DOI: **10.37943/25XRRS8272**

Beibit Abdikenov
PhD, Director of Science and Innovation Center "Artificial Intelligence"
beibit.abdikenov@astanait.edu.kz, orcid.org/0000-0002-0284-0949
Astana IT University, Astana 010000, Kazakhstan
Ayan Kokhan
Master's student, School of Artificial Intelligence and Data Science
242852@astanait.edu.kz, orcid.org/0009-0006-1114-372X
Astana IT University, Astana 010000, Kazakhstan
Temirlan Karibekov
PhD, Director of Science and Innovation Center "MedTech"
t.karibekov@astanait.edu.kz, orcid.org/0009-0008-9801-17749
Astana IT University, Astana 010000, Kazakhstan

A DUAL-PATH MULTI-TASK FRAMEWORK FOR STRICT THREE-CURVE COBB ANGLE ESTIMATION IN IDIOPATHIC SCOLIOSIS

Abstract: Adolescent idiopathic scoliosis management depends on reproducible Cobb angle measurement across three clinically defined spinal regions: proximal thoracic, main thoracic, and thoracolumbar/lumbar. Although manual measurement remains the reference standard, it is observer-dependent and time-consuming, with inter-observer variability exceeding five degrees even among experienced readers. Most automated deep learning approaches target a single dominant curve or use unconstrained outputs, which limits their applicability to structured clinical workflows requiring strict regional assignment. This study presents a dual-path multi-task framework for simultaneous estimation of all three regional Cobb angles from posteroanterior spinal radiographs. The architecture integrates a ConvNeXt-Tiny encoder, vertebral localization heads, direct global angle regression via soft-argmax, and a geometric tilt-aggregation pathway. A learned per-region sigmoid gate fuses the global and geometric pathways, providing a fixed but optimized balance between statistical and anatomical estimation. The model was developed on 21,294 radiographs with leakage-controlled partitioning into training (N = 17,262), validation (N = 2,016), and test (N = 2,016) subsets. Training employed a two-stage curriculum with severity-aware sampling and hard replay for difficult cases. Three independent runs (seeds 42, 52, 62) were ensembled with test-time augmentation. On the primary held-out set (N = 2,015), the ensemble achieved a mean absolute error of 2.24 degrees (proximal thoracic 2.21, main thoracic 1.97, thoracolumbar/lumbar 2.54), with near-zero Bland-Altman bias (0.03 degrees), good-to-excellent intraclass correlation coefficients (0.884–0.971), and 90.4% of predictions within 5 degrees. At the 40-degree treatment threshold, sensitivity was 0.934 and specificity was 0.994. These findings support the feasibility of strict three-curve automation for reader-in-the-loop clinical workflows.

Keywords: Cobb angle estimation; scoliosis; deep learning; multi-task learning; dual-path fusion; medical imaging; curriculum learning; clinical decision support.

Introduction

Adolescent idiopathic scoliosis is the most common spinal deformity in the pediatric population, affecting approximately two to three percent of adolescents worldwide [1, 2]. Clinical management decisions, including observation, bracing, and surgical referral, rely heavily on accurate and reproducible Cobb angle measurement. A five-degree measurement change at critical thresholds can alter the treatment pathway, making measurement consistency a fundamental concern for clinicians [1]. Despite its clinical centrality, the Cobb method remains a manual process susceptible to non-trivial inter- and intra-observer variability, which several reliability studies have documented across different clinical settings and reader populations [1, 3].

The challenge of manual measurement is compounded by the need for longitudinal monitoring, in which repeated measurements must be comparable across time points and readers.

Received: 11.02.2026 Accepted: 25.02.2026 Published: 30.03.2026

In routine clinics, patients may undergo serial imaging over several years, and small apparent changes may represent true progression, measurement noise, or both. This practical burden motivates the development of automated support systems that can provide consistent, reproducible angle estimates without replacing clinician judgment [3, 4].

Recent deep learning methods have advanced automated Cobb angle estimation substantially. Approaches range from landmark-based detection and segmentation-driven pipelines to direct regression models [2, 5, 6]. Representative examples include vertebra localization and tilt-field estimation [7], single major-curve deep regression [8], and SVD-based curve estimation [9]. However, most published systems optimize for a single dominant curve or produce unconstrained angle outputs that do not enforce the clinically structured proximal thoracic, main thoracic, and thoracolumbar/lumbar reporting contract. This distinction is important because Lenke-based clinical decision frameworks rely on explicit identification of curve type and region [4, 10]. A model that achieves low aggregate error without reliable regional assignment may still be inadequate for structured clinical workflows.

