Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Two-Phase Transfer Learning for Automated Measurement and Severity Grading of Scoliosis and Spondylolisthesis on Spinal Radiographs

kumari Bhawana

phdcs10002.20@bitmesra.ac.in

Birla Institute of Technology

I. Mukherjee

Birla Institute of Technology, Mesra

Amritanjali .

Birla Institute of Technology, Mesra

Research Article

Keywords: Scoliosis, Spondylolisthesis, Cobb angle, Meyerding grade, Deep learning, Radiography, Transfer learning, Single-center study

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11118811/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Two-Phase Transfer Learning for Automated Measurement and Severity Grading of Scoliosis and Spondylolisthesis on Spinal Radiographs

kumari Bhawana[1*], I. Mukherjee[2] and Amritanjali[3]

[123*]Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi, 835215, Jharkhand, India.

*Corresponding author(s). E-mail(s): phdcs10002.20@bitmesra.ac.in;

Abstract

Background: Cobb angle and vertebral slip measurements are clinically validated techniques that are time-intensive and prone to inter-observer variation.
Objective: To develop a comprehensive deep learning approach for automated measurements and grading of scoliosis and spondylolisthesis cases using a single-center hospital radiograph dataset.
Methods: In this retrospective study, 2,500 anonymized spinal radiographs from one participating hospital were used, including 1,635 scoliosis and 865 spondylolisthesis radiographs. Researchers had access only to radiographs without any patient identifiers, demographic information, and clinical data. Radiographs were preprocessed by applying Gaussian denoising, CLAHE contrast adjustment, intensity standardization, and scaling to 224 x 224 pixels. Vertebral landmark-based vertebral angles were computed for scoliosis analysis, and vertebral body localization-based slip percentages were calculated for spondylolisthesis detection. Two networks, DenseNet201, and ResNet152, were fine-tuned with two-stage ImageNet transfer learning approach.
Results: DenseNet201 obtained an accuracy of 87.5% for scoliosis and 88.4% for spondylolisthesis on hold-out test subsets. Cobb-angle MAE was 1.845° and slip percentage MAE was 1.8 percentage point. The corresponding ICC values were 0.962 and 0.958, and macro-AUC values were 0.83 and 0.84.
Conclusion: The proposed method demonstrated strong internal validity in terms of spinal deformity measurement and grading within the context of a single-center study. Since patient matching was not possible, patient-level independence of the image split could not be confirmed. These results, therefore, constitute internal validation at the level of the image and need to be externally validated independently.

Keywords: Scoliosis; Spondylolisthesis; Cobb angle; Meyerding grade; Deep learning; Radiography; Transfer learning; Single-center study

1 Introduction

Scoliosis and spondylolisthesis are among the most common structural spinal conditions, where the radiographic assessment of disease severity determines diagnosis, follow-up and management strategies. Scoliosis is usually assessed in terms of Cobb angle, while spondylolisthesis severity is evaluated using the vertebral slip percentage under Meyerding classification [7, 13, 17, 32].

Manually measured Cobb angle and slip percentage involve the identification of bony landmarks and geometric calculations. Although a clinically accepted method, it is rather laborious and associated with a considerable inter- and intra-observer variation, sometimes resulting in severe inconsistencies that could lead to reclassification of severity grades at decision boundaries [15, 17, 30, 35].

Deep learning holds the key to the reproducibility of image-based measurements and grading. Landmark detection, segmentation, and transfer learning using CNNs have shown promising potential for spinal radiography. However, clinical implementation is hindered by heterogeneous data sets, inconsistent validation schemes, constrained external testing, and insufficient incorporation of measurement along with severity grading [9, 19, 43].

There are three reasons why the current study is important. Firstly, many of the techniques deal with either scoliosis or spondylolisthesis. Secondly, architecture comparison is done on various datasets instead of a uniform protocol. Thirdly, single-center studies may offer valuable internal evidence but cannot serve as proof of cross-facility generalizability [2, 7, 22].

Therefore, this investigation will consider a multi-stage workflow in one dataset of hospital radiographs. The purpose is to assess feasibility as well as clear validation, which is composed of condition-specific measurement, operational grading of severity, and a comparison between DenseNet201 and ResNet152 through two-stage ImageNet transfer-learning approach.

Based on the aforementioned research gaps, the current study has the following research aims:

- Derive an end-to-end process for measuring and assessing the severity of scoliosis and spondylolisthesis via radiographic images of the spine.
- Measure the regional Cobb angle for scoliosis and vertebral slip percentage for spondylolisthesis through localization procedures specific to the conditions.
- Conduct a direct comparison between DenseNet201 and ResNet152 with same conditions of preprocessing, training, and evaluation at an image level.
- Analyze the performance of two-stage transfer learning procedure using ImageNet with augmentation and only class-balancing during training.

Contributions of this paper include the following:

- A dual-condition multi-stage method for localizing vertebrae, measuring features, grading severity, and classifying images.
- Controlled comparison of DenseNet201 and ResNet152 models trained on the same single-center dataset using the same training process.
- Two stages of transfer learning with condition-specific augmentations and clear management of class imbalance.
- Internal testing on 2,500 radiographs from a single center using multiple metrics and recognizing lack of patient matching and reference standard metadata.

2 Related Work

Automated assessment of spinal deformities can be classified as automated Cobb angle measurement, spondylolisthesis measurement, vertebral localization, classification of severity, transfer learning, and clinical validation studies. The literature is replete with positive results from an engineering standpoint but insists on validation and standardization of reference measurements.

2.1 Systematic Reviews and Meta-Analyses

As recent review and meta-analysis studies have reported, the accuracy of deep-learning techniques for Cobb angle evaluation has proven clinically relevant; however, heterogeneity of datasets, lack of external validation, and reporting inconsistency pose some ongoing hurdles for translation [9, 19, 41, 43].

2.2 Automated Cobb Angle Measurement for Scoliosis

Cobb angle assessment using automatic systems has evolved from early CNN-based measurements to vertebra detection, landmarks, and multicenter methods. Typical examples provide MAE scores between 2.5° and 4.5° degrees and demonstrate improved reader agreement. The accuracy of these methods heavily depends on localization and study setup [2, 11, 13, 28, 34].

2.3 Automated Slip Percentage Measurement for Spondylolisthesis

There have been fewer studies conducted on automated spondylolisthesis detection. Transfer learning models and vertebral measurement techniques have demonstrated that it is possible to detect slip characteristics from x-ray images, but the studies are fewer and external validations are still lacking [6, 7, 32].

2.4 Vertebral Landmark Detection and Segmentation

Accurate measurement relies on correct vertebral identification. Recent algorithms have used anatomy-based approaches for accurate vertebral localization in cases where there is distortion and overlapping of vertebrae [3, 4, 20, 21, 42, 44].

