Journal of Computer Science

Research Article

Senior Expertise and Peer Consensus: A Comparative Analysis of AI and Clinician Measurements in Multi-Curve Scoliosis Assessment

Frank Ibañez Elijorde[1], Joselito F. Villaruz[2], Ma Beth S. Concepcion[3] and Mylo N. Soriaso[4]

[1]*Division of Information Technology, College of Information and Communications Technology West Visayas State University, Iloilo, Philippines*
[2]*Department of Pediatrics, College of Medicine, West Visayas State University, Iloilo, Philippines*
[3]*Department of Information Systems, College of Information and Communications Technology West Visayas State University, Iloilo, Philippines*
[4]*Department of Orthopedics, West Visayas State University Medical Center, Iloilo, Philippines*

Article history
Received: 07-07-2025
Revised: 22-10-2025
Accepted: 25-11-2025

Corresponding Author:
Frank Ibañez Elijorde
Division of Information Technology, College of Information and Communications Technology, West Visayas State University, Iloilo, Philippines
Email: felijorde@wvsu.edu.ph

Abstract: Given the scarcity in the literature, this study explored the use of AI for multi-curve scoliosis assessment. Its performance was analyzed through comparison against a group of clinicians composed of one senior and five non-senior orthopedic surgeons. The analysis focused on Cobb angle measurement and identification of vertebral endplates across three curve regions, namely Main Thoracic (MT), Proximal Thoracic (PT), and Thoracolumbar/Lumbar (TL/L). As evidenced by the results, there is a strong agreement in the MT region, as shown by low Mean Absolute Differences (MAD) at 2.21 and high interclass correlation coefficients (ICC 0.94 –0.98), suggesting the clinically reliable performance of AI in this area of the spine. Meanwhile, moderate agreement was observed in the TL/L region (ICC 0.74–0.89), although the PT region presented significant challenges, with high MAD values and ICC values near zero. This highlights variations in end vertebra selection due to anatomical and image quality limitations, which significantly affect the respective Cobb angle measurements of the human observers. On the other hand, subjectivity in identifying vertebral landmarks, which is apparent in low-quality radiographs, was revealed through qualitative observations. An interesting finding is that most of AI's measurements aligned more closely with the group consensus of non-senior clinicians than with the senior expert, possibly signifying its inclination towards combined human patterns rather than expert-level preference. Caudal endplate identification showed higher agreement across evaluators than cranial endplates, implying that certain anatomical landmarks are more consistently identifiable. This result is indicative of AI's potential for standardizing scoliosis evaluation, particularly in the MT and TL/L region, despite its underperformance in the PT region. Thus, it concludes that there is a need to emphasize enhanced algorithm development, improved training datasets, and above all, to integrate expert oversight. The alignment of AI with general clinician consensus underlines its potential as a reliable, standardizing tool in clinical practice, but it is imperative that expert input remains a crucial part of the study.

Keywords: Artificial Intelligence and Deep Learning, Scoliosis Assessment, Multi-Curve Scoliosis, Vertebral Endplate Selection, Cobb Angle Measurement

Introduction

Scoliosis is defined as a lateral spinal curvature with a Cobb angle of 10 degrees or more, often accompanied by vertebral rotation (Karpiel et al., 2021). The most widely used measurement of spinal curvature is the Cobb angle, where greater than 10–12° of lateral curvature is considered abnormal. When a scoliotic curve exceeds 40 degrees, it can result in many physiologic symptoms (Kim et al., 2010). It may exist in several locations, like the thoracic, lumbar, or thoracolumbar spine. Scoliosis is frequently divided into structural and nonstructural categories, each of which has unique characteristics. The underlying etiology in the vast majority of instances is unknown; however, it can be the consequence of a congenital or developmental osseous or neurologic abnormality (Trobisch et al., 2010).

To measure the Cobb angle, you need to identify the end vertebrae of the scoliotic curve. These are the vertebrae whose endplates are most tilted toward each other. The angle where these lines intersect is the Cobb angle (Cobb Angle for Scoliosis: Definition and Uses, 2024). Manual Cobb angle measurement has always been the reference standard, with the manual approach's drawbacks including intra- and inter-observer variability as well as the time-consuming nature of the procedure. While it has been the reference standard, manual Cobb angle measurement has shortcomings. For instance, different observers may identify slightly different end vertebrae or draw the lines with slight variations, leading to inconsistent measurements (Langensiepen et al., 2013). Furthermore, manually measuring Cobb angles can be a labor-intensive and time-consuming process (Li et al., 2025).

Accurate and comprehensive scoliosis assessment is crucial for effective patient management. It involves evaluating structural curves, underlying causes, severity, and growth potential (Parr and Askin, 2020). Early diagnosis and referral to specialized care can improve outcomes. By considering a wide range of outcome criteria, clinicians can provide a more patient-oriented approach to scoliosis management, focusing on both present and future needs of the individual (Négrini et al., 2006).

Artificial Intelligence (AI) is revolutionizing spinal imaging and diagnostics, particularly in spinal deformity screening. AI technologies are being applied to improve image quality, automate anatomical measurements, and detect spinal pathologies with expert-level accuracy (Lee et al., 2024; Meng et al., 2022). Given these significant improvements, AI may increase diagnostic precision, speed up the workflow, and help doctors make decisions (Khalifa and Albadawy, 2024). By automating initial image interpretation, AI can change the clinical workflow of radiographic detection and the assessment of the effect of disease and treatment on nearby organs (Bi et al., 2019). Although the use of AI in scoliosis evaluation is still in its early stages, it is expected to transform spine care by offering sophisticated tools for diagnosis, treatment planning, and patient monitoring. While AI presents transformative opportunities for improving patient care, challenges remain in clinical implementation, including model interpretability, generalizability, and data limitations.

Although there has been a lot of progress in the use of AI for scoliosis assessment, there is still a gap in the literature regarding its capacity to evaluate scoliosis across all curve locations (Martín-Noguerol et al., 2023; Zhang et al., 2023). The majority of studies concentrate on a single curve, such as the main thoracic curve; however, the upper and lower curves adjacent to the main curve are also equally important for making treatment decisions (Fuleihan et al., 2024; Retson and Eghtedari, 2023; Wirries et al., 2021). In the treatment of AIS, the utilization of Large Language Models (LLMs), such as advanced AI-based chatbots, has revolutionized patient education and information dissemination. For instance, ChatGPT and Scholar AI demonstrated high accuracy in classifying single-curve scoliosis severity (Fabijan et al., 2023a). However, they showed limitations in Cobb angle assessment and curve direction determination when analyzing radiographic images. In another study, a machine learning model was shown to achieve significant correlation with clinical reports in Cobb angle measurements (Ha et al., 2022). Meanwhile, an AI model that was trained on a considerable amount of data demonstrated excellent performance in automated Cobb angle measurements, with high accuracy in end vertebrae selection and strong agreement with radiologists (Fabijan et al., 2024). Though promising, these studies primarily focused on single-curve scoliosis or general Cobb angle measurements, indicating a potential gap in AI's ability to assess multiple curve locations comprehensively. While AI has shown strong potential in scoliosis evaluation, there remains a gap in the literature regarding its ability to assess scoliosis across all curve locations.