Meta-analytic evidence further complicates direct comparison across studies. A recent systematic review and meta-analysis encompassing 50 studies found an overall circular mean absolute error of approximately three degrees but with substantial heterogeneity across datasets, endpoint definitions, and evaluation protocols [11]. This heterogeneity underscores the need for transparent reporting that distinguishes between different output contracts and evaluation denominators. Against this background, the present study evaluates a dual-path multi-task framework specifically designed for strict three-curve estimation, with emphasis on multi-seed ensemble stability, comprehensive agreement analysis, and clinically interpretable threshold behavior.

The working hypothesis is that combining a direct statistical pathway and an anatomically constrained geometric pathway through a learned per-region fusion gate will improve strict three-region Cobb angle estimation versus relying on a single estimation paradigm. To test this hypothesis, we trained and evaluated a multi-task dual-path model on a leakage-controlled split with multi-seed ensembling, agreement analysis, and clinically relevant threshold metrics.

Literature Review

Automated Cobb angle estimation has evolved through several methodological paradigms. Early approaches relied on classical image processing techniques, including edge detection, Hough transforms, and template matching, which required substantial preprocessing and were sensitive to image quality variations [3]. The introduction of deep learning methods marked a significant shift, enabling end-to-end learning from radiographic images with minimal manual feature engineering.

Within the deep learning paradigm, three principal methodological families can be distinguished. Segmentation-based methods first delineate vertebral bodies and then apply geometric rules to identify the most tilted endplates, following the classical manual workflow in an automated fashion [5, 12]. Landmark-based methods detect specific anatomical keypoints, such as vertebral corners or endplates, and compute angles from their spatial configuration [6, 7]. Direct regression methods bypass intermediate anatomical representation and predict angle values directly from image features [8]. Each approach has characteristic strengths and limitations: segmentation methods maintain anatomical interpretability but require precise pixel-level annotation; landmark methods offer geometric transparency but are sensitive to detection noise on individual points; direct regression is computationally efficient but may lack anatomical coherence.

A notable advance in hybrid architecture is represented by the vertebra localization and tilt estimation network, which combines vertebral centroid detection with a per-vertebra tilt field to derive Cobb angles through geometric aggregation [7]. This approach demonstrated that incorporating structural priors into the estimation pipeline can improve both accuracy and

consistency. However, VLTENet and similar systems typically estimate a single or unconstrained curve output rather than enforcing strict three-region assignment.

Multi-task learning has emerged as a powerful paradigm in medical image analysis, enabling shared feature representations across related tasks with complementary supervision signals [13, 14]. In the context of spinal analysis, multi-task architectures can jointly optimize vertebral localization, segmentation, counting, and angle estimation, potentially improving generalization and robustness through implicit regularization. The combination of multi-task architectures with curriculum learning strategies, which progressively increase training difficulty, has shown promise in handling class imbalance and difficult cases in medical imaging [15, 16].

Despite steady progress, several gaps remain in the literature. First, most studies report performance on unconstrained or single-curve outputs, which limits applicability to clinical workflows requiring strict regional assignment. Second, cross-study comparison is hampered by heterogeneous evaluation protocols, with different studies using different metrics (mean absolute error, circular mean absolute error, symmetric mean absolute percentage error), different denominators, and different endpoint contracts [3, 11]. Third, there is limited reporting of agreement metrics beyond point accuracy, such as Bland-Altman analysis and intraclass correlation, which are standard in clinical measurement sciences but underused in the deep learning literature. Fourth, severity-stratified analysis, which is critical for understanding clinical risk, is rarely presented. The present work addresses these gaps by combining a novel dual-path architectural design with comprehensive, clinically oriented evaluation under a strict three-curve output contract.

Aim and Objectives of the Study

The aim of this study is to develop and evaluate an automated method for strict three-curve Cobb angle estimation from posteroanterior spinal radiographs.

The objectives are:

1. To design a dual-path model that combines direct global regression with geometric tilt aggregation through an adaptive learned fusion gate.