2.5 Classification and Severity Grading Approaches

Apart from measurement-based methods, transfer learning, ensemble, transformer, and multistage approaches have also been explored for scoliosis severity grading and curve classification tasks. Such research shows the applicability of automated scoliosis grading while also highlighting the trade-off between complexity and clinical interpretability of deep neural network architectures [1, 25, 31, 38, 39].

2.6 Transfer Learning and Validation Design

Transfer learning applies in the context of spine radiography due to the small size of the annotated medical datasets compared to natural images dataset. The CNN trained on ImageNet can be used by initially training a head and finetuning some layers of the backbone network, but performance should be validated in the source domain and not just rely on the pretrained model [6, 7, 32].

2.7 Clinical Validation and Implementation Studies

Comparative validation studies between automated and manual measurement show high agreement and greater reproducibility, but stress the significance of having independent readers, separation at a patient level, and external validations prior to implementation [15, 26, 30, 33, 35].

2.8 Comparative Summary of Existing Methods

Table 1 summarizes representative studies and places the present single-center workflow in context.

Table 1 Summary of existing methods

Study	Condition	Method	Dataset	Key Metrics	Limitations
Horng et al. (2019)	Scoliosis	CNN-based Cobb angle measurement	Single private	MAE: 4.5°	Limited dataset; no external validation; early architecture
Rahmaniar et al. (2023)	Scoliosis	Vertebra detection-based Auto-CA	Single dataset	MAE: 3.2°	Single-center data; limited clinical validation
Wu et al. (2025)	Scoliosis	Multi-module full-spine DR framework	Large-scale	MAE: 2.8°	Limited multi-center validation; complex pipeline
Hayashi et al. (2025)	Scoliosis	Deep learning algorithm validation	Multiple datasets	High accuracy confirmed	Limited architectural comparison
Casabo´-Vall´es et al. (2026)	Scoliosis	Fully autonomous multi-center AI	Multi-center	ICC > 0.95	Computational resource intensive
Kato et al. (2024)	Scoliosis	Comparison of three AI algorithms	Multi-disease	Variable performance	Limited generalizability analysis
Gadu & Potala (2026)	Scoliosis	SCANet specialized architecture	AIS dataset	MAE: 2.5–3.5°	Single architecture focus
Lu et al. (2024)	Scoliosis	Center-point AI model	Scoliosis dataset	Superior to VFLDNet	Limited comparison to manual measurement
Zhang et al. (2023)	Scoliosis	Smartphone-based classification	Smartphone images	Classification accuracy reported	Limited precision vs radiography
Yesilmen et al. (2026)	Scoliosis	Vision transformer with attention	Scoliosis dataset	Superior to CNN	Limited clinical validation
Xu et al. (2025)	Scoliosis	Multistage Lenke classification	AIS dataset	Enhanced diagnostic precision	Complex multi-stage pipeline
Tavana et al. (2023)	Scoliosis	Ensemble transfer learning	Scoliosis dataset	High classification accuracy	Limited measurement capability
Fraiwan et al. (2022)	Both	Deep transfer learning	Public + Private	Detection accuracy ∼ 85%	Limited slip % accuracy
Fraiwan & Abukhalaff (2026)	Scoliosis	Dual-source pretrained CNNs	Public + Private	Robust performance	Limited severity grading integration
Wang et al. (2025)	Spinal	Deep learning sagittal parameters	Cervical X-ray	Measurement accuracy	Single-center validation
Ours	**Both**	Integrated multi-stage framework	Single-center hospital	Cobb MAE: 1.845°; slip MAE: 1.8 pp; ICC > 0.95	Internal image-level validation; no external cohort

2.9 Limitations of Existing Research

These gaps in the literature include limited external validation, incomplete integration between quantification and severity grading, inconsistent definition of reference standard, and challenges in comparing architecture across heterogeneous data sets and split designs [9, 19, 43]. Multi-center studies contribute to more diversity but suffer from domain shift, whereas single-center studies demand careful wording about generalizability. This paper thus addresses a more specific question: whether a single workflow can be used for condition-specific measurement and severity grading of scoliosis and spondylolisthesis based on one hospital data set. It poses itself as methodological internal validation, not autonomous clinical decision-making or cross-center validation. The workflow includes localization, quantitative measurement, operational grade, and image-level classification. Two-phase transfer learning and condition-aware augmentation are considered in the context of this data set as engineering decisions, while patient-level validation in external settings is the necessary next step.

3 Methodology

3.1 System Overview

The workflow comprises image preprocessing, condition-specific vertebral localization and quantitative measurement, rule-based severity grading from the measured value, and image-level classification using DenseNet201 or ResNet152 (Figure 1). These modules form an integrated pipeline but are not a single jointly optimized diagnostic network.

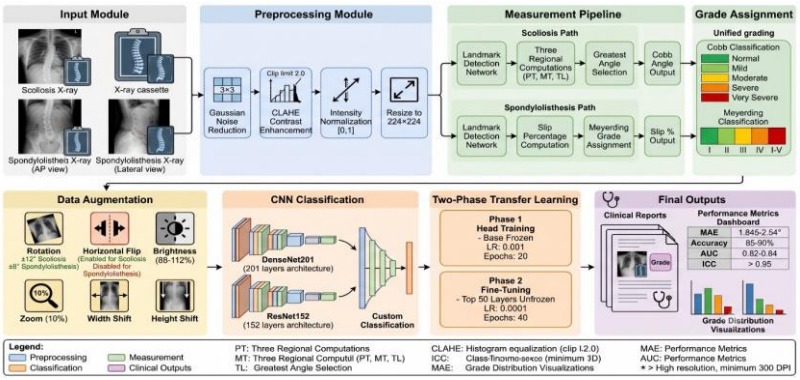


Fig. 1 Integrated system architecture for single-center spinal radiograph analysis.

3.2 Single-Center Clinical Dataset and Study Design

This is a retrospective single-center study that consists of 2,500 anonymized radiographs of spine acquired from one hospital participating in a PhD research project: 1,635 radiographs for scoliosis and 865 radiographs for spondylolisthesis. No public or Kaggle images were used. The distribution of classes is presented in Table 2.

Only image data and corresponding annotations to the studies were provided. There were no identifiers of patients, demographics/clinical data, acquisition information, and longitudinal connections available; hence, unique patient count and repeat studies are unknown, and the given split constitutes image-based internal validation but not independent patient-based validation. The annotation provided included landmarks and region-based Cobb angles for scoliosis and body vertebrae boundaries, translations, and slip percentage for spondylolisthesis. The number, speciality, experience, blinding, and adjudication process of annotators were not provided and should be disclosed before clinical publication. Minority classes were rare, especially extremely severe scoliosis and Grade V spondylolisthesis; class imbalance is taken into account explicitly in training and grading results interpretation.

