Appendix 1 Implementation of the 6-step annotation method

Our 6-step annotation method is a simple and feasible method based on ITK-SNAP software to help radiologists locate SIJs on MRI more quickly and efficiently. As shown in *Figure S1*, the 6-step annotation method only requires six clicks on the screen to obtain the RVOI of SIJs. The detailed implementation steps are as follows:

(I) Import MR images into the ITK-SNAP software;
(II) Scroll the mouse wheel to locate the slice containing the largest SIJ area (S-L), and click the "Polygon Inspector" button in the Main Toolbar to start the annotation;
(III) Confirm the upper-left position of SIJs on S-L and annotate the first point (P1);
(IV) Confirm the upper-right position of SIJs on S-L and annotate the second point (P2). A straight line will be automatically generated between P1 and P2;
(V) Confirm the lower-right position of SIJs on S-L and annotate the third point (P3). A straight line will be automatically generated between P2 and P3;
(VI) Confirm the lower-left position of SIJs on S-L, annotate the fourth point (P4), and press the Enter key, then a straight line will be automatically generated between P3 and P4. In this way, P1, P2, P3, and P4 form a closed rectangular region of interest (RROI);
(VII) Scroll the mouse wheel to the first slice containing SIJs (S-start), press the Enter key, and click the "paste last polygon" button. Then the RROI obtained in Step 6 will be pasted to S-start;
(VIII) Scroll the mouse wheel to the last slice containing SIJs (S-end), and press the Enter key to paste the RROI to the S-end.
(IX) Click the "update" button, and an RVOI with the RROI as the surface and the distance between S-start and S-end as the thickness will be generated.

Appendix 2 Descriptions of development and analysis of NegSpA-AI

Image preprocessing

All MRIs were normalized to a range of [0,1] by a z-score method. For each MRI sequence, we selected the middle slice in the RVOI with its former and next slices and transformed them into a 3-channel image. Next, the 3-channel image was divided into 2 sub-images of 224×224×3 pixels containing the left-side and right-side SIJ, respectively, as shown in *Figure S3*. These sub-images were used as inputs for deep learning (DL) models.

For sub-images on the training set, common data augmentation methods were first performed, including random image rotation, random horizontal flip, and random pixel shift of image location along the x- and y-directions. Then MixCut was introduced for further data augmentation.

Methodology and implementation of MixCut

MixCut was proposed as a local data augmentation method to improve the generalization of models. MixCut includes three steps of augmentation with scribble-level supervision on 2 training samples randomly selected from a mini-batch to generate a new mixed sample. In the first step, the Puzzle Mix algorithm (43) was performed to maximize the saliency, referred to as increments of scribbles. In the second step, a local linear interpolation, namely, the local Mixup algorithm (42), was performed between the 2 images and their labels. In the third step, the Cutout algorithm (48) was deployed to achieve random decrements of scribbles in the mixed images to generate augmented samples. The details of PuzzleMix, local Mixup and Cutout algorithms are explained as follows.

Considering two d-dimensional training images with labels being (x_i, y_i) and (x_j, y_j), the goal of Puzzle Mix is to maximally utilize the saliency information of each input and generate a new training image (x_{ij}, y_{ij}) to train the model with its original loss function. The combining operation is defined as:

$$x_{ij} = M(x_i, x_j) \quad [1]$$

$$y_{ij} = M(y_i, y_j) \quad [2]$$

$$M(a_i, a_j) = (1-z) \odot \prod_i^T a_i + z \odot \prod_j^T a_j \quad [3]$$

where $M(a_i, a_j)$ is the mixup function on a_i and a_j, $\prod_i^T$ and $\prod_j^T$ represent the transportation matrix of dimension $d \times d$; z denotes a mask in [0,1] of dimension d; $\odot$ refers to the element-wise multiplication. The parameter set $\{\prod_i, \prod_j, z\}$ is aimed to maximize the salience of mixed image, which is computed by:

$$\{\prod_i, \prod_j, z\} = \arg\max_{\prod_i, \prod_j, z} [(1-z) \odot \prod_i^T s(x_i) + z \odot \prod_j^T s(x_j)] \quad [4]$$

where $s(x)$ is the saliency of image x and is computed by taking the l_2 norm of the gradient value.

Code-level details are as follows:

https://dx.doi.org/10.21037/qims-24-729

Algorithm puzzle mix

Input: data x_0, x_1, mask z

1: for t = 1,...,T do

2: uniformly sample a mini-batch of training data $B^{(\tau)}$

3: for $(x_i, y_i), (x_j, y_j) \in B^{(\tau)}$ do

4: calculate saliency of x_i and x_j by taking l_2 norm of the gradient value

5: optimize z^* and $\prod_i^*$ in Equation (3)

6: return: $(1-z^*) \odot \prod_0^{*T} x_0 + z^* \odot \prod_j^{*T} x_1$

7: end for

8: update θ

9: end for

Mixup is introduced as a simple and data-diagnostic data augmentation routine that constructs virtual training examples and their corresponding labels to enlarge the support of the training distribution and introduce minimal computation overhead. Considering two random training images with annotations (x_i, y_i) and (x_j, y_j), Mixup generates a new mixed training sample (x, y) by:

$$x = \lambda x_i + (1-\lambda) x_j \quad [5]$$

$$y = \lambda y_i + (1-\lambda) y_j \quad [6]$$

where $\lambda \in [0,1]$ controls the strength of interpolation between feature-target pairs. Code-level details are as follows:

