A clinical classification for radiation-less monitoring of scoliosis based on deep learning of back photographs

Teng Zhang (✉ tgzhang@hku.hk)
The University of Hong Kong https://orcid.org/0000-0002-5310-8766
Chuang Zhu
Yongkang Zhao
Beijing University of Posts and Telecommunications
Moxin Zhao
The University of Hong Kong
Zhihao Wang
Beijing University of Posts and Telecommunications
Ruoning Song
Beijing University of Posts and Telecommunications
Nan Meng
The University of Hong Kong
Alisha Sial
University of New South Wales
Ashish Diwan
University of New South Wales
Jun Liu
Beijing University of Posts and Telecommunications
Jason Pui Yin Cheung
Department of Orthopaedics and Traumatology, The University of Hong Kong https://orcid.org/0000-0002-7052-0875

Article

Keywords:

Posted Date: May 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1655808/v1

Abstract

Routine check-ups for adolescent idiopathic scoliosis are critical to monitor progression and prescribe interventions. AIS is primarily screened via physical examination. If there are features of deformity, radiographs are necessary for diagnosis or follow-up, guiding further management, i.e., bracing for moderate deformity and surgery for severe. However, this subjects children to repetitive radiation and routine practices can be disturbed. Here, we demonstrate a mobile platform powered by *ScolioNet*, being trained and validated by 1780 back photos with paired radiographs, consisting of heterogeneous severities and curve types, manually annotated by experienced deformity surgeons. In further independent testing on 378 patients, the platform recommended follow-up or surgery with area under curves (AUCs) of receiver operating characteristic curves being 0.839, and 0.902, while distinguishing among thoracic, thoracolumbar/lumbar, or mixed curve types with AUCs of 0.777, 0.760 and 0.860 respectively. Our open platform has potential for out-of-hospital accessible applications in managing children with AIS.

Introduction

Adolescent idiopathic scoliosis (AIS) is the most common pediatric spinal disorder manifesting as a 3-dimensional spinal deformity[1]. It is reported to occur in up to 2.2% of boys and 4.8% of girls[2] and with a high prevalence of progression during puberty, thus early detection and proper interventions are critical. Otherwise, the spinal deformity can progress rapidly and reduce the quality of life and mobility of the patient. Cardiopulmonary impairment as well as back pain are associated with severe cases. Therefore, early detection and close follow-up (FU) of the patients to monitor the disease process is critical. Our organisation initiated a school screening program[3] since 1995 as a part of the territory-wide annual comprehensive health assessment scheme for AIS. However, the COVID-19 Pandemic has influenced the practice and indicates an increased need for out-of-hospital (easily accessible) assessment[4].

Current AIS detection and FU assessments require intensive clinician experience and expertise. Clinically, available assessment tools include physical examinations and radiographic (X-ray) assessment[5]. Shoulder height, waist asymmetry, thoracic cavity asymmetry, rib, and breast deformity are recorded. However, the assessments of the external appearance of the back are subjective. Moreover, physical examinations hardly can detect the specific underlying pathology type, thus further X-ray is necessary[6]. However, repeated X-ray examinations, which is the norm for AIS monitoring and management, carries attendant consequences of increased radiation exposure[7] with an increment of 2% in breast cancer[8] and 3% in heritable defect[9], and higher risks of reproductive issues[10] including unsuccessful attempts at pregnancy, spontaneous abortions, infants with congenital malformations and lower birthweight.

Automated detection and classification of AIS using easily accessible smartphone images of the patients is an option for radiation-free and out-of-hospital assessment but is challenging due to mainly two reasons: 1) Though smartphone images have significant advantages in accessibility and being radiation-free, they introduce variability including vibration, angle, lighting, and noisy background, making

classification challenging. 2) The patient's back with spine deformity has variable appearances subjected to different degrees of severity and curve types.

Convolutional neural networks (CNNs) can provide a feasible solution for highly variable input across many classification tasks[11–14]. However, previous work in using deep learning for spine specialists-level classification of AIS[15–17] mainly uses radiographic images to accurately calculate the Cobb angle (CA), and are restricted to the specific machine and location for providing the source images. Additional previous radiation-free approaches in detection includes Moiré topography and/or clinical standardized photographs of the patients' back. Moiré[18] topography's glass panels are difficult to recreate, and its accuracy is arguable. A previous study using images acquired via professional cameras in a controlled environment with clear background[11] lacked accessibility as well as prospective clinical validations and generalization ability. These limitations are largely due to the lack of easily accessible data and focus on the standardized tasks such as controlled medical grade images.

We overcome this challenge by developing a virtual spinal evaluation platform known as Alignment Professionals (AlignProCARE[19]) because there is a need for out-of-hospital spine deformity evaluation, which can help the patients to receive fast assessment out of the hospital and treatment planning[20]. By using the gold standard disease taxonomy, i.e. pathologies classified on real X-rays images as ground truths (GTs) and validated deep neural networks (*ScolioNet*) trained using cohort 1 (1780 patients' back images taken by mobile phones with photographic variability), our system accepts arbitrary scenes, and it is trained end-to-end directly from GT labels and images for automated and mobile scoliosis classification with no introduction of radiation. We further demonstrated the system reliability using an independently collected cohort 2 with 378 AIS patients. This platform can continuously benefit medical practice and research during and after the pandemic, by providing fast and consistent alignments evaluation, as well as standardized medical data management and easily accessible communications. We have made AlignProCARE[19] with *ScolioNet* freely available for academic work at http://aimed.hku.hk/alignprocare.

Results

A summary of collected dataset

Between October 2018 to September 2020, all AIS patients attending a tertiary referral center completed a written consent (ethics approved by the local regulatory body), following the inclusion and exclusion criteria for our study (*Supplementary materials*), Fig. 1a shows the patient recruitment workflow with 1780 patients of the 1924 participants who were eligible for the study to populate cohort 1 (Table 1) (mean age 14 years, range 10-18 years) for the development of the *ScolioNet*. Another independent testing recruited cohort 2 of 378 patients (mean age 14 years, range 10-18 years, Table 1) consecutively attending AIS clinics after the October of 2020 till the March of 2021 were assessed by *ScolioNet*. For both cohorts (n=2158), in total 652 cases (30.2%) required no intervention, 1250 cases (57.9%) required non-surgical interventions with regular FUs and 256 cases (11.9%) were under consideration for surgery.

All 2158 participants had the radiographic data and routine clinical assessment with an extra back image taken (Fig. 1b&c). From the original radiographic images in DICOM format, 4303 radiographic slides in PNG format were exported from the picture archiving and communication system (PACS) of the hospital (Fig. 1d), including 1295 slides (30.1%) for subjects who had no intervention, 2499 slides (58.1%) for those who had non-surgical interventions, and 509 slides (11.8%) for the ones under surgical consideration. Classified on the X-rays by spine surgeons (Fig. 1e), for cohort 1, the severity levels ranged from normal-mild (n=555, 31.2%), moderate (n=1055, 59.3%) to severe (n=170, 9.6%), as well as for cohort 2, normal-mild (n=97, 25.7%), moderate (n=195, 51.6%) and severe (n=86, 22.7%). Different clinical interventions are applied with different AIS severities. For normal-mild deformity, no clinical interventions are provided, while non-surgical interventions (bracing and/or specific exercises) is initiated for moderate cases with regular FUs and surgery may be recommended for severe cases.

The curve types were classified into thoracic (T), thoracolumbar/lumbar (TL/L) and mixed curves with both T and TL/L (Table 1). Despite 54 eligible participants were normal with no curves for cohort 1, the remaining 1726 participants had curve types including T (n=188, 10.9%), TL/T (n=385, 22.3%) and mixed curve (n=1153, 66.8%). For cohort 2, deducting the 13 normal cases, the remaining 365 cases had curve types including T (n=38, 10.4%), TL/T (n=65, 17.8%) as well as mixed curves (n=262, 71.8%). Different curve types (Fig. 1c) demonstrated different appearance features discerned by spine surgeons, with the T curve consisting of rib humps, chest wall deformities and unleveled shoulders, whereas the TL/L curves developing unbalanced pelvic and waistline deformities (Table 1).