Table 2. Single-center hospital dataset composition by condition and severity category

Condition	Severity Category	Number of Radiographs	% of Full Dataset	% Within Condition	Source
Scoliosis	Normal (< 10°)	425	17.0%	26.0%	Single center
	Mild (10°–25°)	570	22.8%	34.9%	Single center
	Moderate (25°–40°)	355	14.2%	21.7%	Single center
	Severe (40°–60°)	205	8.2%	12.5%	Single center
	Very Severe (> 60°)	80	3.2%	4.9%	Single center
	Subtotal	**1,635**	**65.4%**	**100.0%**	**Single center**
Spondylolisthesis	Grade I (0–25%)	320	12.8%	37.0%	Single center
	Grade II (25–50%)	290	11.6%	33.5%	Single center
	Grade III (50–75%)	155	6.2%	17.9%	Single center
	Grade IV (75–100%)	75	3.0%	8.7%	Single center
	Grade V (> 100%)	25	1.0%	2.9%	Single center
	Subtotal	**865**	**34.6%**	**100.0%**	**Single center**
Grand Total	**All categories**	**2,500**	**100.0%**	–	**Single center**

No information about ethics approval/waiver or consent status is provided in the study pack and should be verified with the data provider prior to submission. Investigators obtained de-identified images without any effort to re-identify the patients. The preprocessing stage consisted of Gaussian denoising 3 × 3 kernel, σ = 1.0), CLAHE (clip limit 2.0; 8 × 8 tiles), min-max normalization of intensities, bi-linear resizing to 224 × 224 pixels, and RGB transformation. The same procedure was performed on all images of the single-center dataset.

3.3 Cobb Angle Measurement Pipeline for Scoliosis

In case of scoliosis, vertebral endplate corners on anteroposterior radiographs are detected using a landmark detector. The network is an encoder-decoder framework with skip connections, which outputs landmark heat maps from which the orientation of the superior and inferior endplates is computed. According to the cited source, the PCK is around 85-90%.

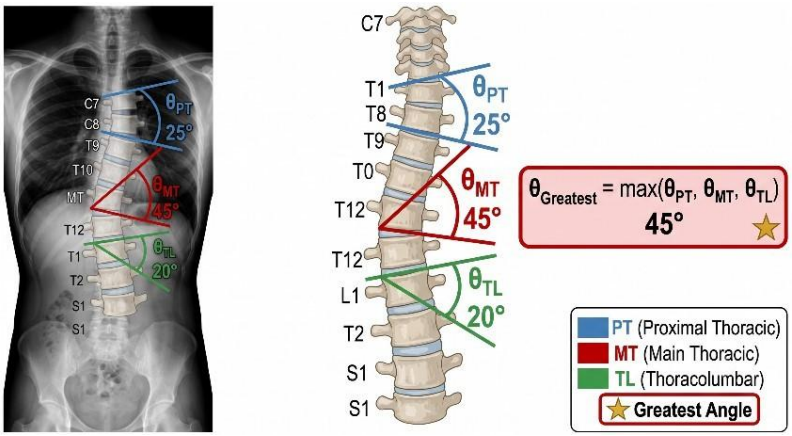


Fig. 2 Cobb Angle Measurement Illustration – Three Regional Curves

The Cobb angle is determined for the proximal thoracic, main thoracic, and thoracolumbar segments based on the angle formed by the selected superior and inferior endplates (see Figure 2). The maximum regional angle is utilized in determining the severity grading [19, 30, 43].

For the calculation of the direction vector, the endplate is specified using two corner points, as per Equation 1.

$$\vec{d} = \frac{(x_2 - x_1,\ y_2 - y_1)}{\|(x_2 - x_1,\ y_2 - y_1)\|} \tag{1}$$

The regional Cobb angle is then obtained from the angle between the selected superior and inferior endplate direction vectors (Equation 2). The maximum of the three regional angles defines the study severity value (Equation 3).

$$\theta_{Cobb} = \arccos\left(\frac{\vec{d}_1 \cdot \vec{d}_2}{\|\vec{d}_1\| \|\vec{d}_2\|}\right) \quad (2)$$

$$\theta_{greatest} = \max(\theta_{PT}, \theta_{MT}, \theta_{TL}) \quad (3)$$

In the source code, there was a backup Cobb angle set to 20° in case of failed localization. In order to prevent misclassification as well as to reduce measurement error, this technique is seen as a drawback, and an operational system needs to identify ungradable exams.

3.4 Slip Percentage Measurement Pipeline for Spondylolisthesis

In case of spondylolisthesis, the vertebral body boundaries are extracted using an image segmentation method using a U-Net-type model with ResNet encoder from lateral X-ray images. The slip percentage is calculated by dividing anterior slip distance by the anteroposterior diameter of the lower vertebra, and multiplying the result by 100 (Equation 4).

$$Slip\ Percentage = \left(\frac{Anterior\ Translation}{Vertebral\ Body\ Width}\right) \times 100 \quad (4)$$

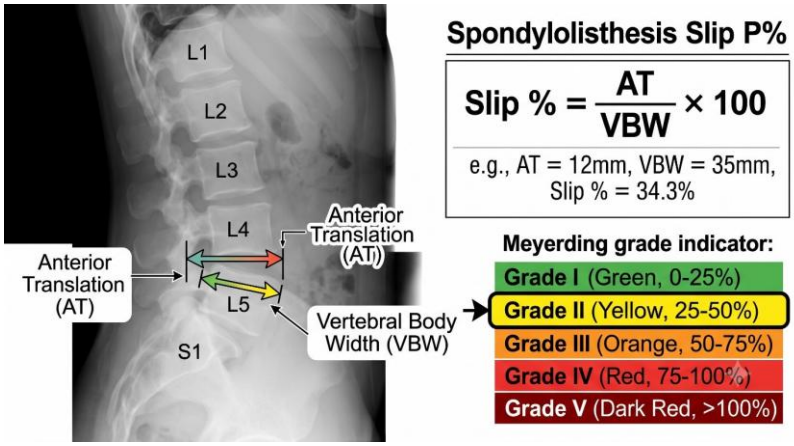


Fig. 3 Slip Percentage Measurement Illustration – Meyerding Classification

The calculated slip percentage is then categorized based on the criteria shown in Table 3 into a Meyerding grade. In the absence of information such as symptoms, instability, age, etiology, among others, the grade is provided for description of severity only.

Table 3 Meyerding grading thresholds used for spondylolisthesis severity description

Grade	Slip Percentage Range	Severity Category	Interpretation in this study
Grade I	0–25%	Mild	Severity description only; clinical management requires additional information.
Grade II	25–50%	Moderate	Severity description only; clinical management requires additional information.
Grade III	50–75%	Severe	Severity description only; clinical management requires additional information.
Grade IV	75–100%	Very Severe	Severity description only; clinical management requires additional information.
Grade V	> 100%	Spondyloptosis	Spondyloptosis category; clinical management not inferred from image grade alone.