Algorithm Mixup

Input: training data $\{x_i, y_i \mid i = 1,...,n\}$; iteration number T; the number of images in each batch m; $\alpha \in (0, \infty)$

Output: a more robust model

1: $\lambda \sim Beta(\alpha, \alpha)$

2: for t = 1,...,T do

3: uniformly sample a mini-batch of training data $B^{(\tau)}$

4: for $(x_i, y_i), (x_j, y_j) \in B^{(\tau)}$ do

5: transform y_i, y_j into one-hot vectors as

6: $x = \lambda x_i + (1-\lambda) x_j$

7: $y = \lambda y_i + (1-\lambda) y_j$

8: end for

9: update θ

10: end for

Cutout is used to randomly drop out the square regions of the mixed images and has been proven effective in enhancing object localization performance. Let (x, y) be the pair of new training data generated from (x_m, y_m), we apply a randomly rotated rectangular area to occlude the image and turn the occluded scribbles into the background:

$$x = (1-B) \odot x_m \quad [7]$$

$$y = (1-B) \odot y_m \quad [8]$$

where B is a binary rectangular mask with a dimension of $d \times d$. In this study, we chose a rectangle with the size of 32×32.

Code-level details are as follows:

Algorithm Cutout

Input: training data $\{x_i, y_i \mid i = 1,...,n\}$; dimension of binary rectangular mask d x d; iteration number T; the number of each batch size m.

Output: a more robust model

1: $r_x = Unif(0, W), r_y = Unif(0, H)$

2: $r_w = r_h = d$

3: $x_1 = r_x - \frac{r_w}{2}, x_2 = r_x + \frac{r_w}{2}, y_1 = r_y - \frac{r_h}{2}, y_2 = r_y + \frac{r_h}{2}$

4: for t = 1,...,T do

5: uniformly sample a mini-batch of training data $B^{(\tau)}$

6: for $(x_i, y_i) \in B^{(\tau)}$ do

7: transform y_i into one-hot vectors as $x_i[:,:,x_1 : x_2, y_1 : y_2] = 0$

8: end for

9: update θ

10: end for

Development of the 3-sequence MRI-based DL models

We established 5 basic convolutional neural network (CNN) architectures, including VGG16_bn, ResNet18, ResNet34, ResNet50, and Resnet101, and modified them into the MRI-based tri-input models to determine the optimal framework. The tri-input models comprised 3 encoders and a classification head. The 3 encoders were constructed by parallelly duplicating the encoder in the

https://dx.doi.org/10.21037/qims-24-729

basic CNN architecture 3 times to allow a simultaneous input of the sub-images from T1-weighted (T1W), T2-weighted (T2W), and fat suppression (FS) sequences. The classification head contained a concatenation layer to fuse features from the 3 encoders, an adaptive average pooling layer, and a fully connected (FC) layer to obtain a multi-modality representation learning. Finally, another FC layer with SoftMax was deployed to predict a probability of axial spondyloarthritis (axSpA). Since each participant has 2 sets of 3-sequence MRI sub-images for the left-side and right-side SIJs, we averaged their predicted probabilities to obtain a bilateral-SIJ score.

Models were trained using the training set with a 5-fold cross-validation. Parameters in the 3 encoders were initialized using those from the pre-trained basic CNN architectures. Parameters in other layers were randomly initialized using the Gaussian distribution algorithm. The Adam algorithm was adopted as an optimizer with a batch size of 64. The binary cross entropy loss was used as the loss function. As a transfer learning strategy, we first froze parameters in the encoders and only trained those in the rest layers for 60 to 100 epochs. Model performance was monitored on the validation set in each epoch, and the best-performing model was picked out for further training. In the second stage, we directly updated all parameters in the best-performing model. The learning rates in the first and second training stages were initialized as 0.0001 and 0.0001/2, which were both decreased by an automatic cosine annealing schedule with the parameters T_0 =30 and T_mult =2 (49).

Construction of NegSpA-AI

According to the Assessment of SpondyloArthritis International Society (ASAS) classification criteria (2009), the 11 SpA features should be combined with images for classification. In this study, since all patients were human leukocyte antigen-B27 (HLA-B27) negative and complained of back pain, we excluded HLA-B27 and inflammatory back pain, and added the remaining 9 SpA features into the MRI-based DL model to construct NegSpA-AI. Specifically, denoting the 9 SpA features as $\{a_i \mid i = 1,2,...,9\}$, where $a_i = 1$ if the feature was positive, otherwise $a_i = 0$, they were projected to a single score $p \in \{0,[0.5,1]\}$ by a conditional linear transformation as:

$$p = \begin{cases} 0, \ if \ a_i = 0, \ i \in \{1,2,....,9\}, \\ \frac{1}{16} * x + \frac{7}{16}, \ if \ a_i = 1, \ i \in \{1,2,....,9\}, \end{cases} \quad [9]$$

where x is the number of positive features. Then p was averaged with the bilateral-SIJ score predicted by the MRI-based model to obtain a patient-level score.

Visualization and clinical stratification analysis of NegSpA-AI

The gradient-weighted class activation mapping (Grad-CAM) method was used to visualize crucial response areas of NegSpA-AI during classification. A clinical stratification analysis was performed for NegSpA-AI on age, sex, disease duration, and structural damage. Patients were divided into different subgroups using the median values of age and disease duration as thresholds and sex (female or male) and structural damage (positive or negative) as binary categorical variables.