An open platform for radiation-free AIS classifications

We trained our model on ordinary back images with the gold standard disease taxonomy (i.e. GTs generated from the X-rays reviewed by spine surgeons) to get accurate and effective features of the spine specialist-level classification shown in Fig. 2a. Overview of the AlignProCARE[19] platform (available via web application, App Store and Google Play) powered by *ScolioNet* for classifying deformity severities and curve types using radiation-free back images (Supplementary Fig. 1) is to exemplify the usefulness of *ScolioNet*. It integrates the smartphone photos, radiographs and GT disease taxonomy to predict clinical interventions (both for severity levels and disease types) of AIS patients, which is made freely open for all clinicians and researchers doing research on AIS. For user convenience, an example (Supplementary Fig. 1) is provided. Entering the platform, users can manage the patient information along with back images and/or radiographs. Cases can be added by clicking the ‘Add patient’ button and images can be taken by clicking the ‘Upload photo’ button. The back images are sent to the server located in our hospital for automated processing. Results are sent back to the mobile device for interpretable analysis with auto-results, with the flexibility for the linked clinician to modify and confirm. If the modified landmarks are not saved, the CNN computed results are saved in the system until further updates. All images are securely and anonymously stored with authorized access under local regulation for research purposes.

Development of *ScolioNet* for comprehensive AIS analysis

ScolioNet comprises of three steps, including the radiation-free back photo-based severity auto-classification, the radiation-free back photo-based disease typing, and intervention recommendations (Fig. 3). We detected the normal-mild deformity and the severe deformity based on the GT X-rays analytical results, to avoid unnecessary radiation to the patients and to activate prompt interventions. The curves are typed into single curve with T or TL/L, or mixed curves with more than one curve, all typed manually by spine specialists on radiographs via consensus. To enable back image-based predictions, we implemented a deep learning framework *ScolioNet* with ResNet50[21] as backbone (Fig. 3b). The original ResNet50 contains 50 layers and 25.5×10^6 parameters. To increase the model robustness and reduce the complexity, we modified the model by adding residual attention block[22] (Fig. 3b) and auxiliary classification branches, with additional attention module and multi-task strategy. *ScolioNet* framework outperformed other deep learning models (Densenet169, ResNet50, ResNeXt50, Vgg16, and Inception V4) when we tested for Spine specialists-level classification of scoliosis (Supplementary Table 1).

The prediction accuracy of *ScolioNet* with independent clinical testing

The completed *ScolioNet* achieved an AUC score of 0.881 on the in-house validation set, details shown in Supplementary Table 1. First, using 1,780 manually labelled radiogaphic results of the disease severity and curve type as GT for the paired smartphone photos. From the ROC curve (Supplementary Fig. 2 and Supplementary Table 1), *ScolioNet* achieved an AUC value of 0.869 in distinguishing normal-mild cases from other cases with back photos recommending no clinical intervention, and an AUC value of 0.958 in predicting severe cases recommending possible surgical interventions (Supplementary Fig. 2 and Supplementary Table 1).

We further conducted independent testing for AIS severity and curve typing predictions using cohort 2 (Fig. 4a and Table 2-4). In *ScolioNet*, we selected a sensitive threshold to maximize the ability to correctly predict severity with corresponding clinical implications. The AUC for predicting the AIS being normal-mild (no special interventions) or severe deformity (considering surgeries: CS) were 0.839 (Table 2) and 0.902 (Table 3) respectively, recommending either FUs or CS, and the remained for no interventions. Confusion matrices were generated to visualize the agreement between actual and predicted results (Fig. 4b). It was found that by using smartphone photos, severities could be correctly recognized (Fig. 4) with the sensitivity (*ScolioNet* vs senior surgeon vs junior surgeon: FU: 84.88% vs 44.19% vs 62.79%; CS: 82.56% vs 20.93% vs 19.76%) and negative predictive values (NPV; *ScolioNet* vs senior surgeon vs junior surgeon: FU: 89.22% vs 71.76% vs 79.35%; CS: 90.00% vs 70.43% vs 70.51%) higher than the severity classification labelled by spine surgeons (Table 2-3).

To train models for curve type predictions, 1,780 photos with a single T curve, a single TL/L curve or mixed curves were used from cohort 1 after eliminating the normal cases with no curves. The AUC value for curve type prediction achieved 0.777, 0.760, 0.860 on the independent testing dataset (Table 4). *ScolioNet* performed with a comparable predictive accuracy with the senior surgeon and an increased accuracy compared to the junior surgeon (*ScolioNet* vs senior surgeon vs junior surgeon: T: 72.51% vs 71.08% vs 65.24%; TL/L 72.93% vs 69.09% vs 65.10%; mixed 74.07% vs 66.95% vs 30.34%). An increased sensitivity was also revealed in detecting the curve types by *ScolioNet* in comparison with the senior surgeon (*ScolioNet* vs senior surgeon: T: 82.31% vs 76.64%; TL/L 81.18% vs 75.49%; mixed 87.32% vs 41.31%), as well as an increased NPV (*ScolioNet* vs senior surgeon: T: 61.97% vs 58.04%; TL/L 62.11% vs 55.56%; mixed 92.52% vs 75.35%). The junior surgeon's sensitivity was 100% with 0 specificity and incomputable NPV due to the fact that all types were selected during manual assessment.

Model interpretability

The interpretable heatmap reflects the areas in the image that are used to support classification decisions. In the heatmap, warm colors are used to describe increased degree of support for classification decisions in different regions in the image. With increased severity, the attention pattern tended to have increased distortions (Fig. 5a-f). For curve typing, we demonstrated that the curve T had the attention in the thoracic region (Fig. 5a, d, g) whereas the curve L had the attention in the lumbar region (Fig. 5b, e, h). Mixed curves had attention in both thoracic and lumbar region (Fig. 5c, f, i).

Discussion

Since March 2020, COVID-19 has been causing disturbance to our usual clinic routine. It is because AIS diagnosis is primarily screened via physical examination followed by radiographic diagnosis if the back appeared to have deformity features and abnormal curvatures. This requires professional clinical expertise, stationary medical imaging machines with controlled environment and physical contact with the patients. Additionally, routine use of X-ray examinations will expose the patient to repetitive radiation, which may affect child development[23]. With growing interest in out-of-hospital management, we developed an open platform AlignProCARE[19] powered by *ScolioNet* (trained on photographs of the patient's back and disease labels) to automatically classify AIS severities and curve types without the need of specific backgrounds or medical machine, as well as without the need for face-to-face contact and exposure to ionizing radiation. *ScolioNet* was integrated into our open platform AlignProCARE[19] and its performance was tested independently. Using the classification done by spine specialists on radiographs (gold standard GT), FU and CS classification as well as curve type classifications results achieved by *ScolioNet* revealed comparable or superior sensitivity and NPV (Table 2-4) in comparison with the classification done by spine surgeons visually assessing the nude back of AIS patient. These two

parameters are desired because clinically our platform needs to be able to promptly detect any disease progression to trigger early interventions, but at the same time to avoid unnecessary interventions.

Our open platform has benefits of lower risk and low-cost easy access. It can effectively assess the severity of the AIS patients and accurately classify the types of scoliosis. This will contribute to further treatment planning for the patient, by providing computer-aided non-contact and real-time assessments for doctors to further diagnose during the difficult time of COVID-19. In the future the open platform can further continuously benefit spine specialists and patients internationally as it can provide them with fully automated, fast, unbiased, and comprehensive analysis of spine malalignment.

Radiographic and back imaging appearance had visual feature associations (Fig. 1b&c) in both severity and curve type. Increased severity leads to increased body deformity and increased necessity for clinical actions including non-surgical treatment (moderate cases) and surgical interventions (severe cases). The severity and curve types of AIS were not evenly distributed in the two cohorts (Table 1, Fig. 1e) with significantly increased numbers of moderate cases and reduced severe cases as well as increased numbers of combined curves. We initially attempted different methods in balancing the dataset[24,25] during the model development. While we adjusted for the data imbalance and during the technical validation the performance improved somewhat, the problem of the main proportion of positive class samples still cannot be completely eliminated.

