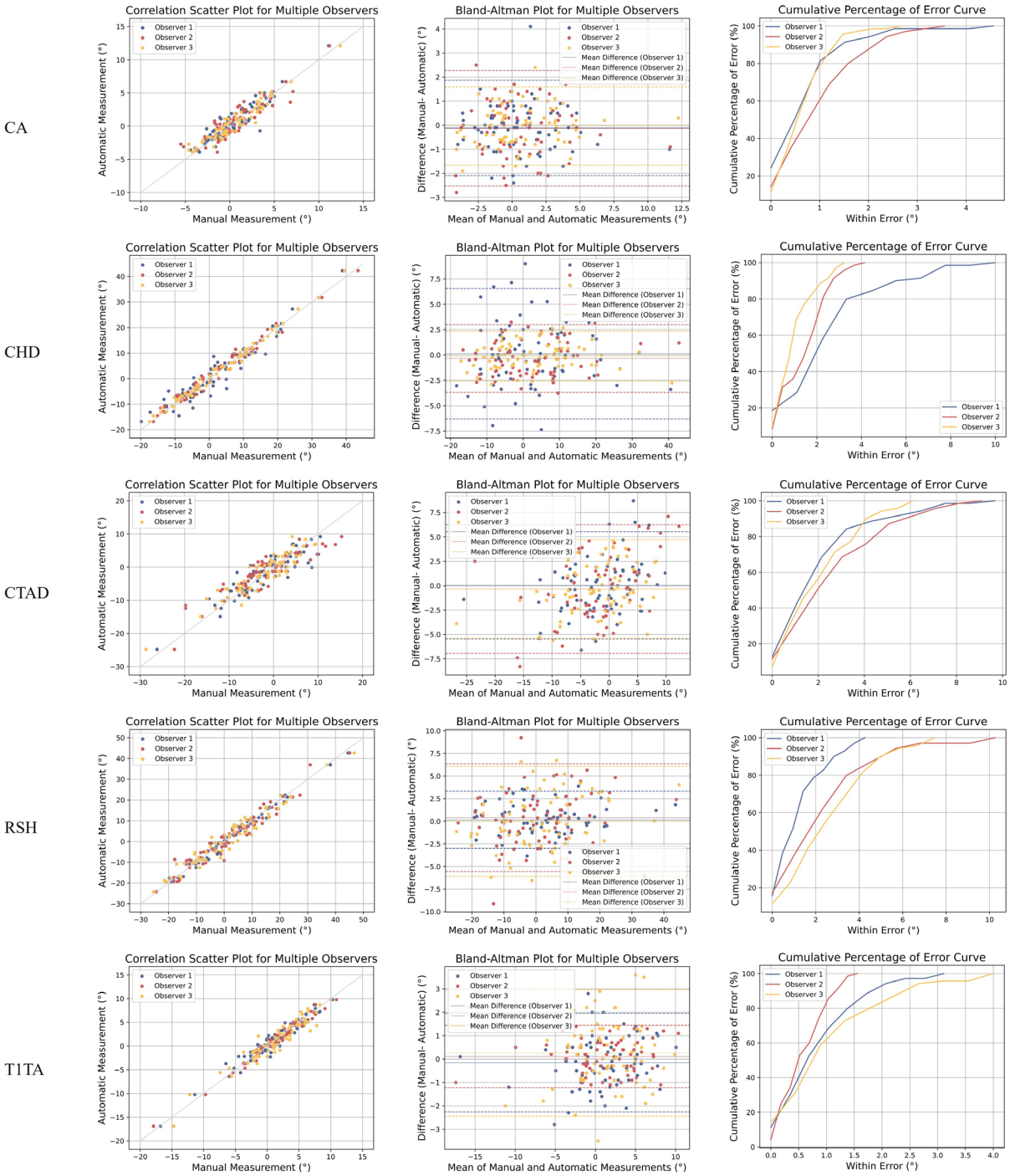

Figure 3. Correlation scatter diagrams (Left), Bland–Altman difference diagrams (Middle) and Cumulative percentage of error curves (Right) for CA, CHD, CTAD, RSH and T1TA measured automatically and by each observer.

Technology (JSRT) dataset (*21*), which includes lung contours, heart contours, and clavicle annotations within the lung fields, has been widely used in studies of clavicle segmentation. van Ginneken *et al*. (*22*) conducted a comparative study that used three supervised segmentation methods on the JSRT dataset: active shape models, active appearance models, and a multi-resolution pixel classification method utilizing Gaussian derivative filters and k-nearest neighbor classification. That method achieved an IoU of 0.736, compared to 0.896 for a human observer. Fully convolutional network (FCN)-based segmentation subsequently improved the reported IoU to 0.868 (*23*). A semi-supervised approach further improved the reported IoU to 0.881 (*24*). Wang *et al*. (*25*) introduced a multi-object segmentation method based on collaborative learning with multiple teacher models. Their model was trained using four heterogeneous partially labeled datasets that included the RCS-CXR dataset, which contains complete clavicle annotations as well as anterior and

posterior rib segmentation labels (*26*). The achieved results, with a Dice coefficient of 0.95 and an IoU of 0.91, outperformed the current state-of-the-art reported in the literature. In comparison, our model achieved Dice coefficients of 0.94 and 0.91–0.92 and IoUs of 0.88–0.89 and 0.84–0.85 for clavicle segmentation in the internal and external validation datasets, respectively.

