An artificial intelligence powered platform for auto-analyses of spine alignment irrespective of image quality with prospective validation

Nan Meng,[a] Jason P.Y. Cheung,[a]* Kwan-Yee K. Wong,[b] Socrates Dokos,[c] Sofia Li,[a] Richard W. Choy,[a] Samuel To,[a] Ricardo J. Li,[a] and Teng Zhang[a]*

[a]Digital Health Laboratory, Queen Mary Hospital, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 5/F, Professorial Block, Pokfulam, Hong Kong, China
[b]Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong, China
[c]Graduate School of Biomedical Engineering, University of New South Wales, Sydney, Australia

EClinicalMedicine 2022;43: 101252
Published online xxx
https://doi.org/10.1016/j.eclinm.2021.101252

Summary

Background Assessment of spine alignment is crucial in the management of scoliosis, but current auto-analysis of spine alignment suffers from low accuracy. We aim to develop and validate a hybrid model named SpineHRNet+, which integrates artificial intelligence (AI) and rule-based methods to improve auto-alignment reliability and interpretability.

Methods From December 2019 to November 2020, 1,542 consecutive patients with scoliosis attending two local scoliosis clinics (The Duchess of Kent Children's Hospital at Sandy Bay in Hong Kong; Queen Mary Hospital in Pok Fu Lam on Hong Kong Island) were recruited. The biplanar radiographs of each patient were collected with our medical machine EOS™. The collected radiographs were recaptured using smartphones or screenshots, with deidentified images securely stored. Manually labelled landmarks and alignment parameters by a spine surgeon were considered as ground truth (GT). The data were split 8:2 to train and internally test SpineHRNet+, respectively. This was followed by a prospective validation on another 337 patients. Quantitative analyses of landmark predictions were conducted, and reliabilities of auto-alignment were assessed using linear regression and Bland-Altman plots. Deformity severity and sagittal abnormality classifications were evaluated by confusion matrices.

Findings SpineHRNet+ achieved accurate landmark detection with mean Euclidean distance errors of 2·78 and 5·52 pixels on posteroanterior and lateral radiographs, respectively. The mean angle errors between predictions and GT were 3·18° and 6·32° coronally and sagittally. All predicted alignments were strongly correlated with GT ($p < 0{\cdot}001$, $R^2 > 0{\cdot}97$), with minimal overall difference visualised via Bland-Altman plots. For curve detections, 95·7% sensitivity and 88·1% specificity was achieved, and for severity classification, 88·6–90·8% sensitivity was obtained. For sagittal abnormalities, greater than 85·2–88·9% specificity and sensitivity were achieved.

Interpretation The auto-analysis provided by SpineHRNet+ was reliable and continuous and it might offer the potential to assist clinical work and facilitate large-scale clinical studies.

Funding RGC Research Impact Fund (R5017-18F), Innovation and Technology Fund (ITS/404/18), and the AOSpine East Asia Fund (AOSEA(R)2019-06).

Keywords: Adolescent idiopathic scoliosis; Spine malalignment; Artificial intelligence; Deep learning; Open platform

Introduction

Spine malalignment is prevalent and in the paediatric population, adolescent idiopathic scoliosis (AIS) is most common.[1] AIS can affect up to 2·2% of boys and 4·8% of girls.[2] Without prompt intervention, the deformity may deteriorate and significantly reduce the quality of

*Corresponding authors.
E-mail addresses: cheungjp@hku.hk (J.P.Y. Cheung), tgzhang@hku.hk (T. Zhang).

Research in context

Evidence before this study

PubMed was searched on July 20, 2021, for articles published in all languages describing the application of deep learning techniques to scoliosis in images using the search terms “artificial intelligence” OR “deep learning” OR “convolutional neural network” AND “scoliosis” OR “spine malalignment” AND “images” without any date restrictions. Cited references in the retrieved articles were further searched. Previous AI studies on scoliosis have mainly focused on a specific task, such as scoliosis severity classification, Cobb Angle (CA) prediction, or vertebrae detection on either posteroanterior or lateral radiographs. Most of these studies suffered from data scarcity. The literature search revealed that no study evaluated the effectiveness of Artificial Intelligence (AI) for scoliosis or spine malalignment in any prospective trial.

Added value of this study

To the best of our knowledge, this study is the first to establish a platform for automatic spine alignment analysis with prospective validation. The performance of the hybrid system was validated via in-house and prospective clinical data, using cases of varying severity, curve type, and radiograph quality. The primary significance of this work compared with previous studies is the prospective data validation and integration into an open platform for clinicians and researchers to obtain fast alignment analysis. The prospective results suggested that our model achieved satisfactory clinical performance.

Implications of all the available evidence

Our study suggests that medical AI has the capacity to assist doctors and clinical researchers via fast and consistent analytical results. However, we should note that the prospective validation of this study was performed at two scoliosis clinics and not at a multi-centre international site. Therefore, caution is required when using SpineHRNet+ directly in other clinical settings. Fine-tuning of the models may be required to adapt to other clinical settings. Further studies are necessary to corroborate our results.

life and mobility.[3] Accurate assessment of spine alignment is crucial for proper treatment. It is based on radiographs with multiple alignment parameters, including but not limited to the coronal Cobb angle (CA) magnitude, curve type (thoracic curve, thoracolumbar and lumbar) as well as the sagittal alignment, i.e., thoracic kyphosis (TK), lumbar lordosis (LL), and sacral slope (SS), which are all measured from specific keypoints (end points of the endplates) of specific vertebrae. Thus, current diagnoses and follow-up assessments require extensive clinical experience and expertise to interpret the alignment parameters manually and assess the patient physical appearance, making fast and accurate alignment analyses challenging.[4]

AI has shown great promise in managing spine disease, including disease detection [5], classification,[6,7] segmentation,[8] and progression prediction,[9] mainly based on medical images. Previous studies on automatic spine alignment analysis[10–12] could directly or indirectly regress CAs from radiographs of the major curve but could not compute heterogeneous curve patterns[13] or investigate the curve types. It is challenging to interpret what the AI algorithm has learned, thus reducing the clinical application. Other AI approaches adopted the clinical gold standard strategy of first locating the vertebral endplates followed by CA calculation,[14],[15] but these focused on screened radiographs from retrospective datasets and lacked prospective validations.[16]

Given that medical AI has significant advantages in speed and consistency, an easily accessible platform using validated AI models can significantly assist clinicians. Particularly when face-to-face contact is restricted, such as under the COVID-19 pandemic, an auto-analysis platform able to tolerate large variations in image quality will greatly assist spine surgeons in making clinical decisions. Thus, we developed an open platform termed AlignProCARE, integrating user applications, a data centre, and a backend AI server powered by deep learning models. Our platform aimed to provide clinicians with faster and reliable spine measures to assist their clinical practice and facilitate communication with patients. The significance of this work is multifaceted. First, the deep learning models were designed and trained using radiographs from a range of sources with varying image quality, including smartphone recaptured radiographs as well as original high-resolution radiographs, to increase the generalizability of our models. Second, AlignProCARE is an open platform that can be freely used upon registration using research institute emails to increase the accessibility of auto-analysis. Third, SpineHRNet+, with the clinical keypoints displayed, can provide interpretable spine alignment analysis for spine surgeons, yielding evidence for how results were generated.