The results of both clinical intervention classification and type of curve classification are comparable. The two binary classification for whether FU or no intervention is needed, and whether CS is required achieved AUC of 0.839 and 0.902, whereas for T, TL/L and mixed type classification achieved AUC of 0.777, 0.760, and 0.860 respectively. Since in the multi-task model, we prefer to consider the spine specialist-level classification of the scoliosis task as the main task and set a smaller weight for the type of scoliosis problem loss. We can find in comparison that the binary classification of ‘T’ performs better than ‘TL/L’. T curve prediction had a better ROC curve and the higher AUC. Therefore, T curve is easier to distinguish than TL/ L for the model.

The performance of *ScolioNet* on the independent cohort 2 was reduced compared to the performance during our internal validation (Table 2-4 vs. supplementary Table 1). We analyzed the independent testing dataset and discovered the severity and curve type distributions were different from the training and validation dataset, due to the differences at the time of data acquisition, camera position and ambient lighting. Thus the prospective dataset might introduce domain shift problems[26-28] to the previously trained model. Furthermore, to test the robustness of *ScolioNet*, we used the collected raw prospective data. Unlike the previous study[11], we did not resample the independent cohort to match the data distribution of the trained dataset. Thus, this independent prospective validation trial demonstrates the adequacy of *ScolioNet* for the triple classification task of scoliosis. However, the prospective study was performed in our local healthcare system, an external multicenter trial to further evaluate the robustness of *ScolioNet* would be desirable.

Outlook

We hope to continuously support the open platform to provide free access for researchers and clinicians handling AIS patients. This special patient population requires regular FUs with minimal radiation and carefully assessment to avoid delay in non-surgical management as well as avoid unnecessary surgical interventions. To avoid errors, AlignProCARE[19] allows researchers and clinicians to modify and confirm *ScolioNet* results to assure the reliability of the severity and type assessment. The modified results will be anonymized and sent back to our institute to improve the existing *ScolioNet*'s performance and to better facilitate research and clinical work in the future.

Methods

Definitions of AIS severity and curve type. The severity of AIS was defined by the degrees of the CA measurement on the coronal radiographs following the clinical gold standard and also being the primary consideration for treatment planning[29]. To measure the CA, the end vertebrae (the most tilted vertebrae from the horizontal apical vertebra) need to be identified at the upper and lower ends of the curve, and the angle formed by lines drawn at the superior and inferior endplates of the upper and lower end vertebrae is measured as the CA (Fig. 1b). AIS contains 3 different severity classes (normal/mild: CA $\leq$ 20°; moderate: 20°<CA $\leq$ 40°; and severe CA > 40°)[30] with accurate CA degrees, end vertebra and apical vertebra labels according to corresponding X-ray images.

The curve types were subsequently decided from the GT labels by the location of the apical vertebra. The majority of population have 12 thoracic vertebrae and 5 lumbar vertebrae. For patients with a single curve (Fig. 1b&c), if the apex locates between the 1st and the 11th thoracic vertebrae it is considered as the thoracic curve (T), whereas if the apex locates between the 12th thoracic and the 5th lumbar vertebrae it is considered as the thoracolumbar/lumbar curve (TL/L)[31]. For patients with more than one curves it is typed as mixed curve (Fig. 1b&c).

Dataset collection and preparation. The collection, use, analyses and prospectively testing the X-rays images with paired back images taken by smartphones were approved by the local ethics committee (UW19-620). All males and females with AIS were recruited. Written consents were completed by the guardian of the participants prior to data collection. Participants were excluded if they were not within the age range of 10–18; diagnosed with or have any signs of psychological disorders that might influence the compliance of the study; have any pre-diagnosed systematic neural disorders that might influence the mobility of the patients (e.g. prior cerebrovascular accident, Parkinson's disease, myopathy); have musculoskeletal diseases including congenital spinal deformities, McCune-Albright syndrome, early-onset scoliosis, previous spine operations and instrumentation performed, trauma that might impair posture and mobility; and/or had severe skin disorders and/or lesions at the back.

All back images were taken voluntarily by AIS patients using smartphones (iPhone X, iPhone 12, Redmi Note 8 Pro). The corresponding radiographs were anonymously archived from the hospital picture archiving and communication system in PNG format, which were taken in the clinic as a routine practice with no extra experimental radiographs taken. Using the gold standard clinical classifications, the ground truth (GT) labels of the cohort 1 and 2 were provided by the same spine surgeon (with over 20 years' experience in AIS management) by annotating the coronal radiographs manually.

The severities and curve types of cohort 2 were also blindly assessed by two spine specialists using visual assessment of the back of the participants. The severity classification, curve types were recorded individually for the two assessors.

Data preprocessing. The back images of cohort 1 were used to develop the technology of *ScolioNet* for classifying scoliosis severity and curve types. Specifically, we randomly divided cohort 1 by 8:2, corresponding to training set and in-house validation set. The in-house validation set was used to evaluate and select different deep learning models. Independent testing cohort 2 were recruited to prospectively test the performance of *ScolioNet*. Back segmentations were firstly performed on the images taken by smartphones. We empirically selected to segment the back with arms to achieve improved classification performance. Data augmentation methods[32–34] included random rotate (-10°~10°), affine transform, crop and pad, gaussian blur, sharpen, change contrast and brightness was introduced. Each method was set to appear with a 50% probability and combined them for each image of the cohort 1.

Performance evaluation. To evaluate the performance of *ScolioNet*, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values were calculated[35,36]. Evaluation metrics included sensitivity (Sn), specificity (Sp), positive predicted value (PPV), negative predicted value (NPV), accuracy (ACC).

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

The ROC curve was plotted based on the final Sn scores and (1 – Sp scores) with different thresholds, and the area under the curve (AUC) was computed from the area of ROC curve which denotes as:

$$AUC = \int_{x=0}^{1} TPR\left(FPR^{-1}\left(x\right)\right) dx = P(X_1 > X_0)$$

Where X_1 and X_0 are the scores for a positive and a negative instance, respectively; TPR represents true positive rate and equals to Sn; FPR represents false positive rate and equals to (1 – Sp).

The Multi-layer CNNs and Model Selection. Several deep convolutional network benchmarks were compared using the data from cohort 1 (Table 1). In our CNN framework, there were 1 input layer, 84 hidden layers, and 1 output layer. In the input layer, the input image was processed by convolution, batch normalization (BN) and pooling operations, to reduce the interference caused by the difference in the range of values of the input data in each dimension.

In the hidden layers, neurons were in each layer were connected and propagated containing internal feature coding and computational outcome. The convolutional layers were used for feature extraction and presentation, and a commonly used rectified linear unit (ReLU) function was selected to active the outcome of a neuron and defined as follows:

$$y = \max(0, x)$$

where *x* was the weighted sum of a neuron and *y* was the output of the activation function.

In the output layer, the weights of all neurons output from the hidden layer were fully connected to obtain the required probability values for classification. Since the final output of the classification task was essentially the probability value of the image input in each category, a softmax function was generally introduced at the end in order to map the output of the fully connected layer to the (0,1) range with a summation value of 1, which is calculated as follows:

$$S_i = \frac{e^{V_i}}{\sum_j e^{V_j}}$$

where V was the input of the softmax layer derived from the fully connected layer and S was the final output. In the models for back image-based severity and type prediction, S_i was a value in the range 0–1 representing the probability of a back photo classified as normal-mild, moderate, or severe with T, TL/L, or mixed type.

For the experiment on model selection, we trained five models, Resnet50[21], Densenet169[37], VGG16[38], InceptionV4[39] and ResNeXt50[40], of which results are shown in Supplementary Fig. 2 and Table 1, where the model with Resnet50 and Densenet169 as backbones performed relatively better. Our initial selection of models was based on the AUC metric to judge the comprehensive performance of the models, and Resnet50 and Densenet169 performed best on the macro-average AUC metric for the severity classification task (Supplementary Table 1). Although, InceptionV4 was comparable to Resnet50 and Desnsenet169 in many metrics in both FU and CS statistics (Supplementary Table 1), we finally chose

Resnet50 and Densenet169 as the backbone considering the size of the model. To further improve performance, we added multi-tasking strategy and attention.