This study uniquely addresses the junction of clinical expertise and digital precision by focusing on two under-examined challenges in multi-curve scoliosis assessment:

a) The discrepancy in endplate selection among human evaluators of varying experience that drives measurement variability
b) The capacity of an AI model to provide consistent, objective endplate selection across all three common curve locations

By utilizing an AI model as a fixed, objective reference marker, we can isolate and quantify the level of subjective endplate selection bias among experts. Thus, the objective of this study is to investigate the difference in curve detection, endplate selection, and measured Cobb angles between an AI system and human clinicians across multiple curve locations, using both high and low-quality radiographs to emphasize the specific challenges faced by human raters.

Materials and Methods

Study Design

The study is intended to assess how AI-based evaluations compare with those of clinicians, including a senior expert, in detecting scoliosis-relevant vertebral landmarks and in measuring spinal curvature in terms of Cobb Angle across the Proximal Thoracic (PT), Main Thoracic (MT), and Thoracolumbar/Lumbar (TL/L) regions. It employs a comparative analytical approach to evaluate agreement levels between AI-assisted vertebral endplate identification and human expert assessments.

Furthermore, it involves six orthopedic surgeons, with one designated as a Senior Expert (SE). The assessment was conducted under four distinct scenarios to investigate the influence of clinician grouping and the presence of a senior expert, which include:

- AI vs. SE: Comparison between the AI system and the Senior Expert
- 2-6 vs. AI: Comparison between the collective measurements of Non-Senior Clinicians (2-6) and the AI system
- 1-6 vs. AI: Comparison between the collective measurements of all six Clinicians (1-6) and the AI system
- 2-6 vs. SE: Comparison between the collective measurements of Non-Senior Clinicians (2-6) and the Senior Expert

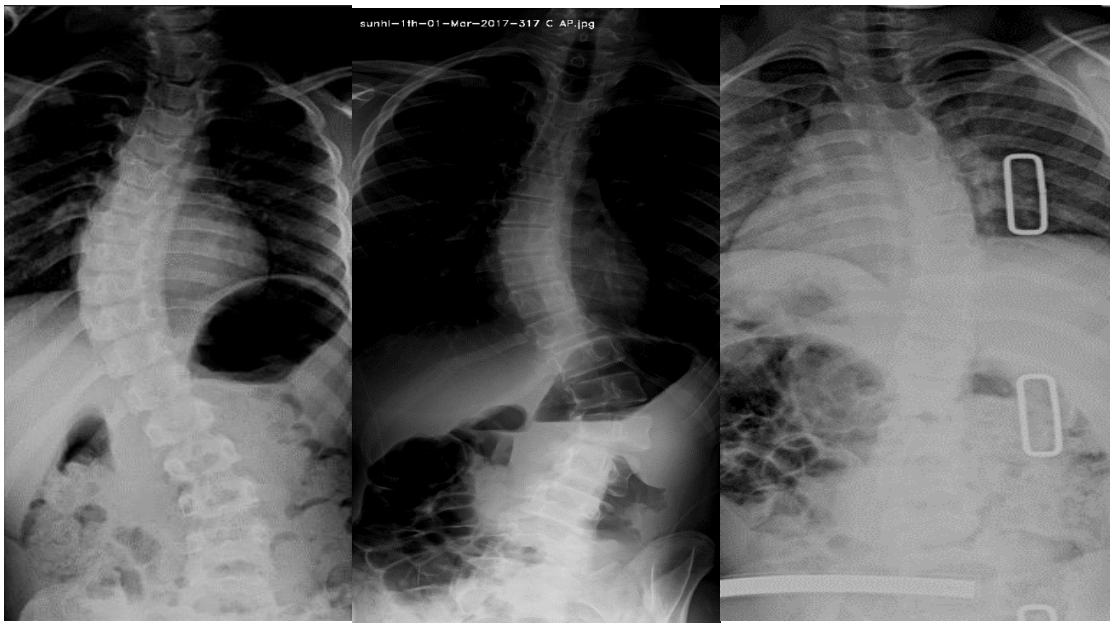
For each radiographic assessment, the following were recorded per spinal curve region, with variations documented for comparative analysis:

- Cranial Endplate (CrE) and Caudal Endplate (CaE) identification
- Variations in endplate identification among assessors
- Cobb angle measurements derived from the identified endplates
- Presence or absence of scoliosis as indicated by spinal curves

The quantitative results were analyzed in order to assess the consistency of AI measurements compared to clinicians, to determine the influence of seniority on agreement, and to determine the consistency of curve identification and measurement on a given spinal region.

Data Collection

The images used in this study were derived from the dataset of the MICCAI 2019 challenge on Accurate Automated Spinal Curvature Estimation, composed of anterior-posterior X-ray images (Hongbo et al., 2018). As shown in the samples in Fig. 1, the quality of images was intentionally varied to simulate the real-world challenges faced by clinicians in dealing with low-quality X-ray images. Some of the difficulties exhibited in the images include a) Image Quality and Artifacts - x-ray images can suffer from poor resolution, motion blur, or artifacts caused by improper positioning, leading to difficulties in endplate selection and CA measurement, which can introduce errors in scoliosis assessment; b) Distortion and Overlapping Structures – due to the two-dimensional nature of X-ray images, diagnosis accuracy can be affected due to difficulty in distinguishing overlapping anatomical structures (Fabijan et al., 2023b).



a) good quality image b) low quality image c) low quality image with artifacts

Fig. 1: Sample images used in the measurement comparison

The 10 samples for assessment were chosen using a maximum variation sampling strategy from the 128-image dataset to maximize the complexity presented to the evaluators. Each image requires the AI and clinicians to perform three interdependent curve detections and measurements (MT, PT, TL/L), where the resultant Cobb angle is highly dependent on the subjective selection of end-vertebral plates by the experts. Therefore, the total number of primary curve measurements analyzed is 30 (10 images x 3 regions). This high level of dependence and subjectivity in endplate selection across multiple curves per image maximizes the clinical variance and challenge within this subset, producing a more vigorous and clinically relevant measure of agreement and disagreement.

