



AUTOMATIC SCOLIOSIS ASSESSMENT USING DEEP LEARNING

Darragh Maguire

This thesis is submitted to University College Dublin in part fulfilment of the requirements for the degree of Master of Biomedical Engineering

School of Electrical & Electronic Engineering

Project Supervisors: Dr John Healy, Dr Kathleen Curran

May 2020

CONTENTS

Abstract	i
Acknowledgements	ii
List of Tables	iii
List of Figures	iv
List of Equations	vi
Glossary of Statistical Metrics	vii
1. Introduction	1
1.1. Context	1
1.2. Motivation	2
1.3. Overall Aim	2
1.4. Specific Objectives	2
1.5. Novel Contribution	3
1.6. Clinical Impact	3
1.7. Thesis Layout	3
2. Literature Review	4
2.1. Background	4
2.1.1. Adolescent Idiopathic Scoliosis	4
2.1.2. Cobb Angle Measurement	5
2.1.3. Lenke Classification	7
2.1.4. Image Analysis Techniques	10
2.2. State-of-the-Art Review	12
2.2.1. Landmark Detection	13
2.2.2. Image Segmentation	16
2.2.3. State-of-the-Art Conclusion	19
2.3. Conclusion	20
3. Methodology	21
3.1. Vertebral Segmentation	22
3.1.1. Datasets	22
3.1.2. Neural Network Architecture	24
3.1.3. U-Net Training	24
3.1.4. Applying the Trained U-Net in Practice	25
3.1.5. Summary of Vertebral Segmentation Phase	26
3.2. Fitting of Endplates	27

3.2.1.	Summary of Fitting of Endplates Phase.....	27
3.3.	Cobb Angle Calculation	28
3.3.1.	Summary of Cobb Angle Calculation Phase	29
3.4.	Lenke Curve Type Probability Analysis	30
3.4.1.	Summary of Lenke Curve Type Probability Analysis Phase.....	34
3.5.	Conclusion.....	35
4.	Results.....	36
4.1.	Vertebral Segmentation.....	37
4.1.1.	U-Net Training.....	37
4.1.2.	Visual Assessment of Performance	37
4.1.3.	Comparison with Ground-truth.....	39
4.2.	Fitting of Endplates	40
4.2.1.	Visual Assessment of Performance	40
4.2.2.	Comparison with Ground-truth.....	41
4.3.	Cobb Angle Calculation	44
4.3.1.	Visual Assessment of Performance	44
4.3.2.	Comparison with Ground-truth.....	45
4.3.3.	Normally Distributed Inter-observer Variation in Cobb angles	47
4.4.	Lenke Curve Type Probability Analysis	49
4.4.1.	Traditional Curve Type Classification.....	49
4.4.2.	Curve Type Probability Analysis.....	51
4.5.	Review of End-to-End Performance	54
5.	Discussion	56
5.1.	Analysis of Findings.....	56
5.2.	Limitations	58
5.3.	Clinical Impact	59
5.3.1.	Fully Automatic Scoliosis Assessment Tool	59
5.3.2.	Lenke Classification Probability Analysis.....	60
5.4.	Future Work	62
6.	Conclusion	63
6.1.	Summary of Work Completed	63
6.2.	Roadmap for Future Work	64
7.	References.....	65

ABSTRACT

Current clinical assessment of scoliosis, a lateral curvature of the spine, involves extensive manual measurements of spinal x-rays to guide intervention. Cobb angles, measured as the angle between vertebrae, are subject to high interobserver variability (3-5° mean absolute difference, MAD). Recently, various systems have been developed to automate the calculation of Cobb angles, using deep learning techniques. However, in this research, clinicians' feedback indicated that advancements in Lenke classification would be more valuable. Lenke classification groups similar types of scoliotic curves, using multiple Cobb angles and additional parameters, in order to guide surgery. Automatic segmentation of spinal x-rays provides a promising path towards robust, reliable, and accurate Lenke classification. However, state-of-the-art deep learning methods for vertebral segmentation have been hindered by limitations in ground-truth datasets.

In this study, the limitations in data were overcome by converting a large dataset of manually labelled spinal landmarks into high-resolution vertebral segmentations. The U-Net, trained using this novel dataset, performed accurately, with subsequent automatically fitted endplates achieving 3.88° MAD, compared to those manually labelled. Algorithms were developed to automatically compute Cobb angles and the Lenke curve type, using these fitted endplates. In addition, a novel probability analysis was developed to account for Cobb angle variability, allowing for objective estimation of the uncertainty in Lenke classifications made by this system. Expanding this proposed automatic system to bending and sagittal view x-rays, along with segmentation of the pedicles and sacrum, would allow for complete Lenke classification. Furthermore, the proposed probability analysis can be applied to aid in clinicians' manual Lenke classification; simply modelling the interobserver variability to provide a confidence level in their assessment and the alternative probable classifications.

ACKNOWLEDGEMENTS

I wish to express gratitude to my supervisors, Dr Kathleen Curran, and Dr John Healy, for their continued support, direction, and insight over the course of this project.

I would also like to recognise the invaluable guidance, provided generously by many clinicians, informing this research, and ensuring its clinical feasibility:

- Mr Michael Dodds, Consultant Orthopaedic Surgeon
- Mr Connor Green, Consultant Orthopaedic Surgeon
- Prof Eoin Kavanagh, Consultant Radiologist
- Dr Rosanne-Sara Lynham, Orthopaedic Registrar
- Dr Ivan Welaratne, Specialist Registrar in Radiology

LIST OF TABLES

TABLE 1: INTEROBSERVER RELIABILITY OF LENKE CLASSIFICATION [14]	9
TABLE 2: GENERATING GROUND-TRUTH SEGMENTATION FROM CORNER LANDMARKS	23
TABLE 3: PROBABILITY OF EACH LENKE CURVE TYPE.....	33
TABLE 4: VERTEBRAL SEGMENTATION ACCURACY	39
TABLE 5: ENDPLATE SLOPES ACCURACY	42
TABLE 6: COBB ANGLES ACCURACY	45
TABLE 7: SHAPIRO-WILK TEST FOR NORMALITY	47
TABLE 8: LENKE CURVE TYPE CLASSIFICATION ACCURACY	50
TABLE 9: LENKE CURVE TYPE PROBABILITY ANALYSIS ACCURACY	52

LIST OF FIGURES

Figure 1. Proximal thoracic, main thoracic and thoracolumbar/lumbar curve Cobb angle measurement [9].....	5
Figure 2. Lenke classification criteria [9].	8
Figure 3. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. [20]	11
Figure 4. Sample of BoostNet performance [21].....	13
Figure 5. Sample of MVE-Net performance (green lines display ground-truth and red indicates test performance) [25].....	14
Figure 6. Sample of prediction of various anatomical parameters [26].....	15
Figure 7. Sample of lumbar spine segmentation [27].	16
Figure 8. Sample of individual vertebra segmentation [28].	17
Figure 9. Sample of DU-Net performance [29].	18
Figure 10. Proposed framework for automatic Cobb angle calculation and Lenke curve type classification.	21
Figure 11. Summary of vertebral segmentation methods.	26
Figure 12. Summary of process to fit endplates to the segmented vertebrae.	27
Figure 13. Summary of automatic Cobb angle calculation from the identified endplates.	29
Figure 14. Probability that one Cobb angle is greater than another.....	31
Figure 15. Probability that a Cobb angle is greater than 25°.	31
Figure 16. Summary of calculation of the probability of each Lenke curve type.	34
Figure 17. U-Net training progress.	37

Figure 18. Vertebral segmentation performance on random sample of images from the Testset.	38
Figure 19. Endplate fitting performance on random sample of images from the Testset.....	40
Figure 20. Accuracy of fitted endplates compared to those manually annotated.	43
Figure 21. Cobb angle selection performance for random sample of images from the Testset.	44
Figure 22. Accuracy of fully automatic Cobb angles compared to ground-truth (those obtained automatically from manually annotated endplates).	46
Figure 23. Q-Q plot of the difference in Cobb angle between prediction and ground-truth, in order to assess validity of the assumption of normally distributed variation.	48
Figure 24. Accuracy of fully automatic Lenke curve type classification compared to ground- truth (that obtained automatically from manually annotated endplates).	50
Figure 25. Lenke curve type probability performance for a random sample from the Testset.	51
Figure 26. Accuracy of fully automatic calculation of Lenke curve type probabilities compared with ground-truth (probabilities calculated automatically using manually annotated endplates).	53
Figure 27. End-to-end performance of the proposed system for a random sample of images from the Testset.....	55
Figure 28. Probability tree explanation of disparity between measured and most probable Lenke curve types.	61

LIST OF EQUATIONS

Equation 1: Normal Probability Density Function	30
Equation 2: Probability of TL/L being the Major Curve	32
Equation 3: Probability of MT being the Major Curve.....	32

GLOSSARY OF STATISTICAL METRICS

The following is a glossary of the statistical metrics referenced in this thesis.

Balanced Accuracy Rate (BAR)	<i>An accuracy metric accounting for class imbalance by reporting the average of the proportions classified correctly for each individual class.</i>
Cohen's Kappa Statistic (κ)	<i>A measure of categorical inter-rater reliability that accounts for the probability of agreement occurring by chance.</i>
Dice Similarity Coefficient	<i>A statistical measure of the overlap between two sets of data.</i>
F_β score	<i>The weighted harmonic mean of precision and recall.</i>
Intraclass Correlation Coefficient (ICC)	<i>An assessment of the consistency of quantitative measurements made by multiple observers.</i>
Mean Absolute Difference (MAD)	<i>The average magnitude of deviation in measurement between two observers.</i>
Pearson Correlation Coefficient	<i>A statistical measure of the linear correlation between two variables.</i>
Percent Agreement	<i>A measure of categorical inter-rater reliability that quantifies the frequency of agreement.</i>
Precision	<i>The proportion of predicted positive samples that are truly positive.</i>
Recall	<i>The proportion of truly positive samples that are predicted positive.</i>

1. INTRODUCTION

1.1. CONTEXT

Scoliosis is a sideways curvature of the spine. Current clinical practice, regarding diagnosis, treatment, and follow-up, involves measuring the extent and progression of this curvature in x-rays. Cobb angles are the standard metric for quantifying spinal curvature, involving manual measurement of the angle between various vertebral endplates. This technique is time-consuming and suffers from high levels of intra- and interobserver variability. Various image processing and deep learning tools have developed in recent years to combat these shortcomings.

Segmentation of vertebrae and subsequent fitting of endplates allows for automatic calculation of Cobb angles. Convolutional neural networks provide a robust mechanism for recognition and segmentation of objects in images. However, these methods require large datasets of labelled images for training. Various attempts to develop vertebral segmentation models have been hindered by limitations in such training datasets.

Systems developed to automate assessment of scoliosis in x-rays have often focused solely on Cobb angles. In this study, however, guidance from clinicians indicated the need for a more comprehensive suite of measurements. The Lenke classification system was cited as requiring more time and effort, but vital in guiding clinical decisions. This system provides a means of guiding surgeons in selective fusion of the scoliotic spine. Furthermore, it allows for comparison of surgical outcomes, grouping similar types of scoliosis curves using various metrics measured in both coronal and sagittal view spinal x-rays. Reliant on the measurements of multiple Cobb angles, this system also suffers from intra- and interobserver variability.

1.2. MOTIVATION

This research is motivated by the shortcomings of current clinical practice and the limitations of state-of-the-art research in scoliosis assessment. The required clinicians' time, along with the limited reliability and accuracy of measurements, do not correlate with modern capabilities in image analysis.

1.3. OVERALL AIM

The aim of this research is to develop robust, reliable, accurate, and above all, useful tools to aid in the clinical assessment of scoliosis in x-rays.

1.4. SPECIFIC OBJECTIVES

The specific objectives of this study, as developed in collaboration with clinicians, are as follows:

- To build a model for accurate and robust vertebral segmentation in spinal x-rays.
- To compute Cobb angles and Lenke curve type classification automatically and reliably, using the developed vertebral segmentation method.
- To quantify the confidence level in predictions made by this system.
- To ensure the research outputs are clinically feasible and useful, through clinical collaboration.
- To develop expandable tools, capable of evolving into comprehensive clinical aids.

1.5. NOVEL CONTRIBUTION

The following components of this study constitute a novel contribution to the field of scoliosis research:

- Conversion of a large publicly available scoliotic spinal landmark dataset into a high-resolution vertebral segmentation dataset.
- Development of methods to quantify the confidence level in Lenke classification, given the interobserver variation of the Cobb angle measurement method.

1.6. CLINICAL IMPACT

The outputs of this research have the following potential clinical implications:

- The developed fully automatic scoliosis assessment tool can be integrated into clinical practice and research for rapid, accurate, and robust calculation of Cobb angles and classification of Lenke curve type in x-rays.
- The proposed system to quantify the uncertainty in Lenke classification can be developed into a tool to aid clinicians' manual assessment of x-rays, providing the confidence level of a given classification and the alternative probable classifications.

1.7. THESIS LAYOUT

This thesis will begin with a literature review, consisting of an overview of the relevant background material and a state-of-the-art review of recently developed systems in this field. The methodology applied in this study will then be described, and the performance of the proposed systems will be analysed in-depth. The limitations and clinical impact of this research will be explored, before addressing areas for improvement in future work. Finally, the project will be summarised with concluding remarks.

2. LITERATURE REVIEW

This literature review presents an overview of the relevant background material, along with a comprehensive review of the current state-of-the-art in automatic scoliosis assessment.

2.1. BACKGROUND

2.1.1. ADOLESCENT IDIOPATHIC SCOLIOSIS

Scoliosis is the most common spinal disorder in children and adolescents, with an overall prevalence of 0.47-5.20% [1]. The condition is characterised by deformation in the coronal, sagittal, and transverse planes, manifesting in lateral curvature, thoracic lordosis (inward rounding of the back), and vertebral rotation, respectively [2]. The Scoliosis Research Society define scoliosis as a lateral deformation $\geq 10^\circ$ in the spinal x-ray [3]. This is deemed idiopathic if other aetiologies, such as congenital, neuromuscular, and mesenchymal, have been ruled out. Adolescent idiopathic scoliosis (AIS), presenting between the ages 11-18 years, accounts for approximately 90% of cases of idiopathic scoliosis in children [1]. For this reason, orthopaedic surgeons consulted in this study suggested that an emphasis be placed on AIS, in the development of clinical tools.

Up to 10% of scoliosis patients require some form of intervention, while only 0.1% require surgery [4]. More conservative treatments include bracing and scoliosis-specific exercise. However, the efficacy of these interventions remains unclear ([5], [6]). Surgery is justified for curves that reach 50° , due to likely progression into adult life, back pain, and cosmetic deformity. Surgical treatments, including restoration using rods and screws, selective fusion, and vertebral tethering, aim to improve cosmesis and function, while limiting complication rates and long-term implications [4].

2.1.2. COBB ANGLE MEASUREMENT

The diagnosis, treatment, and follow-up of scoliosis patients are aided by the measurement of various metrics, including angles and distances, describing anatomical and postural features of the spine. The Cobb method [7], introduced in 1948, remains the standard technique applied for quantifying lateral curvature of the spine. This involves calculation of the angles between the most tilted vertebral endplates in standing anterior-posterior (AP) x-rays, as illustrated in Figure 1. The 33 vertebrae of the spinal column can be divided into the cervical (C1-C7), thoracic (T1-T12), lumbar (L1-L5), sacrum (S1-S5), and coccyx (Co1-Co4). Lateral curvature presents in the thoracic and lumbar spine, and thus the proximal thoracic (PT), main thoracic (MT) and thoracolumbar/lumbar (TL/L) curves are measured in the assessment of scoliosis. While traditionally measured with pencil and protractor, Cobb angles are increasingly calculated digitally [8]. This generally involves manual location of landmarks on the relevant vertebral endplates, and subsequent automatic calculation of angles.

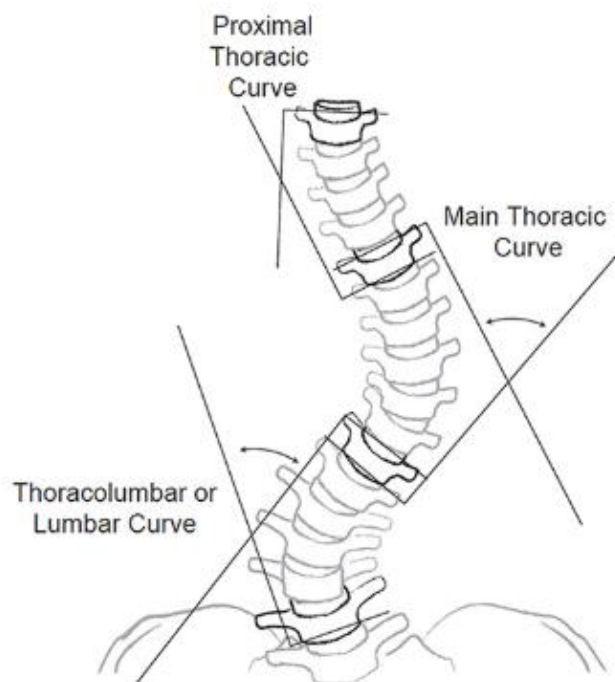


Figure 1. Proximal thoracic, main thoracic and thoracolumbar/lumbar curve Cobb angle measurement [9].

Cobb angles are subject to a high level of intra- and interobserver variability. This can be attributed to variation in selection of end vertebrae, drawing of endplate lines (or digital equivalent), and measurement of angles [10]. This limitation was highlighted by both the radiologists and the orthopaedic surgeons consulted in this study. The literature presents a wide range of recorded magnitudes of intra- and interobserver variability, resulting from variations in study design, such as:

- The capturing of manual or digital methods
- Recording different statistical metrics
- Assessment of the single maximum Cobb angle or the PT, MT, and TL/L angles
- Allowing for or eliminating variability in the selection of vertebrae

Commonly, an interobserver variability of 3-5° mean absolute difference (MAD) is reported ([11], [12], [13]). Additionally, 2D x-rays inherently yield an incomplete view of scoliosis – a 3D deformity. There is evidence that assessment of 2D x-rays leads to an underestimation of the true Cobb angle present [13]. Nonetheless, 2D x-rays remain standard in this clinical assessment, despite technological advancements in 3D imaging.