The Attention Algorithms and Multi-tasking Strategy. The human visual system tends to focus on a certain part of the image that is useful for judgment and ignore other unimportant areas. The attention mechanism[41,42] is a method that allows a network to mimic the human visual system and tend to pay attention to a certain part of the image when performing a classification task. Multi-task strategy[43–45] is designed to improve the model performance by learning multiple tasks in parallel so that the results can interact with each other, either by sharing the weights of feature extraction on the network or by interacting only on some key layers, and finally using classifiers to classify each task.

Different attention algorithms were tested, including SE block[46] and residual attention block[22]. Empirically, residual attention block with the activation function for channel and spatial mixed attention performed the best. The activation function is as follows:

$$f\left(x_{i,c}\right) = \frac{1}{1 + e^{-x_{i,c}}}$$

where *x* was the weighted sum of a neuron, *i* was position of the feature map, *c* was channel of the feature map.

Previous studies reported[22] the residual attention block is an improvement on the residual net, thus we add the attention module to both Resnet50 and Densenet169 for further testing. Resnet50 and Densnet169 consist of 4 residual block and 4 dense block respectively, so we added a total of 4 attention blocks after the input layer and after each submodule respectively. With attention introduced, we empirically discovered that spatial and channel mixed attention improved the performance of the classification task. The model with Resnet50 as backbone with attention module performed better and then was selected to be our final classification model ScolioNet.

Multi-task strategy was tested in addition to the attention to try to improve the generalization ability of the model. We did not change the design of the feature extraction part of the network, shared the feature extraction weights for the severity and type of scoliosis tasks, and only used 3 parallel classifiers for each task at the end. Since our main task is the severity and type classification task, we set the loss to the sum of the three-classification problem loss and set the weight of the type of scoliosis problem to 0.5.

After the selection, our ScolioNet consisted of 1 input layer, 84 hidden layers and 1 output layer, where the hidden layers consisted of 4 residual blocks and 3 attention modules. In the output layer, we used three parallel FC layers to implement multi-task method to finish the classification of the severity, while for classification of type of curve, we only used one FC layer.

Model development and parameter optimization. 1429 images were used for model training and during training, we resized each image to 3×224×224 pixels in order to make it compatible with the original

dimensions of the network architecture. We trained triple classification models for the severity classifications directly and binary classification models for the curve type classification. Mixed curve type was classified if both curve type were presented on one image. ROC curves with each class as a positive class was generated to calculate the AUC.

During model development our validation set consisted of 351 images and the independent test set consisted of 378 images. We performed data augmentation on the in-house validation sets to be proportionally matching with the training dataset. The parameter optimization was stopped when the performance AUC was not improving. The validation set was used for model selection during training, while the independent test set was used to further test the accuracy and robustness of the selected model. The in-house dataset was not used for training, but for the in-house performance testing of the models, before an independent testing study. The prediction of each class performed well with an AUC of around 0.85. All proposed models performed similarly and the ROC curve of *ScolioNet* had the most balanced performance on each classification task (Supplementary Fig. 2).

For model training, we used a workstation configured with an Intel(R) Xeon(R) Gold 5218 2.30GHz central processing unit, 308 GB of RAM, a NVIDIA GeForce GTX 300 core and PyTorch (https://pytorch.org/). Models were trained by minimizing the general cross-entropy loss function between predictions and GT labels,

$$Loss = -\left[y\log\hat{y} + (1-y)\log(1-\hat{y})\right]$$

Where y was the GT and $\hat{y}$ was the prediction.

During training SGD optimizer in PyTorch was adopted and a decay factor γ was used to control the learning rate at each epoch as described below:

$$lr_{new} = lr_{initial} * \gamma^{epoch//stepsize}$$

Where $lr_{initial}$ was the initial learning rate, lr_{new} was the learning rate at i^{th} epoch and step size was the learning rate decay step. Adjustable parameters such as initial rate, decay and batch size were simultaneously adjusted to improve the model performance. Our initial learning rate of 0.01, the momentum of 0.9, weight decay of 0.001, γ of 0.1 and step size of 20, batch size 16 and epoch of 50.

ScolioNet **Model prospective testing and explicability.** The independent testing dataset, 378 images, is a prospectively collection not seen during the model development and in-house testing. No data augmentation or resampling for this part was done to ensure a true validity of the prospective experiments.

With the continuous development of deep learning, the explicability of deep learning models has attracted more and more attention. In order to improve the interpretability of the model and mine the decision logic of the model, Class Activation Mapping (CAM) method[47] was proposed, and some interpretable algorithms based on CAM[48,49] were proposed afterwards. For the explicability of *ScolioNet*, we use the

Score-CAM[50] algorithm to explain the decision of the model. Score-CAM makes use of all the feature maps output by the last convolutional layer to obtain interpretable information and gives the decision-supporting regions in the original image in the form of interpretable heatmap. Features maps from the last convolutional layer contain high-dimensional features and spatial information, and are also the direct input for classification, so they play an important role in explaining the model predictions. Score-CAM overlays each feature map on the original image as an interpretable mask and inputs the masked image into the model to obtain its probability in the predicted label, and the predicted label. The difference between the probability of the masked image and that of the original image in the predicted label is used as the weight of the corresponding feature map. We normalize and up-sample all the feature maps, and then linearly superimpose the feature maps according to their weights to get an interpretable heatmap. Warm color in the heatmap represent the importance of the region in supporting the model to make the classification decision.

Statistical analysis. The Wilcoxon signed-rank test was performed using the stats.wilcoxon() function in SciPy for assess the interrater agreement of the two spine surgeons and *ScolioNet* on using back appearance classifying the scoliosis severity and curve type. The same practice was performed to compare the agreements between different deep learning models (Supplementary Table 2), and $p < 0.0001$ was considered statistically significant.

Declarations

Data availability

All datasets for this study were anonymized to protect patient information. The prospective testing data including the back photo and the paired radiographs are available for research purposes from the corresponding author upon reasonable request via institutional emails.

Code availability

The deep-learning models were developed using standard libraries and the trained model is available to use via our lab website (https://www.aimed.hku.hk/alignprocare). Accesses of the original code are openly available for research purposes upon reasonable request via institutional emails.

References

1. de Seze, M. & Cugy, E. Pathogenesis of idiopathic scoliosis: a review. Annals of physical and rehabilitation medicine **55**, 128–138, doi:10.1016/j.rehab.2012.01.003 (2012).
2. Fong, D. Y. *et al.* A population-based cohort study of 394,401 children followed for 10 years exhibits sustained effectiveness of scoliosis screening. The spine journal: official journal of the North American Spine Society **15**, 825–833, doi:10.1016/j.spinee.2015.01.019 (2015).