Data Analysis

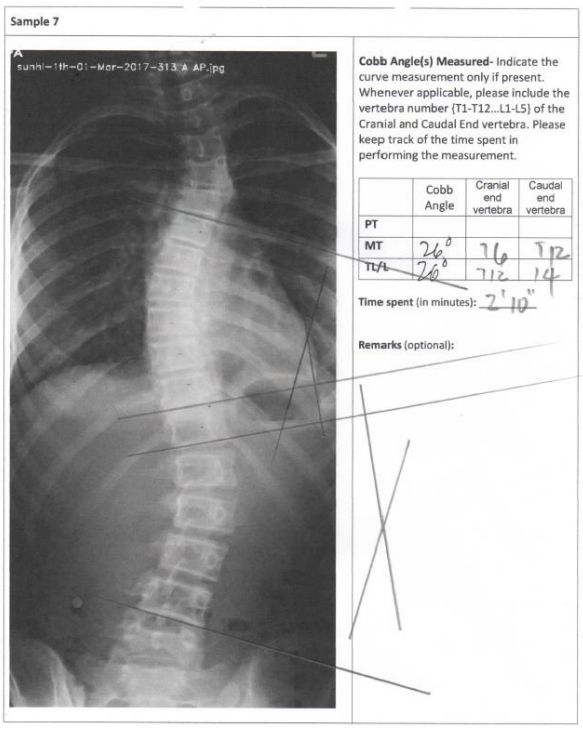
Descriptive methods, which outline consensus and variations, are used to analyze vertebral endplate agreement. According to the observed consensus and few deviations, the overall agreement for vertebral endplates was interpreted as "High" to "Low." The robustness of identifying the vertebral endplate would be indicated by a high overall agreement across curve locations. Mean Absolute Difference (MAD) and Intraclass Correlation Coefficient (ICC) were the two main metrics used to measure the agreement in Cobb angle measurements. Four comparison groups were subjected to these analyses at the MT, PT, and TL/L curve locations. To measure the differences in Cobb Angle measurements between AI and human observers, MAD was derived. Higher MAD values imply measurement variability, whereas lower values show stronger alignment. Additionally, ICC values were calculated to evaluate the reliability of measurements across scenarios, particularly between AI and clinicians. ICC values range from -1 to 1, with values closer to 1 indicating higher reliability and agreement. ICC values were interpreted as follows: Values <0.5 indicated poor reliability, 0.5– 0.75 moderate reliability, 0.75–0.9 good reliability, and >0.9 excellent reliability.

Results and Discussion

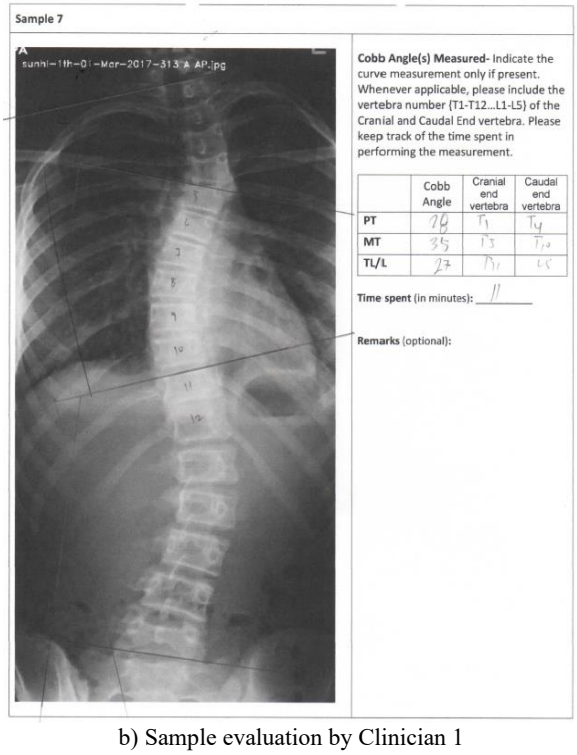
Curve Location and Endplates Selection

Two sample images were taken into consideration to present some interesting observations. These samples highlight differences in both curve identification and endplate selection, which directly impact the measured Cobb angle. For instance, in Fig. 2, Sample 7 was evaluated by the SE and three clinicians. From the given example, a number of observations are prominent. The SE and Clinician 3 both identified two curves within the MT and TL/L area, whereas Clinicians 1 and 3 identified three curve locations. For the more experienced evaluators, they find it unnecessary to mark locations with negligible curvature, such as the PT curve, as opposed to the less experienced ones, who at times would require marking each vertebra for reference.

Although there is an agreement with regard to the possible presence of curves in their respective location, there also exists a difference in the selection of the vertebral endplates, which has a direct impact on the perceived curve measurement by the evaluators.

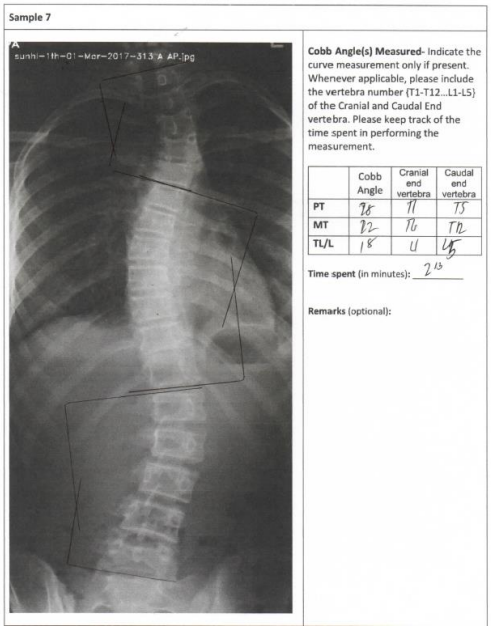


a) Sample evaluation by SE

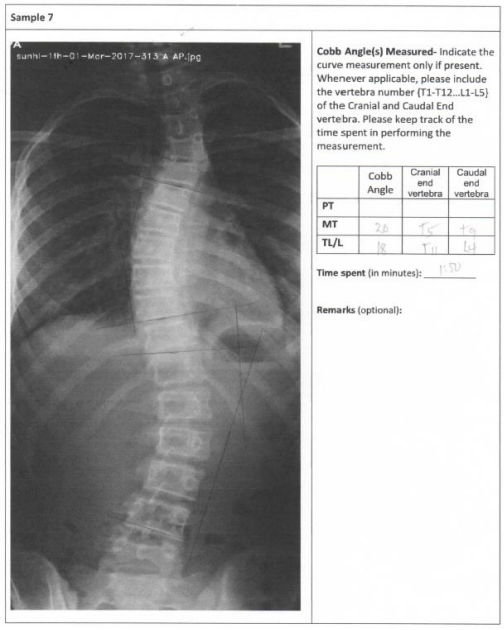


b) Sample evaluation by Clinician 1

DOI: 10.3844/jcssp.2026.886.897



c) Sample evaluation by Clinician 2



d) Sample evaluation by Clinician 3

Fig. 2: A sample image as evaluated by the senior expert and three other clinicians

