Advanced Cobb Angle Measurement with Integrated Uncertainty Scoring Method Using Deep Learnings

Shinyoung Park
Ewha Womans University Engineering Department
sarah2415@ewhain.net

Yeonwoo Nam
Ewha Womans University Engineering Department
yeon430@ewhain.net

Yusun Choi
Ewha Womans University Artificial Intelligence Department
onecatrabbit@ewha.ac.kr

Sang Wouk Cho
Ewha Womans University Engineering Department
swcho24@ ewha.ac.kr

Abstract – The manual measurement of spinal deformities using spine X-ray images—a process that is time-consuming and subject to variability. This study proposes an advanced method for Cobb angle measurement that not only predicts spinal angles but also provides uncertainty scores, enhancing the reliability of the results. By combining YOLOv8 for accurate vertebral detection with Faster R-CNN to reinforce challenging predictions, we significantly improve the consistency and accuracy of Cobb angle measurements, achieving an average IoU score to 0.91. The integration of a scoring system based on IoU and confidence scores further reduces prediction uncertainty, offering clinicians additional insight into the reliability of the AI-generated results. This approach has the potential to streamline clinical workflows and improve patient outcomes by providing more reliable and efficient assessments.

Keywords – AI Trustworthy, Cobb angle, Medical AI, Uncertainty, Vertebrae Detection

I. Background

The spine's natural curvature plays a crucial role in maintaining balance, providing flexibility, and absorbing external shocks, which are essential for posture, mobility, and proper nervous system function. In particular, the lumbar vertebrae bear the body's weight, and repetitive or excessive movements can lead to lower back pain. In most clinical settings, physicians manually measure the degree of spinal deformity using anteroposterior (AP) and lateral (LA) X-ray images. However, this method is time-consuming and poses challenges when the resolution of X-ray images is low, making angle measurement difficult. Additionally, the process of determining spinal tilt can be subjective, leading to inconsistencies in diagnosis (Zhu et al., 2024). This study focuses not only on predicting spinal angles but also on providing uncertainty information about the AI-generated results to enhance reliability. Our model could help users gain confidence in AI predictions, enabling more informed and cautious decision-making.

II. Method

1. Cohort Data Collection

We used the Burapha Spine dataset version 0.1, which was created through a collaboration between Burapha University, Thailand, and the Korea Institute of Oriental Medicine, Republic of Korea. This dataset contains demographic and disease characteristics of 400 individuals, 273 females, and 127 males. Each patient includes two images: AP and LA views.

2. Vertebral Detection Method

1) Model development

We used random stratified sampling to divide the dataset into training (60%), validation (20%), and test sets (20%) for lumbar vertebrae detection with YOLOv8 and Faster R-CNN. Only bounding boxes with a confidence score above 0.60 were used. YOLOv8 showed superior performance and fast processing but occasionally missed fine details, which Faster R-CNN compensated for with more precise detection. By combining YOLOv8's speed and Faster R-CNN's precision, we developed a more accurate vertebra detection system.

2. Cobb Angle and Uncertainty Measurement

In this study, The Cobb angle was measured using the method from Wu et al. (2021), evaluating spinal deformity with both AP and LA radiographic images.

The centroids of each vertebra were calculated, and a smooth curve representing the spine's curvature was generated using spline interpolation. This curve accurately reflects the natural shape of the spine. Tangents and perpendicular lines were computed along the curve. By calculating the angles between all pairs of perpendicular lines, the maximum angle was identified as the Cobb angle. This method was applied consistently to both AP and LA images.

3. Uncertainty Estimation

This system evaluates the reliability of the measurements on a scale from 1 to 5, based on the bounding box predictions from the Faster R-CNN and YOLOv8 models as well as the confidence score from YOLOv8. The system was designed to account for potential scenarios that could arise during the experiments, ensuring these factors are reflected in the reliability assessment. The uncertainty estimation follows two criteria:

1) IoU Comparison Between Faster R-CNN and YOLOv8:

The Intersection over Union (IoU) values of the bounding box predictions for each lumbar vertebra are compared between Faster R-CNN and YOLOv8 models. A threshold of 0.80 is applied; all five lumbar vertebrae in an image must have IoU values greater than 0.80 to be considered reliable.