3. Luk, K. D. *et al.* Clinical effectiveness of school screening for adolescent idiopathic scoliosis: a large population-based retrospective cohort study. Spine **35**, 1607–1614, doi:10.1097/BRS.0b013e3181c7cb8c (2010).
4. Wong, J. S. H. & Cheung, K. M. C. Impact of COVID-19 on Orthopaedic and Trauma Service: An Epidemiological Study. The Journal of bone and joint surgery. American volume **102**, e80, doi:10.2106/JBJS.20.00775 (2020).
5. Brinjikji, W. *et al.* Systematic Literature Review of Imaging Features of Spinal Degeneration in Asymptomatic Populations. AJNR. Am J Neuroradiol **36**, 811–816, doi:10.3174/ajnr.A4173 (2015).
6. Del Grande, F., Maus, T. P. & Carrino, J. A. Imaging the intervertebral disk: age-related changes, herniations, and radicular pain. Radiologic clinics of North America **50**, 629–649, doi:10.1016/j.rcl.2012.04.014 (2012).
7. Levy, A. R., Goldberg, M. S., Hanley, J. A., Mayo, N. E. & Poitras, B. Projecting the lifetime risk of cancer from exposure to diagnostic ionizing radiation for adolescent idiopathic scoliosis. Health Phys **66**, 621–633, doi:10.1097/00004032-199406000-00002 (1994).
8. Doody, M. M. *et al.* Breast cancer mortality after diagnostic radiography: findings from the U.S. Scoliosis Cohort Study. Spine **25**, 2052–2063, doi:10.1097/00007632-200008150-00009 (2000).
9. Bone, C. M. & Hsieh, G. H. The risk of carcinogenesis from radiographs to pediatric orthopaedic patients. J Pediatr Orthop **20**, 251–254 (2000).
10. Goldberg, M. S., Mayo, N. E., Levy, A. R., Scott, S. C. & Poitras, B. Adverse reproductive outcomes among women exposed to low levels of ionizing radiation from diagnostic radiography for adolescent idiopathic scoliosis. Epidemiology **9**, 271–278 (1998).
11. Yang, J. L. *et al.* Development and validation of deep learning algorithms for scoliosis screening using back images. *Commun Biol* **2**, doi:ARTN 390 10.1038/s42003-019-0635-8 (2019).
12. Zeleznik, R. *et al.* Deep convolutional neural networks to predict cardiovascular risk from computed tomography. Nat Commun **12**, 715, doi:10.1038/s41467-021-20966-2 (2021).
13. Zhu, X. *et al.* Hybrid AI-assistive diagnostic model permits rapid TBS classification of cervical liquid-based thin-layer cell smears. Nat Commun **12**, 3541, doi:10.1038/s41467-021-23913-3 (2021).
14. Chen, C. L. *et al.* An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. Nature Communications **12** (2021).
15. Zhang, T., Li, Y., Cheung, J. P. Y., Dokos, S. & Wong, K. Y.-K. Learning-based coronal spine alignment prediction using smartphone-acquired scoliosis radiograph images. IEEE Access **9**, 38287–38295, doi:10.1109/ACCESS.2021.3061090 (2021).
16. Chen, B. *et al.* An Automated and Accurate Spine Curve Analysis System. Ieee Access **7**, 124596–124605, doi:10.1109/Access.2019.2938402 (2019).
17. Horng, M. H., Kuok, C. P., Fu, M. J., Lin, C. J. & Sun, Y. N. Cobb Angle Measurement of Spine from X-Ray Images Using Convolutional Neural Network. *Comput Math Methods Med* 2019, 6357171, doi:10.1155/2019/6357171 (2019).

18. Choi, R. *et al.* Cnn-based spine and cobb angle estimator using moire images. IIEEJ transactions on image electronics and visual computing **5**, 135–144 (2017).
19. Meng, N. *et al.* An artificial intelligence powered platform for auto-analyses of spine alignment irrespective of image quality with prospective validation. EClinicalMedicine **43**, 101252, doi:10.1016/j.eclinm.2021.101252 (2022).
20. Yoon, J. W. *et al.* Remote Virtual Spinal Evaluation in the Era of COVID-19. Int J Spine Surg **14**, 433–440, doi:10.14444/7057 (2020).
21. He, K., Zhang, X., Ren, S. & Sun, J. in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 770–778.
22. Wang, F. *et al.* in *IEEE conference on computer vision and pattern recognition.* 3156–3164.
23. Shah, D. J., Sachs, R. K. & Wilson, D. J. Radiation-induced cancer: a modern view. Br J Radiol **85**, e1166-1173, doi:10.1259/bjr/25026140 (2012).
24. Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw **106**, 249–259, doi:10.1016/j.neunet.2018.07.011 (2018).
25. Thabtah, F., Hammoud, S., Kamalov, F. & Gonsalves, A. Data imbalance in classification: Experimental evaluation. Inform Sciences **513**, 429–441 (2020).
26. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodriguez, R., Chawla, N. V. & Herrera, F. A unifying view on dataset shift in classification. Pattern Recogn **45**, 521–530 (2012).
27. Yan, W. *et al.* in *Medical Image Computing and Computer Assisted Intervention (MICCAI).* 623–631.
28. Ovadia, Y. *et al.* in *33rd Conference on Neural Information Processing Systems (NeurIPS* 2019). 14003–14014.
29. Nachemson, A. L. *et al.* Effectiveness of Treatment with a Brace in Girls Who Have Adolescent Idiopathic Scoliosis - a Prospective, Controlled-Study Based on Data from the Brace Study of the Scoliosis-Research-Society. *The Journal of bone and joint surgery. American volume* **77a**, 815–822, doi:10.2106/00004623-199506000-00001 (1995).
30. Chung, N. *et al.* Spinal phantom comparability study of Cobb angle measurement of scoliosis using digital radiographic imaging. J Orthop Translat **15**, 81–90, doi:10.1016/j.jot.2018.09.005 (2018).
31. Eyvazov, K., Samartzis, D. & Cheung, J. P. The association of lumbar curve magnitude and spinal range of motion in adolescent idiopathic scoliosis: a cross-sectional study. BMC Musculoskelet Disord **18**, 51, doi:10.1186/s12891-017-1423-6 (2017).
32. Chlap, P. *et al.* A review of medical image data augmentation techniques for deep learning applications. J Med Imaging Radiat Oncol **65**, 545–563, doi:10.1111/1754-9485.13261 (2021).
33. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. J Big Data-Ger **6** (2019).
34. Zhang, K., Liang, J., Van Gool, L. & Timofte, R. in *International Conference on Computer Vision.*

35. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**, 427–437 (2009).
36. Wong, T. T. Linear Approximation of F-Measure for the Performance Evaluation of Classification Algorithms on Imbalanced Data Sets. Ieee T Knowl Data En **34**, 753–763 (2022).
37. Huang, G., Liu, Z., Van Der, M. & Lilian, Q. W. in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).* 4700–4708.
38. Simonyan, K. & Zisserman, A. (arXiv:1409.1556, 2014).
39. Szegedy, C., Loffe, S., Vanhoucke, V. & Alemi, A. A. in *Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17).*
40. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. in *IEEE conference on computer vision and pattern recognition.* 1492–1500.
41. Niu, Z. Y., Zhong, G. Q. & Yu, H. A review on the attention mechanism of deep learning. Neurocomputing **452**, 48–62 (2021).
42. Fukui, H., Hirakawa, T., Yamashita, T. & Fujiyoshi, H. in *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10705–10714.
43. Xie, Z. L., Chen, J. L., Feng, Y., Zhang, K. Y. & Zhou, Z. T. End to end multi-task learning with attention for multi-objective fault diagnosis under small sample. J Manuf Syst **62**, 301–316 (2022).
44. Liu, S., Johns, E. & Davison, A. J. in *IEEE/CVF conference on computer vision and pattern recognition.* 1871–1880.
45. Zhao, S., Liu, T., Zhao, S. & Wang, F. in *AAAI Conference on Artificial Intelligence.* 817–824.
46. Hu, J., Shen, L. & Sun, G. in *IEEE conference on computer vision and pattern recognition.* 7132–7141.
47. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2921–2929.
48. Desai, S. & Ramaswamy, H. G. in *IEEE/CVF Winter Conference on Applications of Computer Vision.* 983–991.
49. Jalwana, M. A. A. K., Akhtar, N., Bennamoun, M. & Mian, A. in *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 16327–16336.
50. Wang, H. *et al.* in *IEEE/CVF conference on computer vision and pattern recognition workshops.* 24–25.

Tables

Table 1 | The data characteristics of cohort 1 and cohort 2

Type	Cohort 1		Cohort 2		Clinical implications
Age (10-18)	No. of patients	No. of X-rays	No. of patients	No. of X-rays	
Curve severity and magnitudes					
Normal-mild (CA≤20°)	555	1104	97	191	No intervention required for these mild curves. 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°)	1055	2109	195	390	These patients may require bracing to prevent curve progression if still skeletally immature. No intervention is required at the end of growth. Scoliosis-specific exercises may also be prescribed.
Severe (CA>40°)	170	337	86	172	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 (single curve)	118	374	38	76	Curves that develop rib humps and are more likely to develop chest wall deformities and unleveled shoulders.
Thoracolumbar /Lumbar curve (single curve)	385	767	65	129	Curves more likely to develop pelvic obliquity and waistline deformities.
Mixed curve (more than one)	1153	2303	262	524	Curves that are often more balanced. Unequal curve sizes may lead to more deformities as listed for thoracic major curves or TL/L curves.