2.1.3. LENKE CLASSIFICATION

The Lenke classification system [14] was introduced to improve on the former standard method for classification of AIS, King classification [15]. King classification was developed to guide selective fusion for thoracic idiopathic scoliosis. However, it was found to yield insufficient intra- and interobserver reliability between surgeons [16], failed to recognise all scoliosis curve types, and neglected the 3D nature of the spinal deformity. Lenke aimed to address these shortcomings with a classification system involving curve type, lumbar spine modifier, and sagittal thoracic modifier.

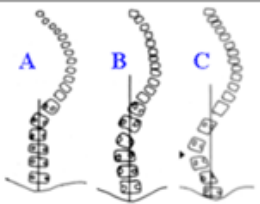
There are 6 curve types, categorised depending on which of the PT, MT and TL/L curves are structural. The major curve (TL/L if it yields the largest Cobb measurement, otherwise MT) is always structural. Structural curves lack normal flexibility, and this can be assessed using supine bending x-rays. The minor curves can be structural or non-structural. The Lenke classification system deems a minor curve to be structural if it is greater than 25° in the standing AP x-ray, and does not ‘bend out’ to less than 25° on bending films. Alternatively, minor curves can be determined as structural if they are greater than 20° in the sagittal plane. The lumbar spine modifier and sagittal thoracic modifiers further categorise the progression of AIS and are critical factors to assess when surgery is being considered. The lumbar modifier (A, B, or C) captures the location of the central sacral vertical line (CSVL – vertical line bisecting the proximal sacrum) in relation to the vertebral pedicles at the apex of the TL/L curve, as illustrated in Figure 2. The sagittal thoracic modifier (-, N, or +) measures the degree of thoracic kyphosis (forward rounding of the back) in the scoliotic spine with respect to the normal level. The complete Lenke classification criteria are summarised in Figure 2, and further analysed in Section 3.4.

CURVE TYPE				
Type	Proximal Thoracic	Main Thoracic	Thoracolumbar/Lumbar	Description
1	Non-Structural	Structural (Major)*	Non-Structural	Main Thoracic (MT)
2	Structural	Structural (Major)*	Non-Structural	Double Thoracic (DT)
3	Non-Structural	Structural (Major)*	Structural	Double Major (DM)
4	Structural	Structural (Major)*	Structural (Major)*	Triple Major (TM) [‡]
5	Non-Structural	Non-Structural	Structural (Major)*	Thoracolumbar/Lumbar (TL/L)
6	Non-Structural	Structural	Structural (Major)*	Thoracolumbar/Lumbar-Main Thoracic (TL/L-MT)

STRUCTURAL CRITERIA (Minor Curves) Proximal Thoracic - Side Bending Cobb $\geq 25^\circ$ - T2-T5 Kyphosis $\geq +20^\circ$ Main Thoracic - Side Bending Cobb $\geq 25^\circ$ - T10-L2 Kyphosis $\geq +20^\circ$ Thoracolumbar/Lumbar - Side Bending Cobb $\geq 25^\circ$ - T10-L2 Kyphosis $\geq +20^\circ$		*Major = Largest Cobb measurement, always structural Minor = All other curves with structural criteria applied [‡] Type 4 - MT or TL/L can be major curve
--	--	--

LOCATION OF APEX (SRS Definition) <table> <tr> <th>CURVE</th><th>APEX</th></tr> <tr> <td>Thoracic</td><td>T2-T11/12 Disc</td></tr> <tr> <td>Thoracolumbar</td><td>T12-L1</td></tr> <tr> <td>Thoracolumbar/Lumbar</td><td>L1/2 Disc-L4</td></tr> </table>		CURVE	APEX	Thoracic	T2-T11/12 Disc	Thoracolumbar	T12-L1	Thoracolumbar/Lumbar	L1/2 Disc-L4
CURVE	APEX								
Thoracic	T2-T11/12 Disc								
Thoracolumbar	T12-L1								
Thoracolumbar/Lumbar	L1/2 Disc-L4								

Modifiers

Lumbar Spine Modifier	CSVL to Lumbar Apex		Thoracic Sagittal Profile T5-T12	
A	CSVL between pedicles		- (Hypo)	< 10°
B	CSVL touches apical body(ies)		N (Normal)	10° - 40°
C	CSVL completely medial		+ (Hyper)	> 40°

Curve Type (1-6) + Lumbar Spine Modifier (A, B, C) + Thoracic Sagittal Modifier (-, N, +) Classification (e.g. 1B+): _____

Figure 2. Lenke classification criteria [9].

The results from an interobserver study of Lenke classification [14], on a randomly chosen independent group of seven surgeons, are summarised in TABLE 1. The interobserver reliability is assessed using Cohen's kappa statistic; a measure of reliability that accounts for the possibility of agreement occurring by chance. The authors report a substantial improvement in reliability when compared to the King classification system. However, while the lumbar and thoracic sagittal modifiers reflect a very strong agreement between the surgeons, significant interobserver variation remains in the classification of some of the curve types (especially types 3, 4 and 6).

TABLE 1: INTEROBSERVER RELIABILITY OF LENKE CLASSIFICATION [14]

	Cohen's Kappa Statistic
<i>Curve Type</i>	
1	0.816
2	0.773
3	0.683
4	0.384
5	1.000
6	0.407
Mean	0.740
<i>Lumbar Modifier</i>	
A	0.763
B	0.738
C	0.880
Mean	0.800
<i>Sagittal Thoracic Modifier</i>	
+	0.901
-	1.000
N	0.930
Mean	0.938

2.1.4. IMAGE ANALYSIS TECHNIQUES

The assessment of scoliosis is heavily reliant on the measurement of various quantities via x-ray imaging. While traditionally captured, annotated and measured using film, advances in technology have led to broad adoption of digital acquisition and assessment [8]. Software aids in manual measurement of scoliosis metrics, for example, through calculation of angles and distances between landmarks entered by the clinician using a mouse. Furthermore, computer-aided and semi-automatic methods have been introduced in order to improve the accuracy and reliability of measurements ([17], [18], [19]). Techniques employed for this purpose include segmentation, curve-fitting, edge detection, and contrast-stretching. Generally, the clinician is required to enter points or select regions of interest when using these semi-automatic tools. More advanced techniques have been introduced, aiming to detect endplates and calculate Cobb angles in a fully automatic process. These methods will be explored further in the following state-of-the-art review.

In recent years, advancements in image analysis techniques have been driven by the application of deep learning methods and the abundance of available data. Deep learning methods have the capacity to eliminate the need for feature engineering and hand-crafted analysis, as their many layers and variables can extract and learn features from raw data. The development of modern systems reliant on image data, such as facial recognition and autonomous vehicles, has motivated the evolution of convolutional neural networks. These networks apply convolutional filters in the extraction of features in images. A convolution refers to the process of conducting a weighted sum of each pixel in an image with its local neighbours, where the weights are defined by a filter (kernel) matrix. Traditionally, the filter would contain hand-crafted weights, in order to produce a desired filtered image. However, in a convolutional neural network, many convolutions are applied to the input image over various

layers, and the weights of each filter are learnt through the processing of labelled sample images during training.

Convolutional neural networks can be applied for image classification and landmark detection, using similar methods. In these networks, the extracted 2-dimensional features are converted into an array of classification targets or landmark coordinates. Alternative convolutional networks have been developed to achieve image segmentation. In this case, the network output is generally a binary segmentation map, i.e. a matrix with the same dimensions as the input image, whereby, each pixel contains a value indicating its class. The U-Net [20], published in 2015, is a segmentation network designed for application in biomedical imaging. To this end, the authors produced a highly accurate system, requiring relatively small training datasets of grayscale images. The U-Net architecture, illustrated in Figure 3, is composed of a contracting path, using convolutions to encode the context of the image, and a symmetric expanding path using transposed convolutions to decode localisation information.

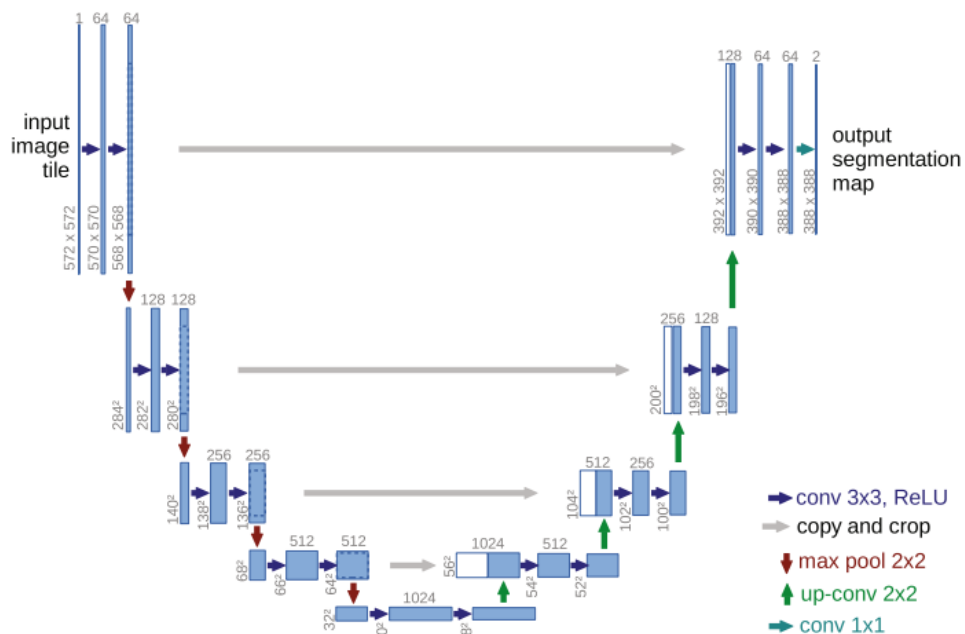


Figure 3. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. [20]

2.2. STATE-OF-THE-ART REVIEW

This section will review recent methods developed for automatic scoliosis assessment. A focus will be placed on 2D spinal x-ray analysis techniques, the current standard imaging modality for clinical assessment of scoliosis. Furthermore, this review will be centred around methods developed within the last three years. Given the recent advancements in the area, these systems invariably utilise deep learning methods, which can be divided into two categories: landmark detection and image segmentation. Each of these approaches will be discussed in turn. Finally, gaps in the current literature will be explored in the state-of-the-art conclusion.

2.2.1. LANDMARK DETECTION

The BoostNet [21] is a neural network designed to detect landmarks on the spine in x-rays. This model automatically annotates the 4 corners of each of 17 lumbar and thoracic vertebrae in AP x-rays, as shown in Figure 4. Robustness was identified as a common pitfall amongst other algorithms tested, due to variation in the images. The authors improve the robustness of the traditional ConvNet, using their developed BoostLayer to remove outlier features, and Spinal Structured Multi-Output Layer to exploit spatial dependencies between the landmarks. The result was an improvement on segmentation and filter-based methods, and previous machine learning frameworks applied in this area. The network achieved a Pearson correlation coefficient of 0.94 between predicted and ground-truth landmarks; and displayed a marginal improvement over the compared random forest [22] and support vector regression methods [23]. Although the network predicted landmarks correlate closely with the outline of the spine, they often deviate from the vertebral corners, as evident in Figure 4. This could potentially impact the resulting Cobb angles, which are calculated as the slope between adjacent vertebral corner landmarks. The accuracy of the resulting predicted Cobb angles was not assessed in the study.



Figure 4. Sample of BoostNet performance [21].

The multi-view correlation network (MVC-Net) [24] and multi-view extrapolation network (MVE-Net) [25] have been designed to measure Cobb angles in both the coronal and sagittal planes. Lateral x-rays, taken in the sagittal plane, pose a greater challenge due to increased obstruction by the ribcage. With the MVC-Net, the authors attempt to overcome this issue by exploiting structural dependencies between the two views using a multi-view convolutional layer. In the MVE-Net, landmarks are learnt both with a joint-view convolution of the two planes, and an independent-view taking each plane separately. Cobb angles calculated from each of these landmark prediction networks are then combined using an inter-error correction layer, combining high-precision calculation with deep learning. The MVE-Net showed significant improvement when compared with the MVC-Net, and the BoostNet. This network was trained with 526 images, equally divided between AP and lateral x-rays. The achieved Cobb angles were compared with those obtained from ground-truth labels, as illustrated in Figure 5.

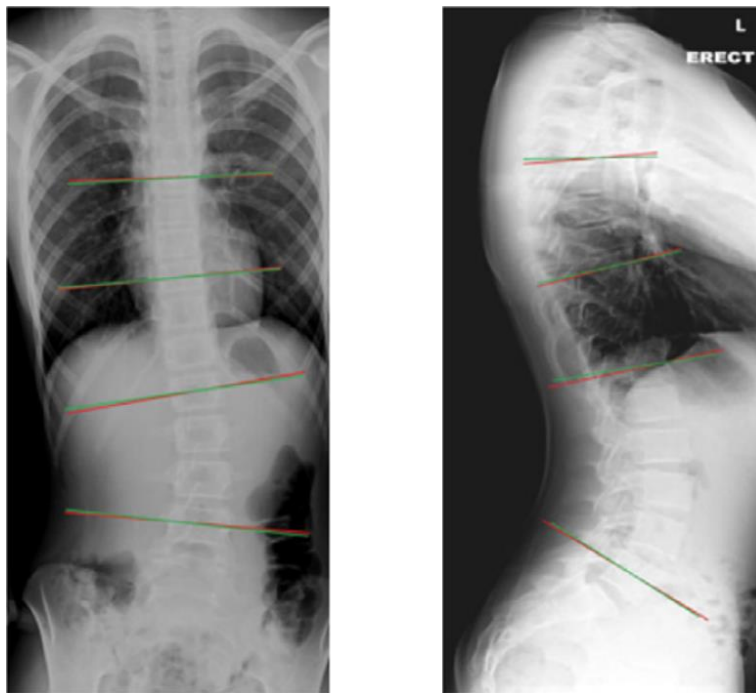


Figure 5. Sample of MVE-Net performance (green lines display ground-truth and red indicates test performance) [25].

A landmark detection network is applied for the calculation of various anatomical parameters in [26]. Angles are calculated for kyphosis, lordosis, pelvic incidence, sacral slope, and pelvic tilt as well as the Cobb angles, as seen in Figure 6. This study highlights the clinical relevance of various metrics in diagnosis, treatment, and follow-up of spinal disorders. The developed model can be used in a variety of subjects, for example, adolescent or elderly, with varying spinal disorders, and with full-body scans or x-rays of the trunk only. A database of biplanar x-rays of 493 patients were used in the study. The predicted parameters strongly correlated with the ground-truth labels; however, the outputs suffer from relatively high standard error. Despite this, the model presents a useful system encompassing a broad range of spinal disorders. Furthermore, in scoliosis alone, various anatomical angles and distances such as trunk shift and sagittal balance are useful in assessment and follow-up. This should therefore be considered in the design of an automatic scoliosis assessment tool.

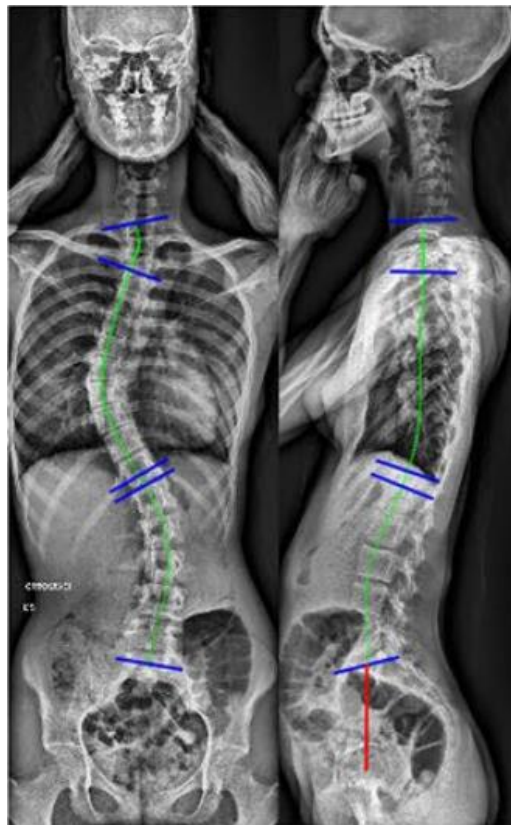


Figure 6. Sample of prediction of various anatomical parameters [26].

2.2.2. IMAGE SEGMENTATION

The U-Net was applied in [27], for pixel level classification as opposed to landmark detection. The output of the developed model is a segmentation map, dividing the image into two classes – ‘vertebra’ and ‘background’ - as shown in Figure 7. The authors developed an algorithm to fit a straight line to each endplate in the segmentation map, using minimum bounding rectangle and least square methods. Cobb angles were automatically calculated by taking the slopes of the fitted lines. Due to data limitations, the U-Net was trained and tested in segmentation of the lumbar spine only. Accurate segmentation was achieved, with a validation accuracy of around 98%. Separately, the automatic measurement of Cobb angles was assessed using manually segmented full-length spine images. Cobb angle calculation in this algorithm was limited to the single maximum angle between endplates, as opposed to calculation of the PT, MT, and TL/L angles. The angles were compared to those obtained by a spinal surgeon, yielding accurate results. A mean deviation of around 1.7° was found between manual calculation and automatic calculation (from the manual segmentation map). Overall, this study revealed promising results, however, its primary shortcoming is the lack of assessment of Cobb angles obtained from network predicted segmentation maps of the full-length spine.

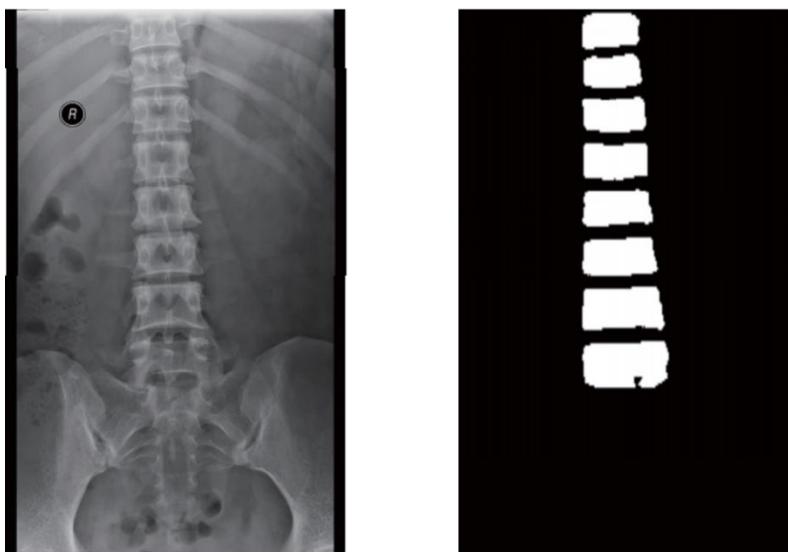


Figure 7. Sample of lumbar spine segmentation [27].