3.5 CNN Architectures

DenseNet201 and ResNet152 were benchmarked through ImageNet initialization. DenseNet201 is a network that incorporates dense features and contains 20 million parameters, while ResNet152 consists of residual bottleneck layers containing 60 million parameters [12, 14]. The two networks employ a similar classification head, which was designed for an unbiased architectural comparison. The classification head consists of global average pooling with dense layers and batch normalization (Figure 4).

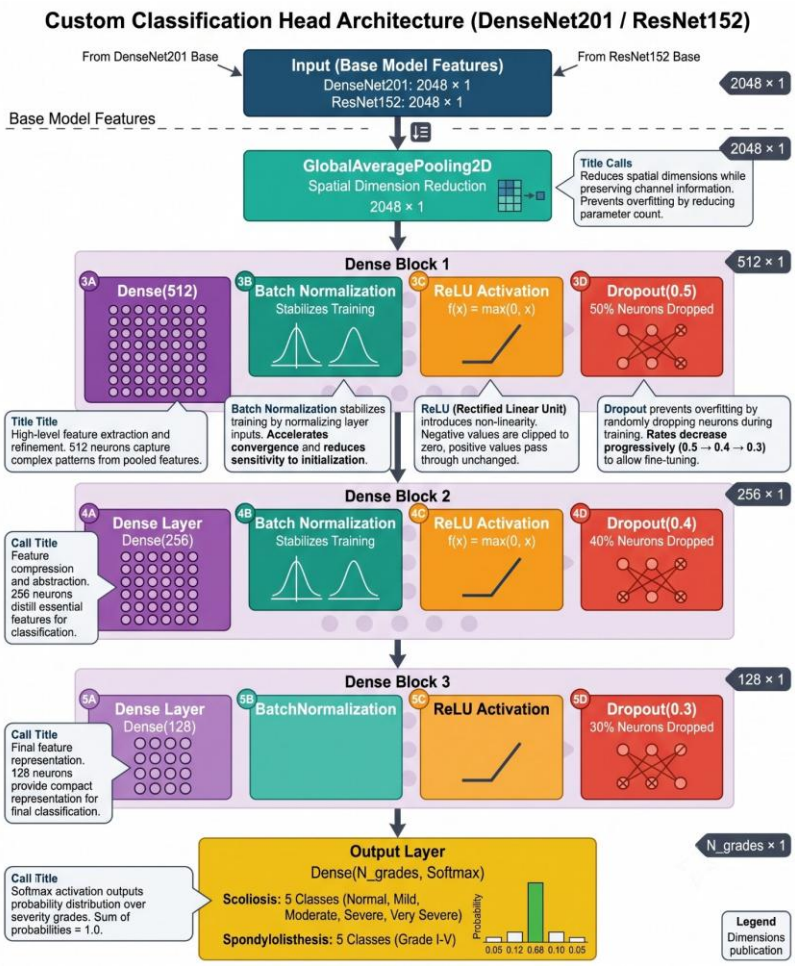


Fig. 4 Custom Classification Head Architecture Diagram

The softmax layer produces probabilities over the five severity categories for each condition.

3.6 Two-Phase Transfer Learning

Training consisted of two stages. In Stage 1, the backbone network was frozen, and only the classification head was fine-tuned, while in Stage 2, the top 50 backbone layers were unfrozen and fine-tuned using ten times lower learning rate. Tables 4 and 5 provide the configuration details. The purpose of this multi-stage strategy was to maintain the generic pretrained features in the first stage and specialize in radiographic features in the second stage.

Categorical cross-entropy was used in both phases; Equations 6–9 summarize the optimization formulation.

Table 4 Phase 1 parameters

Parameter	Value
Base model	Frozen (trainable = False)
Trainable layers	Classification head only
Learning rate	0.001
Max epochs	20
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Loss function	Categorical crossentropy

Table 5 Phase 2 parameters

Parameter	Value
Base model	Top 50 layers unfrozen
Trainable layers	Top 50 layers + classification head
Learning rate	0.0001 (10× reduction from Phase 1)
Max epochs	40
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Loss function	Categorical crossentropy

$$L_{phase1} = -\sum_{c=1}^{N} y_c \log(\hat{y}_c) \tag{5}$$

$$\theta^*_{head} = \arg\min_{\vartheta_{head}} L_{phase1}\left(\theta^{frozen}_{base}, \theta_{head}\right) \tag{6}$$

$$L_{phase2} = -\sum_{c=1}^{N} y_c \log(\hat{y}_c) \tag{7}$$

$$\theta^*_{top50}, \theta^*_{head} = \arg\min_{\vartheta_{top\,50}, \vartheta_{head}} L_{phase2}\left(\theta_{base}, \theta_{head}\right) \tag{8}$$

The ablations compared conventional transfer learning against the staged protocol. As the ablations were conducted using the held-out test set, their findings can be considered to be exploratory in nature rather than confirmatory.

3.7 Data Augmentation Strategy

Condition-dependent enhancement was more evident during the head phase and less so during the fine-tuning phase (Tables 6 and 7). Horizontal flipping was available for scoliosis severity classification but not for spondylolisthesis due to preservation of slip direction [7, 19, 30].

Table 6 Phase 1 augmentation parameters

Parameter	Scoliosis	Spondylolisthesis
Rotation	±12°	±8°
Width shift	10%	8%
Height shift	5%	–
Zoom	10%	8%
Horizontal flip	Enabled	Disabled
Brightness	88–112%	88–112%

Table 7 Phase 2 augmentation parameters

Parameter	Scoliosis	Spondylolisthesis
Rotation	±5°	±4°
Width shift	4%	3%
Height shift	–	–
Zoom	4%	3%
Horizontal flip	Enabled	Disabled
Brightness	–	–

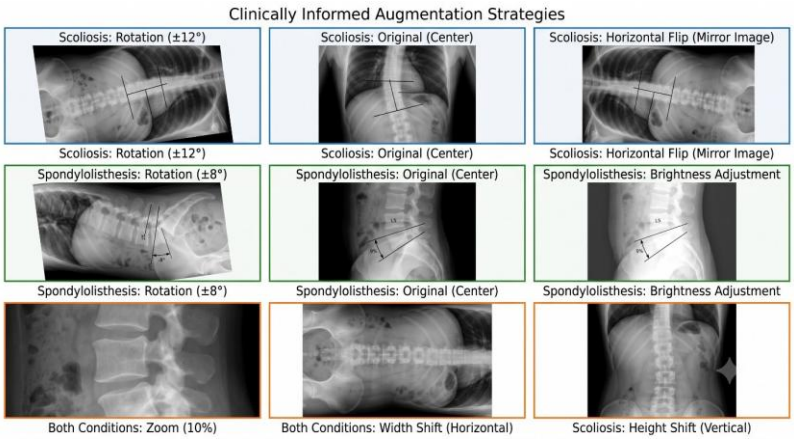


Fig. 5 Data Augmentation Examples for Both Conditions

3.8 Class Balancing

Class balancing was applied only to the training subset through augmentation-based oversampling, with emphasis on severe/very severe scoliosis and Grades III–V spondylolisthesis. Validation and test distributions were left unchanged, and augmented variants were not counted as independent clinical observations.