Table 2 | Comparison of the performance evaluation metrics between *ScolioNet* and surgeons on the independent dataset in distinguishing curves requiring no interventions (<20°) or follow-up (FU) using single back photographs

Method	Sen(%)	NPV(%)	Spe(%)	PPV(%)	ACC(%)	AUC
ScolioNet	84.88[75.54, 91.70]	89.22[84.25, 93.70]	67.44[59.89, 74.38]	56.59[50.81, 62.20]	73.26[67.41, 78.56]	0.839[0.789, 0.882]
Senior Surgeon	44.19	71.76	70.93	43.18	62.02	——
Junior Surgeon	62.79	79.35	71.51	52.43	68.60	——

Table 3 | Comparison of the performance evaluation metrics between *ScolioNet* and surgeons on the independent dataset in distinguishing severe curves that considering surgical interventions (>40°; CS) using single back photographs

Method	Sen(%)	NPV(%)	Spe(%)	PPV(%)	ACC(%)	AUC
ScolioNet	82.56[72.87, 89.90]	90.00[84.95, 93.48]	78.49[71.59, 84.38]	65.74[58.67, 72.18]	79.84[74.42, 84.57]	0.902[0.859, 0.936]
Senior Surgeon	20.93	70.43	94.19	64.29	69.77	——
Junior Surgeon	19.76	70.51	95.93	70.83	70.54	——

Table 4 | Comparison of the performance evaluation metrics between *ScolioNet* and surgeons on the independent dataset in distinguishing curve types using single back photographs

Method		Sen(%)	NPV(%)	Spe(%)	PPV(%)	ACC(%)	AUC
ScolioNet	T	82.31[78.51, 85.70]	61.97[56.45, 67.20]	54.10[47.62, 60.47]	77.10[74.48, 79.52]	72.51[69.04, 75.78]	0.777[0.745, 0.808]
	L	81.18[77.29, 84.66]	62.11[56.85, 67.11]	57.55[51.10, 63.83]	78.11[75.39, 80.60]	72.93[69.48, 76.19]	0.760[0.727, 0.791]
	Mixed	87.32	92.52	68.30	54.55	74.07	0.860[0.834, 0.887]
Senior Surgeon	T	76.64	58.04	60.66	78.52	71.08	——
	L	75.49	55.56	57.14	76.67	69.09	——
	Mixed	41.31	75.35	78.12	45.13	66.95	——
Junior Surgeon	T	100.00	N/A	0	65.24	65.24	——
	L	100.00	N/A	0	65.10	65.10	——
	Mixed	100.00	N/A	0	30.34	30.34	——

Figures

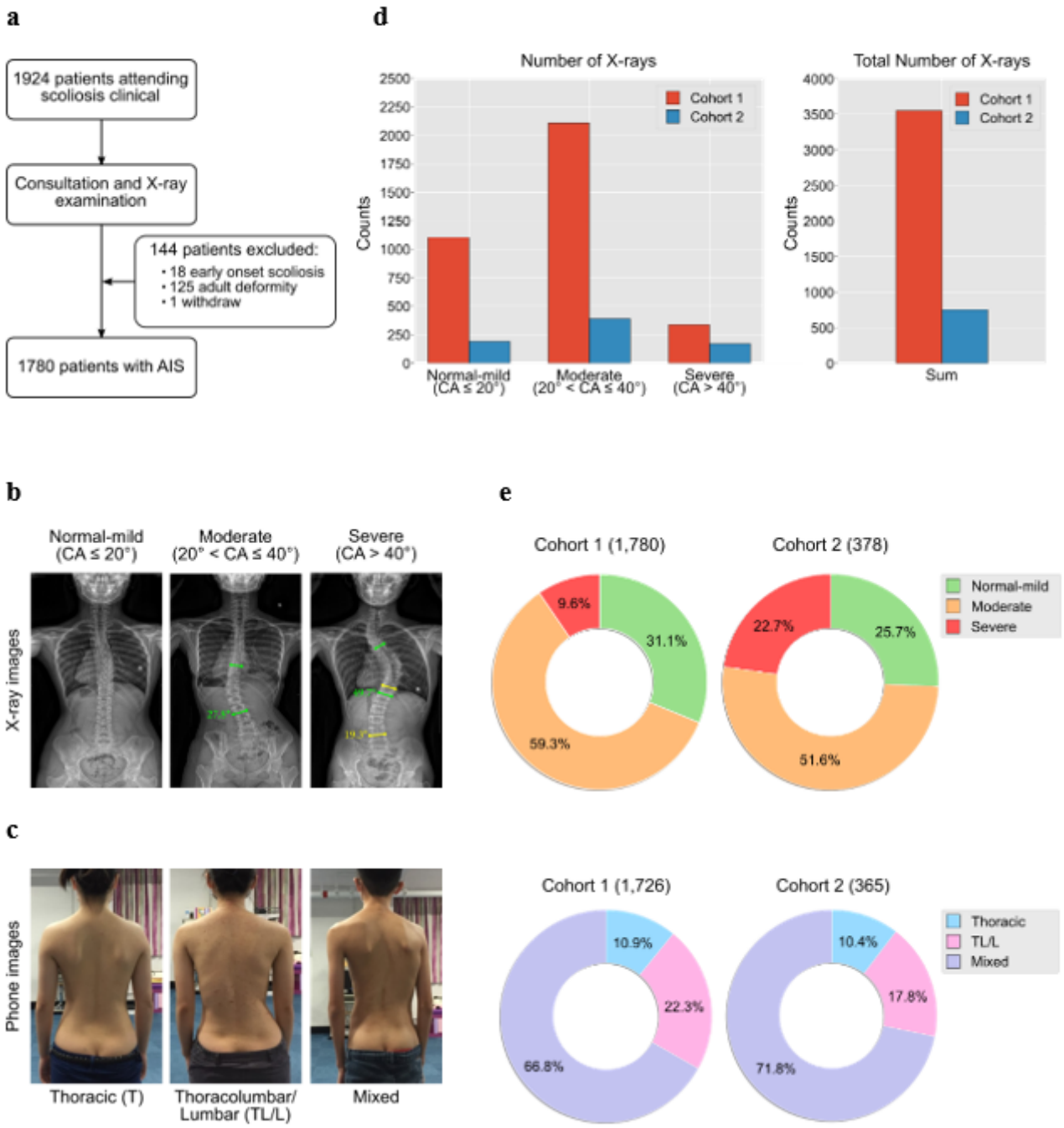

Figure 1

The study data features. a. The flow chart for the patient recruitment process of the cohort 1, and the data were used to develop *ScolioNet*. **b.** Classifications for AIS severity (normal-mild, moderate and severe) and curve types (T, TL/L, and mixed) on radiographs. **c.** The examples for the corresponding bareback images. **d.** Number of the X-ray images with confirmed normal-mild, moderate and severe classifications, as well as the number of X-ray images with confirmed T, TL/L, and mixed curve types in cohort 1 and cohort 2. **e.** Number of the radiographs taken for each severity class.