References

48. DeVries T, Taylor GW. Improved regularization of convolutional neural networks with cutout. arXiv:170804552, 2017.
49. Loshchilov I, Hutter F. SGDR: Stochastic Gradient Descent with Warm Restarts. arXiv:1608.03983, 2017.

https://dx.doi.org/10.21037/qims-24-729

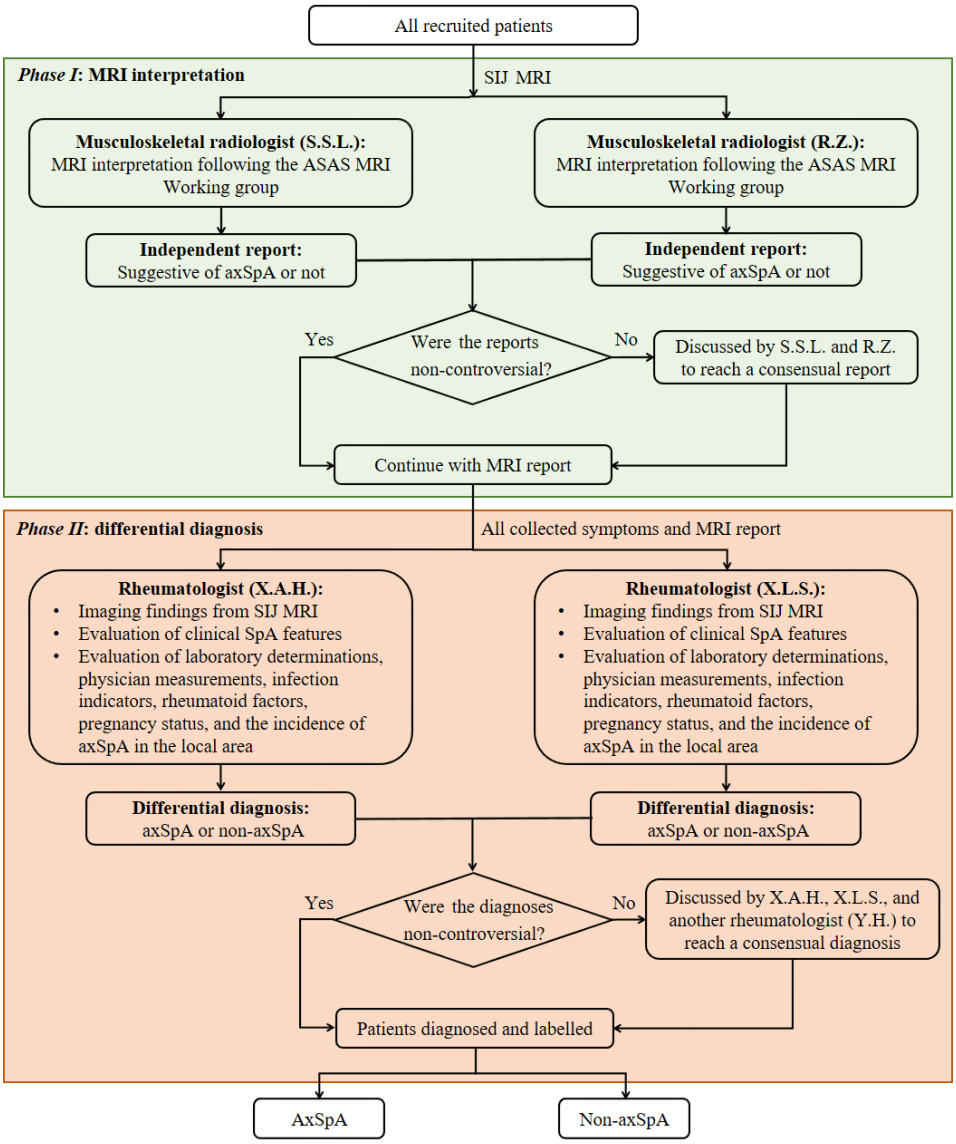

Figure S1 The flowchart for differential diagnosis between axSpA and non-axSpA. MRI, magnetic resonance imaging; SIJ, sacroiliac joint; ASAS, Assessment of SpondyloArthritis International Society; axSpA, axial spondyloarthritis.