2) Confidence Score calculation from YOLOv8:

When using YOLOv8 alone, all classes must be detected with a confidence score above 0.80. The system also checks for overlapping vertebrae by calculating the IoU, flagging predictions as uncertain if the IoU is below 0.50.

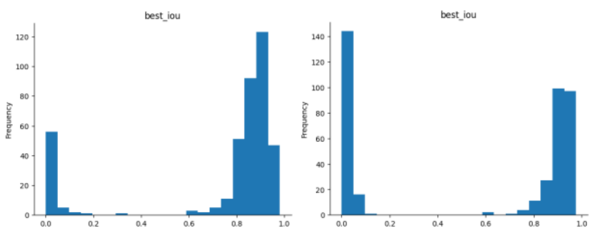

Figure1. Distribution of Best IoU for Object Detection in AP and Lateral X-ray Views

The graphs show the IoU values for Faster R-CNN and YOLOv8 predictions for both AP and LA views, which led to the selection of 0.80 as the uncertainty threshold based on the IoU criterion.

Table1. Uncertainty Scores for Vertebral Detection in X-ray Images

Category	Description	Score
1	Both criteria in ‘AP’ and ‘LA’	5
2	Both criteria in ‘AP’, Criterion 1 in ‘LA’	4
3	Both criteria in ‘AP’, Criterion 2 in ‘LA’	4
4	Both criteria in ‘LA’, Criterion 1 in ‘AP’	4
5	Both criteria in ‘LA’, Criterion 2 in ‘AP’	4
6	Criterion 1 in ‘AP’, Criterion 2 in ‘LA’	3
7	Criterion 2 in both ‘AP’ and ‘LA’	3
8	Criterion 1 in ‘LA’, Criterion 2 in ‘AP’	3
9	Criterion 2 in both ‘AP’ and ‘LA’	3
10	Both criteria in ‘AP’, not in ‘LA’	2
11	Both criteria in ‘LA’, not in ‘AP’	2
12	Else	1

III. Result

1. Basic Clinical Information of Participants

The average age of all participants was 50.21 years, with a standard deviation of 21.57. For female participants, the average age was 49.22 years, with a standard deviation of 22.37, while the average age for male participants was 52.35 years, with a standard deviation of 19.65. Among all participants, 16% had a diagnosed condition, with a disease prevalence of 17.6% in females and 12.6% in males.

2. Detection

The performance evaluation of the YOLOv8 and Faster R-CNN models on lumbar AP and LA images showed that YOLOv8 generally achieved higher IoU scores, demonstrating superior performance. Notably, YOLOv8 recorded the highest mean IoU for L4, with low standard deviation, indicating high consistency in its predictions. In contrast, while Faster R-CNN performed well in detecting bounding boxes, it exhibited higher variability due to frequent errors in correctly labeling the vertebrae, leading to greater inconsistency compared to YOLOv8. Based on these findings, a scoring system was developed with YOLOv8 as the base model, using the results from Faster R-CNN to assess measurement reliability.

Table2. IoU Comparison Across Detection Models

mIoU	AP (Mean ± Std)		Lateral (Mean ± Std)	
	R-CNN	YOLO	R-CNN	YOLO
L1	0.70±0.32	0.76±0.20	0.54±0.42	0.84±0.24
L2	0.71±0.33	0.75±0.19	0.57±0.43	0.85±0.22
L3	0.75±0.32	0.76±0.20	0.57±0.44	0.88±0.21
L4	0.74±0.30	0.75±0.19	0.56±0.43	0.91±0.11
L5	0.77±0.19	0.74±0.17	0.54±0.44	0.90±0.17

3. Uncertainty

In this study, we proposed a novel scoring system and uncertainty criteria to evaluate the reliability of automated Cobb angle measurements from spinal X-ray images. The effectiveness of the system was validated by analyzing the predicted results. The distribution of images satisfying the uncertainty criteria within each scoring category was as follows: Category 1 included 21 images, Category 3 included 17 images, Category 4 had 2 images, Category 5 had 12 images, Category 6 had 1 image, Category 7 had 10 images, and Categories 10, 11, and 12 included 0 images.

Overall, 21 images received the highest confidence score of 5, 29 images were rated with a confidence score of 4, and 11 images were rated with a confidence score of 3. Out of a total of 81 test images, 61 images were given high confidence scores, and alerts were triggered for 20 patients, indicating that only approximately 25% of the X-ray images required expert review.