Methods

Participants and datasets

All deformity patients attending two scoliosis clinics (The Duchess of Kent Children’s Hospital at Sandy Bay in Hong Kong; Queen Mary Hospital in Pok Fu Lam on Hong Kong Island) from December 2019 to November 2020 were recruited. The ethics of this study was approved by the local institutional authority review board (UW15-596), and all participants were required to provide written informed consent before they joined the study. Patients with psychological, systematic neural disorders, congenital deformities, previous spinal operations, any trauma that could impair posture and mobility, and any oncological diseases were excluded. The involved technicians were instructed to recapture the radiographs using a smartphone or take screenshots of the displayed image, with the imaging plane parallel to the screen. No patient identification information was captured. All collected images were anonymised and uploaded to our secure internal server via AlignProCARE. All patient alignment landmarks were manually annotated by spine surgeons using open-source software (ImageJ version 1.52r). The spine alignments recorded as a routine clinical practice were considered as the ground truth (GT). Senior surgeons with more than 20 years' clinical experience manually annotated the end points of all endplates from the 1st thoracic vertebra to the 1st sacrum as key point landmarks. Those annotated landmarks were used as GT landmarks to train our AI model for landmark detection. The surgeons measured alignment parameters were considered as GT angles to evaluate the alignments prediction performance of our model. We tested the inter-rater variability of alignments measurements between two surgeons and discovered a small absolute angle difference of 4°–6° (mean = 4·5° ± SD 0·6). The agreements were satisfactory for clinical practice and thus no third assessor was involved.

Of the 1542 consecutive patients attending the scoliosis clinic during the study period for SpineHRNet+ development, 185 were excluded because of degenerative deformities but not AIS, and another 8 were excluded due to congenital deformities. A total of 1349 cases with biplanar radiographs (both coronal and sagittal) were included, with 1079 cases (74% female; age range 10–18) used to train SpineHRNet+ and 270 used for the in-house validation dataset to evaluate the performance. After the model was developed, it was tested on 337 prospective cases, which were not involved in training and optimising SpineHRNet+. Most of the patients had two radiographs (i.e., posteroanterior and lateral radiographs), while a small proportion of the patients have four biplanar radiographs since they were scanned twice before and after wearing braces. Therefore, the number range of radiographs per patient is 2–4. The average and standard deviation of radiographs per patient for training the model were 2·28 and 0·69, respectively, while for prospective testing were 2·30 and 0·71, respectively. Radiographs of the patients with the brace were eliminated during the data cleaning.

Definition of alignment parameters, severity, and types

For measuring the coronal alignment [5], the end vertebrae must be identified first, which are the most tilted vertebrae from the curve apex. The CA is formed by the angle between the upper endplate of the most cranial vertebra and the lower endplate of the most caudal vertebra. A threshold of 10° was used according to the clinical gold standard to differentiate the presence of a curve at a given location of the spine in the coronal plane.[17]

The severity of the deformity is classified as shown in Table 1. A CA smaller than 20° is considered as normal-mild, 20–40° is considered as moderate and greater than 40° is considered severe. The type of curve is considered as thoracic (T) if the apex of the curve was between the 1st to the 11th thoracic vertebrae; as thoracolumbar (TL) if the apex was located between the 12th thoracic vertebra and the 1st lumbar vertebra; and as lumbar (L) if the apex of the curve was between the 2nd and the 5th lumbar vertebrae. Different clinical interventions and deformity features are associated with various severities and curve types (Table 1).

For the sagittal alignment, the 5th thoracic vertebra, the 12th thoracic vertebra (T5 and T12), the 1st lumbar vertebra (L1) and the 1st sacrum (S1) are landmarks for calculating the TK, LL, and SS. We adopted the previously reported normal ranges for TK (20–40°) [18], LL (20–45°) [19], and SS (32–49°).[20]

Image pre-processing

To minimise input image variance, image sizes were automatically unified by cropping each image to an 896 × 448 pixels patch, containing the entire spine. Subsequently, we adopted data-augmentation to enhance model robustness, including random flipping (probability = 0·5), scaling [0·8, 1·2], rotation [−5°, 5°], horizontal/vertical translation [−10 pixels, 10 pixels], and contrast augmentation [0·8, 1·2]. The generated heatmaps and GT landmarks were scaled accordingly.

Procedures

The AlignProCARE open platform has been deployed to a website (https://aimed.hku.hk/alignprocare). It consists mainly of three parts: the user end, a data centre, and an AI server (Figure 1). Users can utilise the automated alignment analysis function after registration using research institutional emails. They can access the platform and upload new images through both the PC user interface and smartphone application (App Store and Google Play). Analytic results from SpineHRNet+ can be received and parsed directly on the user end with key landmarks visualised on the user interface or smartphone. Alignment degrees and deformity severity can be automatically computed based on the landmarks. Users can modify the landmark positions, and the alignment results are simultaneously re-calculated.

SpineHRNet+

The backend hybrid model is an improved version of our previous framework SpineHRNet [5], thus known as

	Model development		Prospective test		Clinical implications
Age (10–18)	**No. of patients**	**No. of X-rays**	**No. of patients**	**No. of X-rays**	
Curve magnitudes					
Normal-mild (CA≤20°)	502	1138	118	264	No intervention required. For the skeletally immature, regular follow-up is required every 4–6 months to identify curve progression early in which bracing may be recommended.
Moderate (20°<CA≤40°)	670	1532	184	426	These patients may require bracing to prevent curve progression. No intervention may be required at the end of growth. Scoliosis-specific exercises may also be prescribed.
Severe (CA>40°)	177	402	35	86	These severe curves have risk of adulthood progression. Surgical intervention may be required in the form of vertebral body tethering (skeletally immature only) or curve correction and spinal fusion.
Curve types					
Thoracic curve	1067	2422	256	589	Curves that develop rib humps and are more likely to develop chest wall deformities and unlevelled shoulders.
Thoracolumbar/Lumbar curve	969	2198	236	548	Curves more likely to develop pelvic obliquity and waistline deformities.

***Table 1*: Standards of severity level and corresponding clinical interventions for different curve types.**

SpineHRNet+. SpineHRNet+ "mimics" diagnostic procedures of clinicians for AIS via incorporation of both deep learning networks and a rule-based algorithm (Figure 2). It can process and analyse biplanar radiographs of the spine with deformities. The processing pipeline essentially consists of two stages. In stage 1, the algorithm utilises two deep learning models as backbones. HRNet [21] is adopted to detect endplate landmarks of each vertebra and identify the location of the end vertebrae. The detailed architecture of the adopted HRNet is presented in Figure 1.1 in the supplementary materials. The landmarks are visualised to increase result interpretability for users. In parallel, UNet [22] is adopted to segment the spine region, followed by a rule-based algorithm to obtain a spine segmentation map. The detailed architecture of the adopted UNet is presented in Figure 1.2 in the supplementary materials. Details of all the networks are provided in the supplementary materials. In stage 2, SpineHRNet+ integrates the outputs of the previous two stages to perform bad point detection and correction (Figure 2). The corrected landmarks are further used for CA prediction.

For model development, we used a computer with an Intel(R) Xeon(R) Silver 4114 CPU 2.20 GHz central processing unit, 128GB of RAM and an NVIDIA GeForce RTX 2080Ti core. The network was implemented using PyTorch (version 1.6.1) [23] accelerated by Cuda 10.2. The detailed structure of deep learning networks adopted in SpineHRNet+ is provided in the supplementary materials (Section 1.1). Landmark detection and spine segmentation were trained using the training data. HRNet was optimised by minimising a pixel-wise binary cross-entropy loss between the model outputs and GT labels, while UNet was optimised by minimising the cross-entropy loss. During model training, the Adam optimiser was adopted, and a decay factor d was used to control the learning rate at each epoch. The decayed learning rate update follows the equation

$$lr_i = lr \cdot \frac{1}{1 + d \cdot i}, \qquad (1)$$

where lr was the initial learning rate and lr_i was the learning rate at the i^{th} epoch. In our experiments, the Adam optimiser with an initial learning rate of 0·0001, a decay of 0·05, a batch size of 4, and an epoch of 200 was used for landmark detection training. For segmentation training, the Adam optimiser with an initial learning rate of 0·00,005, a decay of 0·03, a batch size of 4, and an epoch of 100 was used, with hyper-parameters decided empirically.