Taking the case of MT curve disagreement, the Cobb angle for the MT curve ranged from 20˚ (Clinician 3) to 35˚ (Clinician 1), showing a 15˚ disparity. This wide range can be directly attributed to the variation in the selection of endplates. For instance, the caudal endplate selection varied substantially from T9 (Clinician 3) to T12 (SE and Clinician 2) - a difference of up to three vertebral levels. As for the cranial endplate, it also showed minor variance between T5 (Clinician 1 and 3) and T6 (SE and Clinician 2). This indicates that despite the clear image, the subjective decision of defining the "most tilted" vertebra significantly impacts the final measurement. As for the TL/L curve agreement and disagreement, it also showed a bimodal distribution of measurements. The SE (26˚) and Clinician 1 (27˚) readings were in close agreement, while Clinician 2 (18˚) and Clinician 3 (18˚) were also in agreement, but with a significant 8-9˚ difference between the two groups. This divergence was again associated with the selection of the caudal endplate, which was split between L4 (SE and Clinician 3) and L5 (Clinician 1 and 2). This difference in vertebral endplate selection has a direct impact on how evaluators perceive curve measurements.

In Fig. 3, a low-quality image, Sample 2, was intentionally included to assess the impact of poor visibility on measurement accuracy and consistency. As can be seen, the thoracolumbar and lumbar sections of the spine are barely visible, which would lead to significant differences in the assessment of the evaluators. All four observers commented on the low quality of the image, which had a direct impact on the duration and accuracy of their assessment. In this example, only one clinician labelled all three curve locations with their respective endplates and corresponding curve measurement; on the other hand, the SE and the other two clinicians only marked a single curve location. Interestingly, although the three of them labelled just one curve location, the SE, clinician 2, and clinician 3 identified different locations in the areas of TL/L, PT, and TL/L, respectively.

First observation, there is an extreme variability in curve identification. The number of identified curves ranged from one (according to SE and Clinician 2, 3) to Clinician 1 identifying three curves. This wide span confirms that low image quality severely degrades the reliability of curve identification itself. Second, the greatest measurement disparity lies along the TL/L curve measurements. In this case, the SE measured the TL/L curve at 19˚, Clinician 3 measured it at 18˚, and, on the contrary, Clinician 1, who identified three curves, measured the TL/L curve at an outlying 6˚. Last is the endplate selection as an indication of difficulty. The difficulty in landmark selection is evident in the cranial endplate selection of the TL/L curve, which varied between T5 (Clinician 3), T7 (SE), and L1 (Clinician 1). A three-vertebra discrepancy for the endplate selection confirms the detrimental effect of low image clarity on assessment accuracy, as described in the general remarks by all four observers.

Although not included in the analysis of this study, there is also an obvious difference in the amount of time spent by each evaluator in the assessment of the X-ray images.

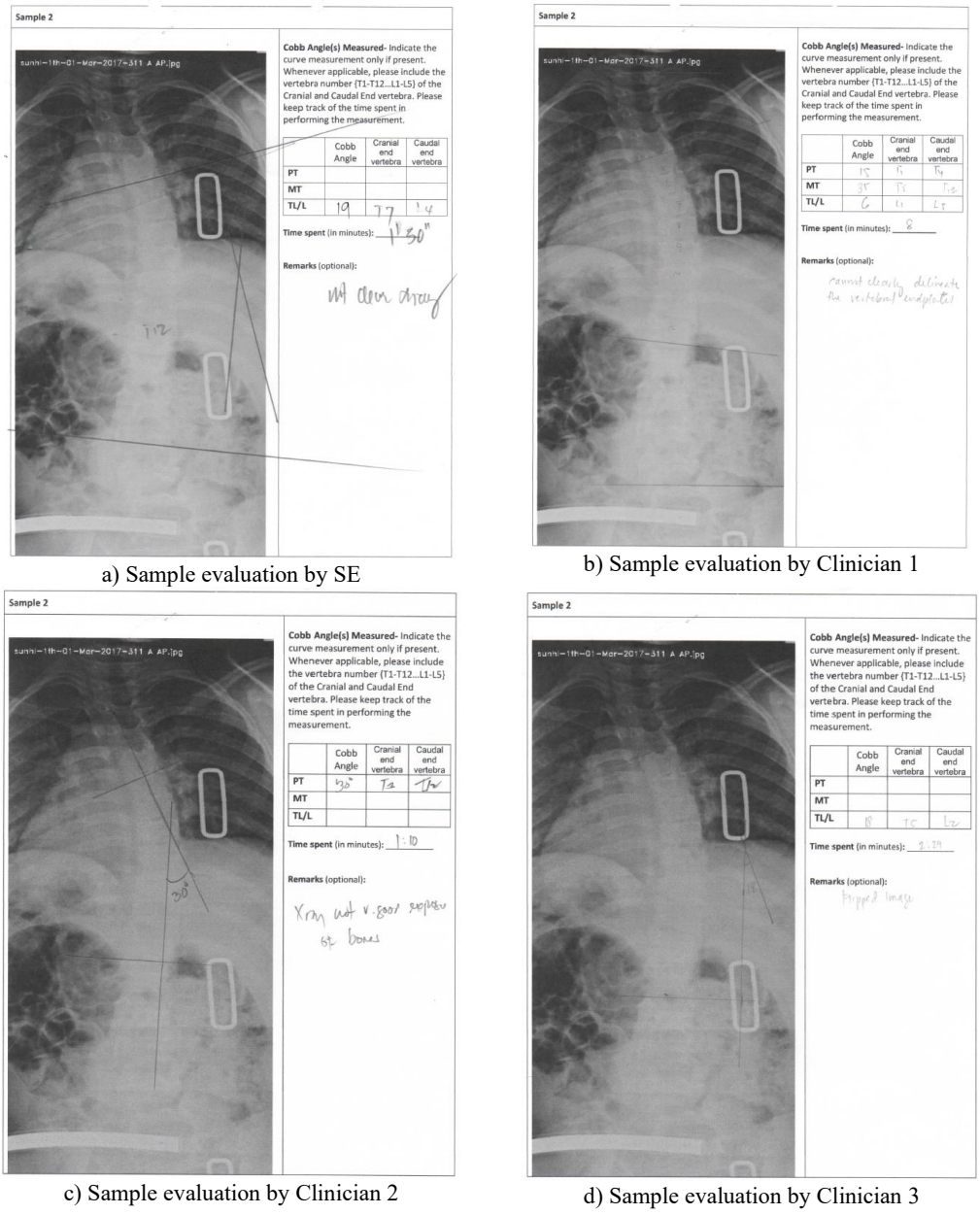


Fig. 3: A low quality sample image as evaluated by the senior expert and three other Clinicians

For instance, in both samples, there is only a moderate time difference between the SE and Clinicians 2-3, while a significant amount of time was required by Clinician 1, who often identified more curves and more extreme endplates to complete the assessment of both images. This suggests that less experience, coupled with poor image quality, substantially increases the duration and inconsistency of the assessment process.