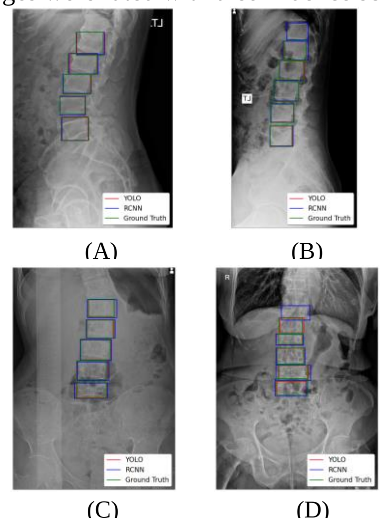

Figure 3. Comparison of YOLO and RCNN Localization Performance on Lateral and AP X-ray Views with Associated Uncertainty Scores - (A) and (B): Lateral views with uncertainty scores of 5 and 4, respectively. (C) and (D): Anteroposterior views with uncertainty scores of 5 and 1, respectively

To evaluate the performance, we compared the IoU values of predicted lumbar regions with ground truth in both AP and LA views. In cases where no alerts were triggered, the system showed strong reliability, with IoU scores of 0.89, 0.88, and 0.87 for confidence levels 5, 4, and 3 in the AP view, respectively. Similarly, in the LA view, confidence scores of 5, 4, and 3 all achieved an IoU of 0.91, reflecting consistent and reliable detection.

For cases in Category 1, which had the highest confidence score of 5, the system demonstrated excellent performance with a mean absolute error (MAE) of 1.55° for the LA view and 1.65° for the AP view, indicating minimal error. The R^2 values were 0.95 for the LA view and 0.94 for the AP view, showing that the model explained the variation in actual Cobb angles well. The mean absolute deviation (MAD) was 6.26° in the LA view and 6.58° in the AP view, further demonstrating consistency.

In conclusion, the scoring system effectively evaluates the accuracy and reliability of Cobb angle measurements. The high accuracy observed in the highest-scoring cases indicates that the system can reliably distinguish between cases suitable for automated measurement and making it a valuable tool for improving clinical measurement efficiency and accuracy.

IV. Discussion

This study advanced existing methods for vertebral detection and Cobb angle measurement by integrating Faster R-CNN with YOLOv8, improving the accuracy and reliability of spinal curvature assessments using AP and LA X-ray images. Building on the Wu et al. (2021), we enhanced vertebral detection and applied a voting strategy to reduce prediction uncertainty, providing more reliable measurements for evaluating spinal deformities.

Unlike previous studies that focused on segmentation or landmark-based models without addressing prediction confidence, this research not only prioritized accuracy but also quantified the uncertainty in Cobb angle measurements. This additional reliability information is crucial for clinical decision-making. The use of a 400-image open-access dataset provided a strong foundation, but future work will aim to expand the dataset to improve model generalization across more diverse clinical scenarios.

V. Conclusion

This study employed a combined approach using YOLOv8 for vertebrae detection and Faster R-CNN to reinforce YOLOv8 where needed, leading to more accurate Cobb angle measurements. By comparing IoU and confidence scores, we improved detection reliability and reduced uncertainty, providing clinicians with valuable information for assessing spinal deformities. Future work will focus on expanding the dataset to include more patients and anatomical variations, aiming to enhance model generalization and applicability in diverse clinical settings.

VI. Reference

Zhu, Y., Yin, X., Chen, Z., Zhang, H., Xu, K., Zhang, J., & Wu, N. (2024). Deep learning in Cobb angle automated measurement on X-rays: A systematic review and meta-analysis. *Spine Deformity*. https://doi.org/10.1007/s43390-024-00954-4

Wu, H., et al. Deep learning automates Cobb angle measurement compared with multi-expert observers. Eur Spine J 30, 2362–2371 (2021). https://doi.org/10.1007/s00586-021-06818-z

Klinwichit, P., Yookwan, W., Limchareon, S., Chinnasarn, K., Jang, J.-S., & Onuean, A. (2023). BUU-LSPINE: A Thai open lumbar spine dataset for spondylolisthesis detection. *Applied Sciences*, 13(15), 8646. https://doi.org/10.3390/app13158646