Performance evaluation

To comprehensively assess our method, we tested the performance of the model on multiple tasks, including

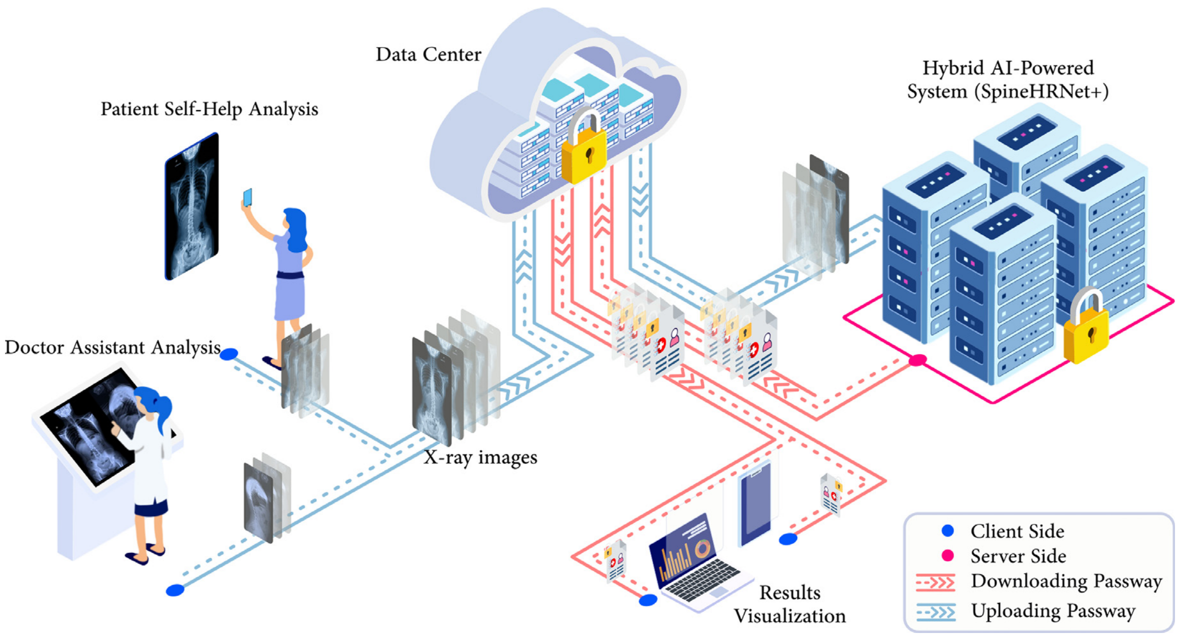


Figure 1. Workflow diagram of our medical AI platform — AlignProCARE. The platform consists of mainly three parts, i.e., the client end, data centre, and server. The workflow is: (1) On the client end, users upload the spine radiographs via PC or smartphone applications. (2) The data centre receives the captures from different users and assigns each one with a specific identity number, and then sends ammonized images to the server with end-to-end encryption. (3) On the server side, a hybrid AI-powered model continuously receives the images, analyses the images using SpineHRNet+, and sends back the results to the data centre. (4) The data centre forwards the results to different devices according to the identity number. (5) The PC or smartphone applications parse the results and visualise them on client devices.

landmark detection, CA detection, and severity prediction (only for coronal images).

For landmark detection, we used two quantitative measurements to examine the difference between predicted results and landmark labels (ground truth). The two measurements are mean Euclidean distance (MED) and mean Manhattan distance (MMD), which are defined as:

$$\mathrm{MED} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{i=1}^{M}\sqrt{\left(x_i - \hat{x}_i\right)^2 + \left(y_i - \hat{y}_i\right)^2}\right), \quad (2)$$

$$\mathrm{MMD} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{i=1}^{M}\left(\left|x_i - \hat{x}_i\right| + \left|y_i - \hat{y}_i\right|\right)\right), \quad (3)$$

where $(\mathrm{x_i}, \mathrm{y_i})$ and $(\hat{x}_i, \hat{y}_i)$ denote the i^{th} predicted landmark coordinates and corresponding label coordinates, respectively. M is the number of landmark coordinates in a single radiograph capture, and N is the number of image samples in the dataset. These two measurements can provide a comprehensive evaluation of the performance of SpineHRNet+ on a landmark detection task. MED measures the 2D Euclidean distance between the predicted landmark coordinates and their GT counterparts, which is a straightforward criterion that can assess the predicted results of each landmark. MMD can measure the distance as well. However, it emphasises the distance on each coordinate dimension. Therefore, by combining these two measurements, the performance of the model can be more accurately evaluated.

Statistical analysis

To demonstrate if the performance of SpineHRNet+ has improved significantly, we conducted a hypothesis test on the two measurements MED and MMD. First, we used the Shapiro-Wilk test, a commonly used normality test, to evaluate the normality of the quantitative error of SpineHRNet and SpineHRNet+ in terms of MED and MMD. Such an evaluation was tested with the stats.shapiro() function of SciPy 1.7.0 in the Python 3.6 environment. A threshold of $p < 0{\cdot}05$ was set as the standard for the input data not following the normal distribution, and the results showed that the error values, regardless of either MED or MMD measurements using either coronal or sagittal captures, did not follow a normal distribution. Therefore, the Wilcoxon signed-rank test [24], the nonparametric equivalent to the paired *t*-test, was performed using the stats.wilcoxon() function in SciPy. Mean value and standard deviation (SD) were calculated, and $p < 0{\cdot}0001$ was considered statistically significant.