In Table 1, the summary data for the Vertebral Endplates Agreement per Curve Location in four different scenarios are shown. Starting with the MT area, as indicated in Scenarios 1 and 2, Clinicians 1-6 and AI generally agree on cranial endplate (CrE) at T4/T5 with minor variations at T3/T4. The same can be observed for the caudal endplate (CaE), which his consistent at T12, with minor variations at either L1 or L2. For scenarios 3 and 4, whether grouped as Clinicians 2-6 or joined by the SE, they consistently identified T5 and T12 as the CrE and CaE, respectively. As for the variations, both scenarios agree at L1 as the CaE, while they are inconsistent in CrE varying at T3, T4, and T6. Considering the PT area, the four scenarios indicate that the majority of assessments agree that there is no scoliosis found, as depicted by the absence of values in both the consensus CrE and CaE.

Table 1: Vertebral Endplates Agreement Per Curve Location

Scenario		Curve Location	Consensus CrE	Variations in CrE	Consensus CaE	Variations in CaE	Overall Agreement
1	Vertebral Endplates Agreement among 6 Clinicians (1-6) and AI	MT	T4	T3	T12	L1, L2	High
		PT	-	T1	-	T5	High
		TL/L	-	L1	-	L5	High
2	Vertebral Endplates Agreement among Non-Senior Clinicians (2-6) and AI	MT	T5	T4	T12	L1	High
		PT	-	T1	-	T5	High
		TL/L	-	L1	-	L5	High
3	Vertebral Endplates Agreement among 6 Clinicians (1-6)	MT	T5	T3, T6	T12	L1	High
		PT	-	T1	-	T5	High
		TL/L	-	L1	-	L5	High
4	Vertebral Endplates Agreement among Non-Senior Clinicians (2-6)	MT	T5	T4	T12	L1	High
		PT	-	T1	-	T5	High
		TL/L	-	L1	-	L5	High

As for the variations in CrE and CaE, all scenarios are consistent in identifying T1 and T5. With regard to the TL/L area, the agreement among the four scenarios remains. The majority of assessments point out the absence of tilted vertebral endplates that would indicate a scoliosis. Variations in CrE and CaE are minimal and are limited only to L1 and L5.

It is apparent that consensus on the caudal endplate is stronger compared to that of the cranial endplate. Taking a closer look, the consensus on CaE appears more consistent than on CrE. All scenarios indicate that the consensus CaE for MT is T12, while it is split between T4 & T5 for CrE. A similar observation also applies to the variations in both endplates. As shown, the CrE values yielded five variations (T1, T3, T4, T6, L1) while CaE is only limted to three (L1, L2, T5). Looking further at PT and TL/L, the consensus in CrE and CaE is consistent across all scenarios. In general, there are minimal differences in CrE identification, particularly for the MT curve location. However, even with these deviations, the overall agreement is still considered high.

At this point, it is worth pointing out that seniority does not significantly impact the collective agreement of the clinicians in identifying the vertebral endplates. Comparing Scenarios 1 and 2 (all clinicians vs. AI and non-senior clinicians vs. AI) and Scenarios 3 and 4 (all six clinicians and non-senior clinicians only) shows minimal differences. This indicates that the level of experience among clinicians in this study does not significantly affect their ability to identify vertebral endplates. Interestingly, Scenario 2 reveals that AI aligns more closely with non-senior clinicians as opposed to Scenario 1, in which the slight variations in CaE (L1, L2 for MT CrE) are present. It can also be observed in Scenario 4 that Clinicians 2-6 tend to have better agreement with each other than with the senior expert; this can be observed in the single variation both in CrE and CaE across all three curves. Meanwhile, Scenario 3 reveals the variations in identifying CrE after the introduction of the SE's readings. These observations could be attributed to differences in expert-level interpretations, whereas the senior expert may have stricter criteria that influence the slight discrepancies between AI and humans. The AI's identification of endplates aligns well with the clinician's consensus. For instance, Scenarios 1 and 3 (all clinicians with AI vs. without AI) show only slight differences in their consensus in the MT curve (T4, T5 for MT CrE, T12 for CaE).

Across all scenarios and curve locations, it is concluded that the overall agreement is consistently "High" as indicated by consensus that is either identifying vertebral endplates that would constitute the scoliosis curve, or conforming to the absence thereof, which implies there are no spinal curvatures. This observation applies regardless of whether AI is included or which clinician group is considered.

Cobb Angle Measurement Variation

In Table 2, the Mean Absolute Difference (MAD) gives insights as to how Cobb Angle measurement varies between clinicians and AI for different spinal curves. Lower values indicate stronger agreement, while higher values suggest greater measurement discrepancies. As can be observed above, AI performs well in MT curve estimation, which can be attributed to the more distinct vertebral landmarks and higher measurement certainties. Based on this observation, when AI is compared to the SE, the MAD value at 3.22 indicates that AI closely matches the senior expert's measurements. It is worth noting that when the five non-senior clinicians are grouped together and compared to AI, the MAD is reduced to 2.81, which is a significant improvement from the previous scenario. Meanwhile, including the SE in the group further reduces the MAD down to 2.21, which is indeed desirable. In addition, comparing the collective assessment of the other clinicians to that of the SE reveals a high MAD at 3.74, which is still acceptable, although significantly higher than the discrepancy levels when comparing both clinician groups to AI. This could be a hindsight to possible disagreements between the readings of the SE

and the five other clinicians due to differences in expertise levels.

Looking at the PT curve, large discrepancies could be observed, especially at a MAD value of 13.22 between AI and the SE, while the lowest value of 5.27 is seen between the non-senior clinicians and AI. Similar to the pattern observed in the MT curve, the collective assessment of the other clinicians is closer to that of AI as opposed to when their assessments were either paired against or put together with the SE's readings. This shows that this particular area of the spine is challenging to measure consistently, irrespective of the scenarios being compared.