Moreover, automated measurement of shoulder balance parameters can enable the screening of large cohorts in a reasonable timeframe with good reliability. The T1TA and RSH demonstrated better reader-agreement in previous studies (*9*), while a considerable variation in the RSH and a reduced variation in the T1TA were evident in our automated measurements. Previous studies reported that the CA and CHD displayed a high level of reliability among observers, with MAE values comparable to those observed in our study (*13*). Previous studies have indicated that shoulder balance should be considered a crucial factor in surgical planning and prognosis. The RSH was significantly correlated with the occurrence of the adding-on phenomenon in the shoulder imbalance group at follow-up (*1*). Moreover, there is a significant correlation between the T1TA just after surgery as well as the CA and the recurrence of shoulder imbalance during the 1-year follow-up in AIS patients. However, there is currently a lack of widely accepted diagnostic and assessment standards for shoulder balance parameters (*27*), and more population-based diagnostic studies need to be conducted to elucidate those parameters.

Several biases and limitations may have influenced the interpretation of our findings. Our dataset was obtained from a single institution, which may not represent the broader population of AIS patients. The data may be subject to selection bias based on the patient demographics, geographical location, or referral patterns. The performance of the deep learning algorithm can be influenced by factors such as image quality, patient positioning, and image artifacts. Additionally, when considering shoulder balance assessment, the positioning of patients during X-ray imaging is of paramount importance (*28*). These methods may yield measurements that do not fully reflect the true shoulder balance, as they are based on radiographic measurements rather than actual photographic measurements (*29*). Accurate measurement of X-ray parameters must be complemented by the identification of anatomical landmarks on the patient's body surface and an overall visual assessment. These factors collectively contribute to the comprehensive evaluation of shoulder balance, emphasizing the need for a holistic approach beyond just numerical measurements. In addition, the empirical acceptability thresholds used in this study were derived from interobserver variability among three senior spinal surgeons rather than from universally established clinical thresholds. Although this approach provides an observer-based reference for measurement acceptability, these thresholds should be further validated in larger multicenter cohorts. Future studies should also evaluate how automated measurement errors may affect shoulder balance classification and treatment decision-making.

In conclusion, our deep learning-based automated method provides a reliable and efficient approach for radiographic shoulder balance assessment in AIS patients. By reducing observer-dependent measurement variability and improving measurement reproducibility, this method may aid in clinical assessment, follow-up evaluation, and large-scale radiographic screening. Further multicenter validation is warranted to determine its impact on shoulder balance classification and treatment decision-making.

Funding: This study was funded by the National Key Research and Development Program of China (Grant No. 2025YFE0214102).

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Cao K, Watanabe K, Hosogane N, Toyama Y, Yonezawa I, Machida M, Yagi M, Kaneko S, Kawakami N, Tsuji T, Matsumoto M. Association of postoperative shoulder balance with adding-on in Lenke type II adolescent idiopathic scoliosis. Spine (Phila Pa 1976). 2014; 39:E705-E712.
2. Chang DG, Kim JH, Kim SS, Lim DJ, Ha KY, Suk SI. How to improve shoulder balance in the surgical correction of double thoracic adolescent idiopathic scoliosis. Spine (Phila Pa 1976). 2014; 39:E1359-E1367.
3. Kuklo TR, Lenke LG, Graham EJ, Won DS, Sweet FA, Blanke KM, Bridwell KH. Correlation of radiographic, clinical, and patient assessment of shoulder balance following fusion versus nonfusion of the proximal thoracic curve in adolescent idiopathic scoliosis. Spine (Phila Pa 1976). 2002; 27:2013-2020.
4. Elsebaie HB, Dannawi Z, Altaf F, Zaidan A, Al Mukhtar M, Shaw MJ, Gibson A, Noordeen H. Clinically orientated classification incorporating shoulder balance for the surgical treatment of adolescent idiopathic scoliosis. Eur Spine J. 2016; 25:430-437.
5. Hong JY, Suh SW, Modi HN, Yang JH, Park SY. Analysis of factors that affect shoulder balance after correction surgery in scoliosis: A global analysis of all the curvature types. Eur Spine J. 2013; 22:1273-1285.
6. Uzümcügil O, Atici Y, Ozturkmen Y, Yalcinkaya M, Caniklioglu M. Evaluation of shoulder balance through growing rod intervention for early-onset scoliosis. J Spinal Disord Tech. 2012; 25:391-400.
7. Kurra S, Cahill PJ, Albanese SA, Betz RR, Toole T, Lavelle WF. Evaluation of shoulder balance in early onset scoliosis after definitive fusion and comparison with adolescent idiopathic scoliosis shoulder balance. Spine Deform. 2022; 10:183-188.
8. Gotfryd AO, Silber Caffaro MF, Meves R, Avanzi O. Predictors for postoperative shoulder balance in Lenke 1 adolescent idiopathic scoliosis: A prospective cohort