2. To optimize training using a two-stage curriculum with severity-aware sampling and hard replay for difficult cases.

3. To quantify performance using angle-level, case-level, agreement, and threshold-based clinical metrics on a leakage-controlled held-out test set.

4. To evaluate model robustness under controlled perturbations and analyze the relationship between predictive uncertainty and error magnitude.

Methods and Materials

1. Data and split integrity

The development cohort comprised 21,294 posteroanterior radiographs. Data were partitioned into training (N = 17,262), validation (N = 2,016), and test (N = 2,016) subsets, with zero patient or image overlap across partitions verified programmatically. Training severity composition was mild (less than 20 degrees): 8,802 cases; moderate (20 to 40 degrees): 5,723 cases; and severe (40 degrees or greater): 1,585 cases. Vertebral annotations followed COCO-format polygon specifications with four corner keypoints per vertebra. Angle supervision provided three regional values (proximal thoracic, main thoracic, and thoracolumbar/lumbar) per image. All images were resized to 512 by 512 pixels for model input.

For transparent reporting, two analysis denominators are explicitly distinguished. Primary predictive metrics are reported on N = 2,015 eligible cases. Paired agreement analyses requiring complete matched triplets across all three regions are reported on N = 2,014 evaluable cases.

2. Network architecture

The model architecture is illustrated in Fig. 1. A ConvNeXt-Tiny encoder [17] pretrained on ImageNet extracts multi-scale features at three resolution levels with 256-dimensional channel representations. Distinct from existing multi-path scoliosis networks that focus on symmetric

feature matching or detection-segmentation fusion, this framework implements a dual-path hybrid combining statistical and geometric estimation paradigms.

Five task-specific heads operate on the shared feature representation: (1) a center head predicting normalized coordinates for up to 24 vertebral centers, regularized by a learned y-axis positional prior that encourages anatomically plausible craniocaudal ordering; (2) a corner keypoint head predicting four positional offsets per vertebra, bounded by a tanh activation and a maximum displacement norm; (3) a count head classifying the number of visible vertebrae and generating a binary mask that gates downstream per-vertebra computations; (4) a global angle head producing 181-bin soft-argmax estimates (spanning minus 90 to plus 90 degrees) for the three regional angles through direct statistical regression; and (5) a tilt head producing per-vertebra angle logits aggregated via dynamic zoning into geometric regional estimates. Dynamic zoning partitions the detected vertebral column into three approximately equal segments (upper, middle, and lower thirds based on the visible vertebra count) and computes the tilt range (maximum minus minimum) within each segment as a regional angle surrogate. Three learned per-region sigmoid parameters fuse the global regression-based and geometric tilt-based pathway outputs. Importantly, these fusion weights are fixed after training and do not vary across input images; rather, they represent the optimal global balance between the two estimation paradigms as determined during optimization.

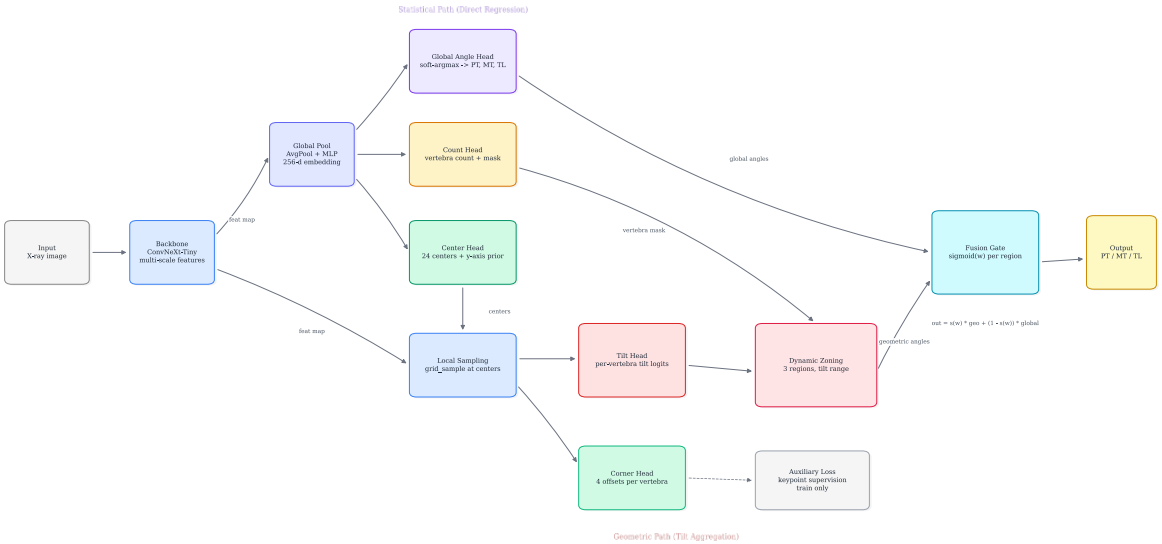

Fig. 1. Dual-path architecture for strict three-curve Cobb angle estimation.