Moving on with the TL/L area, the MAD across all scenarios shows moderate agreement. For instance, a noticeable variation between AI and SE is found at 9.02. Meanwhile, comparing Clinicians 2-6 and Clinicians 1-6 vs. AI yields 5.18 and 5.74, respectively, which shows AI's reasonable alignment with clinicians. Finally, comparing the readings of the other clinicians with that of the SE is at 5.74, which indicates a moderate variability among themselves. It is indicative of a higher level of agreement in this region among the AI and the clinician groups (2-6 and 1-6) as compared to pairing AI and SE only, as well as pairing the other clinicians with the SE (2-6 vs SE).

The significantly low discrepancies in the MT area indicate that there was no difficulty on the part of the clinicians, regardless of the level of expertise. The same observation applies to all scenarios that involve comparison with AI. What contributes to this is the highly distinctive vertebral endplates in the said region, which is the easiest to locate, therefore making the measurement process more straightforward. While the AI vs. SE difference is still notable in the TL/L region, the comparisons between AI and the other observer groups (2-6 and 1-6) show the lowest MAD values for specific curve locations. Interestingly, AI tends to have a lower difference with the collective measurements of clinician groups.

The MAD values for "2-6 vs AI" and "1-6 vs AI" are generally lower than "AI vs SE" and even "2-6 vs SE". This suggests better alignment of the AI's measurements with the collective clinician observations rather than with the senior expert only. Meanwhile, the clinicians, including the senior expert, show slightly more variations across all curve locations, which can be attributed to differences in expertise levels.

As for the implications, these findings suggest that although AI shows a good level of agreement with human observers, there are notable differences, especially when compared only to the senior expert. Throughout all curve locations, the highest MAD value is consistently observed between the AI method and the SE measurements. Such a result suggests that the AI and SE methods of measurement have the most significant disagreements in the assessment of spinal curves indicative of scoliosis. The discrepancies observed raise the need for standardized measurement protocols and further research to determine the reason behind the observed differences. A good start would be to analyze the exact cases where the largest discrepancies occurred.

Cobb Angle Measurement Consistency

The ICC in Table 3 provides valuable insights into the consistency of Cobb angle measurements across four scenarios, specifically comparing AI-based assessments with those of clinicians. Findings show various levels of reliability across different spinal curve locations, which emphasize the strengths and limitations of AI-assisted scoliosis evaluation as compared to human observers.

Table 2: Mean Absolute Difference per curve location (in degrees)

Curve	Scenarios			
	AI vs SE	2-6 vs. AI	1-6 vs AI	2-6 vs SE
Main Thoracic (MT)	3.22	2.80	2.21	3.74
Proximal Thoracic (PT)	13.32	5.27	6.58	8.26
Thoracolumbar/Lumbar (TL/L)	9.02	5.18	5.74	5.74

Table 3: Intraclass Correlation Coefficient (ICC) Per Curve Location

Curve	Scenarios			
	AI vs SE	2-6 vs. AI	1-6 vs AI	2-6 vs SE
Main Thoracic (MT)	.96	.98	.98	.94
Proximal Thoracic (PT)	-8.7E-17	.72	.59	1.38E-16
Thoracolumbar/Lumbar (TL/L)	.74	.89	.88	.89

With regards to the MT curve, it is shown that there is an excellent agreement between the assessments of AI and senior experts, with ICC values ranging from 0.94 to 0.98 in various scenarios. The strong correlation is a good indicator that measurements by AI are highly consistent and reliable, and align closely with evaluations by the clinicians. It can also be noted that when comparing AI and both senior and non-senior clinicians, consistency assures the reliability of AI in assessing said curve location.

Meanwhile, the PT curve introduced significant potential difficulty in measurement consistency. In contrast, the ICC values for AI versus SE assessments are bordering on zero, which is extremely low, indicating no

agreement between AI and expert evaluations. Even looking at the measurement consistency among non-senior clinicians, the agreement with SE remains near zero. When compared to non-senior clinicians, AI shows moderate consistency with an ICC of 0.72; however, it goes down to 0.52 when collectively compared to all six clinicians, including the SE. Even though AI measurements have more internal consistency when compared across different groups of clinicians, the discrepancies between AI and SE imply potential methodological differences in identifying the endpoints of this curve. Finally, the TL/L curve shows moderate to excellent agreement, with ICC values ranging from 0.74 to 0.89. It can be observed that AI's correlation with SE assessments is slightly lower at 0.74; however, when compared with the collective clinician assessments, it shows stronger reliability at 0.88–0.89. As opposed to the PT curve, TL/L measurements do not exhibit extreme variability, highlighting AI's potential applicability for this segment of scoliosis classification.

With these findings, it is shown that the reliability of AI is significantly affected by the location of the curve that is to be identified and measured. As observed, AI performs remarkably well for the MT curve, indicating potential for clinical use. However, there is still a need for AI to be improved when dealing with the PT curve, as the differences between expert and AI evaluations are still noticeable. Further validation might be required before AI can be completely incorporated into clinical decision-making, despite the TL/L curve measurements indicating a reasonable level of agreement. AI exhibited good reliability for TL/L and excellent accuracy for MT; however, methodological inconsistencies could still pose a concern given the poor agreement with PT measurement across scenarios. As such, further exploration is necessary to enhance AI's accuracy in this area, especially since nearly zero ICCs for PT raise the possibility that AI and expert raters are employing different reference points or measurement methods.

Challenges in PT Curve Assessment

The PT curve consistently exhibits the highest MAD values and the lowest ICC values (bordering on zero) across all scenarios. This shows that this particular area of the spine is the most challenging to measure consistently, irrespective of the scenarios being compared. The marked divergence in the PT curve is due to the unique combination of anatomical obstruction (overlapping vertebrae, rib interference, and lower visibility) and subjective endplate selection criteria specific to this area.

On the other hand, the near-zero ICC between the AI and the SE, and between the non-senior clinicians and the SE, strongly suggests that the issue is not only a measurement difference but a fundamental methodological disagreement on the curve's existence and endplate selection. The SE often elects to omit marking a negligible PT curve (an act of clinical expertise), while the AI and less experienced clinicians may measure a small curve. This lack of shared measurement cases causes the ICC to plummet, indicating no agreement rather than just a large difference in value. This highlights the challenges in the consistent measurement of PT curves for both AI and human evaluators. The discrepancies observed, therefore, highlight the need for standardized measurement protocols, particularly focusing on defining endplate criteria for less distinct curves like the PT and TL/L, to synergize the strengths of AI's consistency and human expertise.