https://dx.doi.org/10.21037/qims-24-729

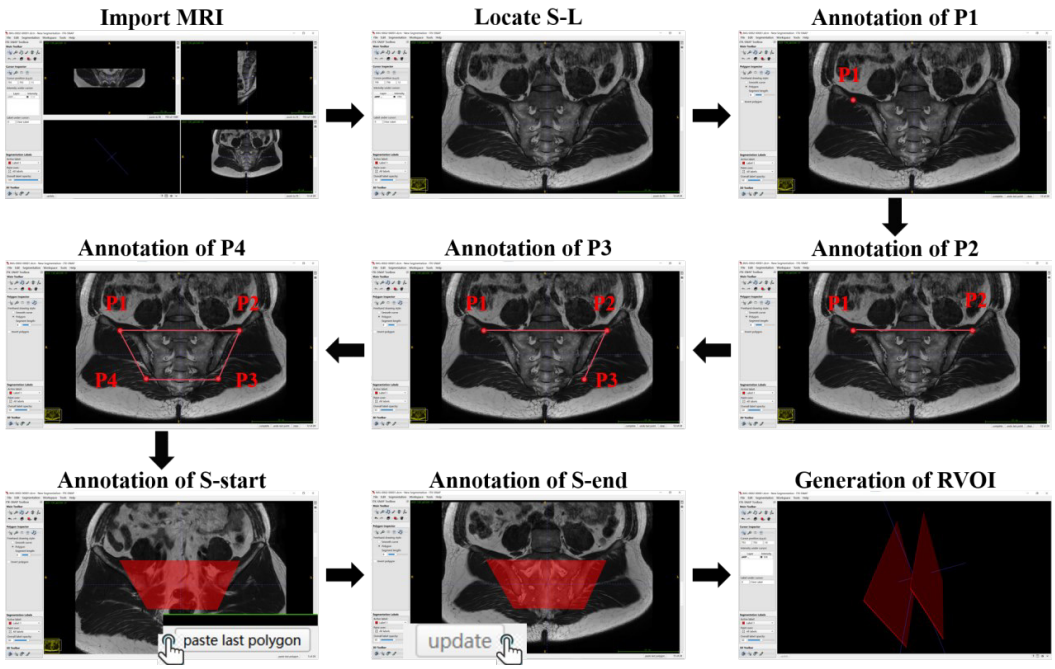

Figure S2 Illustration of the 6-step annotation method. S-L is the slice containing the largest SIJ area in the MRI. S-start and S-end are the first and last slices containing SIJs, respectively. MRI, magnetic resonance imaging; RVOI, rectangular volume of interest; SIJ, sacroiliac joint.

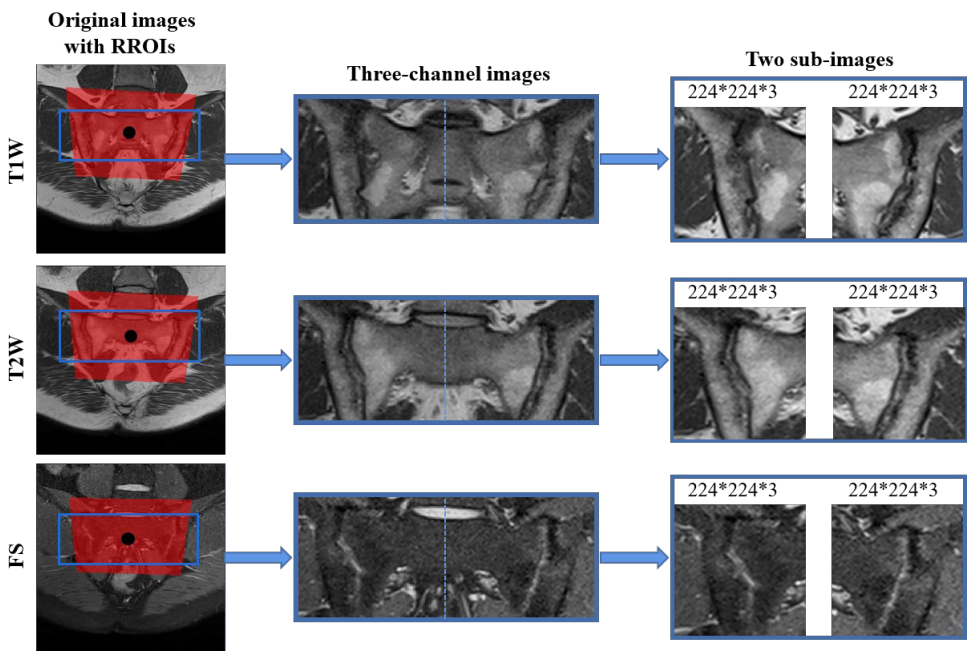

Figure S3 Preprocessing of MRI into sub-images. T1W, T1-weighted; T2W, T2-weighted; FS, fat suppression; RROI, rectangular region of interest; MRI, magnetic resonance imaging.

https://dx.doi.org/10.21037/qims-24-729

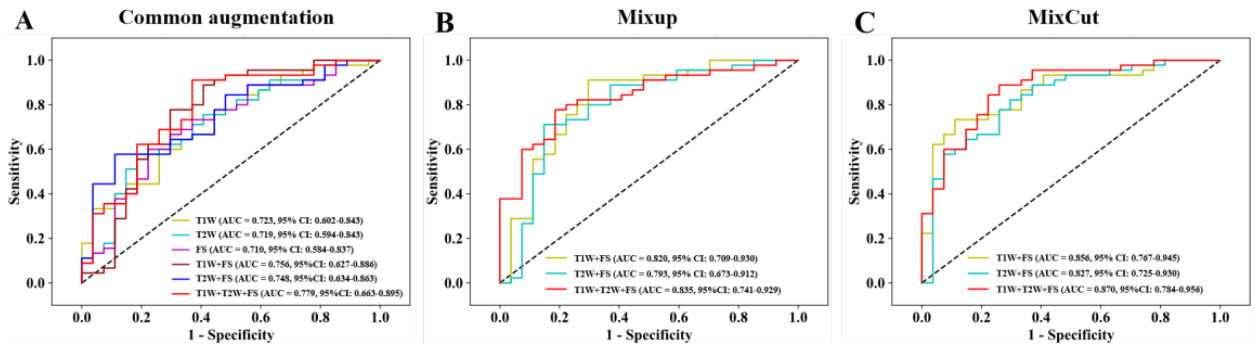

Figure S4 ROC curves for various modified ResNet50 models on the internal test set. (A-C). ROC curves for modified ResNet50 models with different inputs trained with common data augmentation, Mixup, and MixCut, respectively. T1W, T1-weighted; AUC, area under the curve; CI, confidence interval; T2W, T2-weighted; FS, fat suppression; ROC, receiver operating characteristic.