3. Training protocol

Training followed a two-stage protocol. Stage 1 focused on structural localization with angle losses disabled, using up to 120 epochs and an early stopping patience of 25 epochs. Stage 2 initialized from the Stage 1 checkpoint and enabled full-angle supervision on the labeled subset (N = 16,110). Stage 2 employed two curriculum phases: warmup (epochs 1 through 30, with mild severity upsampling at weights [1.0, 1.4, 1.8]) and polish (epochs 31 through 120, with aggressive severity upsampling at weights [1.0, 1.8, 2.6] for mild, moderate, and severe groups, respectively). Loss weights were center 0.2, keypoint 0.7, angle 1.0, auxiliary angle 0.4, and final fused angle 2.0.

Optimization used AdamW [18] with a learning rate of 3 times 10 to the negative fifth power, cosine scheduling, weight decay of 10 to the negative third power, batch size 16, gradient clipping at norm 1.0, and exponential moving average with decay 0.9998 [19]. Loss composition combined Smooth L1 for localization and L1 for angle prediction, with Gaussian negative log-likelihood for uncertainty modeling [20, 21]. Severity-aware sampling used threshold bins at 20

and 40 degrees. Hard replay further upweighted the top 700 highest-error cases per epoch, with a boost factor of 1.8 and per-epoch decay of 0.97. Data augmentation included horizontal flips, rotation of plus or minus 7 degrees, scale-translate, brightness and contrast adjustment, gamma perturbation, additive Gaussian noise, and Gaussian blur.

4. Inference and evaluation

Model selection criterion was the minimum validation mean absolute error for the main thoracic region across three independent training runs (seeds 42, 52, and 62). Final inference used ensemble averaging of the three selected checkpoints with four-view test-time augmentation (original, horizontal flip, and rotations of minus 3 and plus 3 degrees).

The primary endpoint was the overall mean absolute error (MAE). Secondary endpoints included regional mean absolute error for proximal thoracic, main thoracic, and thoracolumbar/lumbar curves; WITHIN_k metrics describing the fraction of angle-level predictions with absolute error at or below k degrees; CASE_WITHIN_k metrics requiring all three predictions within a case to satisfy the threshold simultaneously; and symmetric mean absolute percentage error. Agreement reporting included Bland-Altman bias and limits of agreement, and intraclass correlation coefficient ICC(2,1) per region. Clinical operating characteristics were reported at 10-degree and 40-degree thresholds through sensitivity and specificity with 95% confidence intervals. Error distribution was profiled by median and upper quantiles (P90, P95, P99). Robustness was evaluated through controlled perturbation tests. Uncertainty behavior was assessed via correlation between Monte Carlo dropout uncertainty and absolute error (N = 200 cases, 20 stochastic forward passes).

Results

The multi-seed ensemble improved overall mean absolute error from 2.30 degrees (best single model, seed 62) to 2.24 degrees. Regional values improved from 2.28, 2.03, and 2.60 degrees to 2.21, 1.97, and 2.54 degrees for the proximal thoracic, main thoracic, and thoracolumbar/lumbar regions, respectively (Table 1). Multi-seed validation stability was confirmed by a main thoracic mean absolute error of 2.091 degrees with a standard deviation of 0.038 degrees across runs.

Table 1. Single-model and ensemble performance on the primary held-out analysis set (N = 2,015)

Configuration	MAE	MAE_PT	MAE_MT	MAE_TL	W_5	CW_5	CW_10	SMAPE
Single model	2.30°	2.28°	2.03°	2.60°	89.9%	81.2%	93.5%	31.8%
Ensemble	2.24°	2.21°	1.97°	2.54°	90.4%	82.7%	93.9%	30.8%

W_5 indicates WITHIN_5 at the angle level. CW_5 and CW_10 indicate CASE_WITHIN_5 and CASE_WITHIN_10 at the full three-curve case level.