Conclusion

This study explored the consistency and reliability of AI-assisted scoliosis assessment in comparison to human evaluators composed of a Senior Expert (SE) and non-senior clinicians, with attention to the identification of vertebral endplates and measurement of Cobb angle across different spinal curve locations. The findings emphasized the areas of strong agreement and discrepancies, which provide insights into the potential of AI in clinical practice. From the sample images, qualitative observations were drawn and pointed out the inherent subjectivity in curve identification and selection of endplate, which is particularly evident in low-quality radiographs. Consequently, such variability in endplate selection directly influenced the perceived Cobb angle measurements among clinicians. As the selection of the "most tilted" end vertebrae can be subjective, this can lead to inter-observer variability, which is a well-recognized challenge in scoliosis assessment. Investigation by a number of studies has explored the interrater/intrarater agreement in Cobb angle measurements. Inconsistencies in the definition of the upper and lower end vertebrae were observed, while variability in curvature measurements was reported to range from 3 to 10 degrees (Stott et al., 2024; Ha et al., 2022; Mulford et al., 2024).

As evidenced by the MAD for the MT region, which is considerably within the aforementioned threshold, this suggests that measurement differences remained within clinically acceptable limits. With MAD values ranging within moderate agreement in the TL/L area, this implies that while measurable, this region is also prone to some challenges. From a closer observation, AI tends to have lower MAD values when compared to the collective assessment of clinician groups as opposed to the senior expert ("AI vs SE"). With this, it suggests that the AI system's measurements are closely aligned towards the broader consensus of clinicians, perhaps indicating a more generalizable interpretation. However, the highest MAD values across all scenarios were consistently exhibited by the PT curve, which indicates major challenges in

consistent measurements in this region for both AI and human evaluators.

The ICC further analyzed the reliability of Cobb angle measurements. The high agreement for the MT curve between AI and human observers, including the senior expert, strongly supports the reliability of AI for this specific curve. Although the agreement for TL/L can be considered as moderate-to-excellent, the PT curve is a significant concern due to the extremely low ICC values (close to zero for AI vs SE and non-senior clinicians vs. SE). This lack of agreement highlights an essential methodological discrepancy in identifying the end-vertebrae for this curve. This indicates that AI and human experts' use of different reference points or approaches leads to inconsistencies in measurements. These observations mirror broader clinical practice, pushing forward the need for enhanced algorithms or protocols focused on the PT region.

More experienced evaluators often streamline their marking, avoiding areas with no obvious curvature, contrasting with less experienced observers who might mark every vertebra for reference. The inclusion of low-quality images further exacerbated these challenges, resulting in significant differences in assessment duration and accuracy among all observers. Meanwhile, further analysis of vertebral endplate agreement indicates generally high consensus, particularly for the CaE at T12 in the MT area across all scenarios. Such strong agreement, more evident for CaE than for the CrE at T4/T5, implies that certain anatomical landmarks are more consistently identifiable. The least variations observed in the PT and TL/L areas further suggest a high level of agreement in the absence of obvious scoliosis.

As for the influence of experience on agreement, seniority did not significantly impact the collective agreement among clinicians in identifying vertebral endplates. Interestingly, AI measurements aligned more closely with non-senior clinicians than with the SE in several scenarios. This could be due to AI training, which often imitates averaged patterns instead of expert-level nuance, reflecting a central tendency in curve interpretation. While senior clinicians may adhere to stricter or more selective criteria, AI's alignment with group consensus reinforces its role as a standardizing tool, though it might still require expert supervision. This entails AI's potential to standardize endplate selection by reflecting a collective, evidence-based understanding.

Finally, differences with senior expert evaluations emphasize a further need for human supervision in AI-assisted scoliosis evaluation, even though AI shows internal consistency when compared over multiple assessments. Gaps such as these could be filled and increase AI's dependability regardless of curve types through standardized techniques in scoliosis measurement criteria, as well as improved AI algorithms.

Future research should also put more emphasis on improving algorithms focused on PT curve assessment, possibly utilizing hybrid AI-human validation techniques, intended to increase AI's diagnostic accuracy. Standardization of scoliosis measurement criteria among AI and human methods could reduce discrepancies and improve overall diagnostic consistency. Ultimately, AI-assisted scoliosis detection has good potential and seems promising, but it should be ensured that its integration into clinical practice is methodologically aligned with human expertise to maximize diagnostic reliability.

Acknowledgment

The authors genuinely appreciate the support and contribution of one another, the main and co-authors, for their motivation, continuous development, and enhancement to finalize valuable input and results of this study.

Funding Information

The authors would like to express their sincere gratitude to West Visayas State University, Iloilo, Philippines, for providing the funds for the conduct of this study.

Author’s Contributions

Frank Ibañez Elijorde: Performed the coding and implementation of the AI model used in the study, responsible for the acquisition of the dataset and data analysis.

Joselito F. Villaruz: Responsible for the approval of the study’s protocol, contributed to the structured introduction, data analysis, and results sections.

Ma Beth S. Concepcion: Conceptualized the design of the study and contributed to structuring the methodology and results section.

Mylo N. Soriaso: Provided ideas and domain-specific aspects of the methodology; contributed to the analysis of results for discussion.

We solemnly declare that each author has had the ability and opportunity to make substantive contributions to this research work, and we mutually acknowledge and respect the contributions of each author. Collaboration and teamwork have been critical factors in our success in this research endeavor.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and that no ethical issues are involved.

References

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I. F., Mak, R. H., Tamimi, R. M., Tempany, C. M., Swanton, C., Hoffmann, U., Schwartz, L. H., Gillies, R. J., Huang, R. Y., & Aerts, H. J. W. L. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*, *69*(2), 127–157. https://doi.org/10.3322/caac.21552

Cobb Angle for Scoliosis: Definition and Uses. (2024). https://www.healthline.com/health/scoliosis/cobb-angle-scoliosis

Fabijan, A., Fabijan, R., Zawadzka-Fabijan, A., Nowosławska, E., Zakrzewski, K., & Polis, B. (2023a). Evaluating Scoliosis Severity Based on Posturographic X-ray Images Using a Contrastive Language–Image Pretraining Model. *Diagnostics*, *13*(13), 2142. https://doi.org/10.3390/diagnostics13132142

Fabijan, A., Polis, B., Fabijan, R., Zakrzewski, K., Nowosławska, E., & Zawadzka-Fabijan, A. (2023b). Artificial Intelligence in Scoliosis Classification: An Investigation of Language-Based Models. *Journal of Personalized Medicine*, *13*(12), 1695. https://doi.org/10.3390/jpm13121695

Fabijan, A., Zawadzka-Fabijan, A., Fabijan, R., Zakrzewski, K., Nowosławska, E., & Polis, B. (2024). Artificial Intelligence in Medical Imaging: Analyzing the Performance of ChatGPT and Microsoft Bing in Scoliosis Detection and Cobb Angle Assessment. *Diagnostics*, *14*(7), 773. https://doi.org/10.3390/diagnostics14070773