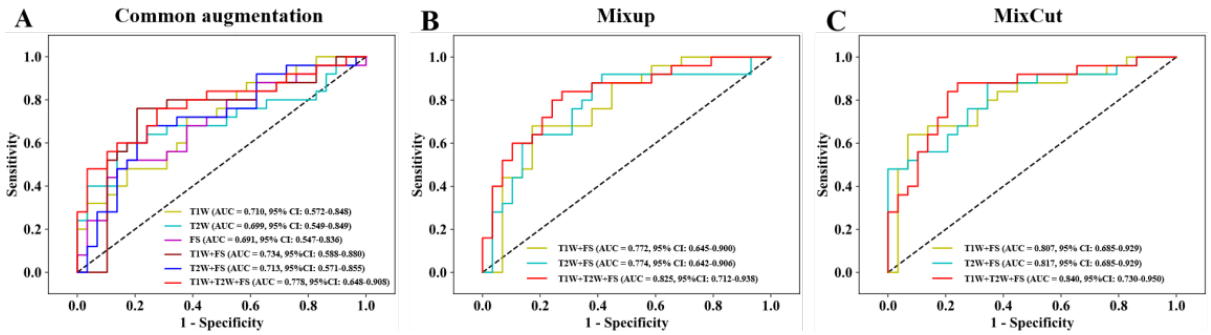

Figure S5 ROC curves for various modified ResNet50 models on the external test set. (A-C). ROC curves for modified ResNet50 models with different inputs trained with common data augmentation, Mixup, and MixCut, respectively. T1W, T1-weighted; AUC, area under the curve; CI, confidence interval; T2W, T2-weighted; FS, fat suppression; ROC, receiver operating characteristic.

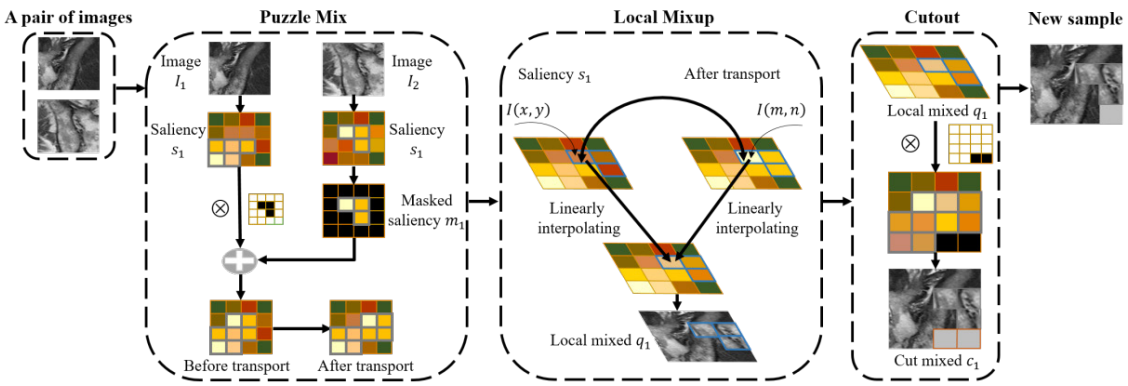

Figure S6 Illustration of MixCut for data augmentation.

https://dx.doi.org/10.21037/qims-24-729

Table S1 Detailed disease subtypes of patients with non-axial spondyloarthritis

Disease subtypes	Training set & internal test set (n=217)	External test set (n=29)	Prospective test set (n=34)
Undifferentiated sacroiliitis	57 (26.27)	15 (51.72)	1 (2.94)
Rheumatoid arthritis	5 (2.30)	0 (0.00)	3 (8.82)
Condensing osteitis	13 (5.99)	0 (0.00)	3 (8.82)
Gout	4 (1.84)	2 (6.90)	0 (0.00)
Degenerative arthritis	46 (21.20)	0 (0.00)	0 (0.00)
Juvenile idiopathic arthritis	4 (1.84)	0 (0.00)	3 (8.82)
Infection of SIJs	7 (3.23)	0 (0.00)	1 (2.94)
Malignant tumors	4 (1.84)	0 (0.00)	0 (0.00)
Tuberculous rheumatism	1 (0.46)	0 (0.00)	0 (0.00)
Osteoarthritis	7 (3.23)	1 (3.45)	5 (14.71)
Non-specific LBP	66 (30.41)	11 (37.93)	18 (52.95)

Data in parentheses are percentage. n, number; SIJ, sacroiliac joint; LBP, low back pain.