A similar vertebral segmentation technique was applied in [28]. In contrast to the previous study, the authors apply the U-Net for segmentation of vertebrae individually, as shown in Figure 8, following detection of the location of each vertebra. The proposed process begins with isolation of the spinal column, from the skull, limbs, and hips, using analysis of pixel intensity histograms in the vertical and horizontal directions. The location of each vertebra is then detected, applying polynomial fitting and further intensity histogram analysis. The local area of each vertebra is then input into the neural network for segmentation. The U-Net, Residual U-Net, and Dense U-Net were compared in segmentation of single vertebrae, for this final step. The minimum bounding rectangle approach was then applied, in order to obtain endplate slopes from the vertebral segmentation. Unfortunately, this method alone cannot allow for different slopes on the upper and lower endplate of a given vertebra. Furthermore, the Cobb angle calculation in this system was also limited to the single maximum angle, rather than the comprehensive PT, MT, and TL/L analysis. The networks were trained and tested using just 35 x-rays, divided into single vertebra images. This may limit the generalisation of learnt features, for real-world application. Nonetheless, the system achieved accurate vertebral segmentation and resulting Cobb angles that were highly correlated to manual assessment by clinicians.

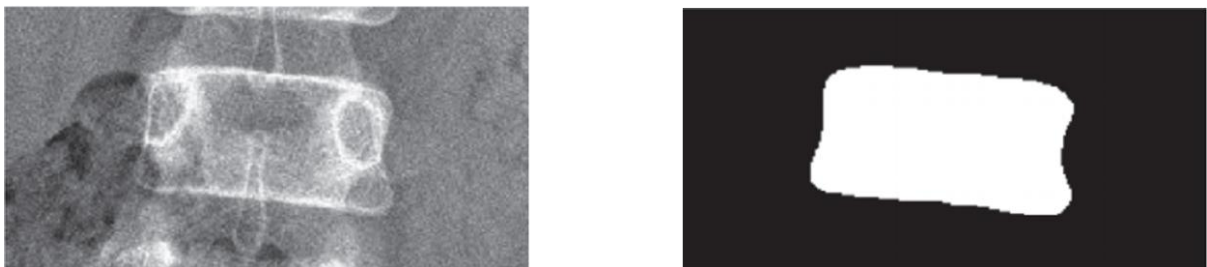


Figure 8. Sample of individual vertebra segmentation [28].

The DU-Net [29] was proposed to segment the spinal column rather than each vertebra, as shown in Figure 9. The proposed model combines an algorithm for detection of the spine with the U-Net segmentation network. 100 images were used in training the network and 10 for testing. Predicted segmentation results were compared with manual annotations using the Dice coefficient, among other metrics. The proposed DU-Net yielded better results than the standard U-Net under each criterion. A 6th order polynomial was fitted to the spine contour in the resulting segmentation map, allowing for Cobb angle calculation using tangents to the curve. This method of Cobb angle calculation is less consistent with current practice than the previously mentioned methods, as the vertebral endplates are not incorporated. Nonetheless, the algorithm displayed promising results. Cobb angles were computed for the testing dataset and compared with those measured by an orthopaedic specialist. Again, a single Cobb angle was measured for each x-ray, rather than the PT, MT, and TL/L angles. The automatic and manually measured angles aligned closely, with the largest deviation observed being 5.4° .

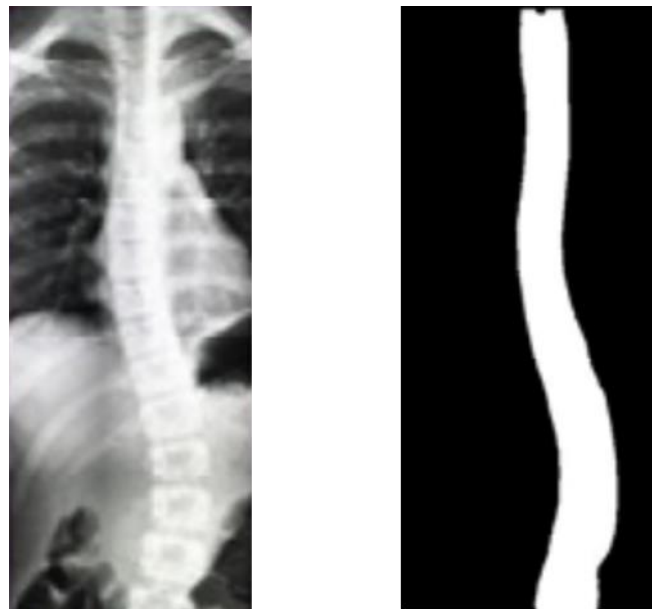


Figure 9. Sample of DU-Net performance [29].

2.2.3. STATE-OF-THE-ART CONCLUSION

The reviewed systems each displayed strong potential in the application of neural networks in this domain. Both the landmark detection methods, and vertebral segmentation, provide a comparable Cobb angle measurement technique to that performed clinically using vertebral endplates. Despite achieving accurate performance, the DU-Net for full spinal column segmentation is limited due to the deviation from current clinical practice, calculating Cobb angles without the use of endplates. In addition to the technique of angle measurement, the reviewed systems differed in their approaches to Cobb angle selection. Most authors opted for the simpler method of locating the single largest Cobb angle on the spine. This limits assessment, and the identification of a patient's curve type.

Each of the landmark detection systems that incorporated sagittal x-rays into their design provided a more comprehensive assessment, with additional metrics beyond the PT, MT, and TL/L Cobb angles. Guidance from surgeons, in this research, indicated that many measurements are considered in the assessment of scoliosis, and thus, it is important that automatic systems can be expanded to achieve a complete analysis. It was noted that small deviations in detected landmarks can lead to large errors in the resulting Cobb angles and other measurements. In contrast, the vertebral segmentation techniques can provide a more robust and comprehensive calculation of clinical metrics. Rather than describing a vertebra by its four corners, a high-resolution segmentation can capture much of the information present in the x-ray, in a format that can be leveraged for simple computation of many anatomical measurements. The reviewed vertebral segmentation techniques were limited by the availability of datasets. In addition, neither of these techniques incorporated sagittal x-rays into the analysis. This can be explained by the fact that ground-truth vertebra segmentations are tedious to annotate manually, requiring many hours for relatively small datasets.

2.3. CONCLUSION

Automatic assessment of spinal x-rays is a growing domain, approaching broad adoption in practice. In order to produce relevant and clinically impactful research in this project, improvements will be sought in automatic vertebral segmentation. To achieve this, the issue of limited ground-truth data will be addressed. In addition, a method will be developed for comprehensive calculation of the PT, MT, and TL/L Cobb angles, allowing for automatic Lenke curve type classification. This study will strive to develop a system capable of expansion into the segmentation of sagittal x-rays and the measurement of many clinical metrics.

3. METHODOLOGY

The proposed framework for fully automatic assessment of scoliosis based on the AP spinal x-ray consists of four phases, as illustrated in Figure 10. The clinical endpoints of this process are the PT, MT, and TL/L Cobb angles, as well as the Lenke curve type classification and its associated confidence level. Firstly, a U-Net is applied to segment the vertebrae in the AP spinal x-ray. Using this segmentation map, endplates are fit to each of the identified vertebrae. Following calculation of the slopes of these endplates, an algorithm is applied to identify the locations of, and calculate the magnitude of, the three Cobb angles. This allows for automatic classification of the Lenke curve type present in the x-ray. Furthermore, modelling the variation in measured Cobb angles with a normal distribution, the probability distribution of the true Cobb angles is estimated. Using this, the confidence level in the Lenke curve type classification is calculated, as well as the probability associated with each curve type for a given set of measured Cobb angles. Each of these phases are outlined in the following sections.

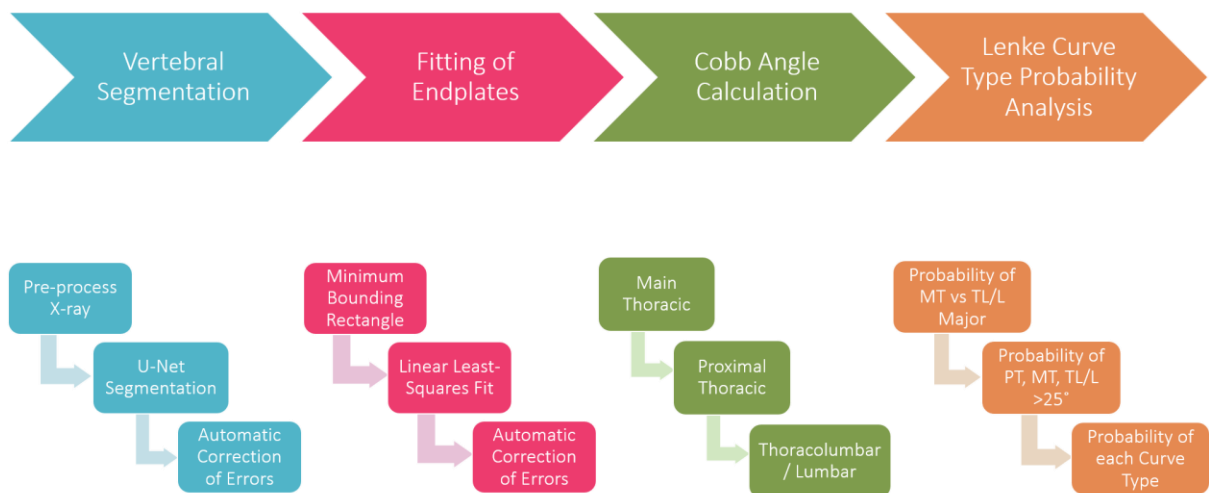


Figure 10. Proposed framework for automatic Cobb angle calculation and Lenke curve type classification.

3.1. VERTEBRAL SEGMENTATION

3.1.1. DATASETS

3.1.1.1. *BOOSTNET DATASET*

One of the primary novelties of this study stems from the leveraging of a publicly available dataset of manually annotated vertebra landmarks, to automatically generate high resolution ground-truth vertebra segmentations. The landmark dataset was constructed in the development of the BoostNet [21] for detection of the corners of vertebrae in x-rays. The dataset consists of 609 spinal AP x-ray images showing signs of scoliosis – divided into 481 training/validation (*Trainset*) and 128 testing (*Testset*) images, such that no patient is present in both sets. The four corners of 17 thoracic/lumbar vertebrae were manually annotated by the authors, resulting in 68 landmark coordinates associated with each image. In this study, an algorithm has been developed to locate the entire vertebral perimeter from these corner landmarks, thus converting ground-truth landmarks to ground-truth segmentations for training.

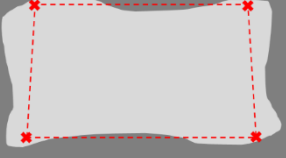
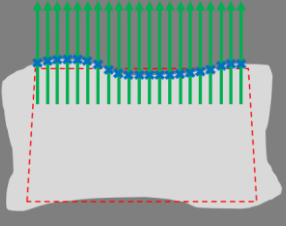
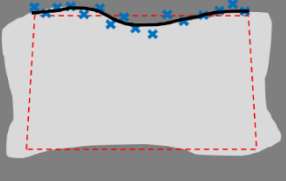
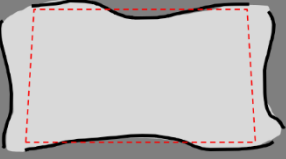
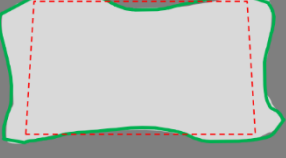
3.1.1.2. *CORRECTION OF LANDMARK ERRORS*









Some errors were encountered in the BoostNet dataset, including, inconsistent ordering of the landmark coordinates and labelling of a single endplate on a vertebra instead of the 4 corners. These data collection errors would not be immediately obvious when plotting the landmarks on the spinal x-rays, however, they manifested as clear problems when converting to vertebra segmentations. Around 10% of the images were found to contain issues such as this, all of which were manually corrected. It is worth noting that these errors would have negatively impacted performance of the BoostNet, and other models developed using the same dataset.

3.1.1.3. *CONVERSION OF GROUND-TRUTH LANDMARKS TO SEGMENTATIONS*

The segmentation of each vertebra was generated automatically from the 4 corner landmarks, following the process outlined in TABLE 2.

TABLE 2: GENERATING GROUND-TRUTH SEGMENTATION FROM CORNER LANDMARKS

<i>Illustration</i>	<i>Step</i>	<i>Description</i>
	1	The ‘initial perimeter’ is defined as the quadrilateral connecting the 4 manually labelled corner landmarks.
	2	For each pixel on an edge of the initial perimeter, a vector perpendicular to that edge is defined, extending from a point within the initial perimeter to an exterior point. The length of this vector is scaled by the distance between landmarks. Each vector is traversed to find the maximum change in brightness from bright to dark, i.e. locating the edge of the vertebra.
	3	These points of maximum change in brightness define a new estimate of the vertebral edge. This estimate is usually noisy, due to variation in the x-rays. To combat this, a 4 th order polynomial is fit to the points.
	4	This process is repeated for each edge of the vertebra.
	5	The ground-truth segmentation is completed by forming a continuous perimeter defined by the 4 quartic edges. In this case, MATLAB’s <i>boundary</i> function [30] was applied, with a shrink factor of 0.85, for this final step.

	Background		Vertebra		Corner label		Initial perimeter
	Vector to search for maximum brightness change		Location of maximum brightness change		Fitted polynomial		Final segmentation

3.1.1.4. DATA AUGMENTATION

Employing data augmentation, the *Trainset* was increased 14-fold from 481 to 6734 images. The following augmentations were performed for each image: rotation $\pm 5^\circ$, rotation $\pm 10^\circ$, gamma correction with $\gamma = 0.5$ and $\gamma = 1.5$, as well as horizontal mirroring of each of the previous augmentations and the original.

3.1.2. NEURAL NETWORK ARCHITECTURE

The U-Net [20] was applied for semantic segmentation of vertebrae in the x-rays. This involves pixel-level classification, identifying each pixel as either ‘vertebra’ or ‘background’. The applied network comprises 4 down-samples in the encoder section, from a resolution of 128x256 to 16x32. The first encoder stage contains 16 output channels, and the number of output channels doubles for each subsequent encoder stage. 3x3 kernels were used in the convolutions, with zero padding applied, such that the input and output feature maps are the same size. The decoder section mirrors these dimensions and specifications with transposed convolutions and up-samples.

3.1.3. U-NET TRAINING

The network was trained for 500 epochs, with a mini-batch size of 64 images. The *Trainset* was divided into a training:validation split of 85:15. Tversky loss [31] was applied in training – a loss function derived from a generalisation of the Dice similarity coefficient and the F_β scores. Recall was weighed higher than precision, by setting $\beta = 0.7$, to alleviate difficulties arising from the imbalance of vertebra pixels relative to background pixels (around 10% and 90% respectively). Adam optimisation [32] was employed, and the weights from the epoch with the lowest validation loss were used. The model was trained using the Keras Deep Learning Library [33] with Tensorflow [34] backend using a GPU in Colaboratory [35].

3.1.4. APPLYING THE TRAINED U-NET IN PRACTICE

X-rays should be resized to 128x256 to suit the input dimensions of the trained U-Net. It is worth noting that the training images used were usually limited to showing roughly the C7 to L5 section of the spine, without the skull, pelvis, and limbs in view. Therefore, for accurate real-world application of the trained network, input x-rays should be cropped to this area. However, the need for this step can be eliminated by applying transfer learning of the network on a dataset of x-rays with a more varied field of view, or through development of an additional automatic step to extract a region of interest similar to that seen in training.

Post-processing is applied to the U-Net output, exploiting the consistent structure of the spinal column, to automatically find and correct for errors made in the neural network prediction. This processing includes removal of outlier objects mistakenly classified by the network as vertebrae, including spatial outliers and objects that are significantly smaller than other predicted vertebrae. A combination of erosion, watershed processing, and dilation of the objects classified as vertebrae is also applied. This process is tailored to separate vertebrae that are mistakenly merged in the network prediction.

3.1.5. SUMMARY OF VERTEBRAL SEGMENTATION PHASE

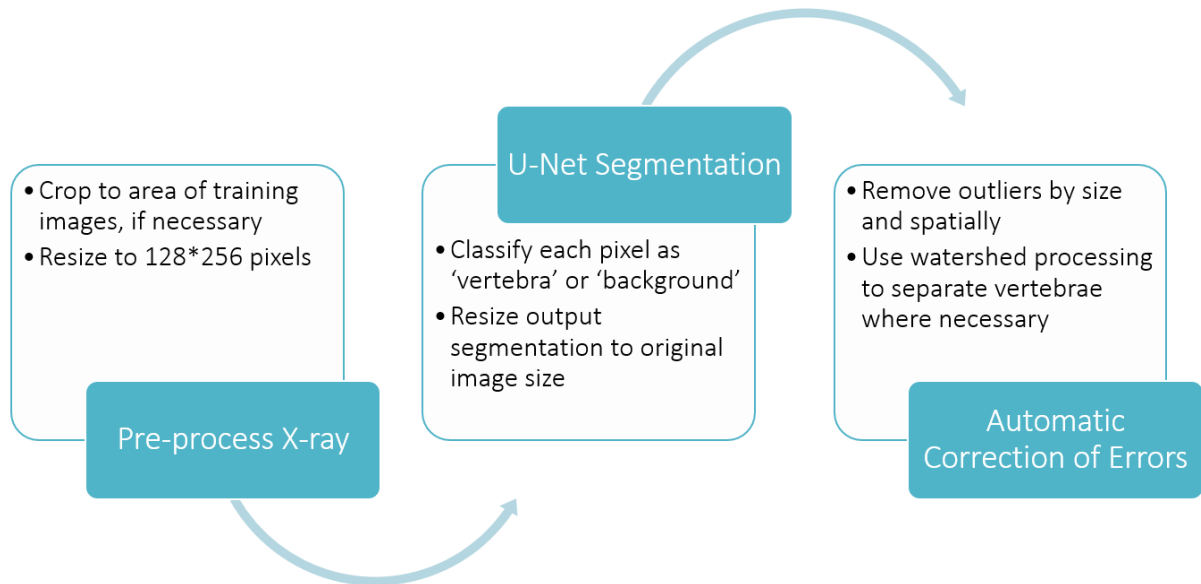


Figure 11. Summary of vertebral segmentation methods.