Bland-Altman analysis on N = 2,014 evaluable paired cases (6,042 angle-level pairs across three regions) yielded a mean bias of 0.03 degrees with limits of agreement from minus 8.28 to plus 8.35 degrees and a negligible proportional trend (slope minus 0.026), with regional Bland-Altman plots for proximal thoracic (Fig. 2, *a*), main thoracic (Fig. 2, *b*), and thoracolumbar/lumbar (Fig. 2, *c*) regions, plus a pooled overview (Fig. 2, *d*). Intraclass correlation coefficients were 0.884 for the proximal thoracic region (95% CI: 0.87 to 0.89), 0.971 for the main thoracic region (95% CI: 0.97 to 0.97), and 0.904 for the thoracolumbar/lumbar region (95% CI: 0.90 to 0.91), indicating good-to-excellent agreement.

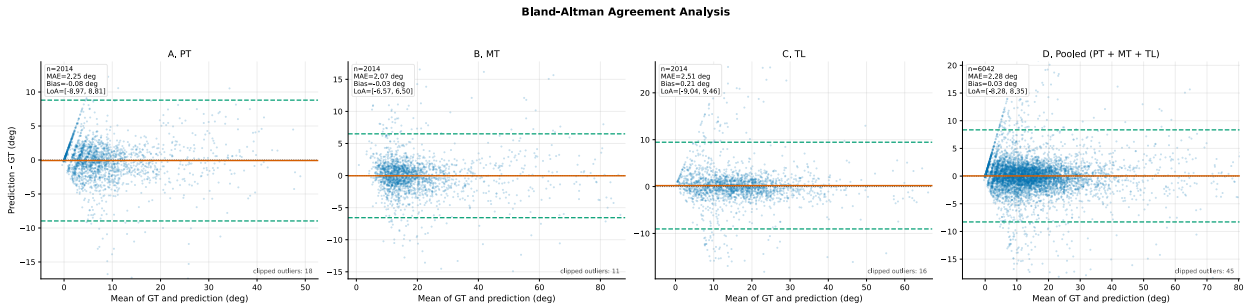

Fig. 2. Bland-Altman signed-error analysis: *a* – proximal thoracic; *b* – main thoracic; *c* – thoracolumbar/lumbar; *d* – pooled predictions with mean bias and 95% limits of agreement.

At the 40-degree treatment threshold, sensitivity was 0.934 (95% CI: 0.891 to 0.961) and specificity was 0.994 (95% CI: 0.989 to 0.997). At the 10-degree screening threshold, sensitivity was 0.976 (95% CI: 0.968 to 0.982) and specificity was 0.743 (95% CI: 0.666 to 0.807).

Mean absolute error increased with severity from 1.89 degrees in mild cases (N = 1,103) to 2.37 degrees in moderate cases (N = 713) and 4.09 degrees in severe cases (N = 198). The per-case error profile includes the histogram (Fig. 3, *a*), severity-stratified MAE (Fig. 3, *b*), error spread by severity (Fig. 3, *c*), and empirical CDF (Fig. 3, *d*), indicating strong central performance with a limited but clinically relevant high-error tail.

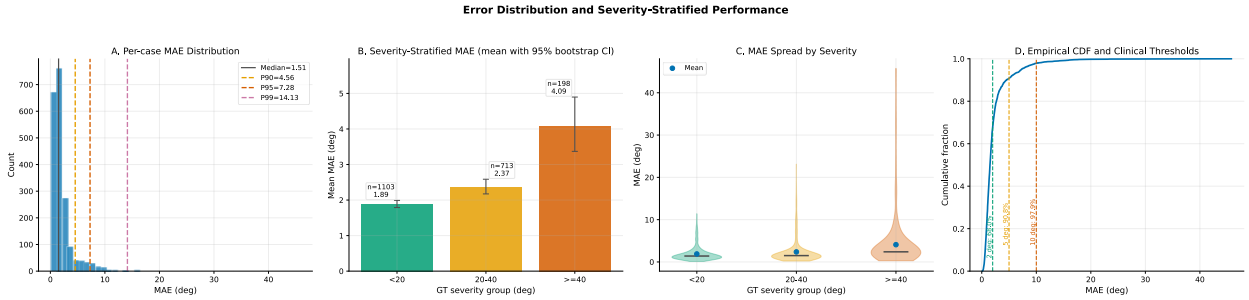

Fig. 3. Error distribution and severity-stratified performance: *a* – per-case histogram; ***b*** **–** severity-stratified mean absolute error with 95% bootstrap confidence interval; *c* – error spread by severity; *d* – empirical cumulative distribution with clinical thresholds.