https://dx.doi.org/10.21037/qims-24-729

Table S2 Detailed acquisition parameters of magnetic resonance imaging

Hos.	Scanner	Sequence	TR/TE (ms)	Slice thickness (mm)	Spacing between slices (mm)	FOV (mm)	Matrix size
TAH	Philips 1.5T (Achieva)	Coronal T1W	500/18	4.0	4.5	143	396×318
		Axial T2W	3,000/100	4.0	5.0	140	418×320
		Coronal T2W	5,228/100	4.0	4.5	139	400×310
		Coronal SPAIR T2W	3,000/100	4.0	5.0	126	418×320
		Axial PDW SPAIR	2,586/30	6.0	7.0	140	336×264
		Coronal PDW SPAIR	2,898/30	4.0	4.5	132	348×270
		Coronal STIR	2,400/60	3.0	3.3	150	300×238
	Philips 3.0T (Ingenia)	Axial T1W	664/20	5.0	5.5	175	476×351
		Coronal T1W	550/22	3.0	3.5	140	380×336
		Axial T2W	4,565/90	5.0	5.5	133	400×308
		Coronal T2W	2,300/85	3.0	3.5	139	400×392
		Axial SPAIR T2W	4,230/70	6.0	7.0	175	412×405
		Coronal SPAIR T2W	2,929/70	3.0	3.5	139	400×382
NHH	Philips 1.5T (Achieva)	Coronal T1W	468/10	4.0	3.5	140	400×310
		Axial T2W	3,200/100	4.0	3.5	175	420×320
		Coronal T2W	5,500/95	4.0	3.5	139	400×310
		Coronal SPAIR T2W	3,000/60	4.0	3.5	175	420×320
		Coronal SPAIR PDW	2,900/30	4.0	3.5	175	350×270
		Axial STIR	2,500/30	5.0	5.0	175	340×264
		Coronal STIR	2,500/60	4.0	3.5	140	300×240
	Philips 3.0T (Elition)	Axial T1W	450/8	4.0	5.0	141	460×350
		Coronal T1W	506/10	3.0	3.5	175	380×340
		Axial T2W	5,000/100	4.0	5.0	140	400×310
		Coronal T2W	2,500/90	3.0	3.5	175	400×390
		Axial SPAIR T2W	5,100/80	5.0	5.0	175	412×405
		Coronal SPAIR T2W	3,000/70	3.0	3.5	175	400×380

Hos., hospital; TR, time of repetition; TE, time of echo; ms, millisecond; mm, millimeter; FOV, field of view; TAH, the Third Affiliated Hospital of Southern Medical University; T1W, T1-weighted; T2W, T2-weighted; SPAIR, spectral attenuated inversion recovery; PDW, proton density weighted; STIR, short tau inversion recovery; NHH, Nanhai Hospital.

https://dx.doi.org/10.21037/qims-24-729

Table S3 Detailed clinical characteristics of patients on the training set

Characteristics	AxSpA	Non-axSpA	P value
Age* (years)	30.7±12.9	41.4±18.1	2.07E−08
Sex			0.041
Female	51 (37.0)	93 (48.9)	
Male	87 (63.0)	97 (51.1)	
Disease duration# (M)	24.0 (6.0, 60.0)	12.0 (4.0, 60.0)	0.099
Structural damage			0.002
Positive	89 (64.5)	88 (46.3)	
Negative	49 (35.5)	102 (53.7)	
Arthritis			5.26E−06
Positive	21 (15.2)	74 (38.9)	
Negative	117 (84.8)	116 (61.1)	
Heel enthesitis			NA
Positive	0 (0.0)	0 (0.0)	
Negative	138 (100)	190 (100)	
Uveitis			NA
Positive	0 (0.0)	0 (0.0)	
Negative	138 (100)	190 (100)	
Dactylitis			NA
Positive	0 (0.0)	0 (0.0)	
Negative	138 (100)	190 (100)	
Psoriasis			NA
Positive	0 (0.0)	1 (0.5)	
Negative	138 (100)	189 (98.5)	
Crohn's disease or ulcerative colitis			0.344
Positive	2 (1.4)	0 (0.0)	
Negative	136 (98.6)	190 (100)	
Good response to NSAIDs			2.24E−15
Positive	131 (94.9)	103 (54.2)	
Negative	7 (5.1)	87 (45.8)	
Family history of axSpA			8.93E−08
Positive	25 (18.1)	2 (1.1)	
Negative	113 (81.9)	188 (98.9)	
Elevated CRP concentration			0.718
Positive	50 (36.2)	64 (33.7)	
Negative	88 (63.8)	126 (66.3)	

Data* are means ± standard deviations; data# are present using median (Q1, Q3). Except where specified, data are numbers of patients, with percentages in parentheses. AxSpA, axial spondyloarthritis; M, months; NA, not available; NSAIDs, nonsteroidal anti-inflammatory drugs; CRP, C-reactive protein.