Fuleihan, A. A., Menta, A. K., Azad, T. D., Jiang, K., Weber-Levine, C., Davida, A. D., Hersh, A. M., & Theodore, N. (2024). Navigating artificial intelligence in spine surgery: implementation and optimization across the care continuum. *Artificial Intelligence Surgery*, *4*(4), 288–295. https://doi.org/10.20517/ais.2024.39

Ha, A. Y., Do, B. H., Bartret, A. L., Fang, C. X., Hsiao, A., Lutz, A. M., Banerjee, I., Riley, G. M., Rubin, D. L., Stevens, K. J., Wang, E., Wang, S., Beaulieu, C. F., & Hurt, B. (2022). Automating Scoliosis Measurements in Radiographic Studies with Machine Learning: Comparing Artificial Intelligence and Clinical Reports. *Journal of Digital Imaging*, *35*(3), 524–533. https://doi.org/10.1007/s10278-022-00595-x

Hongbo, W., Bailey, C., Rasoulinejad, P., & Li, S. (2018). Automated comprehensive Adolescent Idiopathic Scoliosis assessment using MVC-Net. *Medical Image Analysis*, *48*, 1–11. https://doi.org/10.1016/j.media.2018.05.005

Karpiel, I., Ziębiński, A., Kluszczyński, M., & Feige, D. (2021). A Survey of Methods and Technologies Used for Diagnosis of Scoliosis. *Sensors*, *21*(24), 8410. https://doi.org/10.3390/s21248410

Khalifa, M., & Albadawy, M. (2024). AI in diagnostic imaging: Revolutionising accuracy and efficiency. *Computer Methods and Programs in Biomedicine Update*, *5*, 100146. https://doi.org/10.1016/j.cmpbup.2024.100146

Kim, H., Kim, H. S., Moon, E. S., Yoon, C.-S., Chung, T.-S., Song, H.-T., Suh, J.-S., Lee, Y. H., & Kim, S. (2010). Scoliosis Imaging: What Radiologists Should Know. *RadioGraphics*, *30*(7), 1823–1842. https://doi.org/10.1148/rg.307105061

Langensiepen, S., Semler, O., Sobottke, R., Fricke, O., Franklin, J., Schönau, E., & Eysel, P. (2013). Measuring procedures to determine the Cobb angle in idiopathic scoliosis: a systematic review. *European Spine Journal*, *22*(11), 2360–2371. https://doi.org/10.1007/s00586-013-2693-9

Lee, S., Jung, J.-Y., Mahatthanatrakul, A., & Kim, J.-S. (2024). Artificial Intelligence in Spinal Imaging and Patient Care: A Review of Recent Advances. *Neurospine*, *21*(2), 474–486. https://doi.org/10.14245/ns.2448388.194

Li, K., Gu, H., Colglazier, R., Lark, R., Hubbard, E., French, R., Smith, D., Zhang, J., McCrum, E., Catanzano, A., Cao, J., Waldman, L., Mazurowski, M. A., & Alman, B. (2025). Deep learning automates Cobb angle measurement compared with multi-expert observers. *BJR|Artificial Intelligence*, *2*(1), ubaf009. https://doi.org/10.1093/bjrai/ubaf009

Martín-Noguerol, T., Oñate Miranda, M., Amrhein, T. J., Paulano-Godino, F., Xiberta, P., Vilanova, J. C., & Luna, A. (2023). The role of Artificial intelligence in the assessment of the spine and spinal cord. *European Journal of Radiology*, *161*, 110726. https://doi.org/10.1016/j.ejrad.2023.110726

Meng, N., Cheung, J. P. Y., Wong, K.-Y. K., Dokos, S., Li, S., Choy, R. W., To, S., Li, R. J., & Zhang, T. (2022). An artificial intelligence powered platform for auto-analyses of spine alignment irrespective of image quality with prospective validation. *E Clinical Medicine*, *43*, 101252. https://doi.org/10.1016/j.eclinm.2021.101252

Mulford, K. L., Regan, C. M., Todderud, J. E., Nolte, C. P., Pinter, Z., Chang-Chien, C., Yan, S., Wyles, C., Khosravi, B., Rouzrokh, P., Maradit Kremers, H., & Larson, A. N. (2024). Deep learning classification of pediatric spinal radiographs for use in large scale imaging registries. *Spine Deformity*, *12*(6), 1607–1614. https://doi.org/10.1007/s43390-024-00933-9

Négrini, S., Grivas, T. B., Kotwicki, T., Maruyama, T., Rigo, M., & Weiss, H. R. (2006). Why do we treat adolescent idiopathic scoliosis? What we want to obtain and to avoid for our patients. SOSORT 2005 Consensus paper. *Scoliosis*, *1*(1), 1–4. https://doi.org/10.1186/1748-7161-1-4

Parr, A., & Askin, G. (2020). Paediatric scoliosis: Update on assessment and treatment. *Australian Journal of General Practice*, *49*(12), 832–837. https://doi.org/10.31128/ajgp-06-20-5477

Retson, T. A., & Eghtedari, M. (2023). Expanding Horizons: The Realities of CAD, the Promise of Artificial Intelligence, and Machine Learning's Role in Breast Imaging beyond Screening Mammography. *Diagnostics*, *13*(13), 2133. https://doi.org/10.3390/diagnostics13132133

Stott, S. M., Wu, Y., Hosseinpour, S., Chen, C., Namdar, K., Amirabadi, A., Shroff, M., Khalvati, F., & Doria, A. S. (2024). Correlative Assessment of Machine Learning-Based Cobb Angle Measurements and Human-Based Measurements in Adolescent Idiopathic and Congenital Scoliosis. *Canadian Association of Radiologists Journal*, *75*(4), 751–760. https://doi.org/10.1177/08465371241231577

Trobisch, P., Suess, O., & Schwab, F. (2010). Idiopathic Scoliosis. *Deutsches Ärzteblatt International*, *107*(49), 875–883.

Wirries, A., Geiger, F., Hammad, A., Redder, A., Oberkircher, L., Ruchholtz, S., Bluemcke, I., & Jabari, S. (2021). Combined Artificial Intelligence Approaches Analyzing 1000 Conservative Patients with Back Pain—A Methodological Pathway to Predicting Treatment Efficacy and Diagnostic Groups. *Diagnostics*, *11*(11), 1934. https://doi.org/10.3390/diagnostics11111934

Zhang, H., Huang, C., Wang, D., Li, K., Han, X., Chen, X., & Li, Z. (2023). Artificial Intelligence in Scoliosis: Current Applications and Future Directions. *Journal of Clinical Medicine*, *12*(23), 7382. https://doi.org/10.3390/jcm12237382