Vertebral keypoint localization quality was assessed as an auxiliary diagnostic. Across 138,208 predicted keypoints, the mean pixel error was 8.81 pixels with a median of 1.34 pixels, P90 20.38 pixels, and P99 69.90 pixels; approximately 19.1 percent of keypoints exceeded 10 pixels error, indicating that most localizations were accurate, but a subset of anatomically challenging vertebrae drove elevated pixel-level residuals.

Controlled robustness tests showed that mild photometric perturbations produced changes in mean absolute error within plus or minus 0.075 degrees, whereas center-cropping caused larger degradation (plus 0.68 degrees at 80 percent crop). Predicted-mask versus ground-truth-mask evaluation showed a negligible gap of minus 0.015 degrees. Monte Carlo dropout uncertainty correlated with absolute error (r = 0.467), with high-uncertainty cases showing a mean absolute error of 6.75 degrees compared to 2.44 degrees for low-uncertainty cases.

Contextual comparison with representative published studies is provided in Table 2. Direct numerical ranking across studies should be interpreted cautiously because datasets, endpoint contracts, and evaluation conventions vary substantially [11].

Table 2. Contextual comparison with representative published reports.

Study	Endpoint contract	Cohort	Reported metric
AASCE2019 challenge [2]	Major-curve challenge	609 images	SMAPE 21.71%, CMAE ~3.9°
VLTENet [7]	Automated non-strict	Not matched	CMAE 3.51°
Shi et al. [9]	SVD-based estimation	630 radiographs	MAE 2.55°
Suri et al. [6]	Hardware-invariant	Not matched	MAE 2.96°
Wang et al. [8]	Single major-curve	N = 297	MAE 1.97°
Present study	Strict PT/MT/TL	N = 2,015	MAE 2.24°

Discussion

The results demonstrate that strict three-curve Cobb angle automation can achieve low internal error while maintaining agreement characteristics relevant for clinical decision support. The convergence of three complementary evidence types, including low mean absolute error, near-zero Bland-Altman bias, and good-to-excellent intraclass correlation, supports coherent model behavior rather than metric-specific optimization artifacts. This multi-faceted evaluation approach aligns with recommendations for reporting artificial intelligence performance in medical imaging [22, 23].

A key architectural contribution is the dual-path design. The direct global regression pathway provides resilience when individual anatomical landmarks are noisy or partially occluded, as it learns holistic spatial patterns from the entire radiograph. In contrast, the geometric tilt aggregation pathway maintains explicit anatomical structure by computing angles from per-vertebra tilt estimates through dynamic zoning, similar in spirit to the tilt-field approach of VLTENet [7] but applied within a strict three-region contract. A set of three learned per-region sigmoid parameters determines the balance between these pathways. These weights are fixed after training and do not vary per image, meaning the model learns a globally optimal trade-off between statistical and geometric evidence for each spinal region. This fusion paradigm differs from existing multi-path scoliosis networks that often focus on symmetric feature matching or detection-segmentation fusion rather than combining fundamentally different mathematical estimation paradigms.

Error analysis of the 50 worst-performing cases (per-case mean absolute error ranging from 9.2 to 26.3 degrees) revealed two primary error drivers: severe deformities with ground-truth angles exceeding 50 degrees, and high keypoint localization errors (mean pixel error above 10 pixels). Several of these cases exhibited both characteristics simultaneously, suggesting that extreme vertebral rotation and structural overlap degrade both estimation pathways. This analysis reinforces the importance of confidence-based flagging for severe and anatomically ambiguous presentations.