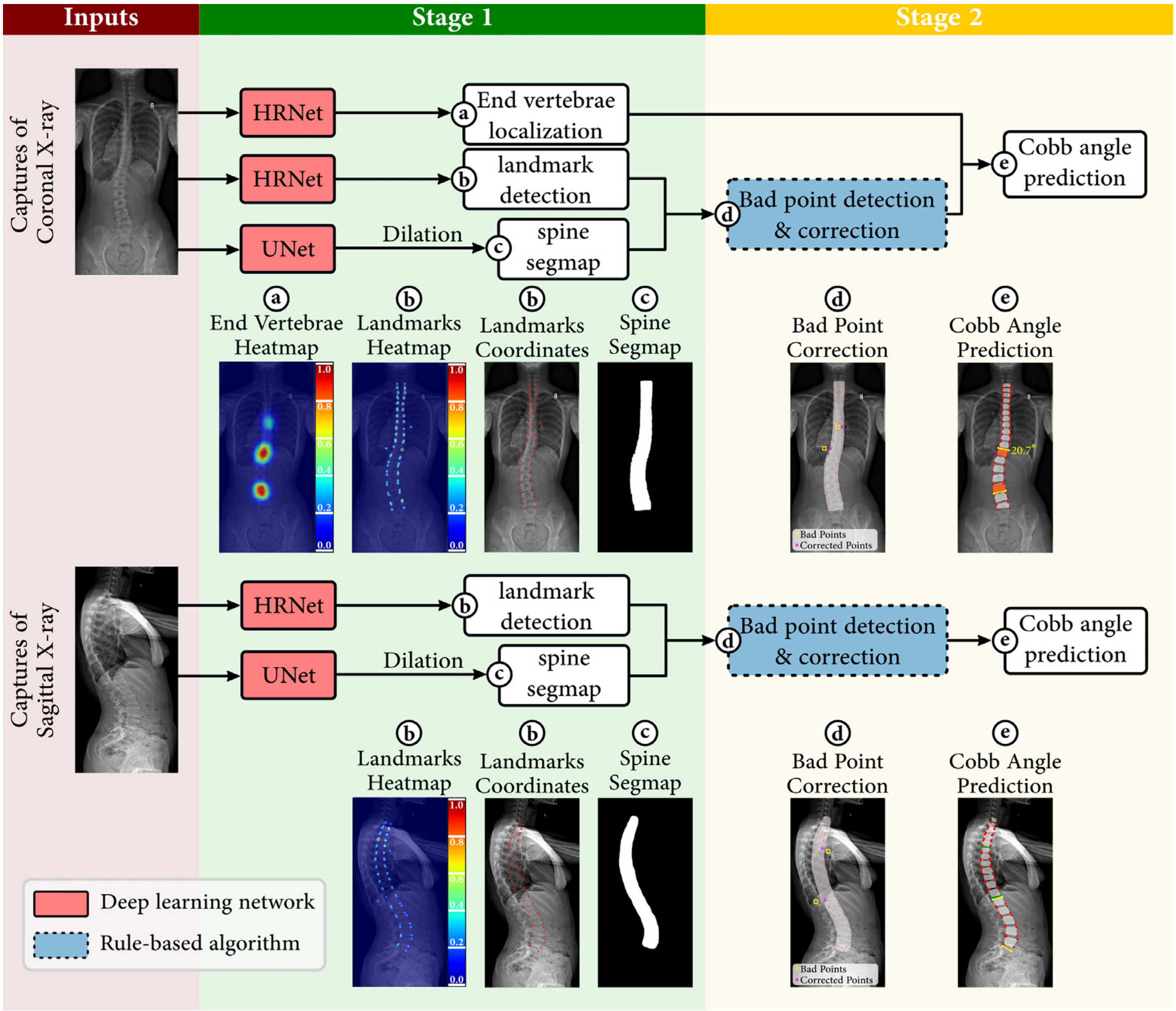


Figure 2. Overview processing pipeline of the SpineHRNet+. The model consists of two stages. The stage 1 (denoted with the green panel) adopts deep learning models for endplate and vertebral landmark detection and spine segmentation map (segmap) generation. The stage 2 (yellow panel) uses rule-based techniques for bad point detection and correction. The corrected landmarks are used for alignment prediction. In each panel, we visualize the outputs in the corresponding stage and use digits (a–e) to concatenate each output with the corresponding processing step. Compared with end-to-end models, the designed staged pipeline increases our model interpretability. For coronal X-ray captures, stage 1 predicts the end vertebra (a) and endplate landmarks (b) indicated by the heatmaps. The heatmap suggests the probability of the appearance of end vertebrae or landmarks within the image. High probability regions are indicated with warm colour while low probability regions are indicated with cold colour. Another output is a binary map indicating the spine region with white pixels (c). Such binary map is useful for bad point detection subsequently.

CA presence detection is a binary classification task to distinguish a curve on the spine radiograph, and we evaluated the model performance on posteroanterior radiographs. Coronal CA location (3 types: T, TL and L) detection at different regions of the spine and the severity (3 types: normal-moderate, moderate, and severe) were assessed. To assess classification performance, true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values were counted. Five descriptive quantitative assessments were calculated, namely, sensitivity (Sn), specificity (Sp), precision (Pe), negative prediction value (NPV), and accuracy (Acc), as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{4}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \tag{5}$$

$$\mathrm{Pe} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{6}$$

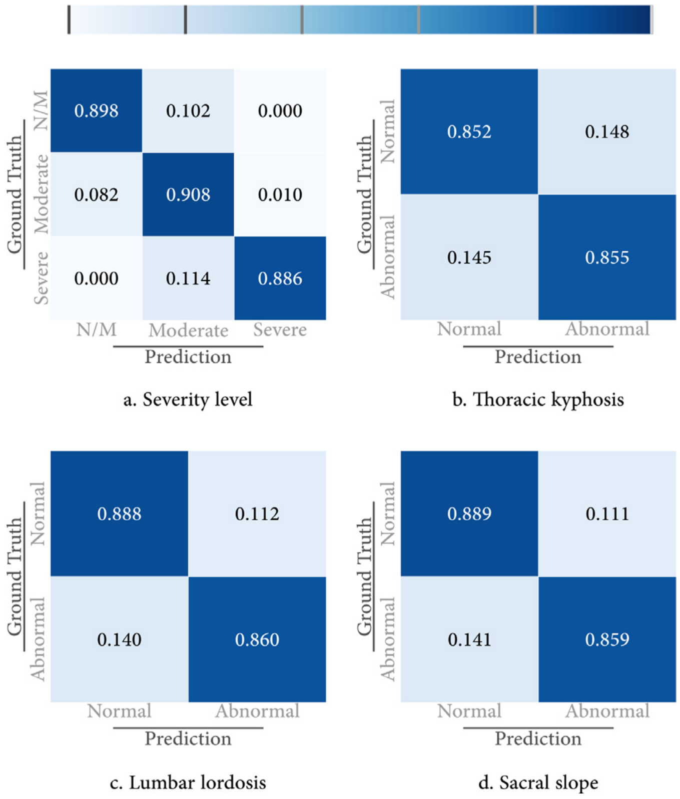


Figure 3. Confusion matrices for the severity classification and sagittal CA detection on the prospective test data. The first Figure (a) presents the confusion matrix for severity classification. "N/M" denotes the class Normal-mild. Figure (b–d) present the confusion matrix for three sagittal CAs (TK, LL, and SS), respectively.

$$\mathrm{NPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{7}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}. \tag{8}$$

A confusion matrix analysis was conducted for the severity and type prediction on coronal alignments. For sagittal alignment assessment, we distinguish the patients from healthy controls according to the normal range of different parameters. Confusion matrices were generated for these sagittal parameters to visualise the agreement between GT and predicted results.

To quantitatively assess the validity of our medical AI system, linear regression and Bland-Altman analysis [25] were conducted on several clinically relevant parameters defined in Section 2.2. That is, for coronal cases, thoracic CA, thoracolumbar CA, and lumbar CA were considered, and for sagittal cases, TK, LL, and SS were considered. For each clinical parameter, linear regression was conducted between the predicted value and GT. The regression line (blue line), the 95% confidence interval of the predictions (green dashed line) and the perfect correspondence (red line) between the predictions and GT are shown in Figure 3 as well. The Bland-Altman analysis was performed between the mean of the predicted and GT values ((prediction+GT)/2) and their residual ((prediction-GT)/2) to examine the agreement of CAs between the predictions of SpineHRNet+ and the GTs. All statistical analyses in this study were performed using SciPy (version 1.7.0) [26] and scikit-learn (version 0.23.2) [27] python packages.