https://dx.doi.org/10.21037/qims-24-729

Table S4 Detailed clinical characteristics of patients on the 3 test sets

Characteristics	Internal test set (n=72)			External test set (n=54)			Prospective test set (n=87)		
	AxSpA	Non-axSpA	P value	AxSpA	Non-axSpA	P value	AxSpA	Non-axSpA	P value
Age* (years)	32.1±10.7	44.4±18.1	0.001	34.2±8.5	38.6±15.4	0.385	35.1 ± 13.9	43.0 ± 14.1	0.003
Sex			0.267			0.003			0.111
Female	16 (35.6)	14 (51.9)		2 (8.0)	14 (48.3)		27 (50.9)	24 (70.6)	
Male	29 (64.4)	13 (48.1)		23 (92.0)	15 (51.7)		26 (49.1)	10 (29.4)	
DD# (months)	12.0 (5.8, 47.3)	8.0 (2.5, 93.0)	0.558	13.0 (3.0, 60.0)	6.0 (0.3, 12.3)	0.044	12.0 (5.3, 45.0)	24.0 (3.5, 54.0)	0.579
Structural damage			0.583			5.76E-06			0.015
Positive	26 (57.8)	13 (48.1)		25 (100)	11 (37.9)		37 (69.8)	14 (41.2)	
Negative	19 (42.2)	14 (51.9)		0 (0.0)	18 (62.1)		16 (30.2)	20 (58.8)	
Arthritis			0.018			0.025			0.037
Positive	6 (13.3)	11 (40.7)		23 (92.0)	18 (62.1)		11 (20.8)	15 (44.1)	
Negative	39 (86.7)	16 (59.3)		2 (8.0)	11 (37.9)		42 (79.2)	19 (55.9)	
Heel enthesitis			NA			0.082			0.91
Positive	0 (0.0)	0 (0.0)		18 (72.0)	13 (44.8)		2 (3.8)	2 (5.9)	
Negative	45 (100)	27 (100)		7 (28.0)	16 (55.2)		51 (96.2)	32 (94.1)	
Uveitis			NA			NA			NA
Positive	0 (0.0)	0 (0.0)		0 (0.0)	1 (3.4)		0 (0.0)	0 (0.0)	
Negative	45 (100)	27 (100)		25 (100)	28 (96.6)		53 (100)	34 (100)	
Dactylitis			NA			0.535			0.95
Positive	0 (0.0)	0 (0.0)		7 (28.0)	5 (17.2)		1 (1.9)	1 (2.9)	
Negative	45 (100)	27 (100)		18 (72.0)	24 (82.8)		52 (98.1)	33 (97.1)	
Psoriasis			NA			NA			NA
Positive	0 (0.0)	1 (3.7)		0 (0.0)	0 (0.0)		1 (1.9)	0 (0.0)	
Negative	45 (100)	26 (96.3)		25 (100)	29 (100)		52 (98.1)	34 (100)	
Crohn's disease or ulcerative colitis			NA			0.94			NS
Positive	0 (0.0)	0 (0.0)		1 (4.0)	0 (0.0)		0 (0.0)	1 (2.9)	
Negative	45 (100)	27 (100)		24 (96.0)	29 (100)		53 (100)	33 (97.1)	
Good response to NSAIDs			0.001			0.526			0.251
Positive	42 (93.3)	16 (59.3)		8 (32.0)	6 (20.7)		20 (37.7)	8 (23.5)	
Negative	3 (6.7)	11 (40.7)		17 (68.0)	23 (79.3)		33 (62.3)	26 (76.5)	
Family history of axSpA			0.068			0.004			0.668
Positive	15 (33.3)	3 (11.1)		8 (32.0)	0 (0.0)		4 (7.5)	1 (2.9)	
Negative	30 (66.7)	24 (88.9)		17 (68.0)	29 (100)		49 (92.5)	33 (97.1)	
Elevated CRP concentration			0.795			0.465			0.382
Positive	14 (31.1)	8 (29.6)		13 (52.0)	19 (63.5)		6 (11.3)	7 (20.6)	
Negative	31 (68.9)	19 (70.4)		12 (48.0)	10 (34.5)		47 (88.7)	27 (79.4)	

Data* are means ± standard deviations; data# are present using median (Q1, Q3). Except where specified, data are numbers of patients, with percentages in parentheses. AxSpA, axial spondyloarthritis; n, number; DD, disease duration; NA, not available; NSAIDs, nonsteroidal anti-inflammatory drugs; CRP, C-reactive protein.

https://dx.doi.org/10.21037/qims-24-729

Table S5 Performance of various modified ResNet50 models on the internal test set

Aug.	Framework	AUC	Accuracy	Sensitivity	Specificity	F1-score
Com.	O-ResNet50* {T1W}	0.723 (0.602–0.843)	65.28 (53.14–76.12)	71.11 (55.69–83.63)	55.56 (35.33–74.52)	0.667
	O-ResNet50* {T2W}	0.719 (0.594–0.843)	66.67 (54.57–77.34)	57.78 (42.15–72.34)	81.48 (61.92–93.70)	0.667
	O-ResNet50* {FS}	0.710 (0.584–0.837)	66.67 (54.57–77.34)	60.00 (44.33–74.30)	77.78 (57.74–91.38)	0.605
	Dual-ResNet50^ {T1W+FS}	0.756 (0.627–0.886)	69.44 (57.47–79.76)	68.89 (53.35–81.83)	70.37 (49.82–86.25)	0.738
	Dual-ResNet50^ {T2W+FS}	0.748 (0.634–0.863)	69.44 (57.47–79.76)	57.78 (42.15–72.34)	88.89 (70.84–97.65)	0.703
	Tri-ResNet50# {T1W+T2W+FS}	0.779 (0.663–0.895)	72.22 (60.41–82.14)	77.78 (62.91–88.80)	62.96 (42.37–80.60)	0.778
Mixup	Dual-ResNet50^ {T1W+FS}	0.820 (0.709–0.930)	77.78 (66.44–86.73)	82.22 (67.95–92.00)	70.37 (49.82–86.25)	0.822
	Dual-ResNet50^ {T1W+FS}	0.793 (0.673–0.912)	76.39 (64.91–85.60)	71.11 (55.69–83.63)	85.19 (66.27–95.81)	0.790
	Tri-ResNet50# {T1W+T2W+FS}	0.835 (0.741–0.929)	79.17 (67.98–87.84)	77.78 (62.91–88.80)	81.48 (61.92–93.70)	0.824
MixCut	Dual-ResNet50^ {T1W+FS}	0.856 (0.767–0.945)	79.17 (67.98–87.84)	73.33 (58.06–85.40)	88.89 (70.84–97.65)	0.815
	Dual-ResNet50^ {T2W+FS}	0.827 (0.725–0.930)	77.78 (66.44–86.73)	82.22 (67.95–92.00)	70.37 (49.82–86.25)	0.822
	Tri-ResNet50# {T1W+T2W+FS}	0.870 (0.784–0.956)	83.33 (72.70–91.08)	88.89 (75.95–96.29)	74.07 (53.72–88.89)	0.870