3.2. FITTING OF ENDPLATES

Linear endplates were automatically fit to each of the vertebral segmentations in order to calculate the Cobb angles. A minimum bounding rectangle function [36] was applied to the perimeter of each vertebra in order to determine its orientation, before applying a linear least-squares fit to the relevant edges. Again, here, the consistent structure of the spinal column was leveraged in order to automatically find and adjust for errors present in the initial fitted endplates. This includes removal of endplates with an outlier length; and replacement of an endplate with a weighted average of its neighbours, if an outlier change in slope is observed between consecutive vertebrae/endplates.

3.2.1. SUMMARY OF FITTING OF ENDPLATES PHASE

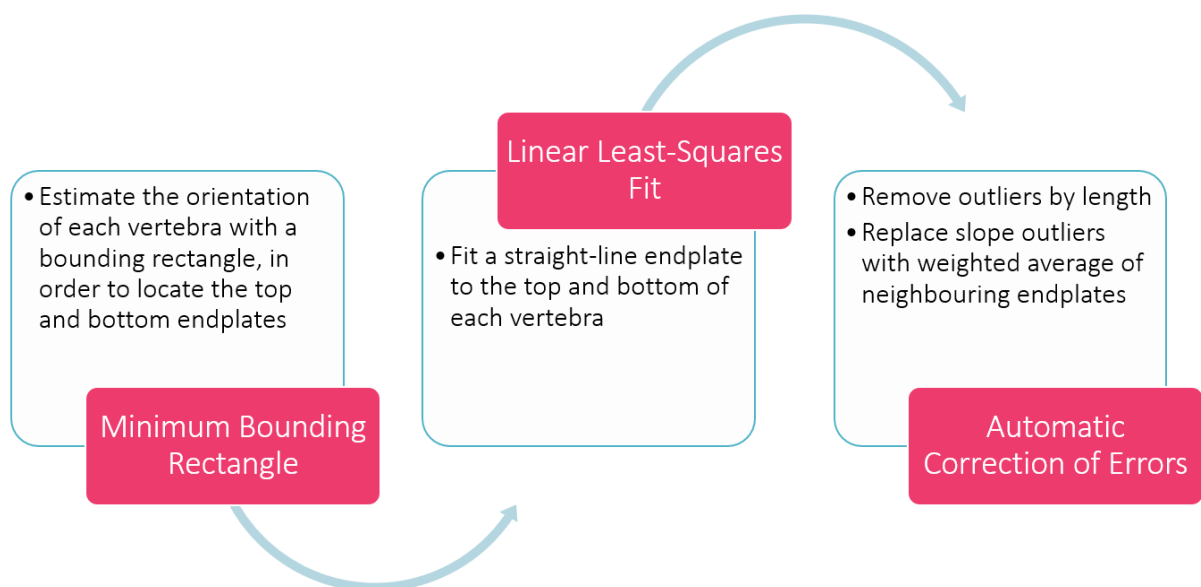


Figure 12. Summary of process to fit endplates to the segmented vertebrae.

3.3. COBB ANGLE CALCULATION

The PT, MT, and TL/L curves were automatically located and measured using the vertebral segmentations and fitted endplates.

The relevant curves are located as follows:

- Firstly, the apex of the MT curve is located as the vertebra between T6 and T11 whose centroid is most horizontally deviated from the average centroid of T1-T11.
- The most tilted vertebra between T4 and the identified MT apical vertebra is selected as the superior vertebra of the MT curve and, therefore, the inferior vertebra of the PT curve.
- The superior vertebra of the PT curve is identified as the vertebra, superior to the MT curve, whose slope makes the greatest angle with that of the associated inferior vertebra.
- The vertebrae between the MT apical vertebra and L2 are then traversed, and the inferior vertebra of the MT curve (and, consequently, the superior vertebra of the TL/L curve) is selected as the vertebra whose slope makes the greatest angle with the corresponding superior vertebra.
- Finally, the inferior vertebra of the TL/L curve is located as the vertebra, inferior to the MT curve, whose slope makes the greatest angle with that of the associated superior vertebra.

Each of the Cobb angles are then calculated as the angle between the slope of the superior endplate of the superior vertebra and the inferior endplate of the inferior vertebra, of the relevant curve.

3.3.1. SUMMARY OF COBB ANGLE CALCULATION PHASE

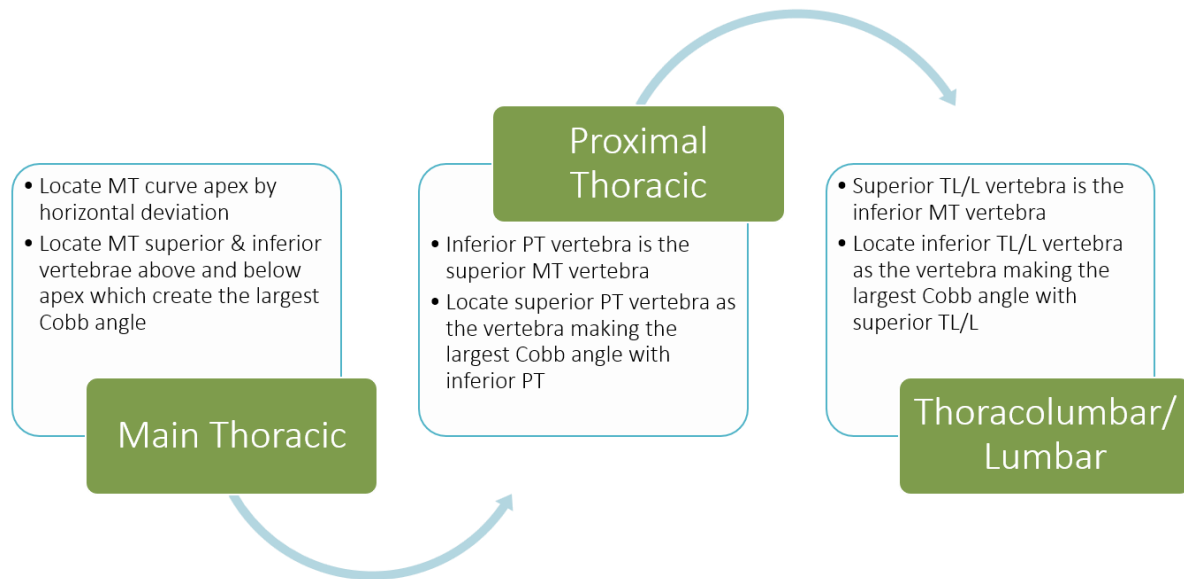


Figure 13. Summary of automatic Cobb angle calculation from the identified endplates.

3.4. LENKE CURVE TYPE PROBABILITY ANALYSIS

Each spinal x-ray with scoliosis can be categorised by a curve type, in accordance with Lenke classification [37]. The curve type is established by determining which of the PT, MT and TL/L curves are structural, and which one constitutes the major curve. The major curve is the curve with the largest Cobb angle measurement (with the exception of PT), and this curve is always structural. The other (minor) curves are defined as structural if their Cobb angle is 25° or greater. Further minor curve structural/non-structural criteria are defined for bending and sagittal-view spinal x-rays. However, solely standing AP x-rays will be considered, for the purposes of this study. This classification can be achieved automatically, given the PT, MT, and TL/L Cobb angles.

As well as automatically classifying the Lenke curve type based on the AP x-ray criteria, in this study, a novel approach has been developed to quantify the uncertainty in this classification. The variation in Cobb angle measurements using the proposed system can be modelled using a normal distribution with a standard deviation of 6.8°, as evidenced in Section 4.3.3. This model of variation constitutes a probability distribution of the true Cobb angle, given its measured value.

Integrating over the normal probability density function (pdf), Equation 1, allows for calculation of the area under sections of the curve, and thus, the probability that the function will take a value in a particular range. Applying this, we can assess the probability that the true Cobb angle will lie in a range of values, given its measured value.

Equation 1: Normal Probability Density Function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

In this case: μ = measured Cobb angle; σ = estimated standard deviation

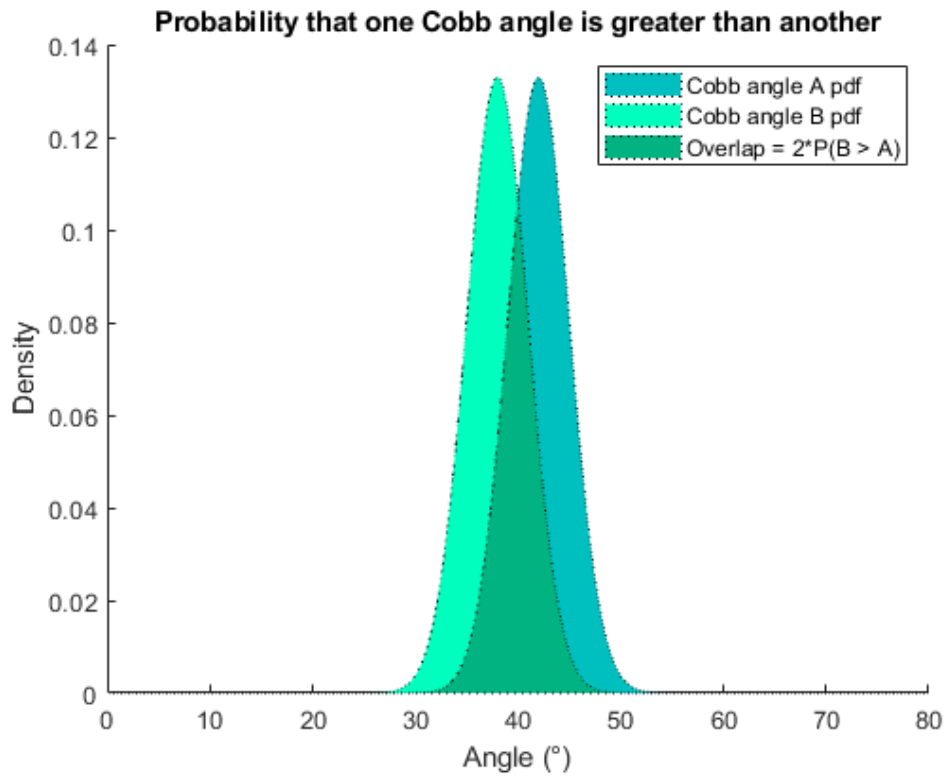


Figure 14. Probability that one Cobb angle is greater than another.

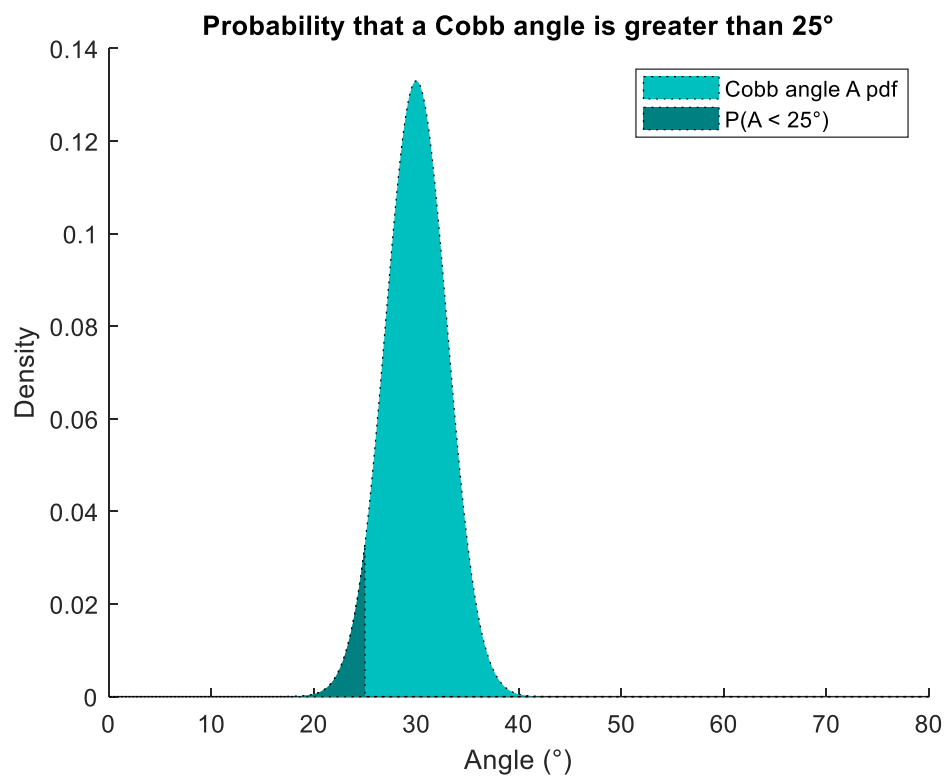


Figure 15. Probability that a Cobb angle is greater than 25°.

The probability that the TL/L curve is major (denoted $P(TL/L \text{ Major})$) can be calculated as the probability that the TL/L curve is greater than both the PT and MT curves:

Equation 2: Probability of TL/L being the Major Curve

$$P(TL/L \text{ Major}) = P(TL/L > PT) * P(TL/L > MT)$$

The guidelines assert that MT should be assigned the major curve if PT yields the largest Cobb angle measurement. Thus, if the TL/L curve is not major, then the MT curve must be.

Equation 3: Probability of MT being the Major Curve

$$P(MT \text{ Major}) = 1 - P(TL/L \text{ Major})$$

It is necessary to calculate the probability that the TL/L curve is greater than each other curve, in order to evaluate Equation 2. The overlap of the TL/L probability distribution with each of the other distributions can be used to calculate this probability, as shown in Figure 14. Furthermore, the probability that a Cobb angle is 25° or greater can be simply calculated as the portion of the probability distribution greater than 25°, as shown in Figure 15.

Applying this analysis, the probability of each curve type can be calculated, and thus the uncertainty in classification can be quantified. The associated probabilities are outlined in TABLE 3. This table accounts for all possible permutations of the three angles that result in each Lenke curve type. There is a single combination that remains undefined under the Lenke system. This is the case in which TL/L is the major curve and PT is also structural, but the MT curve is non-structural. This case is unlikely, given the structure of the spine. If both PT and TL/L are structural, it is likely that the MT curve will be structural as its angle is measured between the PT inferior and TL/L superior vertebrae. Nonetheless, it is possible and would most likely result from a triple major curve. For the purposes of this study, however, it is not added into the probability of any curve type.

TABLE 3: PROBABILITY OF EACH LENKE CURVE TYPE

<i>Curve</i>	<i>Probability</i>
<i>Type 1:</i>	$[P(PT < 25) * P(MT > 25) * P(TL/L < 25)] + [P(PT < 25) * P(MT < 25) * P(TL/L < 25) * P(MT \text{ Major})]$
<i>Type 2:</i>	$[P(PT > 25) * P(MT > 25) * P(TL/L < 25)] + [P(PT > 25) * P(MT < 25) * P(TL/L < 25)]$
<i>Type 3:</i>	$[P(PT < 25) * P(MT > 25) * P(TL/L > 25) * P(MT \text{ Major})]$
<i>Type 4:</i>	$[P(PT > 25) * P(MT > 25) * P(TL/L > 25)] + [P(PT > 25) * P(MT < 25) * P(TL/L > 25) * P(MT \text{ Major})]$
<i>Type 5:</i>	$[P(PT < 25) * P(MT < 25) * P(TL/L > 25)] + [P(PT < 25) * P(MT < 25) * P(TL/L < 25) * P(TL/L \text{ Major})]$
<i>Type 6:</i>	$[P(PT < 25) * P(MT > 25) * P(TL/L > 25) * P(TL/L \text{ Major})]$
<i>Undefined:</i>	$[P(PT > 25) * P(MT < 25) * P(TL/L > 25) * P(TL/L \text{ Major})]$

This analysis can be extended to reflect inter-observer variation by adjusting the standard deviation accordingly, and the probability calculations will hold for manually measured Cobb angles. Furthermore, bending and kyphosis Cobb angles can easily be incorporated into the analysis, for a more complete assessment of the uncertainty in Lenke curve type classification.

3.4.1. SUMMARY OF LENKE CURVE TYPE PROBABILITY ANALYSIS

PHASE

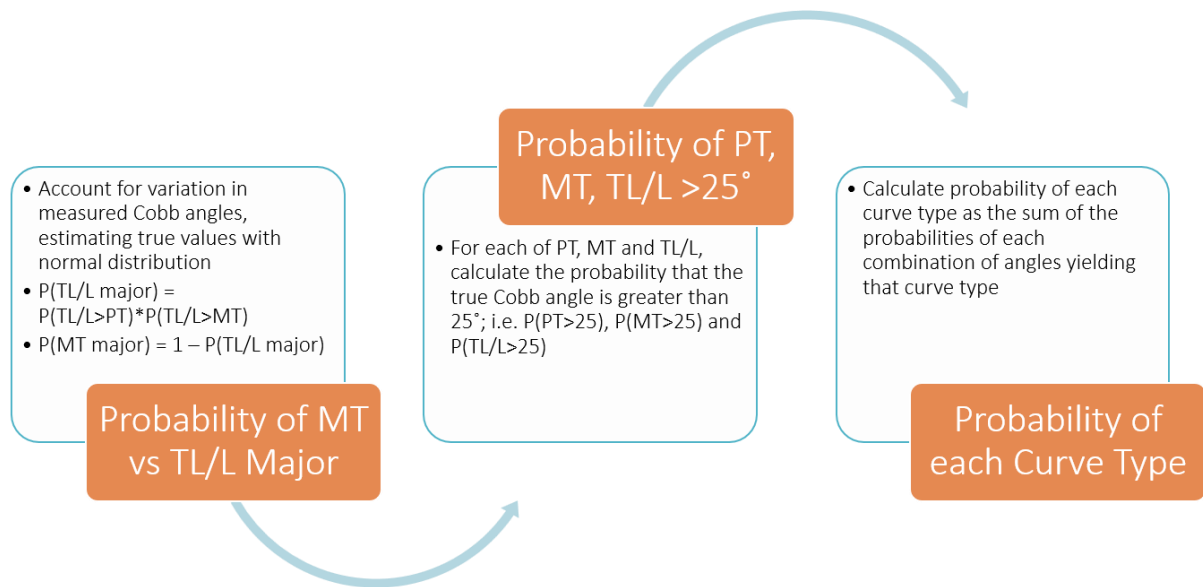


Figure 16. Summary of calculation of the probability of each Lenke curve type.