Regional asymmetry in agreement performance is clinically plausible and interpretable. The main thoracic region achieved the highest intraclass correlation, likely reflecting stronger anatomical delineation and larger absolute angle magnitudes in the mid-thoracic spine. The proximal thoracic and thoracolumbar/lumbar regions showed lower agreement, consistent with known imaging challenges, including upper-thoracic ambiguity, shoulder and rib overlap, pelvic overlap, and positioning variability. The ensemble gain of minus 0.06 degrees in mean absolute error was consistent across all regions, suggesting that independently initialized models capture partially independent error modes amenable to simple averaging, a finding consistent with ensemble learning theory [24].

The severity-stratified degradation from 1.89 to 4.09 degrees is clinically interpretable. Severe curves involve greater vertebral rotation, rib cage deformity, and structural overlap that reduce landmark visibility and increase annotation uncertainty. The curriculum training strategy with progressive severity upsampling was designed to mitigate this effect, but the residual

performance gap suggests that anatomy-driven information loss imposes a performance floor that cannot be fully overcome by sampling strategy alone. This finding has practical implications: approximately 90 percent of cases fall within clinically acceptable error tolerance, while the remaining 10 percent, concentrated on severe and anatomically complex presentations, should be flagged for expert review.

From a clinical workflow perspective, these results suggest two concrete deployment modes. First, a screening module that auto-clears mild-to-moderate cases with high reliability, reducing routine measurement burden. Second, a flagging module that assigns confidence scores to each prediction, routing cases with high uncertainty or severity above 40 degrees for priority expert review. The 40-degree threshold performance (sensitivity 0.934, specificity 0.994) is particularly relevant for surgical planning pipelines [4]. This supports a confidence-aware reader-in-the-loop paradigm rather than fully autonomous reporting.

Robustness testing revealed stability under photometric variation but vulnerability to field-of-view truncation, motivating explicit image-quality checks in deployment pipelines. The uncertainty-error correlation ($r = 0.467$) supports triage-level risk stratification but is insufficient for fully calibrated prediction intervals. Further calibration research is needed before uncertainty estimates can be used for formal clinical risk quantification.

The study has several limitations that must be acknowledged. All evaluations were performed on internal data without external multi-center validation, which limits generalizability claims. Formal multi-rater benchmarking against multiple clinicians with varying experience levels was not included. Component-level ablation for the gate design, curriculum schedule, severity sampling weights, and hard replay parameters is ongoing. Cross-domain shift analysis, including scanner manufacturer differences, acquisition protocol variation, and population demographics, remain to be conducted. Additionally, the uncertainty analysis, while directionally informative, is associative rather than formally calibrated.

Conclusion

A dual-path multi-task ensemble achieved a mean absolute error of 2.24 degrees for strict proximal thoracic, main thoracic, and thoracolumbar/lumbar Cobb angle estimation. The model demonstrated good-to-excellent intraclass correlation (0.884 to 0.971), near-zero Bland-Altman bias, sensitivity of 0.934, and specificity of 0.994 at the 40-degree surgical threshold, and stable performance across three independently initialized training runs. Together with severity-stratified analysis and uncertainty profiling, these findings support the technical feasibility of strict three-curve automation for reader-in-the-loop scoliosis workflows under supervised clinical use. External validation, component-level ablation, and multi-rater benchmarking remain essential next steps before clinical deployment.

Acknowledgment

This research is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993145 “Artificial intelligence technologies for analyzing multimodal big data for breast cancer diagnosis and prognosis”)

References

[1] Prestigiacomo, C., Hulsbosch, M. H. H. M., Bruls, V. E. J., & Nieuwenhuis, J. J. (2022). Intra- and inter-observer reliability of Cobb angle measurements in patients with adolescent idiopathic scoliosis. *Spine Deformity, 10*(1), 79-86. https://doi.org/10.1007/s43390-021-00398-0

[2] Wang, L., Xie, C., Lin, Y., Zhou, H.-Y., Chen, K., Cheng, D., Kocev, D., Yap, C. H., Staring, M., & de Brujine, M. (2021). Evaluation and comparison of accurate automated spinal curvature estimation algorithms with spinal anterior-posterior X-ray images: The AASCE2019 challenge. *Medical Image Analysis, 72*, 102115. https://doi.org/10.1016/j.media.2021.102115

[3] Jin, C., Wang, S., Yang, G., Li, E., & Liang, Z. (2022). A review of the methods on Cobb angle measurements for spinal curvature. *Sensors, 22*(9), 3258. https://doi.org/10.3390/s22093258