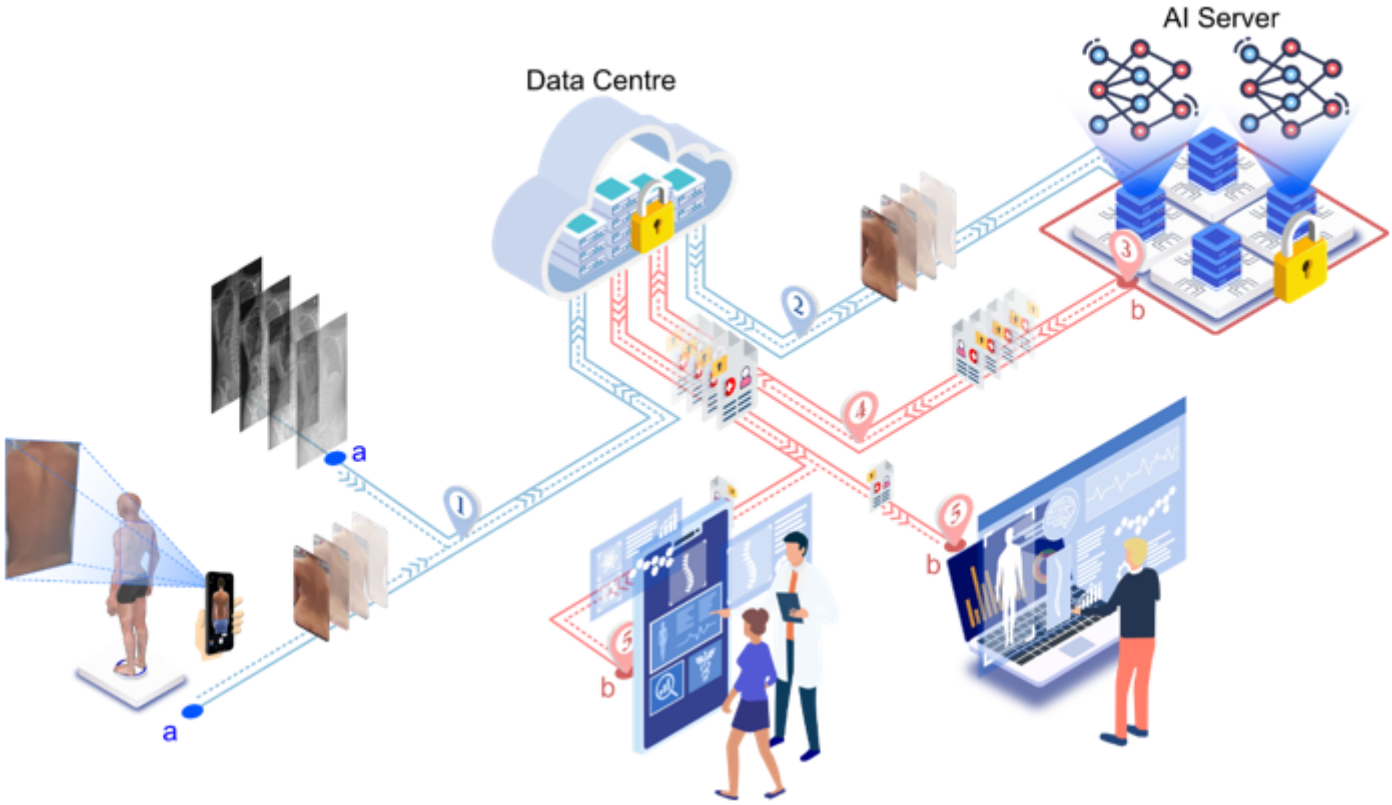

Figure 2

Overview of the data and the proposed deep learning diagnosing flow. a, Collection of the dataset. The training dataset was composed of the captured phone images from the involved patients and securely stored in our data center. To obtain the deformity severity for each patient, radiographs were also scanned, and further analyzed by the specialists. The diagnosis results, including the degrees of the Cobb angles and the curve types, were used as the ground truth labels for the collected phone images. **b,** Online AIS checking flow. The proposed multi-task attention-based CNN model (*ScolioNet*) was deployed on the backend AI server after training for further independent testing. Practically, the AIS checking could be conducted in steps: 1) back photographs are captured and transferred to the data center together with X-ray images; 2) the encrypted and deidentified back photographs are send to the AI server for diagnosis via the network; 3) the diagnosing server performs the classification of AIS severity and curve types; 4) the diagnosing results are returned to the patients through the network; 5) the analytic results are visualized at the client end equipment, such as the smartphone and laptop computer.

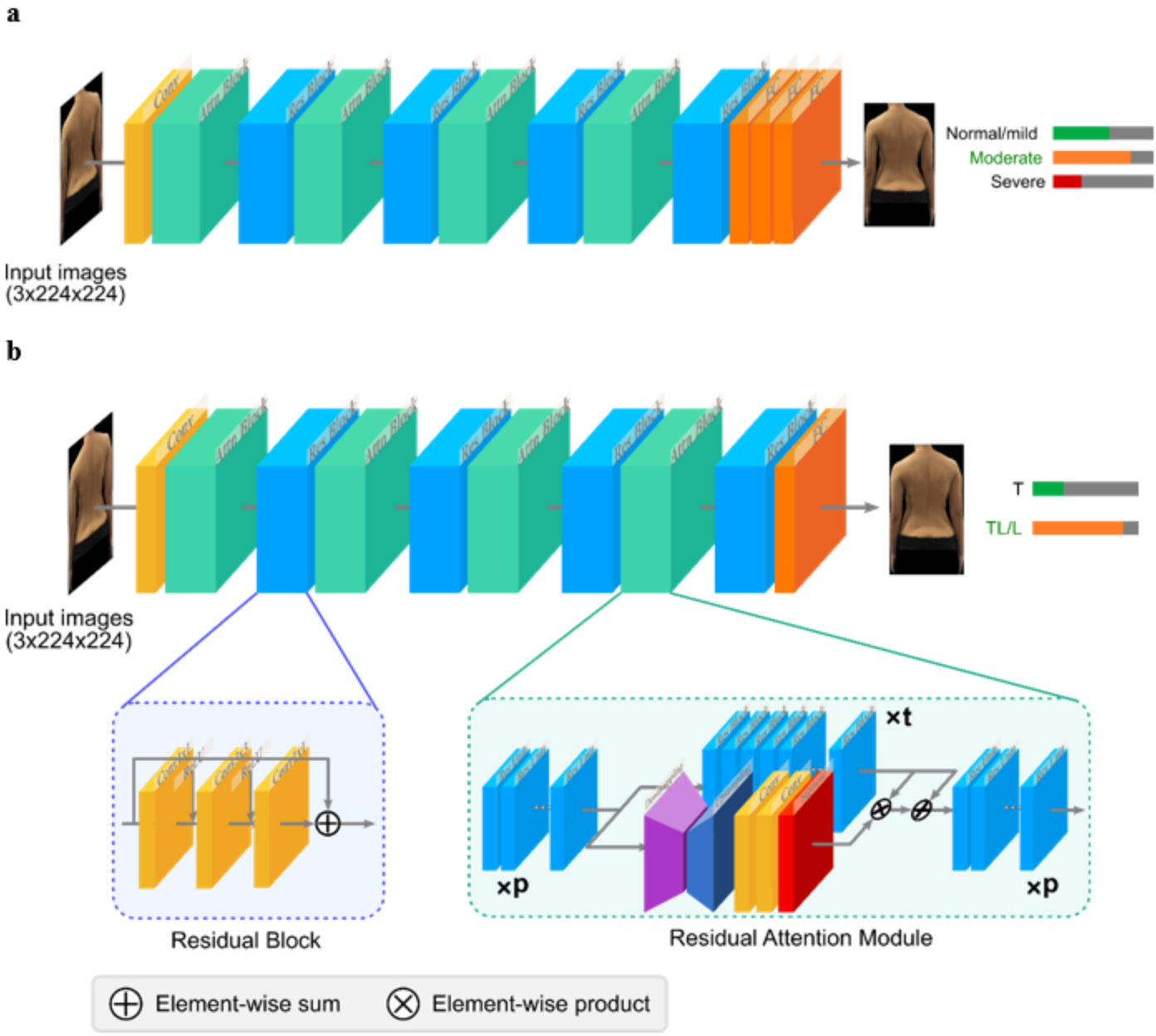

Figure 3

The detailed architecture of *ScolioNet* for severity grading (a) and curve type classification (b). The input size is resized to 3×224×224 for both a and b. To develop *ScolioNet*, Resnet50 with residual attention module is used as feature extractor. Residual blocks and attention blocks are alternately connected in series. Three parallel fully connected layers are used as classifiers at last. The residual block consists of several residual units. And the residual unit is just a sample, the number of convolutional layers and parameters vary for different blocks of units. One residual attention block consists of two branches called trunk branch and mask branch. The number of residual units in a residual attention block is controlled by the parameters '**p**' and '**t**'. Different attention stages have different depth of layers. The output of a is the severity of deformity and of b is the type of the curve.