3.5. CONCLUSION

The four phases presented in this chapter combine to provide a fully automatic end-to-end scoliosis assessment tool. The proposed system integrates novel methodology with prior art and traditional assessment techniques. The following chapter will address the functioning of the developed models in practice, with a detailed evaluation of performance.

4. RESULTS

In this chapter, the results from the four phases of the proposed system will be reported and analysed. The accuracy of the output at each phase was evaluated exclusively using the 128 image *Testset*, both through visual inspection and using relevant performance metrics in comparison with the ground-truth expected output. Some of the ground-truth comparisons are suboptimal, due to the limited availability of desired, expertly measured, segmentations and clinical metrics. The manually labelled vertebral corners in the *Testset*, with the necessary corrections applied (see Section 3.1.1.2), were used as the ground-truth endplates. The ground-truth segmentations, Cobb angles, and Lenke curve type classifications were each generated automatically from these manually annotated endplates using the algorithms described in Sections 3.1.1.3, 3.3, and 3.4, respectively.

The limitations in ground-truth comparisons can lead to both over- and underestimation of performance. For example, if a set of automatically generated ground-truth metrics are 90% accurate, then a perfect system evaluated against these ground-truths would yield an accuracy of 90%. On the other hand, if the developed algorithm is subject to the same inherent flaws as the ground-truth set, it will produce similar errors and thus its accuracy will be falsely inflated. For a more robust evaluation of the proposed system, each phase should be compared against the average of multiple clinical experts using their current methods on an entirely different dataset. However, this clinical evaluation was outside the scope of the current study.

Visualisation of the performance will be achieved by overlaying the computed outputs on random samples of x-rays from the *Testset*, in order to combat the limitations outlined above. The sections are organised so that each type of overlaid output will be visualised with 3 separate random samples of 5 x-rays in total.

4.1. VERTEBRAL SEGMENTATION

4.1.1. U-NET TRAINING

The chosen U-Net parameters and training configuration resulted in steady improvement of the Tversky loss over the course of training, as shown in Figure 17. The network reached optimal performance on the validation set within 100 epochs.

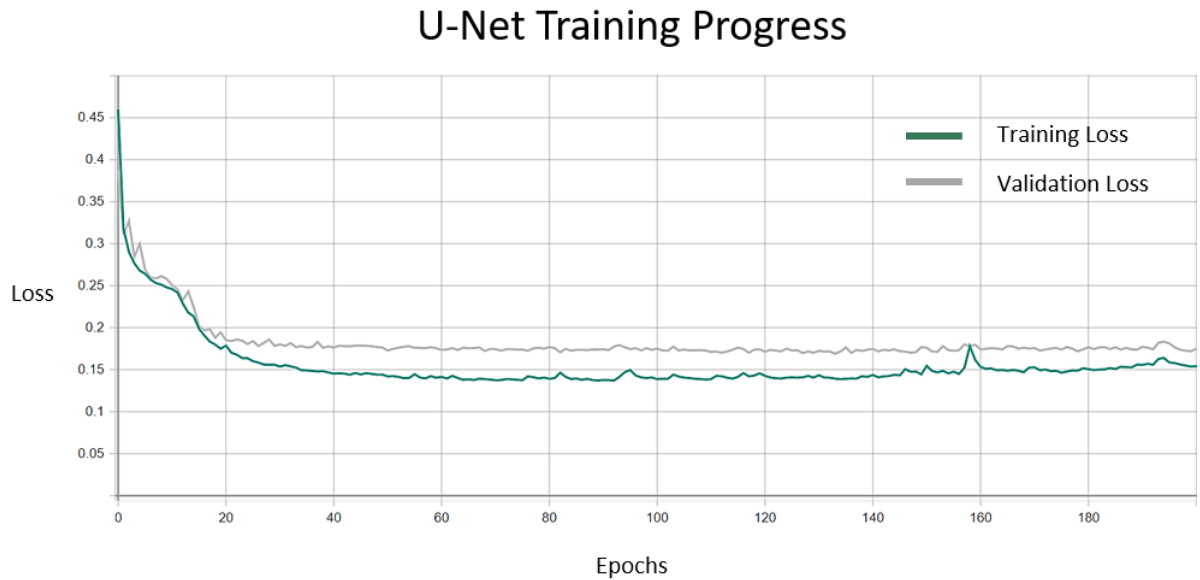


Figure 17. U-Net training progress.

4.1.2. VISUAL ASSESSMENT OF PERFORMANCE

As evidenced by the sample of images in Figure 18, the U-Net successfully learnt to segment the vertebrae accurately. As the network was trained to segment vertebrae T1-L5, the output segmentation map is usually confined to this area. However, occasionally, the network segments more or less vertebrae. The developed methods for estimation of endplates and Cobb angles are robust to this scenario. Furthermore, the network responds well to images containing 6 lumbar vertebrae, an anomaly occurring in approximately 10% of the population [38]. In

contrast, landmark detection methods are limited to a defined number of outputs. The network is also robust to cases with image artefacts, such as overlaid text, surgical pins, or bracing.

The network is further aided by the automatic detection and removal of the occasional errors in prediction. Errors that could be automatically identified by position or size were eliminated; and vertebrae that were mistakenly merged in the U-Net output were automatically separated, to allow for more accurate fitting of endplates. The result is a highly accurate depiction of the location and shape of each vertebra in the input image, allowing for identification of each of the desired endplates.

Random Sample of Images & Corresponding Vertebral Segmentations

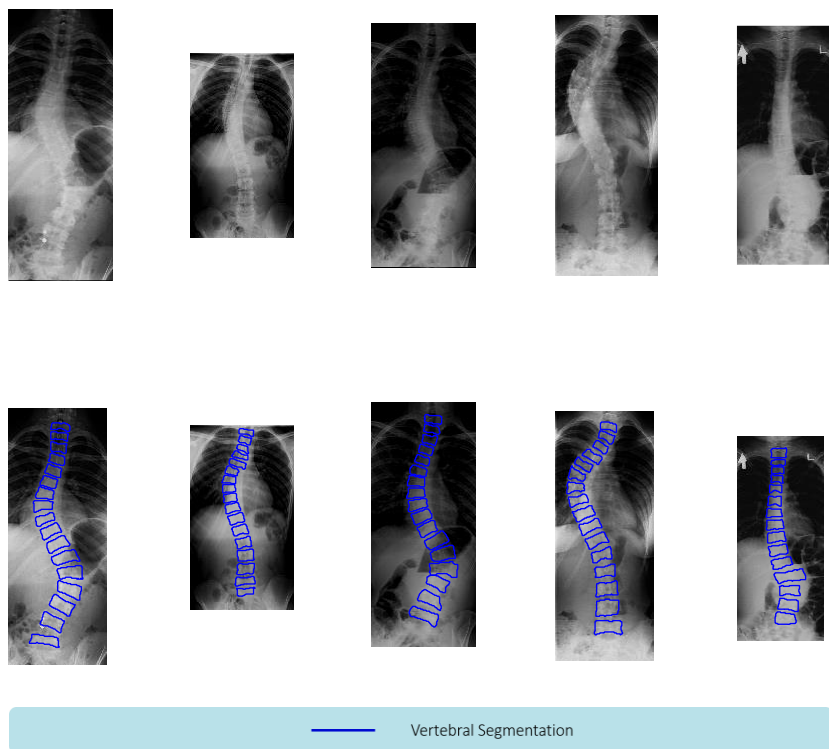


Figure 18. Vertebral segmentation performance on random sample of images from the Testset.

4.1.3. COMPARISON WITH GROUND-TRUTH

Vertebral segmentation performance was evaluated in comparison to the ground-truth segmentations, generated automatically using the algorithm outlined in Section 3.1.1.3. Accuracy metrics were chosen carefully to avoid impact of the imbalance in class, between ‘vertebra’ and ‘background’ pixels. The Dice similarity coefficient and balanced accuracy rate were selected; the former is a measure of the overlap between predicted and ground-truth segmentations, and the latter, the average of the proportions classified correctly for each individual class. These metrics are displayed in TABLE 4, both for the raw prediction of the U-Net, and this prediction after errors have automatically been corrected with post-processing.

The balanced accuracy rate suggests similar, very accurate, performance in both the initial U-Net segmentation and the processed output. However, the Dice score demonstrates the significance of this post-processing step. Dice appears more sensitive to additional objects in the segmentation, incorrectly classified as vertebrae, as well as the merging of borders between adjacent vertebrae. Both of these issues are successfully alleviated automatically with post-processing. Additionally, the Dice score is likely sensitive to the shortcomings of the ground-truth segmentations. The boundaries of the vertebrae in the ground-truth segmentation, while sufficiently accurate to train the U-Net, will inevitably contain errors. These errors negatively impact the computed accuracies below. Furthermore, the developed vertebral segmentation method may result in the segmentation of more vertebrae than the consistent 17 of the ground-truth. This will further hinder performance evaluation.

TABLE 4: VERTEBRAL SEGMENTATION ACCURACY		
FULLY AUTOMATIC VS. AUTOMATICALLY GENERATED FROM MANUAL ENDPLATES		
	<i>Raw U-Net Prediction</i>	<i>After Automatic Error Correction</i>
<i>Dice Similarity Coefficient</i>	0.74	0.84
<i>Balanced Accuracy Rate</i>	0.92	0.93

4.2. FITTING OF ENDPLATES

4.2.1. VISUAL ASSESSMENT OF PERFORMANCE

The proposed method for fitting endplates to the vertebral segmentation performs accurately. This is made clear in the examples of Figure 19. The algorithm successfully determines the endplate location and slope for each of the segmented vertebrae. This developed system accounts for possible errors in the vertebral segmentation, by automatically correcting for endplates with an outlier change in length/slope in comparison to neighbouring endplates. In conjunction with the automatic correction of errors in the vertebral segmentation itself, this results in a highly robust extraction of endplate information from an input x-ray.

Random Sample of Vertebral Segmentations & Corresponding Fitted Endplates

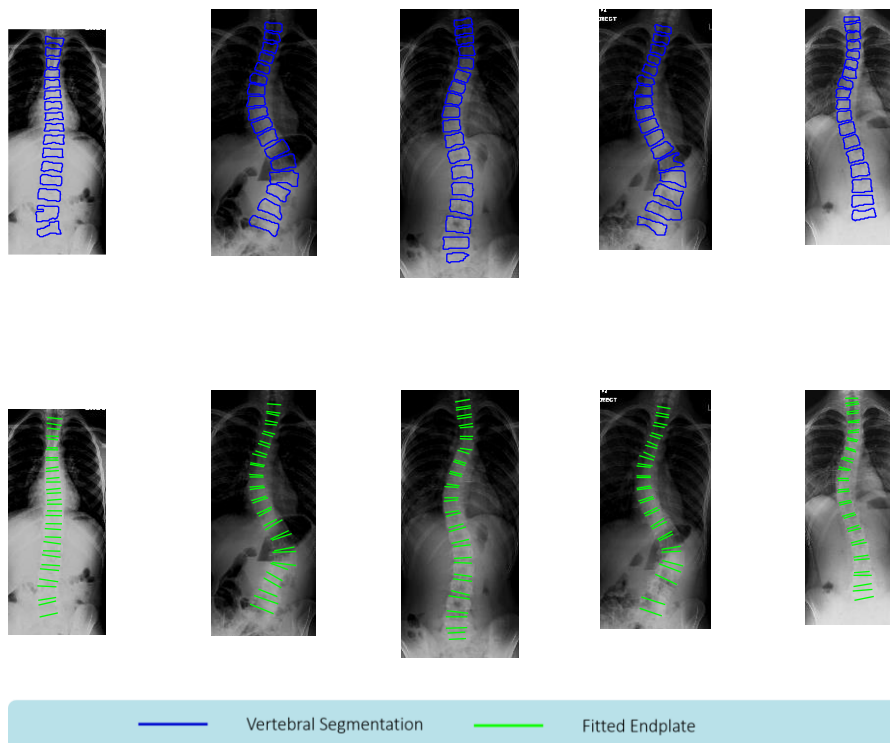


Figure 19. Endplate fitting performance on random sample of images from the Testset.

4.2.2. COMPARISON WITH GROUND-TRUTH

The fitted endplate slopes were directly compared to that of the manually annotated endplates. As described in Section 4.1.2, the proposed fully automatic process can result in a different number of segmented vertebrae to the fixed 17 vertebrae of the ground-truth data. Consequently, there may also be a different number of fitted endplates in some cases. This did not present an issue in subsequent Cobb angle calculation. However, for the purposes of this endplate fitting performance evaluation, the fitted endplates were adjusted to contain 34 endplates – the consistent number of ground-truth endplates.

In order to achieve this, the fitted endplate slopes for each x-ray were treated as an array, and its alignment with that of the ground-truth equivalent was evaluated by summing the absolute difference between the two arrays. If there were more than 34 fitted endplates, then the 34 contiguous endplates that aligned most closely with the ground-truth equivalent were used, and the rest discarded. In the rarer occasions with less than 34 fitted endplates, they were positioned with the best alignment of contiguous endplates, and gaps at either end of the array were filled with the nearest available fitted slope so that 34 values were achieved.

The metrics chosen to compare the automatically obtained endplate slopes against the manually measured values were the mean absolute difference (MAD), Pearson correlation coefficient, and intraclass correlation coefficient (ICC). The MAD offers a valuable insight into the magnitude of the expected deviation between the two methods. The Pearson correlation coefficient reflects linear correlation between the methods. The ICC is commonly used in the comparison of medical assessment methods, as it provides a measure of the agreement and inter-rater reliability between methods. The three metrics offer a rounded view of the agreement between the predicted and ground-truth methods. Consequently, they were applied

for evaluation of each of the continuous outputs of the proposed system, namely, the endplate slopes, Cobb angles and Lenke curve type probabilities.

In addition to these statistical metrics, the data were plotted in various configurations in order to assess any inherent bias or unusual characteristics between the methods. A scatter plot was generated, in order to visualise the relationship between the predicted and ground-truth methods. The distribution of differences between methods was plotted to assess its shape. A Bland-Altman plot was generated, to visualise these differences as a function of the mean slope between the methods. This combination of visualisation techniques provides a robust analysis of any bias or pitfalls in the developed model. Analogous to the statistical metrics, these plotting techniques were also applied in the evaluation of each of the continuous outputs of the proposed system.

The computed metrics for the fitted endplates are displayed in TABLE 5. Both the Pearson correlation coefficient and the ICC suggest excellent agreement between the fully automatic and manual methods. Additionally, the MAD reflects a relatively small deviation in slope between the two methods.

TABLE 5: ENDPLATE SLOPES ACCURACY FULLY AUTOMATIC VS. MANUAL	
<i>Mean Absolute Difference (MAD)</i>	3.88°
<i>Pearson Correlation Coefficient</i>	0.92
<i>Intraclass Correlation Coefficient (ICC)</i>	0.92

The three plots for the endplate slope data are shown in Figure 20. The scatterplot displays a strong linear correlation with no obvious bias. The distribution of errors is very close to a normal distribution. As desired, this distribution is approximately centred at zero with a

relatively low variance. The Bland-Altman plot cements this assessment, displaying an acceptable 95% confidence interval for the expected variation between the two methods of $0.36 \pm 10.61^\circ$. Although there are outlier errors of up to roughly 40° , the proposed method generally agrees with the manually annotated slopes and there is no evidence of systematic error.

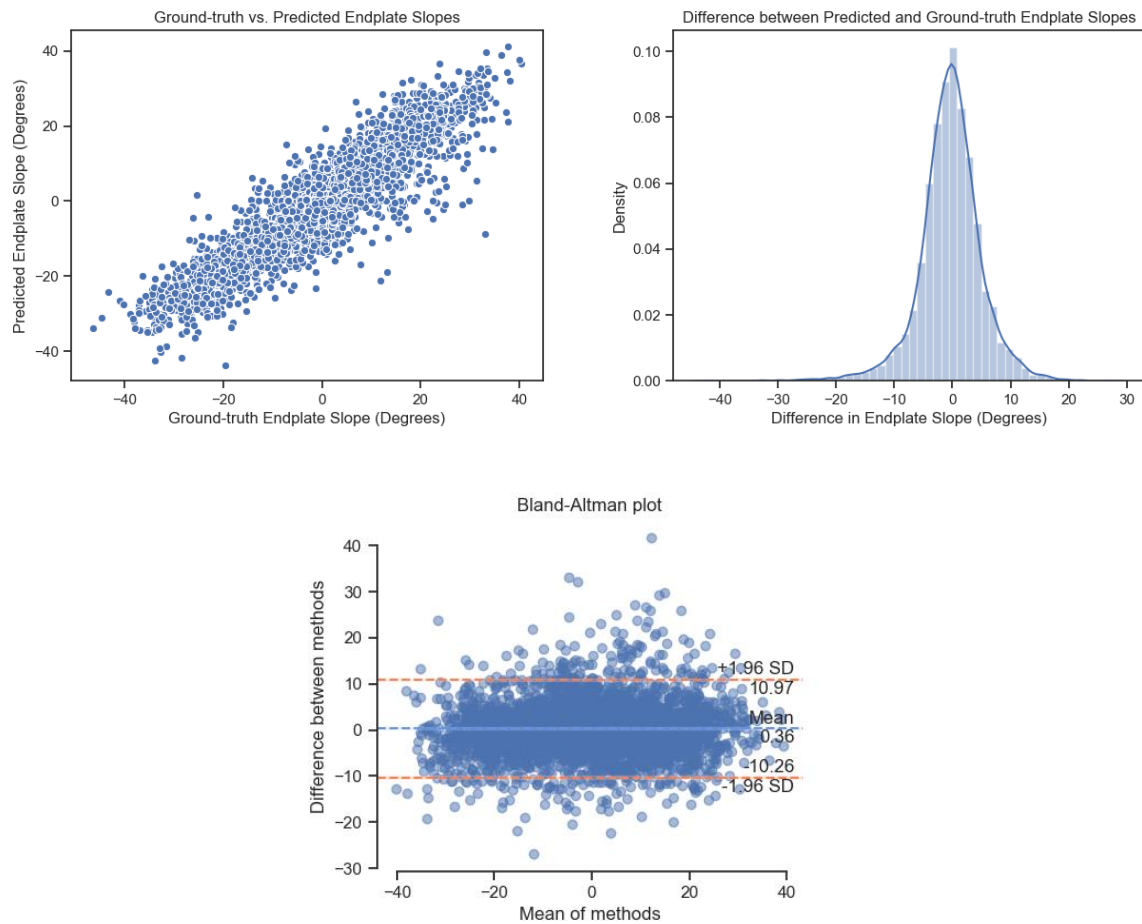


Figure 20. Accuracy of fitted endplates compared to those manually annotated.