[4] d'Astorg, H., Faron, M., Jolivet, E., Folinais, D., & Bertheau, R. C. (2023). Comparison of Cobb angle measurements for scoliosis assessment using different imaging modalities: A systematic review. *EFORT Open Reviews, 8*(6), 489-498. https://doi.org/10.1530/EOR-23-0032

[5] Horng, M.-H., Kuok, C.-P., Fu, M.-J., Lin, C.-J., & Sun, Y.-N. (2019). Cobb angle measurement of spine from X-ray images using convolutional neural network. *Computational and Mathematical Methods in Medicine, 2019*, 6357171. https://doi.org/10.1155/2019/6357171

[6] Suri, A., Tang, S., Kargilis, D., Taratuta, E., Kneeland, B. J., Choi, G., Fritz, J., & Carrino, J. A. (2023). Conquering the Cobb angle: A deep learning algorithm for automated, hardware-invariant measurement of Cobb angle on radiographs in patients with scoliosis. *Radiology: Artificial Intelligence, 5*(4), e220158. https://doi.org/10.1148/ryai.220158

[7] Zou, L., Guo, L., Zhang, R., Ni, L., Chen, Z., He, X., & Wang, J. (2023). VLTENet: A deep-learning-based vertebra localization and tilt estimation network for automatic Cobb angle estimation. *IEEE Journal of Biomedical and Health Informatics, 27*(6), 3002-3013. https://doi.org/10.1109/JBHI.2023.3258361

[8] Wang, M. X., Kim, J. K., Choi, J.-W., Park, D., & Chang, M. C. (2024). Deep learning algorithm for automatically measuring Cobb angle in patients with idiopathic scoliosis. *European Spine Journal, 33*(11), 4155-4163. https://doi.org/10.1007/s00586-023-08024-5

[9] Shi, C., Meng, N., Zhuang, Y., Cheung, J. P. Y., Zhao, M., Huang, H., Lin, Y., & Zhang, T. (2025). Accurate Cobb angle estimation via SVD-based curve detection and vertebral wedging quantification. *IEEE Journal of Biomedical and Health Informatics, 29*(12), 8607-8614. https://doi.org/10.1109/JBHI.2025.3600647

[10] Chen, R., Xi, Y., Wang, T., Wang, A., Ma, Z., Liang, M., Yuan, S., Zang, L., & Fan, N. (2026). Advances in fusion level selection and surgical approaches for adolescent idiopathic scoliosis based on the Lenke classification system: A narrative review. *BMC Surgery, 26*(1), 102. https://doi.org/10.1186/s12893-025-03481-9

[11] Zhu, Y., Yin, X., Chen, Z., Zhang, H., Xu, K., Zhang, J., & Wu, N. (2025). Deep learning in Cobb angle automated measurement on X-rays: A systematic review and meta-analysis. *Spine Deformity, 13*(1), 19-27. https://doi.org/10.1007/s43390-024-00954-4

[12] He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, 2961-2969. https://doi.org/10.1109/ICCV.2017.322

[13] Zhang, Y., & Yang, Q. (2022). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering, 34*(12), 5586-5609. https://doi.org/10.1109/TKDE.2021.3070203

[14] Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*. https://doi.org/10.48550/arXiv.2009.09796

[15] Soviany, P., Ionescu, R. T., Rota, P., & Sebe, N. (2022). Curriculum learning: A survey. *International Journal of Computer Vision, 130*(6), 1526-1565. https://doi.org/10.1007/s11263-022-01611-x

[16] Jimenez-Sanchez, A., Mateus, D., Kirchhoff, S., Navab, N., & Ballester, M. A. G. (2019). Medical-based deep curriculum learning for improved fracture classification. *Proceedings of Medical Image Computing and Computer-Assisted Intervention*, 694-702. https://doi.org/10.1007/978-3-030-32226-7_77

[17] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976-11986.

[18] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *Proceedings of the International Conference on Learning Representations*.

[19] Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems, 30*.

[20] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems, 30*.

[21] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 1050-1059.

[22] Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., & Kahn, C. E. (2024). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence, 6*(4), e240300. https://doi.org/10.1148/ryai.240300

[23] Mongan, J., Moy, L., & Kahn, C. E. Jr. (2020). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiology: Artificial Intelligence, 2*(2), e200029. https://doi.org/10.1148/ryai.2020200029

[24] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems, 30*.