Role of the funding source

Our study was funded by the RGC Research Impact Fund (R5017-18F), Innovation and Technology Fund (ITS/404/18), and the AOSpine East Asia Fund (AOSEA(R) 2019-06). The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

The 1349 cases for model development consisted of 502 normal-mild deformity curves, 670 moderate curves,

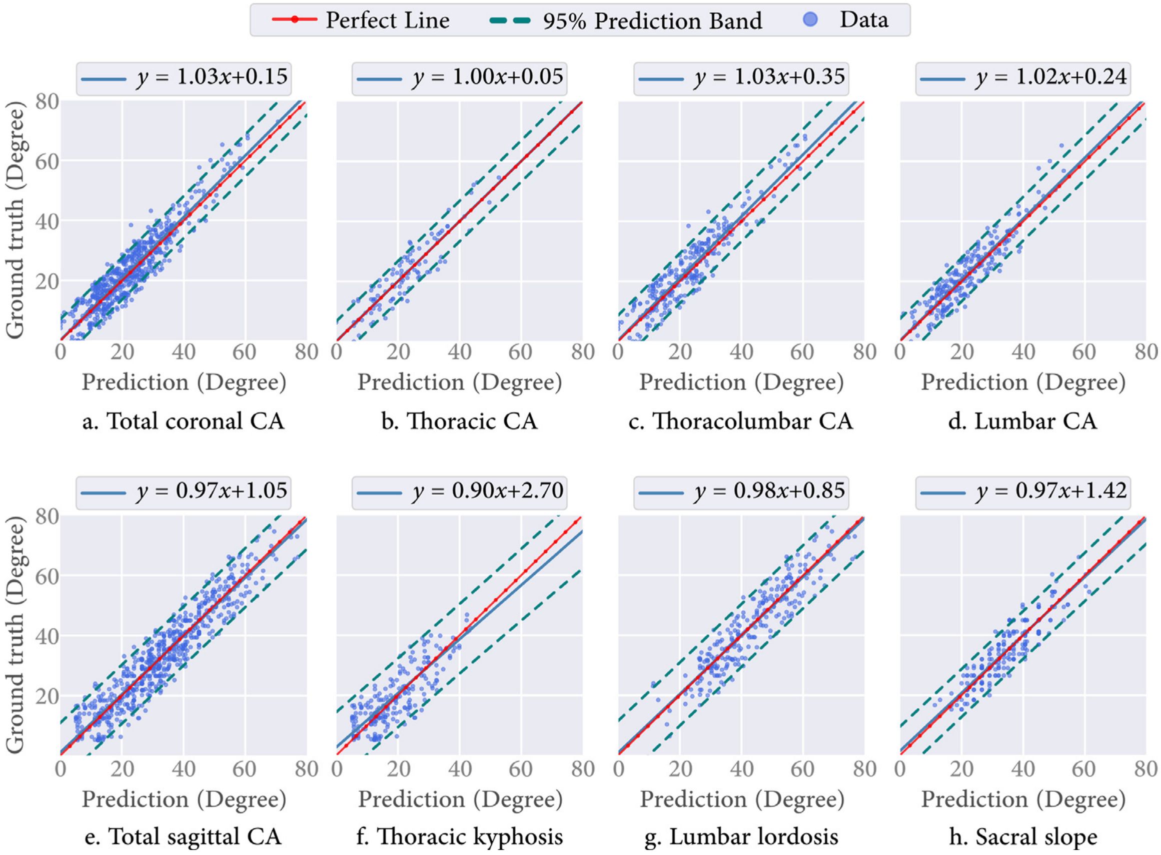
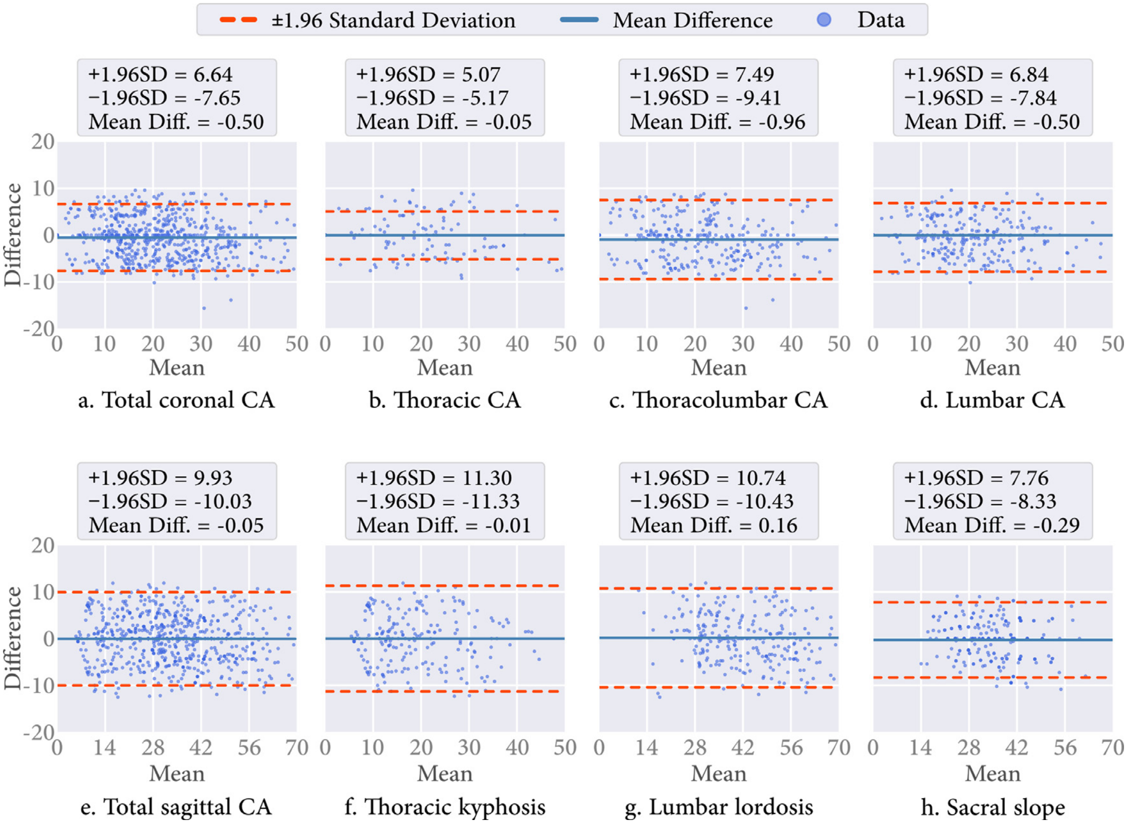


Figure 4. Linear regression analysis of three coronal (T, TL, and L CAs) and three sagittal (TK, LL, and SS) clinical parameters on the prospective test data. The x-axis denotes the values predicted by SpineHRNet+, while the y-axis refers to the GT alignments. The first row presents the regression results for coronal parameters: Figure (a) counts all the coronal parameters while Figure (b–d) present linear regression results for T, TL, and L CAs, respectively. The second row presents the regression results for sagittal parameters: Figure (e) counts all the sagittal parameters while Figure (f–h) present linear regression results for TK, LL, and SS, respectively. The 95% confidence interval of the predictions (green dashed lines), the regression line (blue line) and the perfect alignment line (red line) between GT and predictions are shown. The coefficients for the regression line are presented in the legend of each figure. All units in the figure are in degrees.

177 severe curves, 1067 thoracic curves, 582 thoracolumbar curves, and 387 lumbar curves. The prospective cohort consisted of 118 normal-mild deformity curves, 184 moderate curves, 35 severe curves, 256 thoracic curves, 158 thoracolumbar curves, and 82 lumbar curves (Table 1). The detailed demographics of each studied cohort were presented in Table 2.

For vertebral endplate landmark predictions, we compared the performance of SpineHRNet+ with its previous version (SpineHRNet).[5] By combining both the rule-based techniques and deep learning methods, SpineHRNet+ achieved increased accuracy in terms of two quantitative measurements for landmark predictions. Table 3 compared the quantitative performance of the two models on the prospective test data (quantitative results on the in-house validation data were presented in Table 2.1 in the supplementary materials). As shown, SpineHRNet+ reduced both MED and MMD loss by at least 26·3% and 25·3%, respectively. Results of the Wilcoxon signed-rank test suggested that the performance of SpineHRNet+ was significantly improved (all $p < 0·0001$) compared with our previous version. In addition, several nonparametric statistical descriptors of two quantitative measurements, such as the median and interquartile range (IQR) were calculated to comprehensively evaluate the performance. The quantitative results were presented in Table 4, where SpineHRNet+ outperformed SpineHRNet with lower median value and smaller IQR value.