4.3. COBB ANGLE CALCULATION

4.3.1. VISUAL ASSESSMENT OF PERFORMANCE

As shown in Figure 21, the Cobb angle calculation phase of the proposed system is successful in locating the PT, MT and TL/L curves. The selected vertebrae appear appropriate in all cases.

Random Sample of Fitted Endplates & Corresponding Cobb Angles

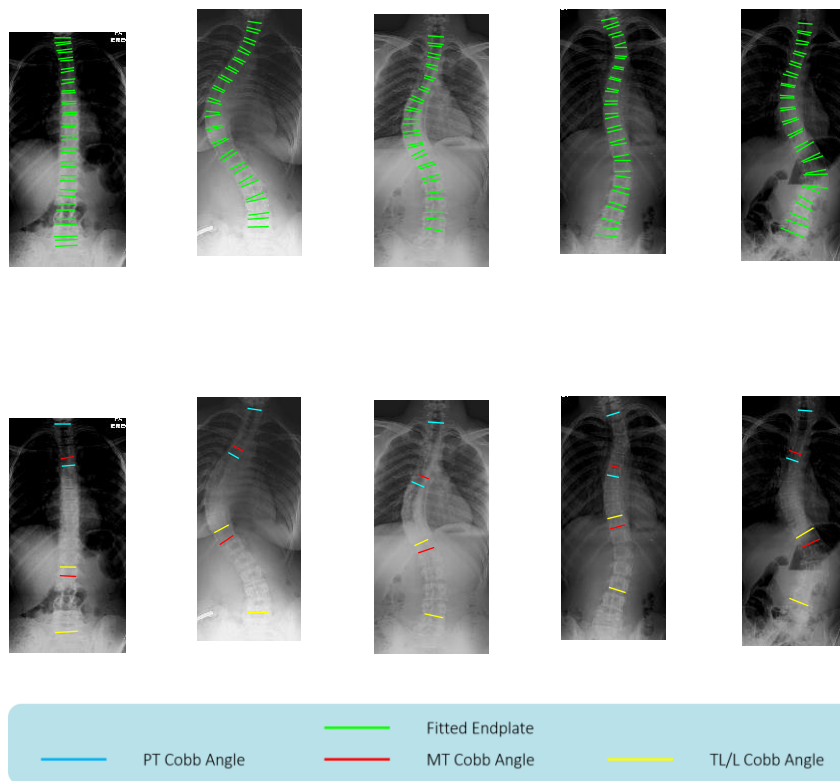


Figure 21. Cobb angle selection performance for random sample of images from the Testset.

4.3.2. COMPARISON WITH GROUND-TRUTH

The Cobb angles measured by the proposed fully automatic system were compared with ground-truth angles, generated automatically from the manually labelled endplates using the algorithm described in Section 3.3. The resulting statistical metrics are outlined in TABLE 6. Similar to the fitted endplate performance, the Pearson correlation coefficient and ICC indicate excellent agreement between the two methods and the MAD displays a relatively low magnitude of deviation between the methods. However, the computed metrics indicate slightly more variation between the predicted and ground-truth Cobb angles, when compared to that of the endplates. This is expected, as the variation in multiple endplates can compound into larger variation in the resulting Cobb angles. Furthermore, there is increased opportunity for variation in the selection of superior and inferior vertebrae for each curve.

It must also be considered that the designed ground-truth Cobb angles are not a true real-world comparison. In assessing the accuracy of the proposed method for vertebral selection in Cobb angle calculation, it would be desirable to compare against clinicians' selection and manual measurement. This suboptimal ground-truth alternative somewhat limits the interpretation of the comparison, as the PT, MT and TL/L curves are located using the same algorithm in both cases. Despite this, the comparison offers an insight into the magnitude of the variation between Cobb angles resulting from the proposed fully automatic method and those obtained through manual endplate measurement.

TABLE 6: COBB ANGLES ACCURACY	
FULLY AUTOMATIC VS. AUTOMATIC CALCULATION WITH MANUAL ENDPLATES	
<i>Mean Absolute Difference (MAD)</i>	5.08°
<i>Pearson Correlation Coefficient</i>	0.90
<i>Intraclass Correlation Coefficient (ICC)</i>	0.90

The plots of predicted and ground-truth Cobb angle data are displayed in Figure 22. Again, the results are similar to those observed for the endplate slope data. The scatter plot displays the desired linear relationship. The distribution of errors is approximately normal, and the Bland-Altman plot identifies the 95% confidence interval of the expected variation of $-0.38 \pm 13.47^\circ$. Although there are outliers with large differences in Cobb angles between the methods, there does not seem to be evidence of a systematic bias in the proposed system.

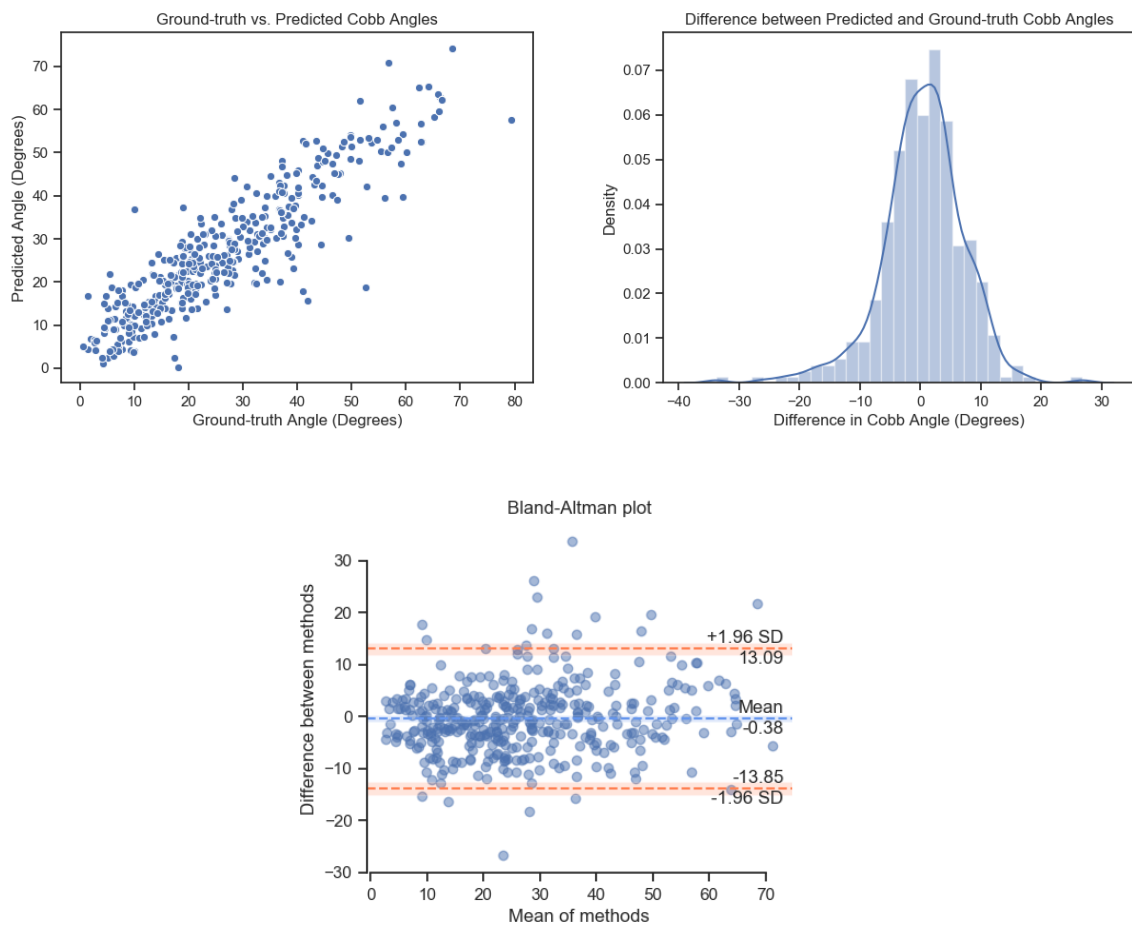


Figure 22. Accuracy of fully automatic Cobb angles compared to ground-truth (those obtained automatically from manually annotated endplates).

4.3.3. NORMALLY DISTRIBUTED INTER-OBSERVER VARIATION IN COBB ANGLES

The inter-observer variation between the proposed fully automatic Cobb angles and the ground-truth Cobb angles was explored further, in order to assess the validity of the assumption of normality for the purposes of the Lenke curve type probability calculations in Section 3.4. The Shapiro-Wilk test was applied to formally test this assumption, see TABLE 7. The table shows that the variation between the two methods of Cobb angle calculation failed the test for normality. At the 95% confidence level, this test concludes that we should reject the null hypothesis that the data follows an underlying normal distribution.

TABLE 7: SHAPIRO-WILK TEST FOR NORMALITY DIFFERENCE BETWEEN PREDICTED AND GROUND-TRUTH COBB ANGLES	
<i>W</i>	0.96
<i>p-value</i>	2.26×10^{-8}
<i>Normally Distributed</i>	False – Reject Null Hypothesis

However, this test can be sensitive to very minor deviations from normality, especially if overpowered with a large sample size. In this case, the sample size is 3 Cobb angles in each of 128 images (384 datapoints in total). Given the outcome of the Shapiro-Wilk test, it is useful to apply additional techniques in order to investigate the effect size. A Q-Q plot, Figure 23, was applied to assess the magnitude of deviation of the measured quantiles from those expected of a normal distribution. It is clear from this plot that the observed variations follow a normal distribution very closely. The computed R^2 value asserts that roughly 96% of the variation in differences between the two methods is accounted for by a normal distribution. Thus, although there are minor deviations from a normal distribution, the assumption of normality remains valid for the required probability calculations.

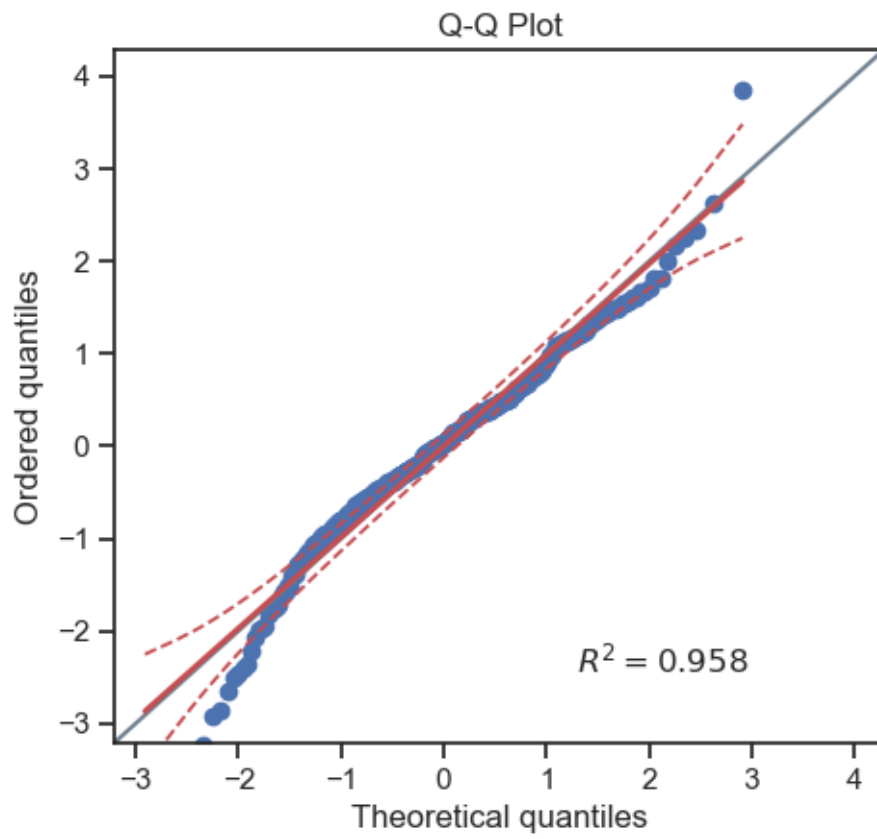


Figure 23. Q-Q plot of the difference in Cobb angle between prediction and ground-truth, in order to assess validity of the assumption of normally distributed variation.

4.4. LENKE CURVE TYPE PROBABILITY ANALYSIS

In this section, both traditional Lenke curve type classification and the newly developed probability analysis performance will be evaluated for the proposed fully automatic system. It is worth noting that Lenke classification is designed for AIS assessment only, whereas the images used in this study cover a broad range of scoliosis patients. In real-world application, the developed model may well yield improved performance, using AIS images only.

4.4.1. TRADITIONAL CURVE TYPE CLASSIFICATION

The performance of the proposed fully automatic Lenke classification method was compared to the Lenke classification achieved using the manually annotated endplates and the algorithms outlined in Sections 3.3 and 3.4. It should be noted that, bending and sagittal-view x-rays would usually be utilised for complete Lenke curve type classification. However, for the purposes of this analysis, the standing AP x-ray criteria will be naively assumed conclusive in determining the curve type. Different statistical metrics were applied here, as the Lenke curve type is a nominal variable. In this case, the percent agreement and Cohen's kappa statistic were used. The percent agreement indicates the proportion of cases in which both methods classified the same curve type. Cohen's kappa statistic is a measure of the inter-rater reliability between the two methods, analogous to the ICC.

The computed metrics are displayed in TABLE 8. Both measures indicate moderate agreement between the predicted and ground-truth methods. This is a significant decline in performance, considering the excellent agreement between the Cobb angles on which the classification is based. This poor performance can be explained by the volatile nature of the Lenke curve type classification. For each x-ray, if the variation between the two methods results in any of the three Cobb angles crossing the relevant 25° threshold or causes a different angle to be major, then the methods' resulting curve type classifications will disagree.

TABLE 8: LENKE CURVE TYPE CLASSIFICATION ACCURACY FULLY AUTOMATIC VS. AUTOMATIC CLASSIFICATION WITH MANUAL ENDPLATES	
<i>Percent Agreement</i>	59.38%
<i>Cohen's Kappa Statistic (κ)</i>	0.48

The scatter plot in Figure 24 illustrates the agreement between the predicted and ground-truth curve types. The most common disparity between the methods appears to be disagreement between curve types 1 and 5. In this case, all minor curves are deemed non-structural and the sole difference between the two curve types is the selection of the major curve. Distinguishing between curve types 3 and 4 also presents a pitfall. Again, in this case, the two curve types disagree in one area only – whether or not the PT angle is structural.

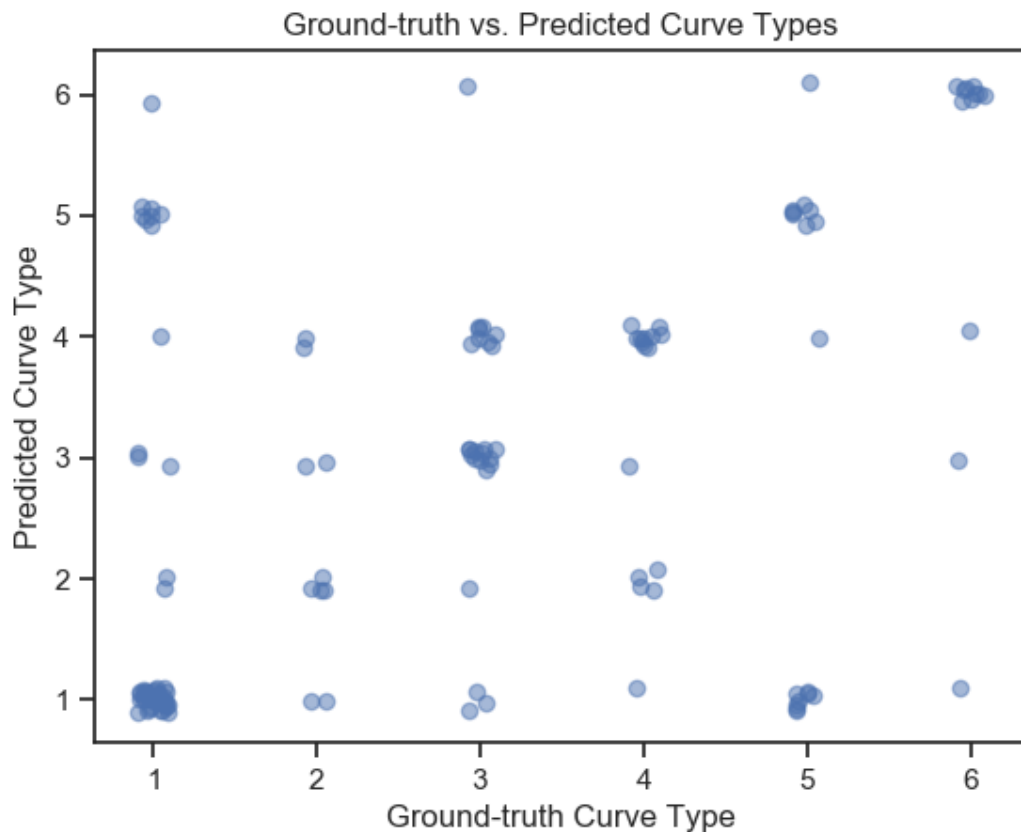


Figure 24. Accuracy of fully automatic Lenke curve type classification compared to ground-truth (that obtained automatically from manually annotated endplates).