3.9 Training Protocol

The 2,500 images were then stratified into 1,750 for training, 250 for validation, and 500 internal hold-out test images (Table 8). The selection of the model, early stopping, and reducing learning rate were all based on validation loss. The lack of patient linkage made separation possible only through images.

Table 8 Corrected image-level training, validation, and internal test distribution

Split	Percentage	Scoliosis Samples	Spondylolisthesis Samples	Total
Training	70.0%	1,145	605	1,750
Validation	10.0%	163	87	250
Internal test	20.0%	327	173	500
Total	**100%**	**1,635**	**865**	**2,500**

3.10 Evaluation and Statistical Analysis

The performance of the model was evaluated using MAE in the case of continuous measurement errors, accuracy, precision, recall, F1 score, and specificity in the case of severity classification, macro AUC-ROC for discrimination, and ICC with 95% confidence intervals in agreement with the provided reference values. As far as DenseNet201 is concerned, Wilson 95% confidence intervals were calculated for the two major accuracies: 87.5% for scoliosis (83.4-90.6%) and 88.4% for spondylolisthesis (82.8-92.4%).

4 Experimental Results

4.1 Dataset Statistics and Class Distribution

The dataset comprised 1,635 scoliosis and 865 spondylolisthesis radiographs. Table 9 summarizes class distribution and training-only balancing; rare categories were very severe scoliosis (80 images) and Grade V spondylolisthesis (25 images).

Table 9 Class distribution and training-only balancing policy

Condition	Grade	Full-dataset N	% within condition	Training-set handling
Scoliosis	Normal	425	26.0%	Standard training augmentation
Scoliosis	Mild	570	34.9%	Standard training augmentation
Scoliosis	Moderate	355	21.7%	Standard training augmentation
Scoliosis	Severe	205	12.5%	Training-only targeted oversampling
Scoliosis	Very Severe	80	4.9%	Training-only targeted oversampling
Spondylolisthesis	Grade I	320	37.0%	Standard training augmentation
Spondylolisthesis	Grade II	290	33.5%	Standard training augmentation
Spondylolisthesis	Grade III	155	17.9%	Training-only targeted oversampling
Spondylolisthesis	Grade IV	75	8.7%	Training-only targeted oversampling
Spondylolisthesis	Grade V	25	2.9%	Training-only targeted oversampling

4.2 Training Performance Analysis

Both architectures converged under the two-phase protocol with early stopping and learning-rate reduction based on validation loss. Optimization behavior is descriptive and does not exclude overfitting because patient-level split independence could not be verified.

4.3 Quantitative Results for Scoliosis Assessment

Tables 10–12 summarize scoliosis classification, Cobb-angle error, and per-grade performance. DenseNet201 achieved 87.5% accuracy on 327 test images (95% CI 83.4–90.6%) and Cobb-angle MAE of 1.845°; the reported ensemble MAE was 1.650°.

Table 10 Overall performance metrics for scoliosis assessment

Metric	DenseNet201	ResNet152	Improvement (DenseNet vs. ResNet)
Accuracy	87.5%	86.5%	+1.0%
Weighted Precision	87.2%	85.9%	+1.3%
Weighted Recall	87.5%	86.5%	+1.0%
Weighted F1-Score	87.5%	86.2%	+1.3%
Macro Specificity	88.3%	87.1%	+1.2%
Macro AUC-ROC	0.83	0.82	+0.01

The source analysis reported 87% of Cobb-angle predictions within ±2° of the supplied reference measurement. Per-grade performance was lowest for the Severe

Table 11 Cobb angle measurement accuracy

Model	MAE (degrees)	Observed Range	Evaluation Scope
DenseNet201	1.845°	0.8°–3.2°	Internal source analysis
ResNet152	2.540°	1.2°–4.1°	Internal source analysis
Ensemble (DenseNet + ResNet)	1.650°	0.7°–2.9°	Internal source analysis

category; estimates for the Very Severe category should be interpreted cautiously because only 16 test images were available.

Table 12 Per-grade performance for scoliosis classification

Grade	Precision	Recall	F1-Score	Support (Test Set)
Normal (< 10°)	91.2%	89.8%	90.5%	85
Mild (10°–25°)	88.5%	87.9%	88.2%	114
Moderate (25°–40°)	85.3%	86.1%	85.7%	71
Severe (40°–60°)	82.7%	81.4%	82.0%	41
Very Severe (> 60°)	88.9%	91.3%	90.1%	16
Weighted Average	**87.2%**	**87.5%**	**87.5%**	**327**

4.4 Quantitative Results for Spondylolisthesis Assessment

Tables 13–15 summarize spondylolisthesis classification, slip-percentage error, and per-grade performance. DenseNet201 achieved 88.4% accuracy on 173 test images (95% CI 82.8–92.4%) with slip-percentage MAE of 1.8 percentage points. Grade IV had the lowest reported F1-score; Grade V estimates are highly uncertain because only five test images were available.

Table 13 Overall performance metrics for spondylolisthesis assessment

Metric	DenseNet201	ResNet152	Improvement (DenseNet vs. ResNet)
Accuracy	88.4%	86.7%	+1.7%
Weighted Precision	87.9%	86.1%	+1.8%
Weighted Recall	88.4%	86.7%	+1.7%
Weighted F1-Score	88.1%	86.4%	+1.7%
Macro Specificity	88.9%	87.3%	+1.6%
Macro AUC-ROC	0.84	0.83	+0.01

Table 14 Slip percentage measurement accuracy

Model	MAE (percentage points)	Observed Range	Evaluation Scope
DenseNet201	1.8	0.5%–3.5%	Internal source analysis
ResNet152	2.5	0.8%–4.2%	Internal source analysis
Ensemble	1.5	0.4%–3.0%	Internal source analysis

Table 15 Per-grade performance for spondylolisthesis classification

Grade	Precision	Recall	F1-Score	Support (Test Set)
Grade I (0–25%)	92.1%	90.6%	91.3%	64
Grade II (25–50%)	89.3%	88.1%	88.7%	58
Grade III (50–75%)	85.7%	87.1%	86.4%	31
Grade IV (75–100%)	83.3%	81.8%	82.5%	15
Grade V (> 100%)	90.0%	93.3%	91.6%	5
Weighted Average	**87.9%**	**88.4%**	**88.1%**	**173**

4.5 Comparison with Existing Methods

Performance was contextualized against representative prior studies summarized in Table 1. Because datasets, reference standards, and split protocols differ, the numerical comparisons are descriptive rather than head-to-head evidence of superiority.