To assess the predictive accuracy of the coronal CAs, we calculated the statistics of the errors between the predicted CAs and GT CAs (Table 5). The mean (±SD) errors of the predicted T, TL, and L CAs from SpineHRNet+ were 1·2° (± 2·3°), 3·2° (± 2·9°), and 2·6° (± 2·7°), respectively. For sagittal parameters TK, LL, and SS, the errors were 6·3° (± 6·1°), 5·9° (± 6·9°), and



a. Total coronal CA

b. Thoracic CA

c. Thoracolumbar CA

d. Lumbar CA

e. Total sagittal CA

f. Thoracic kyphosis

g. Lumbar lordosis

h. Sacral slope

Figure 5. Bland-Altman plots assessing the agreement of alignments between the SpineHRNet+ predictions and the GT (The order of the subfigures corresponds to those in Figure 5) on the prospective test data. The Y-axis indicates the angle difference (in degree) between predicted results and the GT (i.e., predictions-GT). The X-axis represents the average degree of them (i.e., (predictions+GT)/2). Both the mean and standard deviation values are reported in the legend of each subfigure. Figure (a) presents the Bland-Altman plot for all the coronal parameters while Figure (b–d) count on T, TL, and L CA, respectively. Figure (e) presents the Bland-Altman plot for all the sagittal parameters while Figure (f–h) count on TK, LL, and SS, respectively.

4·1° (± 5·4°), respectively. Table 5 compared the performance of SpineHRNet and SpineHRNet+ on the prospective test data (the comparison results on the in-house validation data were presented in Table 2.2 in the supplementary materials). SpineHRNet+ outperformed SpineHRNet on all the coronal and sagittal parameters. The overall prediction of coronal CAs was more accurate than the prediction of sagittal measurements with smaller mean degree error and smaller SD.

For coronal CA detection at different regions of the spine, Table 6 displayed the results of statistics (results of statistics on the in-house validation data were presented in Table 2.3 in the supplementary materials). The Sn (95·7–97·4%), Sp (88·1–98·4%), NPV (87·1–98·8%), and Acc (93·8–97·9%) exhibited high scores in L curve detection and low scores in T curve detection, however Pe (95·4–96·0%) had the highest score for T curve detection. For AIS severity classification (Figure 3(a)), our model achieved a specificity of no less than 88·6% on all three groups, with the highest performance in the moderate cases but lowest performance in the severe cases. For sagittal alignment, confusion matrices (Figure 3 (b–d)) demonstrated high performance (>85·2%) for all sagittal parameters. The severity classification results on the in-house validation data were presented in Figure 2.1 in the supplementary materials.

The prospective reliability of the SpineHRNet+ assessed by linear regression analysis between the predicted and GT (Figure 4), indicated a strong correlation with the GT (Table 7: $p < 0·001$) in predicting all alignment parameters, with an overall R^2 of 0·970 for coronal parameters and 0·982 for sagittal parameters. The slope of the regression line was 45·73° for all coronal parameters and 44·13° for the sagittal parameters. Both were close to the ideal value of 45°, which indicates the perfect agreement between predicted CA and GT. Table 2.4 and Figure 2.2 in the supplementary materials showed the regression results on the in-house validation data. Bland-Altman plots (Figure 5 in the main text and Figure 2.3 in the supplementary materials) visualised the difference versus average degree values between the predicted and GT CAs. The overall mean difference between the GT and the predicted CAs was minimal at −0·5° for coronal and −0·05° for sagittal parameters.

Severity	Total number	Male	Female	Average CA (degree)	Average age	Curve type			
	of subjects				(years old)	N	T	TL/L	Mixed
Training cohort									
Normal-mild (CA≤20°)	381	125	256	14·1	14·3	68	119	82	112
Moderate (20°<CA≤40°)	551	111	440	27·9	14·2	–	100	71	380
Severe (CA>40°)	147	29	118	54·3	14·8	–	20	3	124
Validation cohort									
Normal-mild (CA≤20°)	123	33	90	14·1	14·2	17	32	37	37
Moderate (20°<CA≤40°)	115	22	93	28·5	14·1	–	20	3	92
Severe (CA>40°)	32	6	26	52·2	15·3	–	4	1	27
Prospective cohort									
Normal-mild (CA≤20°)	115	43	72	13·7	14·2	23	30	36	26
Moderate (20°<CA≤40°)	190	45	145	29·1	14·3	–	41	21	128
Severe (CA>40°)	32	3	29	52·2	15·6	–	7	1	24

***Table 2*: Demographics of different cohorts.**
CA: Cobb angle; N: normal; T: thoracic; TL: thoracolumbar; L: lumbar; Mixed: both appear T and TL/L.

	MED (Pixels)		MMD (Pixels)	
	Coronal	Sagittal	Coronal	Sagittal
SpineHRNet	3·8	7·6	4·2	8·7
SpineHRNet+	2·8	5·5	3·1	6·5
Loss reduction (%)	26·3%	27·6%	26·2%	25·3%
p-value*	< 0·0001	< 0·0001	< 0·0001	< 0·0001

***Table 3*: Quantitative error results in terms of two distance measurements between predicted landmarks and GT landmarks on the prospective test data.**
* Significant improvement ($p < 0·0001$).
MED: mean Euclidean distance; MMD: mean Manhattan distance.

The visualisation of landmark detection and CA prediction (Figure 6) provides interpretable alignment analysis for users, while our platform allows users to modify the landmarks with the alignment re-computed concurrently. The difference between predicted landmarks and GTs was minimal (MED: 1·88–3·69 pixels; MMD: 2·06–3·97 pixels). The performance of landmark detection and CA prediction were both satisfactory and clinically interpretable.

Discussion

In this study, we developed a hybrid AI-powered model (SpineHRNet+) linking with our open platform AlignProCARE for easily accessible auto-alignment analysis. We prospectively tested its validity on multiple analytic tasks, including endplate landmark detection with the prospective evaluation of CA prediction, end vertebrae localisation, and severity classification in real-world trials. With this open platform, users can upload their radiographs by taking photos with a smartphone and instantly access the analysis results with the flexibility of further modification. Both the images and analytic results are encrypted during the uploading and downloading, with all processing and analysis performed on our server with SpineHRNet+.

Considering that spine alignment analysis is critical in the diagnosis and further management planning of scoliosis, assessment reliability and speed are important. Previously published studies have reported the use of machine learning models for measuring CAs [28–30]. Although these studies have shown the potential value

	Euclidean distance				Manhattan distance			
	Coronal Median	IQR	Sagittal Median	IQR	Coronal Median	IQR	Sagittal Median	IQR
SpineHRNet	3·2	1·8	5·7	5·1	3·5	1·9	6·3	5·7
SpineHRNet+	2·5	0·7	3·4	2·7	2·8	0·8	3·9	3·3

***Table 4*: Median and IQR results in terms of two distance measurements between predicted landmarks and GT landmarks on the prospective test data.**
IQR: interquartile range.

Alignment parameters		Mean (degree)		Standard deviation	
		SpineHRNet	SpineHRNet+	SpineHRNet	SpineHRNet+
Coronal	T curve	1·3	1·2	2·6	2·3
	TL curve	3·3	3·2	3·2	2·9
	L curve	2·7	2·6	3·2	2·7
Sagittal	Thoracic kyphosis	6·8	6·3	6·6	6·1
	Lumbar lordosis	6·7	5·9	8·3	6·9
	Sacral slope	5·1	4·1	7·5	5·4

***Table 5*: Angle difference between predicted alignments using SpineHRNet+ and GT alignments on the prospective test data.**
T: thoracic; TL: thoracolumbar; L: lumbar.