4.4.2. CURVE TYPE PROBABILITY ANALYSIS

4.4.2.1. VISUAL ASSESSMENT OF PERFORMANCE

Applying the Lenke curve type probability analysis outlined in Section 3.4, the performance of the proposed fully automatic method to quantify the probability of each curve type was evaluated. A sample of the resulting curve type probability distributions are displayed in Figure 25. This highlights the utility of the probability analysis in identifying uncertainty in predictions. While some cases yield a high level of confidence in the curve type classification, others display significant uncertainty, identifying considerable likelihood of two or more curve types.

Random Sample of Cobb Angles & Corresponding Lenke Curve Type Probabilities

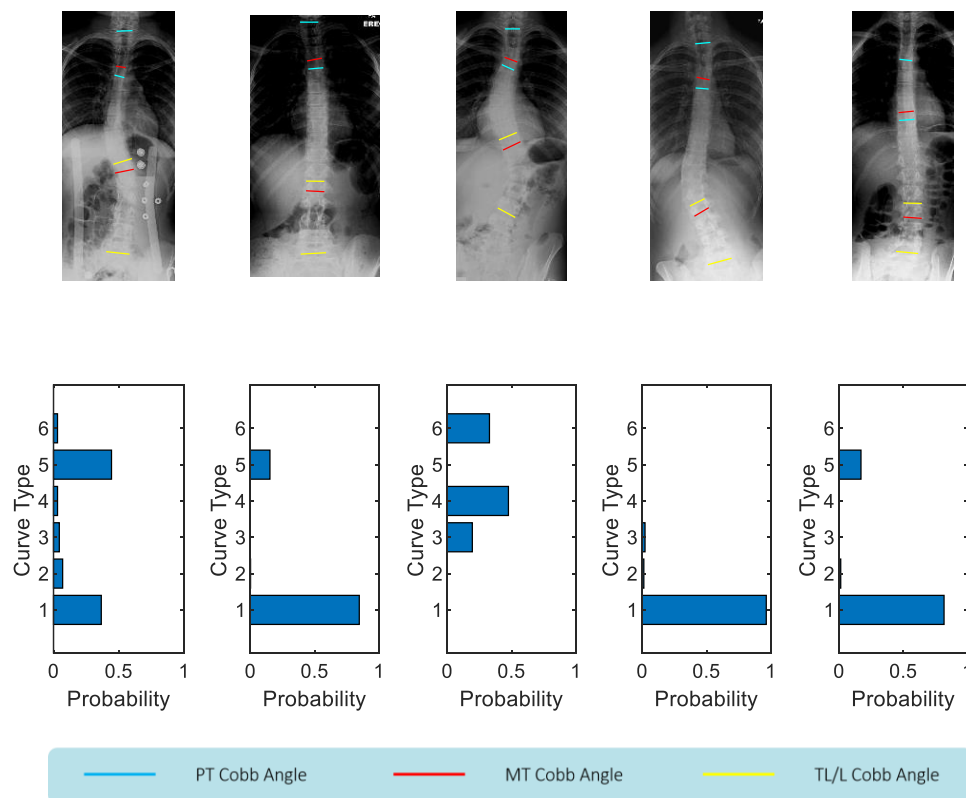


Figure 25. Lenke curve type probability performance for a random sample from the Testset.

4.4.2.2. COMPARISON WITH GROUND-TRUTH

The curve type probabilities were compared to those of the ground-truth method. This allowed for a more nuanced analysis of the agreement between the two methods. As opposed to studying the single curve type classification, the likelihood of each curve type could be assessed for each x-ray. The resulting metrics are outlined in TABLE 9. The Pearson correlation coefficient and ICC suggest good agreement between the two methods, and the MAD shows a reasonable expected deviation between the two probability predictions.

TABLE 9: LENKE CURVE TYPE PROBABILITY ANALYSIS ACCURACY FULLY AUTOMATIC VS. AUTOMATIC CALCULATION WITH MANUAL ENDPLATES	
<i>Mean Absolute Difference (MAD)</i>	0.10
<i>Pearson Correlation Coefficient</i>	0.76
<i>Intraclass Correlation Coefficient (ICC)</i>	0.75

Plotting the data, Figure 26, provides a further insight into the variation between predicted and ground-truth methods. The scatter plot indicates that there is generally low variation in unlikely curve types, i.e. the predicted and ground-truth methods often agree that multiple curve types are highly unlikely given the x-ray data. However, there is a high degree of variation for probabilities greater than approximately 0.2. This indicates that the methods often disagree on the magnitude of the uncertainty in the curve types that are deemed probable. The Bland-Altman plot supports this interpretation, indicating very low deviation with the mean of the methods less than 0.2 and substantial deviation thereafter. The Bland-Altman plot also identifies the 95% confidence interval of the expected deviation between methods as 0.00 ± 0.32 . The distribution of differences between the methods is less close to a normal distribution than those observed previously. The density of differences contains a large peak around ± 0.05 , however, the tails remain substantial up to roughly ± 0.3 .

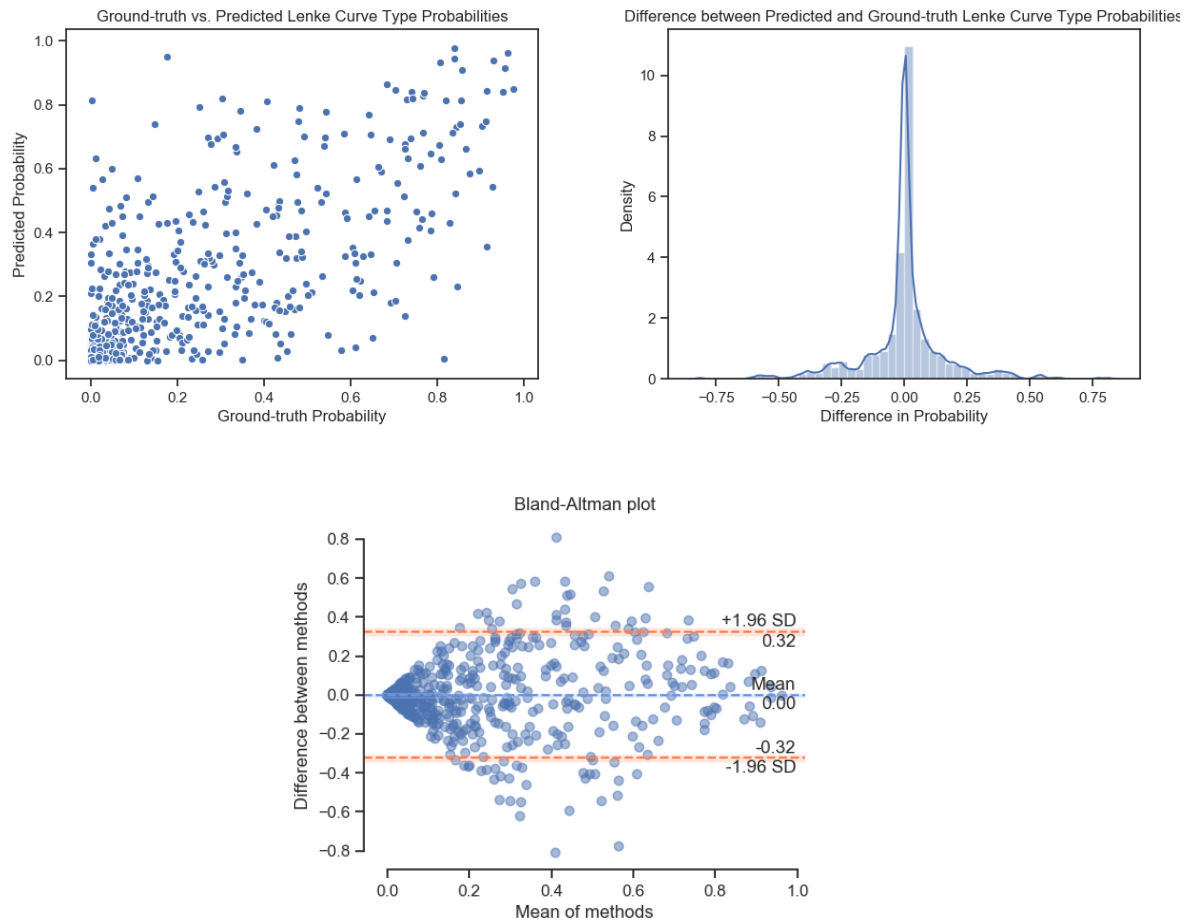


Figure 26. Accuracy of fully automatic calculation of Lenke curve type probabilities compared with ground-truth (probabilities calculated automatically using manually annotated endplates).

4.5. REVIEW OF END-TO-END PERFORMANCE

This chapter has demonstrated the efficacy of the proposed fully automatic scoliosis assessment tool. The end-to-end performance, through each phase of the system, is summarised in Figure 27. An in-depth analysis and exploration of the impact of these novel methods will follow.

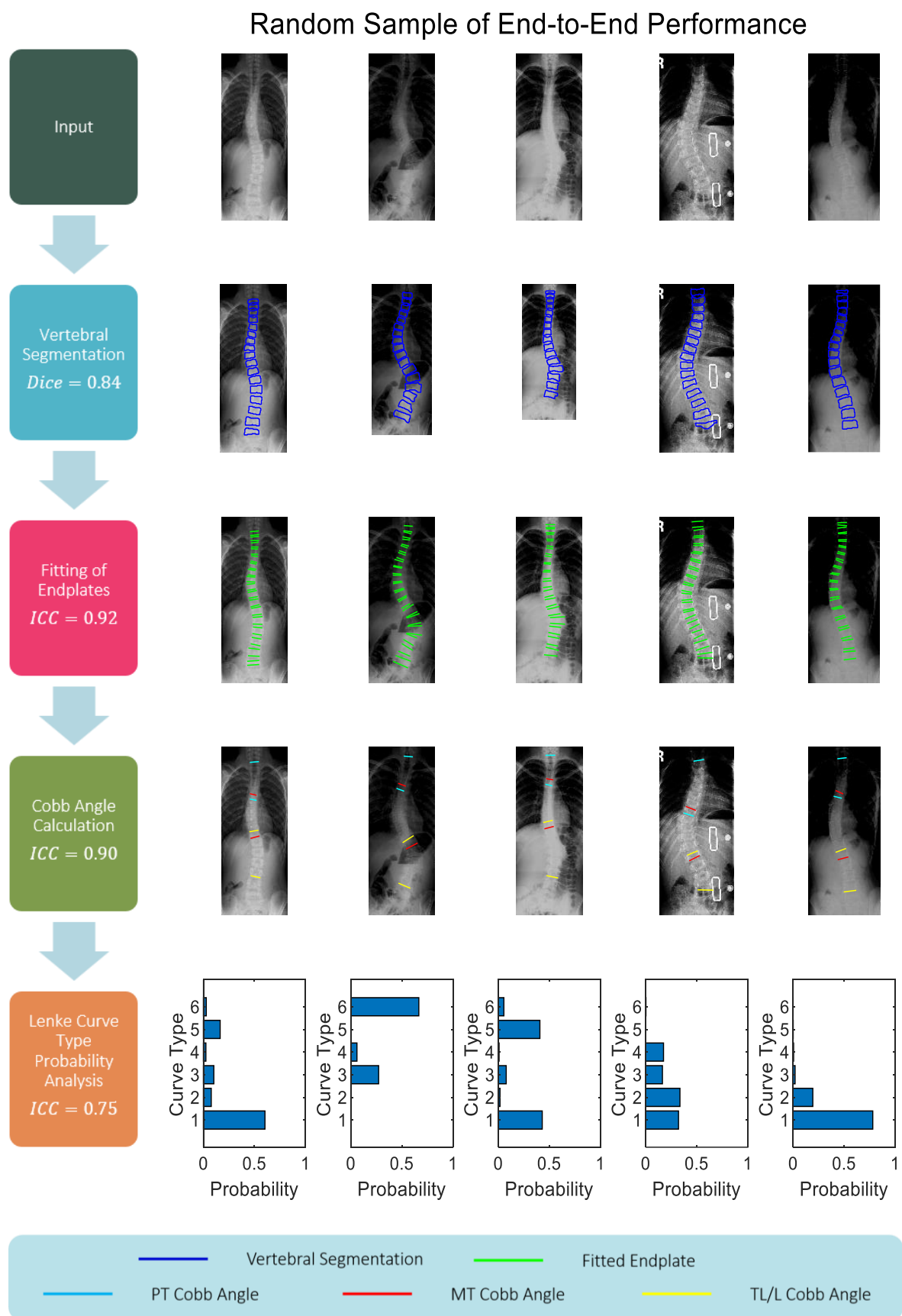


Figure 27. End-to-end performance of the proposed system for a random sample of images from the Testset.

5. DISCUSSION

5.1. ANALYSIS OF FINDINGS

The fitted endplates achieved comparable results to those annotated manually, achieving a MAD of 3.88° , very strong correlation, and no evidence of systematic error. This is, perhaps, the most appropriate measure for evaluation of the proposed system; given the direct comparison of the developed automatic tool with a truly manual alternative. In contrast, interpretation of the vertebral segmentation, Cobb angle calculation, and Lenke curve type classification performance is somewhat limited. This is due to the potential omission of systematic error, as a result of comparison with ground-truth data generated, in part, using the automatic algorithms under test. Nonetheless, the supporting samples, visually displaying the proposed system's performance, provide a clear means of evaluation for these outputs. In addition, the overall performance of the proposed system is comparable to that observed in the state-of-the-art systems reviewed in section 2.2. However, direct comparison with these methods is limited by variation in experimental design.

The conversion of ground-truth corner landmarks to ground-truth segmentation maps for training was successful. The trained U-Net, coupled with automatic removal of errors using *a priori* information, resulted in highly accurate and robust vertebral segmentations. This is demonstrated in the various sample images provided. It appears that the computed Dice score of 0.84 is underestimating the true segmentation performance, due to inaccuracies in the compared ground-truth segmentations. However, it is clear that the automatically generated ground-truth segmentations were of sufficient quality for training of the U-Net.

The developed methods for obtaining the PT, MT, and TL/L Cobb angles from the vertebral segmentations functioned well. As described previously, the fitted endplates proved to be accurate and reliable. Furthermore, the algorithm for selection of vertebrae in the calculation of the Cobb angles performed as desired. The proposed fully automatic Cobb angles yielded a MAD of 5.08° , when compared to angles generated from the manually annotated endplates. This aligns closely with, while at the higher end of, the expected interobserver Cobb angle variation, identified in the literature review as a MAD of $3\text{-}5^\circ$. The Lenke curve type classification resulting from fully automatic and ground-truth Cobb angles resulted in 59.38% agreement and a kappa statistic of 0.48. This is significantly lower than the mean interobserver agreement in Lenke curve type classification found in the literature review ($\kappa = 0.74$). Clinicians' feedback indicated that this may be explained by the presence of non-AIS images in the dataset used in this study. Despite this, the subsequent probability analysis achieved a reasonably high agreement, with an ICC of 0.75. It was concluded that the automatic and ground-truth systems generally disagree on relatively uncertain curves. This uncertainty also, in some part, explains the lower interobserver agreement in manual classification for curve types 3, 4, and 6 identified in the literature review ($\kappa = 0.68, \kappa = 0.38$, and $\kappa = 0.41$, respectively).

The developed Lenke classification probability analysis constitutes a clear and useful indicator of the uncertainty in classification. In some cases, this analysis yielded a probability (or confidence level) of over 90% for a given curve type, while other cases, with Cobb angles close to the relevant structural thresholds, indicated substantial probability of multiple curve types. In addition, by taking interobserver variation into account, the most probable underlying Lenke curve type can be uncovered; there were some cases, with high uncertainty, whereby this most probable curve type differed from the measured curve type. An example of this phenomenon, and a discussion of its clinical implications, will follow in Section 5.3.

5.2. LIMITATIONS

The primary shortcoming of this study stems from the lack of available datasets.

- In the absence of appropriate data, clinical validation of the proposed fully automatic method for scoliosis assessment could not be performed. The vertebral segmentation and endplate fitting methods appear highly accurate, however, the resulting Cobb angles and Lenke curve types must be evaluated against the average of multiple clinicians before being introduced as a clinical aid.
- Without bending and sagittal view x-ray datasets, this portion of scoliosis assessment could not be included in the development of the proposed tool. Additional examination of these x-rays is required in Lenke classification in order to interpret the 3D nature of the spinal deformity. Thus, the Lenke analysis presented in this study is incomplete.
- The x-rays used in this study were limited to a fixed field of view that did not include the skull, pelvis, or limbs, as some x-rays would. The proposed system could potentially perform poorly when encountering images with a wider scope, unseen in training. This shortcoming could be overcome using transfer learning on a broader dataset, or by developing an automatic tool to extract an appropriate region of interest.
- Lacking clear indication of which x-rays contained patients with AIS, the evaluation of the developed Lenke classification methods is limited.

Conversely, the identified limitations in ground-truth vertebral segmentation datasets motivated the novel approach developed for their automatic generation. This method could equivalently be applied to a dataset of bending or sagittal view x-rays, with corner landmarks annotated, such as that employed in development of the reviewed MVE-Net. Additionally, in the absence of a large dataset of solely AIS x-rays, the system produced in this study has achieved robust performance, capable of application with x-rays of any scoliosis variant.

5.3. CLINICAL IMPACT

The clinical implications of this study will be discussed, firstly, with reference to the developed automatic scoliosis assessment tool, and subsequently, the proposed Lenke classification probability analysis.

5.3.1. FULLY AUTOMATIC SCOLIOSIS ASSESSMENT TOOL

Although the developed system requires further validation in advance of application in a clinical setting, influencing surgical decisions, the methods are aided by their interpretable design. The predicted Cobb angle endplates can be overlaid onto the x-ray under consideration, allowing for detection of errors by the clinician. In this way, clinical interpretability can offer continued validation of the automatic system. Additionally, in case of erroneous outputs, clinicians could be provided with the option of manually overriding parts of the automatic process, such as the selection of endplates used in the Cobb angle calculation.