4.6 Clinical Validation: Intraclass Correlation Coefficient (ICC)

Table 16 summarizes ICC values for both conditions.

Table 16 ICC analysis for both conditions

Condition	Model	ICC Value	95% Confidence Interval	Interpretation
Scoliosis	DenseNet201	0.962	0.951–0.972	Excellent
	ResNet152	0.951	0.938–0.963	Excellent
	Ensemble	0.974	0.965–0.982	Excellent
Spondylolisthesis	DenseNet201	0.958	0.945–0.969	Excellent
	ResNet152	0.947	0.932–0.960	Excellent
	Ensemble	0.971	0.961–0.979	Excellent

DenseNet201 ICC was 0.962 (95% CI 0.951–0.972) for scoliosis and 0.958 (0.945–0.969) for spondylolisthesis; ResNet152 and the ensemble also showed excellent internal agreement under Koo and Li (2016) guidance. These estimates indicate agreement with the supplied reference measurements within this dataset but should be interpreted alongside the unavailable annotator/adjudication details and single-center image-level validation design.

4.7 AUC-ROC Analysis

Macro AUC was 0.83 for DenseNet201 and 0.82 for ResNet152 in scoliosis, and 0.84 and 0.83, respectively, in spondylolisthesis; the ensemble values were 0.85 and 0.86. These results are internal discrimination estimates and are already included in the principal performance tables.

4.8 Ablation Studies

Exploratory ablations are reported in Tables 17 and 18. Because the source analysis evaluated alternative configurations on the held-out test subsets, the results may contain model-selection bias and are not treated as confirmatory component-level evidence.

Table 17 Ablation study results for scoliosis assessment

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC	Cobb MAE (°)
Baseline Models						
ResNet50 (no transfer)	68.2	67.5	68.2	67.8	0.65	4.72
ResNet50 + ImageNet transfer	74.5	73.8	74.5	74.1	0.71	3.85
Architecture Comparison						
ResNet152 + conventional transfer	82.3	81.6	82.3	81.9	0.78	3.12
DenseNet201 + conventional transfer	84.1	83.4	84.1	83.7	0.79	2.85
Proposed Components						
DenseNet201 + CLAHE	84.8	84.0	84.8	84.4	0.80	2.62
DenseNet201 + CLAHE + augmentation	86.0	85.3	86.0	85.6	0.81	2.38
DenseNet201 + CLAHE + augmentation + two-phase TL	87.5	87.2	87.5	87.5	0.83	1.845
Ensemble						
DenseNet201 + ResNet152 ensemble	88.2	87.9	88.2	88.0	0.85	1.650

Table 18 Ablation study results for spondylolisthesis assessment

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC	Slip (%)	MAE
Baseline Models							
ResNet50 (no transfer)	65.3	64.8	65.3	65.0	0.62	5.80	
ResNet50 + ImageNet transfer	72.8	72.1	72.8	72.4	0.69	4.25	
Architecture Comparison							
ResNet152 + conventional transfer	80.9	80.3	80.9	80.6	0.76	3.45	
DenseNet201 + conventional transfer	83.2	82.6	83.2	82.9	0.78	3.02	
Proposed Components							
DenseNet201 + CLAHE	84.4	83.8	84.4	84.1	0.80	2.75	
DenseNet201 + CLAHE + augmentation	86.1	85.5	86.1	85.8	0.82	2.40	
DenseNet201 + CLAHE + augmentation + two-phase TL	88.4	87.9	88.4	88.1	0.84	1.80	
Ensemble							
DenseNet201 + ResNet152 ensemble	89.0	88.6	89.0	88.8	0.86	1.50	

The exploratory results suggest incremental benefit from ImageNet pretraining, CLAHE, condition-aware augmentation, and staged fine-tuning; a publication-grade confirmatory rerun should perform all ablation/model selection on training-validation data and reserve an untouched patient-level test cohort for final evaluation.

5 Discussion

5.1 Interpretation of Key Findings

This single-center study demonstrates the ability of a multimodal framework to achieve promising measurement and severity classification performance in cases of scoliosis and spondylolisthesis on one radiographic dataset from a single hospital. DenseNet201 was slightly better compared to ResNet152 in numbers, while a low MAE and ICC > 0.95 are indicative of agreement with the provided reference labels. The Cobb angle MAE value of 1.845° and slip percentage MAE of 1.8 % points are comparable with several other studies [7, 13, 28, 34]. Directly comparing the findings and claiming superiority is impossible due to differences in case mix, acquisition, annotation, and validation procedures. The key limitation of the study is the validation design. It was conducted on data from one center, patient-level independence was not checked, annotators' information was incomplete; and no external or prospective cohort is available. Accordingly, the results support methodological feasibility rather than cross-center clinical generalizability.

5.2 Clinical Significance and Advantages

The clinical usefulness of the proposed method of automated spinal measurements lies in increased repeatability and faster quantification and not in automated interpretation of the images. Inference time from the source study was approximately 0.35 seconds per image, implying computational feasibility, though this was not evaluated on other computing platforms. Severity values must continue to be used only as decision-support information. Scarring and spondylolisthesis are treated based on symptoms, patient age, progression, neurologic condition, cause, and instability. These parameters are not known from this database of images only. Hence, any image with poor quality or no measurement should be noted and reviewed manually.

5.3 Advantages of the Unified Approach

One of the engineering strengths of the approach is the common preprocessing/classifier architecture that uses condition-specific geometry. Condition-aware data augmentation maintains the distinct orientation needs of the two tasks; nevertheless, the contributions of individual elements are still experimental since ablations from the source used the test set.

5.4 Limitations

The primary drawbacks are as follows. Class imbalance: there were very few independent images of very severe scoliosis and Grade V spondylolisthesis; the class-level statistics have high sampling uncertainty; augmentation does not generate any new observations from clinical practice. Single-center image-level validation: patient ID data and longitudinal linkage were not available; the lack of patient-level stratification cannot be validated. There may be repeated testing among splits, and external patient-level validation is necessary. Limited metadata and reference-standard annotation: demographic/clinical information, imaging acquisition parameters, annotator details, blinding, and adjudication process were not available; therefore, subgroup analysis, fairness, and multi-reader analysis cannot be performed. Measurement-failure and model-selection bias: the original implementation used a constant 20° Cobb-angle fallback, and ablation study used the same test set; future research should label non-gradable instances and maintain a novel patient-level test set. External validity: there are no independent institution, prospective study, calibration study, reader assistance experiment, and workflow study available; current results are not enough to make autonomous diagnoses and treatment decisions.