Performance metrics	CA location T	TL	L
Sensitivity	0·965	0·957	0·974
Specificity	0·984	0·881	0·903
Precision	0·954	0·960	0·958
NPV	0·988	0·871	0·939
Accuracy	0·979	0·938	0·953

***Table 6*: Performance metrics of SpineHRNet+ on coronal alignment detection on the prospective test data.**
NPV: negative predictive value; T: thoracic; TL: thoracolumbar; L: lumbar.

of AI techniques in determining spine shape, the datasets used to build the models were relatively small with debatable reliability. Instead of directly predicting CAs, some recent studies promoted the use of convolutional architectures to detect vertebral landmarks, followed by the calculation of CAs [12],[14],[31]]. Most of these approaches were trained in an end-to-end manner, directly outputting the landmark coordinates. In this regard, the AI model was used as a “blackbox” in these studies, making it hard to interpret how the predictions were made.

SpineHRNet+ can visualise the heatmap of landmarks as outputs, indicating the location of each endplate landmark, and the landmark coordinates are determined by the region with the highest probability value (Figure 2). The design of the image processing

Parameter	R [2]	*p*-value	Regression line Slope in Degree	Standard error of the prediction difference
Thoracic CA	0·964	< 0·001	45·02°	2·55°
Thoracolumbar CA	0·972	< 0·001	45·89°	4·22°
Lumbar CA	0·970	< 0·001	45·46°	3·67°
Total	0·970	< 0·001	45·73°	3·57°
Thoracic kyphosis	0·929	< 0·001	41·98°	5·66°
Lumbar lordosis	0·987	< 0·001	44·33°	5·29°
Sacral slope	0·988	< 0·001	44·05°	4·02°
Total	0·982	< 0·001	44·13°	4·99°

***Table 7*: Regression analysis of correlation between GT alignments and those predicted by SpineHRNet+ on the prospective test data.**
GT: ground truth; CA: Cobb angle.

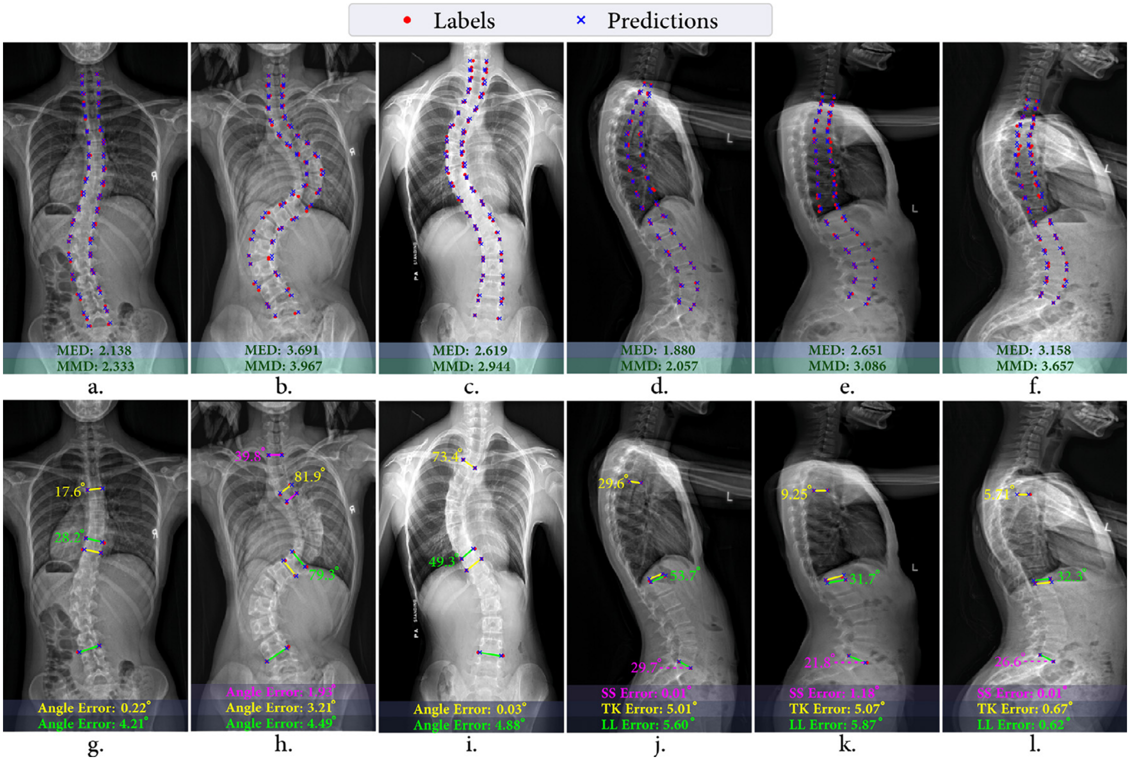


Figure 6. Visual results of landmark detection and CA prediction using our SpineHRNet+ model on the prospective test data. Three coronal and three sagittal radiograph captures were randomly selected as input and the results obtained from our model were displayed on these images. The first row exhibits the landmark detection results: Figure (a–c) present the landmark detection results on coronal radiographs; Figure (d–f) present the landmark detection results on sagittal radiographs. Red points denote the GT landmarks while the blue crosses represent the predicted position of vertebral landmarks. Both MED and MMD are reported on the bottom of each image. The second row presents the CA prediction results: Figure (g–i) present the CA prediction results on coronal radiographs; Figure (j–l) present the CA prediction results on sagittal radiographs. For coronal radiographs (first three), CAs with different types are visualized using different colours. Some have two curves while the second one has three curves. For sagittal radiographs (last three), we calculate the three CAs, i.e., TK, LL, and SS. The predicted degrees of CAs are printed beside the endplate of the top end vertebra, and on the bottom, we count the angle difference (degree) of each CA between the predicted results and ground truth.

pipeline adopted in SpineHRNet+ was inspired by the AIS clinical diagnosis procedure. We designed the pipeline into two stages, and in each stage, the model fulfils a sub-task, outputting the intermediate results. Using this approach, model interpretability is increased, also facilitating the supervision of model operation.

Our AI-powered model was trained and evaluated on captured biplanar radiographs with variable image quality to improve both the usefulness and generalisability of different user scenarios. Considering the resolution of a smartphone camera, environmental disturbances, and user handshake, the quality of user-uploaded images is usually much worse than the original high-resolution radiographs. However, even with low-quality images, our model can still achieve competitive and even better results compared with other AI models trained directly on the actual radiographs. In this regard, this work suggested that smartphone-captured radiographs are informative enough for the AI model to learn the spine alignment features.

Alignment prediction performed better for the coronal alignment and L curve detection since the radiographs demonstrate clearer individual vertebrae with the best intensity contrast in the L region (Figure 6). For the severity classifications, the moderate curves were predicted most accurately due to the larger number of such samples in the training dataset (Table 1). Moderate cases are the most prevalent in scoliosis clinics. Furthermore, the severe cases were significantly fewer compared to moderate cases, and thus the performance was slightly decreased. The imbalance in curve severities is a feature of this population. We attempted to balance the training data using re-sampling during the model development stage, but this did not improve the performance during the in-house training. Thus, we kept the original dataset distribution.