In its current form, the proposed system provides a useful research tool, offering rapid identification of the Cobb angles and Lenke curve type present in a given x-ray. Following further validation, however, this system could be integrated into current clinical practice, to improve workflows. Clinicians' feedback indicated that the proposed automatic method would be beneficial if deployed in the referral process between peripheral sites and a central orthopaedic surgical hospital. This would allow for automatic triage, determining the urgency of care required, using the Cobb angles and curve type identified in each x-ray. A system such as this, facilitating prioritisation, would alleviate issues surrounding the progression of a curve between booking and presentation in the operating room, by expediting this process.

5.3.2. LENKE CLASSIFICATION PROBABILITY ANALYSIS

5.3.2.1. AN AID TO MANUAL ASSESSMENT

The proposed Lenke classification probability analysis presents a useful aid to current clinical practice. This analysis can be easily adjusted to reflect interobserver error in manual assessment, by modelling the Cobb angle measurement variation with a standard deviation obtained from a review of the literature. In addition, this analysis can be extended to include the probability distributions of bending and sagittal view Cobb angle criteria. Following these minor adjustments, the proposed Lenke probability analysis would provide an accurate, case-by-case, confidence level associated with the Lenke curve type and sagittal thoracic modifier, given a clinically measured set of Cobb angles. The aim of this system would be to highlight, to surgeons, x-rays with uncertain Lenke classification, so that this can be considered in surgical decisions.

5.3.2.2. DISPARITY BETWEEN MEASURED AND MOST PROBABLE CURVE TYPES

In cases where the measured Cobb angles were close to the structural thresholds, the probability analysis occasionally indicated that the most probable underlying curve type was not the same as the measured curve type. Assuming a standard deviation of 3° interobserver variation and limiting analysis to standing AP x-ray criteria as before, the following analysis was conducted to demonstrate this effect. If the Cobb angles for a given patient are measured as PT = 24°, MT = 35°, TL/L = 36°; the probability of each Lenke curve type is calculated, as follows:

$$\text{Type 1} = 0.0001, \text{Type 2} = 0.0000, \text{Type 3} = 0.2815, \text{Type 4} = 0.3693, \text{Type 5} = 0.0003, \text{Type 6} = 0.3487$$

Although the measured angles indicate a curve type 6, when interobserver measurement variation is considered, it is more likely that the patient is truly a type 4. This case may appear more alarming if the equivalent inverse is considered. If a patients' true Cobb angles are PT = 24°, MT = 35°, TL/L = 36°, their true curve type is 6 but they are more likely to be clinically misclassified as a type 4.

This phenomenon is best illustrated with a probability tree diagram:

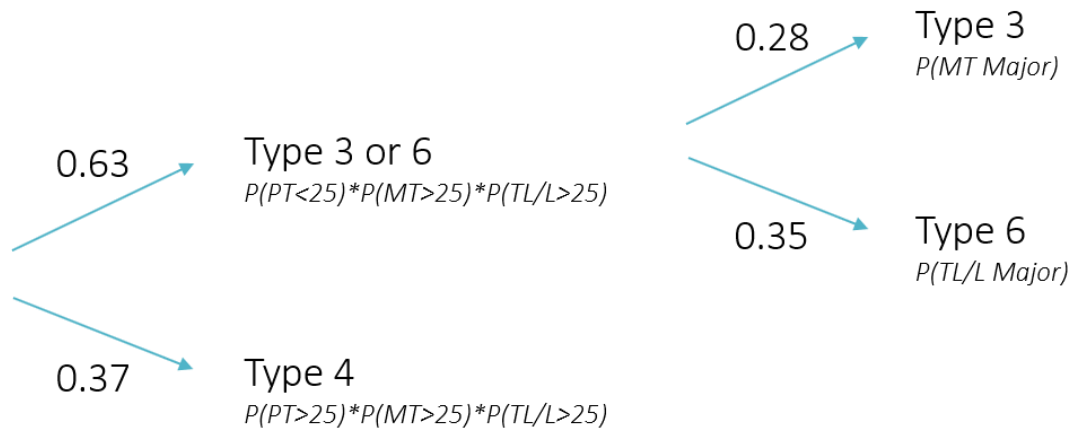


Figure 28. Probability tree explanation of disparity between measured and most probable Lenke curve types. This patient's curve is a type 4 if all Cobb angles are greater than 25°. However, for the curve to be a type 6, the PT angle must be less than 25°, the MT and TL/L angles greater than 25° and the TL/L angle must be major.

Although unusual, this effect is not a limitation of the Lenke curve type classification system. It could be eliminated by subdividing curve types so that they are more readily comparable; curve type 4, in this case, would be divided into two subtypes accounting for each variation of major curve. If this were the case, patients presenting with the true angles given above will be more likely to be clinically classified as their correct subtype than any other *individual* subtype. However, nothing will have been gained, as the patient will remain just 35% likely to be classified correctly, i.e. they will still be more likely to be misclassified.

Any case that presents a disparity between measured and most probable curve type is likely to be very uncertain, and neither curve type should be accepted without in-depth consideration of alternatives. Nonetheless, it is useful that the proposed probability analysis is capable of discerning the most probable underlying curve type from the measured curve type. This analysis may also shed some light on the reduced interobserver agreement for curve types 3, 4, and 6 identified in the literature review.

5.4. FUTURE WORK

The developed protocol for automatic scoliosis assessment is readily expandable to a more comprehensive examination. The following steps should be taken in order to achieve this:

- Transfer learning of the U-Net should be applied on a broader set of AP x-rays, to ensure robustness with x-rays containing a different field of view and bending x-rays.
- The automatic endplate fitting and vertebra selection techniques should be tested with bending x-ray data.
- Additional segmentation of the sacrum (to identify the CSVL), vertebral pedicles, and trunk would allow for automatic assignment of the lumbar modifier and calculation of trunk shift.
- A similar fully automatic process should be tailored to sagittal view x-rays, for a more complete curve type classification and further sagittal thoracic modifier assignment. The process for automatic generation of ground-truth vertebral segmentations could be applied in this case.
- Novel visual aids should be developed for clinical investigation of curve progression. Rather than comparing Cobb angles between x-rays taken over time, the high-resolution vertebral segmentations can be availed of for a more detailed visual analysis of changes in the scoliotic spine.

In addition to this expansion of the developed fully automatic assessment tool, the proposed Lenke classification probability analysis should be extended to include the sagittal view x-ray criteria. As described previously, this would allow for objective and accurate determination of the uncertainty in Lenke classification, to aid in current manual assessment by clinicians.

6. CONCLUSION

This final chapter summarises the outcomes of this study to date and presents a roadmap for future work that is underway.

6.1. SUMMARY OF WORK COMPLETED

In this study, difficulties with current clinical practice and limitations in state-of-the-art research in scoliosis assessment have been alleviated, using novel approaches.

- A large dataset of ground-truth vertebral segmentations was generated from publicly available resources, overcoming the prior insufficiency of data to train a robust deep learning tool for vertebral segmentation.
- A system for automatic scoliosis assessment was developed, with the aim of providing a comprehensive Lenke classification tool, in response to guidance from clinicians.
- A novel analysis of Cobb angle interobserver variability was established, quantifying the confidence level for a given Lenke classification, by computing the probability of each category.

The systems resulting from this study can advance current clinical practice and research in the field of scoliosis. The data and code generated in this project will be made open-source, and the areas identified for future work will be pursued, in consultation with leading clinicians in this field, as described in the following section.

6.2. ROADMAP FOR FUTURE WORK

In collaboration with clinicians, the development of further datasets is underway, in order to complete the discussed future work. Furthermore, the systems designed to date, and future methods, will be made open-source, and publications will be sought, in order to achieve the maximum impact in advancing the assessment and treatment of patients with scoliosis. The following journal articles are in preparation:

- Automatic Vertebral Segmentation and Cobb Angle Measurement using Novel Ground-truth Database
- Lenke Classification Uncertainty Analysis: Knowing Your Limits
- Automatic Segmentation of the Vertebrae, Pedicles, Sacrum, and Trunk in Spinal X-rays for Comprehensive Assessment of Scoliosis
- Intra- and Interobserver Study of Lenke Classification and Comparison with Fully Automatic Method

The articles will be submitted to various journals, such as ‘Radiology: Artificial Intelligence’, ‘Nature Methods’, and ‘The Spine Journal’, with the aim of reaching radiologists, orthopaedic surgeons, and researchers of machine learning in healthcare alike.

7. REFERENCES

- [1] M. R. Konieczny, H. Senyurt, and R. Krauspe, ‘Epidemiology of adolescent idiopathic scoliosis’, *J. Child. Orthop.*, vol. 7, no. 1, pp. 3–9, Feb. 2013, doi: 10.1007/s11832-012-0457-4.
- [2] ‘The Pathogenesis of Adolescent Idiopathic Scoliosis: Review of the Literature | Ovid’. <https://oce-ovid-com.ucd.idm.oclc.org/article/00007632-200812150-00011/HTML> (accessed Apr. 30, 2020).
- [3] ‘Definitions & Causes | Scoliosis Research Society’. <https://www.srs.org/patients-and-families/common-questions-and-glossary/frequently-asked-questions/general-spinal-deformity-faqs> (accessed Apr. 30, 2020).
- [4] A. D. Tambe, S. J. Panikkar, P. A. Millner, and A. I. Tsirikos, ‘Current concepts in the surgical management of adolescent idiopathic scoliosis’, *Bone Jt. J.*, vol. 100-B, no. 4, pp. 415–424, Apr. 2018, doi: 10.1302/0301-620X.100B4.BJJ-2017-0846.R2.
- [5] M. Płaszewski and J. Bettany-Saltikov, ‘Non-Surgical Interventions for Adolescents with Idiopathic Scoliosis: An Overview of Systematic Reviews’, *PLoS One San Franc.*, vol. 9, no. 10, p. e110254, Oct. 2014, doi: <http://dx.doi.org.ucd.idm.oclc.org/10.1371/journal.pone.0110254>.
- [6] J. Y. Thompson, E. M. Williamson, M. A. Williams, P. J. Heine, and S. E. Lamb, ‘Effectiveness of scoliosis-specific exercises for adolescent idiopathic scoliosis compared with other non-surgical interventions: a systematic review and meta-analysis’, *Physiotherapy*, vol. 105, no. 2, pp. 214–234, Jun. 2019, doi: 10.1016/j.physio.2018.10.004.
- [7] J. COBB, ‘Outline for the study of scoliosis’, *Instr Course Lect AAOS*, vol. 5, pp. 261–275, 1948.
- [8] ‘Comparison of Cobb Angle Measurement of Scoliosis Radiographs With Preselected End Vertebrae: Traditional Versus Digital Acquisition | Ovid’. <https://oce-ovid-com.ucd.idm.oclc.org/article/00007632-200701010-00016/HTML> (accessed Apr. 30, 2020).
- [9] ‘Lenke Calculator’, *Harms Study Group*. <http://hsg.settingscoliosisstraight.org/lenke-calculator/> (accessed Apr. 30, 2020).
- [10] M. Gstoettner, K. Sekyra, N. Walochnik, P. Winter, R. Wachter, and C. M. Bach, ‘Inter- and intraobserver reliability assessment of the Cobb angle: manual versus digital measurement tools’, *Eur. Spine J.*, vol. 16, no. 10, pp. 1587–1592, Oct. 2007, doi: 10.1007/s00586-007-0401-3.
- [11] S. Langensiepen *et al.*, ‘Measuring procedures to determine the Cobb angle in idiopathic scoliosis: a systematic review’, *Eur. Spine J.*, vol. 22, no. 11, pp. 2360–2371, Nov. 2013, doi: 10.1007/s00586-013-2693-9.
- [12] M. C. Tanure, A. P. Pinheiro, and A. S. Oliveira, ‘Reliability assessment of Cobb angle measurements using manual and digital methods’, *Spine J.*, vol. 10, no. 9, pp. 769–774, Sep. 2010, doi: 10.1016/j.spinee.2010.02.020.
- [13] R. Lechner, D. Putzer, D. Dammerer, M. Liebensteiner, C. Bach, and M. Thaler, ‘Comparison of two- and three-dimensional measurement of the Cobb angle in scoliosis’, *Int. Orthop.*, vol. 41, no. 5, pp. 957–962, May 2017, doi: 10.1007/s00264-016-3359-0.
- [14] L. G. Lenke, R. R. Betz, J. Harms, K. H. Bridwell, and *et al.*, ‘Adolescent idiopathic scoliosis: A new classification to determine extent of spinal arthrodesis’, *J. Bone Jt. Surg. Am. Vol. Needham*, vol. 83, no. 8, pp. 1169–81, Aug. 2001.

- [15] H. King, J. Moe, D. Bradford, and R. Winter, 'The Selection of Fusion Levels in Thoracic Idiopathic Scoliosis'. The Journal of Bone and Joint Surgery, 1983.
- [16] L. G. Lenke, R. R. Betz, K. H. Bridwell, D. H. Clements, and et al, 'Intraobserver and interobserver reliability of the classification of thoracic adolescent idiopathic scoliosis', *J. Bone Jt. Surg. Am. Vol. Needham*, vol. 80, no. 8, pp. 1097–106, Aug. 1998.
- [17] J. Zhang, H. Li, L. Lv, and Y. Zhang, 'Computer-Aided Cobb Measurement Based on Automatic Detection of Vertebral Slopes Using Deep Neural Network', *Int. J. Biomed. Imaging*, vol. 2017, 2017, doi: 10.1155/2017/9083916.
- [18] Y. Pan *et al.*, 'Evaluation of a computer-aided method for measuring the Cobb angle on chest X-rays', *Eur. Spine J.*, Aug. 2019, doi: 10.1007/s00586-019-06115-w.
- [19] A. Safari, H. Parsaei, A. Zamani, and B. Pourabbas, 'A Semi-Automatic Algorithm for Estimating Cobb Angle', *J. Biomed. Phys. Eng.*, vol. 9, no. 3Jun, Jun. 2019, doi: 10.31661/jbpe.v9i3Jun.730.
- [20] O. Ronneberger, P. Fischer, and T. Brox, 'U-Net: Convolutional Networks for Biomedical Image Segmentation', May 2015, Accessed: Nov. 30, 2018. [Online]. Available: <https://arxiv.org/abs/1505.04597>.
- [21] H. Wu, C. Bailey, P. Rasoulinejad, and S. Li, 'Automatic Landmark Estimation for Adolescent Idiopathic Scoliosis Assessment Using BoostNet', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, 2017, pp. 127–135.
- [22] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu, 'Regression Forests for Efficient Anatomy Detection and Localization in CT Studies', in *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, vol. 6533, B. Menze, G. Langs, Z. Tu, and A. Criminisi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 106–117.
- [23] H. Sun, X. Zhen, C. Bailey, P. Rasoulinejad, Y. Yin, and S. Li, 'Direct Estimation of Spinal Cobb Angles by Structured Multi-output Regression', in *Information Processing in Medical Imaging*, 2017, pp. 529–540.
- [24] H. Wu, C. Bailey, P. Rasoulinejad, and S. Li, 'Automated comprehensive Adolescent Idiopathic Scoliosis assessment using MVC-Net', *Med. Image Anal.*, vol. 48, pp. 1–11, Aug. 2018, doi: 10.1016/j.media.2018.05.005.
- [25] L. Wang, Q. Xu, S. Leung, J. Chung, B. Chen, and S. Li, 'Accurate automated Cobb angles estimation using multi-view extrapolation net', *Med. Image Anal.*, vol. 58, p. 101542, Dec. 2019, doi: 10.1016/j.media.2019.101542.
- [26] F. Galbusera *et al.*, 'Fully automated radiological analysis of spinal disorders and deformities: a deep learning approach', *Eur. Spine J.*, vol. 28, no. 5, pp. 951–960, May 2019, doi: 10.1007/s00586-019-05944-z.
- [27] Z. Tan *et al.*, 'An Automatic Scoliosis Diagnosis and Measurement System Based on Deep Learning', in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Kuala Lumpur, Malaysia, Dec. 2018, pp. 439–443, doi: 10.1109/ROBIO.2018.8665296.
- [28] M.-H. Horng, C.-P. Kuok, M.-J. Fu, C.-J. Lin, and Y.-N. Sun, 'Cobb Angle Measurement of Spine from X-Ray Images Using Convolutional Neural Network', *Comput. Math. Methods Med.*, vol. 2019, pp. 1–18, Feb. 2019, doi: 10.1155/2019/6357171.
- [29] Y. Tu, N. Wang, F. Tong, and H. Chen, 'Automatic measurement algorithm of scoliosis Cobb angle based on deep learning', *J. Phys. Conf. Ser.*, vol. 1187, no. 4, p. 042100, Apr. 2019, doi: 10.1088/1742-6596/1187/4/042100.
- [30] 'Boundary of a set of points in 2-D or 3-D - MATLAB boundary - MathWorks United Kingdom'. <https://uk.mathworks.com/help/matlab/ref/boundary.html> (accessed Apr. 27, 2020).

- [31] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, 'Tversky loss function for image segmentation using 3D fully convolutional deep networks', *ArXiv170605721 Cs*, Jun. 2017, Accessed: Mar. 23, 2020. [Online]. Available: <http://arxiv.org/abs/1706.05721>.
- [32] '[1412.6980] Adam: A Method for Stochastic Optimization'. <https://arxiv.org/abs/1412.6980> (accessed Apr. 01, 2020).
- [33] 'GitHub - keras-team/keras: Deep Learning for humans'. <https://github.com/keras-team/keras> (accessed Apr. 01, 2020).
- [34] 'TensorFlow'. <https://www.tensorflow.org/> (accessed Apr. 01, 2020).
- [35] 'Colaboratory – Google'. <https://research.google.com/colaboratory/faq.html> (accessed Apr. 01, 2020).
- [36] 'A suite of minimal bounding objects - File Exchange - MATLAB Central'. <https://uk.mathworks.com/matlabcentral/fileexchange/34767> (accessed Apr. 02, 2020).
- [37] L. G. Lenke, 'Lenke classification system of adolescent idiopathic scoliosis: treatment recommendations', *Instr. Course Lect.*, vol. 54, pp. 537–542, 2005.
- [38] K. Yokoyama *et al.*, 'Spinopelvic alignment and sagittal balance of asymptomatic adults with 6 lumbar vertebrae', *Eur. Spine J.*, vol. 25, no. 11, pp. 3583–3588, Nov. 2016, doi: 10.1007/s00586-015-4284-4.