6 Conclusion

6.1 Main Findings

This retrospective study of a single-center nature investigated the framework proposed by the authors to automatically measure and grade scoliosis and spondylolisthesis. On the internal test datasets based on images only, the accuracy of DenseNet201 for scoliosis was 87.5%, whereas the accuracy of ResNet152 for scoliosis was 88.4%. Cobb angle MAE was 1.845°, slip percentage MAE was 1.8%, macro AUC was 0.83/0.84, and ICC was 0.962/0.958, respectively. However, the findings can only imply internal agreement with the provided reference labels but do not provide a small numerical edge for DenseNet201 over ResNet152, as no patient-level split independence, external and prospective validation, and complete metadata of the reference standard were conducted.

6.2 Future Improvements and Research Directions

Future efforts must focus on patient-related external validation at separate hospitals, prospective pre-specified assessment, failure analysis of localization/segmentation, calibration and sub-population analysis, and release of code/evaluation methods when

appropriate. In conclusion, the paper presents a short and concise methodological starting point for dual disease spinal X-ray analysis. For clinical translation, an untainted patient-based test set and ethics/reference standard documentation is required.

7 Declarations

- Ethics approval and consent to participate: The study was reviewed by the Institutional Review Board of USHA Hospital, which determined that the study was exempt from formal IRB approval because it involved retrospective secondary analysis of existing anonymized/de-identified spinal radiographs. The researchers had access only to radiographic images, and no personally identifiable patient information was available. The study involved no direct interaction or intervention with human participants. The requirement for informed consent was therefore waived/exempted by the Institutional Review Board/Institutional Ethics Committee.
- Clinical trial registration: Not applicable. This study was retrospective and did not involve a clinical trial or prospective intervention.
- Consent for publication: Not applicable. The study used anonymized/de-identified radiographic images and no identifiable patient information is presented.
- Code availability:The source code and trained models are not publicly available due to data-sharing but may be made available from the corresponding author upon reasonable request, subject to institutional approval and applicable data-use conditions.
- Data availability: The radiographic dataset used in this study contains anonymized hospital imaging data and is not publicly available due to institutional and data-governance restrictions. Access may be considered upon reasonable request to the corresponding author and subject to approval by the data-providing institution and applicable ethical/data-sharing requirements.
- Funding: no specific funding was received for this research.
- Competing interests: The authors declare that they have no known competing financial or non-financial interests that could have influenced the work reported in this manuscript.

References

[1] Amin, A., Abbas, M., Salam, A.A.: Automatic detection and classification of scoliosis from spine X-rays using transfer learning. In: 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), pp. 1–6. IEEE (2022).

[2] Casabó-Vall´es, G., Aldecoa, R., P´erez-Machado, G., Egea-Gámez, R.M., Sánchez-Raya, J., Vilalta, I., et al.: A fully autonomous AI system for accurate and reproducible Cobb angle measurement in adolescent idiopathic scoliosis: a multicenter study. The Spine Journal (2026).

[3] Dagli, M.M., Sussman, J.H., Gujral, J., Budihal, B.R., Kerr, M., Yoon, J.W., et al.: Development and validation of a keypoint region-based convolutional neural network to automate thoracic Cobb angle measurements using whole-spine standing radiographs. Acta Neurochirurgica 167(1), 227 (2025).

[4] Du, X., Wang, H., Jiang, L., Lv, C., Xi, Y., Yang, H.: Adjacent point aided vertebral landmark detection and Cobb angle measurement for automated AIS diagnosis. Computerized Medical Imaging and Graphics 121, 102496 (2025).

[5] Feng, C.K., Chen, Y.J., Dinh, Q.T., Tran, K.T., Liu, C.Y.: Multi-task learning for classification and prediction of adolescent idiopathic scoliosis based on fringe-projection three-dimensional imaging. European Spine Journal 34(11), 5321–5333 (2025).

[6] Fraiwan, L., AbuKhalaf, H.Y.: Robust AI-based Cobb angle measurement using dual-source X-ray radiographic datasets and pretrained CNNs. Cluster Computing 29(4), 201 (2026).

[7] Fraiwan, M., Audat, Z., Fraiwan, L., Manasreh, T.: Using deep transfer learning to detect scoliosis and spondylolisthesis from X-ray images. PLoS ONE 17(5), e0267851 (2022).

[8] Gadu, S.R., Potala, C.S.: SCANet: automated Cobb angle measurement using deep learning model in X-ray images for adolescent idiopathic scoliosis patients. Biomedical Signal Processing and Control 112, 108750 (2026).

[9] Goldman, S.N., Hui, A.T., Choi, S., Mbamalu, E.K., Tirabady, P., Eleswarapu, A.S., et al.: Applications of artificial intelligence for adolescent idiopathic scoliosis: mapping the evidence. Spine Deformity 12(6), 1545–1570 (2024).

[10] Hassan, M., Gonzalez Ruiz, J.M., Mohamed, N., Burke, T.N., Mei, Q., Westover, L.: Ensemble learning of deep CNN models and two-stage level prediction of Cobb angle on surface topography in adolescents with idiopathic scoliosis. Medical Engineering & Physics 140(1), 104332 (2025).

[11] Hayashi, D., Regnard, N.E., Ventre, J., Marty, V., Clovis, L., Lim, L., et al.: Deep learning algorithm enables automated Cobb angle measurements with high accuracy. Skeletal Radiology 54(7), 1469–1478 (2025).

[12] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[13] Horng, M.H., Kuok, C.P., Fu, M.J., Lin, C.J., Sun, Y.N.: Cobb angle measurement of spine from X-ray images using convolutional neural network. Comput. Math. Methods Med. 2019, 6357171 (2019).

[14] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708 (2017). https://doi.org/10.1109/CVPR.2017.243

[15] Huang, X., Luo, M., Liu, L., Wu, D., You, X., Deng, Z., et al.: The comparison of convolutional neural networks and the manual measurement of Cobb angle in adolescent idiopathic scoliosis. Global Spine Journal 14(1), 159–168 (2024).

[16] Jia, Y., Lu, Y., Wang, H., Yin, R., Ma, R.: CVS-Spine: integrating curve learning and vertebra segmentation for spine Cobb angle estimation. Biomedical Signal Processing and Control 125, 110679 (2026).

[17] Kato, S., Maeda, Y., Nagura, T., Nakamura, M., Watanabe, K.: Comparison of three artificial intelligence algorithms for automatic Cobb angle measurement using teaching data specific to three disease groups. Scientific Reports 14(1), 17989 (2024).

[18] Koo, T.K., Li, M.Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine 15(2), 155–163 (2016). https://doi.org/10.1016/j.jcm.2016.02.012

[19] Kumar, R., Gupta, M., Abraham, A.: A critical analysis on vertebra identification and Cobb angle estimation using deep learning for scoliosis detection. IEEE Access 12, 11170–11184 (2024).