Accuracy, sensitivity, and specificity are expressed as percentages. Data in brackets are 95% confidence intervals. Frameworks* are the basic single-input CNN architectures; frameworks^ are the modified dual-input models using two-sequence MRI as inputs; frameworks# are the modified tri-input models; words in curly brackets are the detailed sequences of MRI as inputs for each model. The common data augmentations referred to random rotation, random horizontal flip, and random pixel shift along the x- and y-directions. Aug., data augmentation method; AUC, area under the curve; Com., common; T1W, T1-weighted; T2W, T2-weighted; FS, fat suppression.

Table S6 Performance of various modified ResNet50 models on the external test set

Aug.	Framework	AUC	Accuracy	Sensitivity	Specificity	F1-score
Com.	O-ResNet50* {T1W}	0.710 (0.572–0.848)	66.67 (52.52–78.91)	72.00 (50.61–87.93)	62.07 (42.26–79.31)	0.667
	O-ResNet50* {T2W}	0.699 (0.549–0.849)	72.22 (58.36–83.54)	60.00 (38.67–78.87)	82.76 (64.22–94.15)	0.667
	O-ResNet50* {FS}	0.691 (0.547–0.836)	68.52 (54.45–80.48)	52.00 (31.31–72.20)	82.76 (64.26–94.15)	0.605
	Dual-ResNet50^ {T1W+FS}	0.734 (0.588–0.880)	69.44 (57.47–79.76)	68.89 (53.35–81.83)	70.37 (49.82–86.25)	0.681
	Dual-ResNet50^ {T2W+FS}	0.713 (0.571–0.855)	72.22 (58.36–83.54)	64.00 (42.52–82.03)	79.31 (60.28–92.00)	0.694
	Tri-ResNet50# {T1W+T2W+FS}	0.778 (0.648–0.908)	74.07 (60.35–85.04)	76.00 (54.87–90.64)	72.41 (52.76–87.27)	0.731
Mixup	Dual-ResNet50^ {T1W+FS}	0.772 (0.645–0.900)	75.93 (62.36–86.51)	68.00 (46.50–85.05)	82.76 (64.26–94.15)	0.723
	Dual-ResNet50^ {T1W+FS}	0.774 (0.642–0.906)	74.07 (60.35–85.04)	92.00 (73.97–99.01)	58.62 (38.94–76.48)	0.767
	Tri-ResNet50# {T1W+T2W+FS}	0.825 (0.712–0.938)	77.78 (64.40–87.96)	84.00 (63.92–95.46)	72.41 (52.76–87.27)	0.778
MixCut	Dual-ResNet50^ {T1W+FS}	0.807 (0.685–0.929)	79.63 (66.47–89.37)	64.00 (42.52–82.03)	93.10 (77.23–99.15)	0.744
	Dual-ResNet50^ {T2W+FS}	0.817 (0.701–0.932)	75.93 (62.36–86.51)	88.00 (68.78–97.45)	65.51 (45.67–82.06)	0.772
	Tri-ResNet50# {T1W+T2W+FS}	0.840 (0.730–0.950)	81.48 (68.57–90.75)	88.00 (68.78–97.45)	75.86 (56.46–89.70)	0.815

Accuracy, sensitivity, and specificity are expressed as percentages. Data in brackets are 95% confidence intervals. Frameworks* are the basic single-input CNN architectures; frameworks^ are the modified dual-input models using two-sequence MRI as inputs; frameworks# are the modified tri-input models; words in curly brackets are the detailed sequences of MRI as inputs for each model. The common data augmentations referred to random rotation, random horizontal flip, and random pixel shift along the x- and y-directions. Aug., data augmentation method; AUC, area under the curve; Com., common; T1W, T1-weighted; T2W, T2-weighted; FS, fat suppression; CNN, convolutional neural network.

https://dx.doi.org/10.21037/qims-24-729

Table S7 Clinical characteristics of patients with accurate AI predictions but inaccurate clinician classifications

Data sets	Clinicians	Sex		Age		Disease duration	
		Female	Male	≥28 years	<28 years	≥24 months	<24 months
Internal test set	Rad1	9	4	9	4	7	6
	Rad2	10	4	10	4	7	7
	Rad3	4	5	5	4	4	5
	Rad4	1	4	3	2	3	2
External test set	Rad1	10	5	12	3	8	7
	Rad2	8	3	8	3	4	7
	Rad3	8	3	8	3	5	6
	Rad4	6	2	6	2	3	5
Prospective test set	Rad2	16	10	20	6	11	15
	Rad3	13	4	12	5	7	10
	Rheu1	18	8	21	5	12	14
	Rheu2	11	4	11	4	7	8

All data are expressed as the number of patients. Rad1 and Rad2 are junior radiologists. Rad3 and Rad4 are senior radiologists. Rheu1 and Rheu2 are junior and senior rheumatologists, respectively. AI, artificial intelligence; Rad, radiologist; Rheu, rheumatologist.

https://dx.doi.org/10.21037/qims-24-729