study. Spine Deform. 2017; 5:66-71.
9. Bagó J, Carrera L, March B, Villanueva C. Four radiological measures to estimate shoulder balance in scoliosis. J Pediatr Orthop B. 1996; 5:31-34.
10. Chiu CK, Chan CYW, Tan PH, Goh SH, Ng SJ, Chian XH, Ng YH, Ler XY, Chandren JR, Chung WH, Kwan MK. Conformity and changes in the radiological neck and shoulder balance parameters throughout 3-year follow-up period: Do they remain the same? Spine (Phila Pa 1976). 2020; 45:E319-E328.
11. Luhmann SJ, Sucato DJ, Johnston CE, Richards BS, Karol LA. Radiographic assessment of shoulder position in 619 idiopathic scoliosis patients: Can T1 tilt be used as an intraoperative proxy to determine postoperative shoulder balance? J Pediatr Orthop. 2016; 36:691-694.
12. Akel I, Pekmezci M, Hayran M, Genc Y, Kocak O, Derman O, Erdogan I, Yazici M. Evaluation of shoulder balance in the normal adolescent population and its correlation with radiological parameters. Eur Spine J. 2008; 17:348-354.
13. Hong JY, Suh SW, Yang JH, Park SY, Han JH. Reliability analysis of shoulder balance measures: Comparison of the 4 available methods. Spine (Phila Pa 1976). 2013; 38:E1684-E1690.
14. Meng N, Cheung JPY, Wong KYK, Dokos S, Li S, Choy RW, To S, Li RJ, Zhang T. An artificial intelligence powered platform for auto-analyses of spine alignment irrespective of image quality with prospective validation. EClinicalMedicine. 2022; 43:101252.
15. Wang L, Xie C, Lin Y, *et al*. Evaluation and comparison of accurate automated spinal curvature estimation algorithms with spinal anterior-posterior X-Ray images: The AASCE2019 challenge. Med Image Anal. 2021; 72:102115.
16. Zhang K, Xu N, Guo C, Wu J. MPF-net: An effective framework for automated Cobb angle estimation. Med Image Anal. 2022; 75:102277.
17. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J. RepVGG: Making VGG-style ConvNets great again. CVPR 2021. 2021; 13733-13742.
18. Mun C, Lee S, Uh Y, Choe J, Byun H. Small objects matters in weakly-supervised semantic segmentation. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, Waikoloa, HI, USA, 2024; 414-423.
19. Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. BMC Res Notes. 2022; 15:210.
20. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging. 2015; 15:29.
21. Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, Matsui M, Fujita H, Kodera Y, Doi K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR Am J Roentgenol. 2000; 174:71-74.
22. van Ginneken B, Stegmann MB, Loog M. Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database. Med Image Anal. 2006; 10:19-40.
23. Novikov AA, Lenis D, Major D, Hladuvka J, Wimmer M, Buhler K. Fully convolutional architectures for multiclass segmentation in chest radiographs. IEEE Trans Med Imaging. 2018; 37:1865-1876.
24. Bortsova G, Dubost F, Hogeweg L, Katramados I, de Bruijne M. Semi-supervised medical image segmentation *via* learning consistency under transformations. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, Khan A, eds.). Springer International Publishing, Cham, 2019; pp. 810-818.
25. Wang H, Zhang D, Feng J, Cascone L, Nappi M, Wan S. A multi-objective segmentation method for chest X-rays based on collaborative learning from multiple partially annotated datasets. Information Fusion. 2023; 102:102016.
26. Wang W, Feng H, Bu Q, Cui L, Xie Y, Zhang A, Feng J, Zhu Z, Chen Z. MDU-Net: A convolutional network for clavicle and rib segmentation from a chest radiograph. J Healthc Eng. 2020; 2020:2785464.
27. Clement RC, Anari J, Bartley CE, Bastrom TP, Shah R, Talwar D, Upasani VV. What are normal radiographic spine and shoulder balance parameters among adolescent patients? Spine Deform. 2020; 8:621-627.
28. Marks MC, Stanford CF, Mahar AT, Newton PO. Standing lateral radiographic positioning does not represent customary standing balance. Spine (Phila Pa 1976). 2003; 28:1176-1182.
29. Qiu X, Ma W, Li W, Wang B, Yu Y, Zhu Z, Qian B, Zhu F, Sun X, Ng BKW, Cheng JCY, Qiu Y. Discrepancy between radiographic shoulder balance and cosmetic shoulder balance in adolescent idiopathic scoliosis patients with double thoracic curve. Eur Spine J. 2009; 18:45-51.

Received May 15, 2026; Revised June 11, 2026; Accepted June 15, 2026.

Released online in J-STAGE as advance publication June 20, 2026.

[§]*These authors contributed equally to this work.*
**Address correspondence to*:
Yan Yu, Division of Spine, Department of Orthopaedics, Tongji Hospital, Tongji University School of Medicine, Tongji University, 389 Xincun Road, Putuo District, Shanghai 200065, China.
E-mail: yyu15@tongji.edu.cn