[20] Lang, S., Jokeit, M., Kim, J.H., Urbanschitz, L., Fisler, L., Torrez, C., et al.: Anatomical landmark detection on bi-planar radiographs for predicting spinopelvic parameters. Spine Deformity 13(2), 423–431 (2025).

[21] Li, L., Zhang, T., Lin, F., Li, Y., Wong, M.S.: Automated 3D Cobb angle measurement using U-Net in CT images of preoperative scoliosis patients. Journal of Imaging Informatics in Medicine 38(1), 309–317 (2025).

[22] Lu, Q., Ni, L., Zhang, Z., Zou, L., Guo, L., Pan, Y.: Superior performance of a center-point AI model over VFLDNet in automated Cobb angle estimation for scoliosis assessment. European Spine Journal 33(12), 4710–4719 (2024).

[23] Minotti, M., Negrini, S., Cina, A., Galbusera, F., Zaina, F., Bassani, T.: Deep learning prediction of curve severity from rasterstereographic back images in adolescent idiopathic scoliosis. European Spine Journal 33(11), 4164–4170 (2024).

[24] Mohamed, N., Ruiz, J.M.G., Hassan, M., Costa, O.A., Burke, T.N., Mei, Q., Westover, L.: Three-dimensional markerless surface topography approach with convolutional neural networks for adolescent idiopathic scoliosis screening. Scientific Reports 15(1), 8728 (2025).

[25] Nasrabadi, A.M., Moammer, G., McPhee, J.: Automated Lenke classification for preoperative spine surgery by extracting anatomical landmarks from X-ray images using a deep learning approach. North American Spine Society Journal (NASSJ) 22, 100697 (2025).

[26] Ni, L., Zhang, Z., Zou, L., Wang, J., Guo, L., Qian, W., et al.: Clinical application of automatic assessment of scoliosis Cobb angle based on deep learning. Current Medical Imaging 20(1), e15734056278130 (2024).

[27] Prasad, K.S., Jana, S.: DeepSDC: An efficient deep neural network based spinal disease classification using ResNet152 model. In: 2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), pp. 1–6. IEEE (2025).

[28] Rahmaniar, W., Suzuki, K., Lin, T.L.: Auto-CA: automated Cobb angle measurement based on vertebrae detection for assessment of spinal curvature deformity. IEEE Transactions on Biomedical Engineering 71(2), 640–649 (2023).

[29] Shao, Z., Yuan, Y., Ma, L., Zhu, X.: CurvNet: Latent contour representation and iterative data engine for curvature angle estimation. Pattern Recognition, 112546 (2025).

[30] Sun, Y., Xing, Y., Zhao, Z., Meng, X., Xu, G., Hai, Y.: Comparison of manual versus automated measurement of Cobb angle in idiopathic scoliosis based on a deep learning keypoint detection technology. European Spine Journal 31(8), 1969–1978 (2022).

[31] Tavana, P., Akraminia, M., Koochari, A., Bagherifard, A.: An efficient ensemble method for detecting spinal curvature type using deep transfer learning and soft voting classifier. Expert Systems with Applications 213, 119290 (2023).

[32] Wang, S., Li, K., Zhang, S., Zhang, D., Hao, Y., Zhou, Y., et al.: Development of a deep learning model for measuring sagittal parameters on cervical spine X-ray. European Spine Journal 34(9), 3798–3808 (2025).

[33] Wong, J., Reformat, M., Lou, E.: Applying machine learning and point-set registration to automatically measure the severity of spinal curvature on radiographs. IEEE Journal of Translational Engineering in Health and Medicine 12, 151–161 (2023).

[34] Wu, H., Zheng, S., Du, W., Zhang, J., Wang, Z., Lu, L., et al.: A fully automatic Cobb angle measurement framework of full-spine DR images based on deep learning. European Spine Journal 34(9), 3756–3766 (2025).

[35] Xie, K., Zhu, S., Lin, J., Li, Y., Huang, J., Lei, W., Yan, Y.: A deep learning model for radiological measurement of adolescent idiopathic scoliosis using biplanar radiographs. Journal of Orthopaedic Surgery and Research 20(1), 236

(2025).

[36] Xu, L., Liu, C., Zhong, W., Xu, K., He, N., Hu, J., et al.: A fully automated multistage deep learning system for Lenke classification: enhanced diagnostic precision in adolescent idiopathic scoliosis. The Journal of Bone and Joint Surgery. American Volume 108(12), 914 (2025).

[37] Yao, Y., Yu, W., Gao, Y., Dong, J., Xiao, Q., Huang, B., Shi, Z.: W-transformer: accurate Cobb angles estimation by using a transformer-based hybrid structure. Medical Physics 49(5), 3246–3262 (2022).

[38] Ye¸silmen, N., Danacı, Ç ., Baydoğan, M.P., Tuncer, S.A., Ç ınar, A., Tuncer, T.: Enhanced vision transformer with custom attention mechanism for automated idiopathic scoliosis classification. Journal of Imaging Informatics in Medicine 39(2), 1122–1139 (2026).

[39] Zhang, B., Chen, K., Yuan, H., Liao, Z., Zhou, T., Guo, W., et al.: Automatic Lenke classification of adolescent idiopathic scoliosis with deep learning. JOR Spine 7(2), e1327 (2024).

[40] Zhang, H., Huang, C., Wang, D., Li, K., Han, X., Chen, X., Li, Z.: Artificial intelligence in scoliosis: current applications and future directions. Journal of Clinical Medicine 12(23), 7382 (2023).

[41] Zhang, T., Zhu, C., Zhao, Y., Zhao, M., Wang, Z., Song, R., et al.: Deep learning model to classify and monitor idiopathic scoliosis in adolescents using a single smartphone photograph. JAMA Network Open 6(8), e2330617 (2023).

[42] Zhao, Y., Zhang, J., Li, H., Gu, X., Li, Z., Zhang, S.: Automatic Cobb angle measurement method based on vertebra segmentation by deep learning. Medical & Biological Engineering & Computing 60(8), 2257–2269 (2022).

[43] Zhu, Y., Yin, X., Chen, Z., Zhang, H., Xu, K., Zhang, J., Wu, N.: Deep learning in Cobb angle automated measurement on X-rays: a systematic review and meta-analysis. Spine Deformity 13(1), 19–27 (2025).

[44] Zou, S., Hsieh, S.Y., Huang, K.Y., Chou, H.H.: Automatic spinal Cobb angle measurements from X-ray images using a novel vertebra centroid radiation landmark detection network. Biomedical Signal Processing and Control 112, 108712 (2026).

[45] Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: Heckbert, P.S. (ed.) Graphics Gems IV, pp. 474–485. Academic Press (1994).