We prospectively validated SpineHRNet+ for auto-spine analysis and deployed the first automatic AI-powered platform. Exiting open platforms for this purpose require clinicians to manually annotate the radiographs, which is laborious, time-consuming, and suffers from inter-rater variation. Our model and open platform can facilitate clinicians and researchers in handling large volumes of alignment assessment requests. However, femoral head detections in lateral radiographs were not included, thus pelvic incidence and pelvic tilt cannot be provided. It is because in our training dataset, a collection of the radiographs had the femoral head not included in the imaging field. We are in the process of collection recaptured radiographs with femoral heads and improve the sagittal alignments auto-analyses further. It must be noted the system was tested in two centres followed the same procedure to collect radiographs with the same clinical assessment standard to evaluate the spine alignments. The performance of the auto-alignment may reduce when directly apply in another centre. We are planning an international multi-centre trial to further assess the reliability of Spine-HRNet+. Another limitation is no post-operative data were collected thus the auto-alignments cannot be performed on radiographs with instrumentation.

In summary, we deployed the first prospectively validated auto-alignment analysis model for spine curve classification using an open platform. The AI-powered hybrid system, SpineHRNet+ was trained and assessed with radiographs with variable qualities and sources, but still exhibited improved performance on multiple spine disease scenarios compared with our previous version. On further multi-centre validation in future, our platform can better assist clinicians and clinical research in large volumes.

Declaration of interests

All authors declare there is no conflict of interest.

Funding

We sincerely appreciated the support from the RGC Research Impact Fund (R5017–18F), Innovation and Technology Fund (ITS/404/18), and the AOSpine East Asia Fund (AOSEA(R) 2019–06).

Contributors

NM implemented the hybrid deep learning and rule-based model, deployed the model on the server, and managed the server. JPC supervised the clinical parameters and significance, as well as was responsible for the annotations and labelling. KKW and SD co-supervised the pipeline design. PHL, WCC, CHT and JRL contributed to data collections at two spine centres. TZ designed the study, searched the literature, supervised the data collection and model implementation, was responsible for results assessments, as well as developed the open platform. NM and TZ drafted the manuscript, while all the authors reviewed the manuscript for important intellectual content and approved the final manuscript.

Data sharing statement

The test data is available from the corresponding authors on a reasonable request and under the related institution policy.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eclinm.2021.101252.

References

1 De Se₹ze M, Cugy E. Pathogenesis of idiopathic scoliosis: a review. *Ann Phys Rehabil Med* 2012;**55**(2):128–38.

2 Fong DY, Cheung KM, Wong YW, Wan YY, Lee CF, Lam TP, et al. A population-based cohort study of 394,401 children followed for 10 years exhibits sustained effectiveness of scoliosis screening. *Spine J* 2015;**15**(5):825–33.

3 Cheung JPY, Cheung PWH, Samartzis D, Luk KD. Curve progression in adolescent idiopathic scoliosis does not match skeletal growth. *Clin Orthop Relat Res* 2018;**476**(2):429–36.

4 Zhang L, Wang H, Li Q, Zhao M-H, Zhan QM. Big data and medical research in China. *BMJ* 2018: 360.

5 Zhang T, Li Y, Cheung JPY, Dokos S, Wong KYK. Learning-based coronal spine alignment prediction using smartphone-acquired scoliosis radiograph images. *IEEE Access* 2021;**9**:38287–95.

6 Meng N, Lam EY, Tsia KK, So HKH. Large-scale multi-class image-based cell classification with deep learning. *IEEE J Biomed Health Inform* 2018;**23**(5):2091–8.

7 Meng N, So HK, Lam EY. Computational single-cell classification using deep learning on bright-field and phase images. In: In: Proceedings of the IAPR international conference on machine vision applications, IEEE; 2017.

8 Kuang X, Cheung JP, Wu H, Dokos S, Zhang T. MRI-SegFlow: a novel unsupervised deep learning pipeline enabling accurate vertebral segmentation of MRI images 2020. In: In: Proceedings of the 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC), IEEE; 2020.

9 Zhang J, Cheuk KY, Xu L, Wang Y, Feng Z, Sit T, et al. A validated composite model to predict risk of curve progression in adolescent idiopathic scoliosis. *EClinicalMedicine* 2020;**18**:100236.

10 Horng MH, Kuok CP, Fu MJ, Lin CJ, Sun YN. Cobb angle measurement of spine from X-ray images using convolutional neural network. *Comput Math Methods Med* 2019;**2019**.

11 Pan Y, Chen Q, Chen T, Wang H, Zhu X, Fang Z, et al. Evaluation of a computer-aided method for measuring the Cobb angle on chest X-rays. *Eur Spine J* 2019;**28**(12):3035–43.

12 Wu H, Bailey C, Rasoulinejad P, Li S. Automated comprehensive adolescent idiopathic scoliosis assessment using MVC-Net. *Med Image Anal* 2018;**48**:1–11.

13 Zhang J, Li H, Lv L, Zhang Y. Computer-aided Cobb measurement based on automatic detection of vertebral slopes using deep neural network. *Int J Biomed Imaging* 2017;**2017**.

14 Galbusera F, Niemeyer F, Wilke HJ, Bassani T, Casaroli G, Anania C, et al. Fully automated radiological analysis of spinal disorders and deformities: a deep learning approach. *Eur Spine J* 2019;**28**(5):951–60.

15 Liu J, Yuan C, Sun X, Sun L, Dong H, Peng Y. The measurement of Cobb angle based on spine X-ray images using multi-scale convolutional neural network. *Phys Eng Sci Med* 2021: 1–13.

16 Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI

interventions: the STARD-AI steering group. *Nat Med* 2020;**26**(6):807–8.

17 Mak T, Cheung PWH, Zhang T, Cheung JPY. Patterns of coronal and sagittal deformities in adolescent idiopathic scoliosis. *BMC Musculoskelet Disord* 2021;**22**(1):1–10.

18 Fon GT, Pitt MJ, Thies AC. Thoracic kyphosis: range in normal subjects. *Am J Roentgenol* 1980;**134**(5):979–83.

19 Lin R, Jou IM, Yu CY. Lumbar lordosis: normal adults. *J Formos Med Assoc* 1992;**91**(3):329–33.

20 Legaye J. The femoro-sacral posterior angle: an anatomical sagittal pelvic parameter usable with dome-shaped sacrum. *Eur Spine J* 2007;**16**(2):219–25.

21 Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 2020.

22 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: In: Proceedings of the international conference on medical image computing and computer-assisted intervention, Springer; 2015.

23 Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. *Neural Inf Process Syst* 2019;**32**:8026–37.

24 Wilcoxon F. Individual comparisons by ranking methods. Breakthroughs in statistics. Springer; 1992. p. 196–202.

25 Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *J R Stat Soc Ser D Stat* 1983;**32**(3):307–17.

26 Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.

27 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.

28 Sun H, Zhen X, Bailey C, Rasoulinejad P, Yin Y, Li S. Direct estimation of spinal cobb angles by structured multi-output regression. In: In: Proceedings of the international conference on information processing in medical imaging, Springer; 2017.

29 Kokabu T, Kanai S, Kawakami N, Uno K, Kotani T, Suzuki T, et al. An algorithm for using deep learning convolutional neural networks with three dimensional depth sensor imaging in scoliosis detection. *Spine J* 2021;**21**(6):980–7.

30 Zhang J, Lou E, Le LH, Hill DL, Raso JV, Wang Y. Automatic Cobb measurement of scoliosis based on fuzzy Hough transform with vertebral shape prior. *J Digit Imaging* 2009;**22**(5):463.

31 Wu H, Bailey C, Rasoulinejad P, Li S. Automatic landmark estimation for adolescent idiopathic scoliosis assessment using BoostNet. In: In: Proceedings of the international conference on medical image computing and computer-assisted intervention, Springer; 2017.