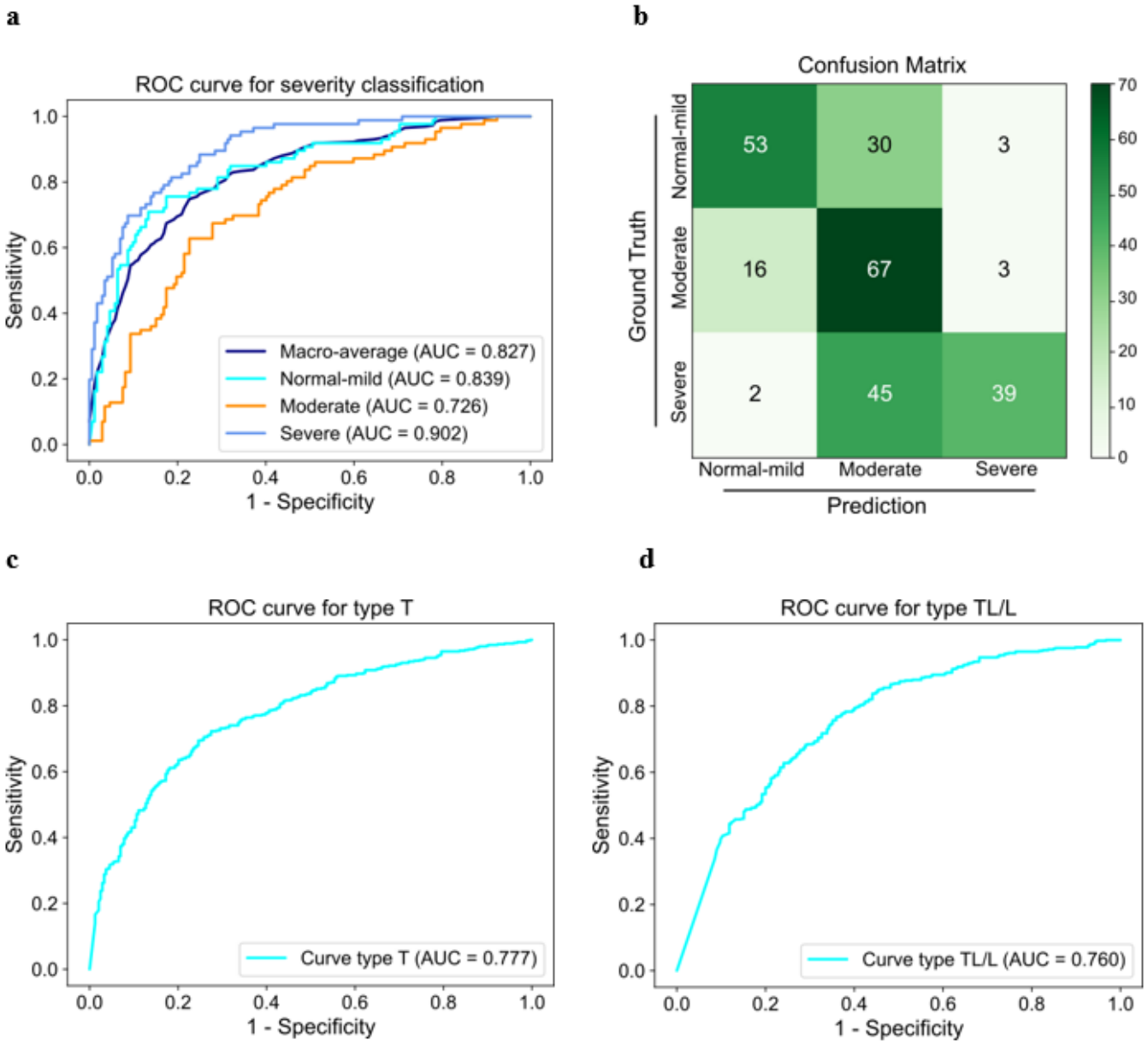

Figure 4

ROC curve and confusion matrix on independent test. a, The ROC curve of *ScolioNet* on severity classification task. The curves of the 3 different dichotomous results are represented by different colors, and the macro-average is taken to obtain the trichotomous average roc curve. b, The confusion matrix correspond to our final severity classification results. c, The ROC curve on curve type classification task, which predicts if the patient has a thoracic (T) curve. d, The ROC curve on curve type classification task, which predicts if the patient has a thoracolumbar/lumbar (TL/L) curve.

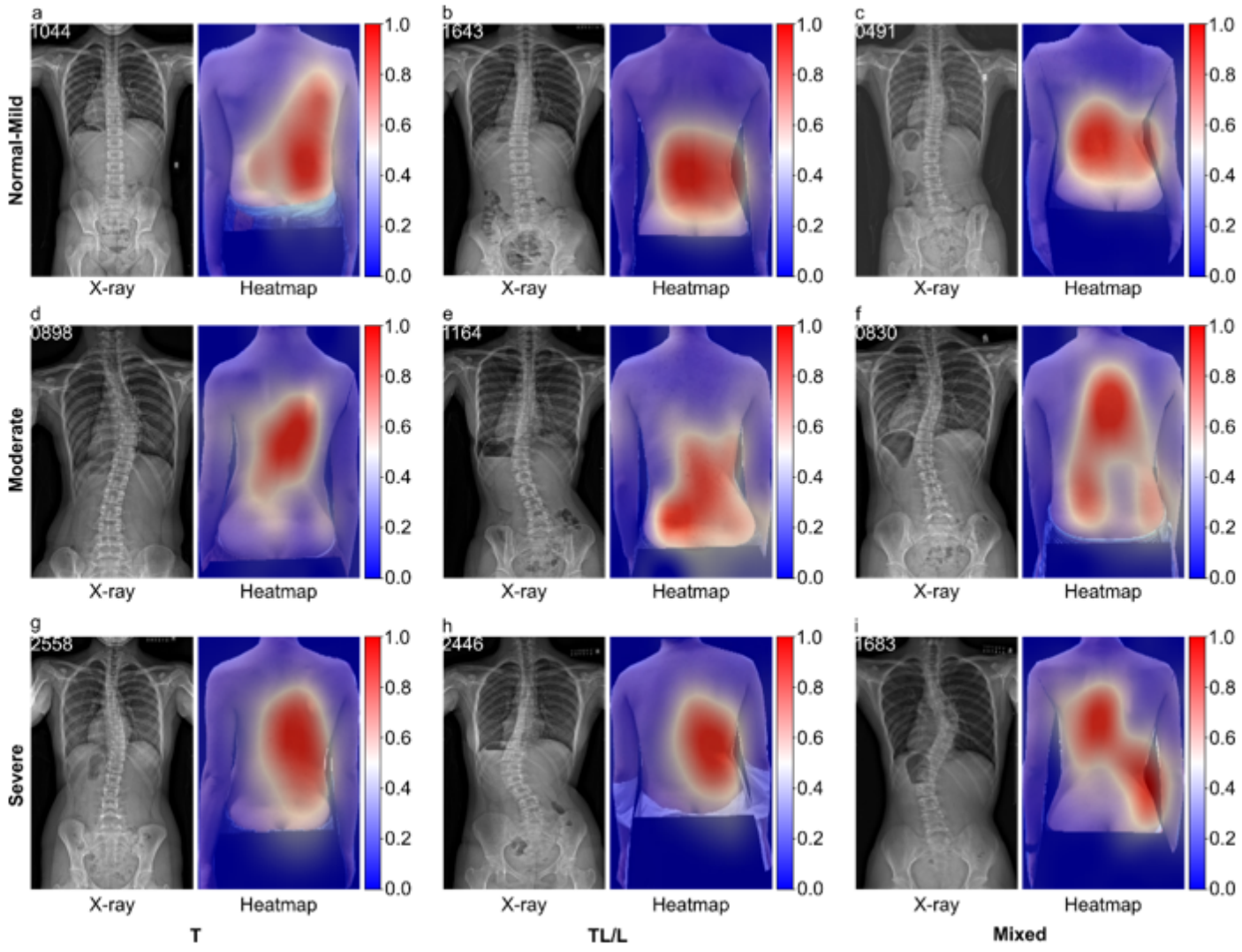

Figure 5

The interpretable heatmaps. The interpretable heatmap reflects the areas in the image that are used to support classification decisions. In the heatmap, different colors are used to describe the degree of support for classification decisions in different regions in the image. From blue to red, the degree of support for classification decisions is increasing. Patient examples with different severities and curve types of spinal deformity were visualized based on the thermographic decision area.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.docx