


MUSCULOSKELETAL



MRI-based automated multitask deep learning system to evaluate supraspinatus tendon injuries

Ming Ni^{1†}, Yuqing Zhao^{1†}, Lihua Zhang¹, Wen Chen¹, Qizheng Wang¹, Chunyan Tian^{1*†} and Huishu Yuan^{1*†} 

Abstract

Objective To establish an automated, multitask, MRI-based deep learning system for the detailed evaluation of supraspinatus tendon (SST) injuries.

Methods According to arthroscopy findings, 3087 patients were divided into normal, degenerative, and tear groups (groups 0–2). Group 2 was further divided into bursal-side, articular-side, intratendinous, and full-thickness tear groups (groups 2.1–2.4), and external validation was performed with 573 patients. Visual geometry group network 16 (VGG16) was used for preliminary image screening. Then, the rotator cuff multitask learning (RC-MTL) model performed multitask classification (classifiers 1–4). A multistage decision model produced the final output. Model performance was evaluated by receiver operating characteristic (ROC) curve analysis and calculation of related parameters. McNemar's test was used to compare the differences in the diagnostic effects between radiologists and the model. The intra-class correlation coefficient (ICC) was used to assess the radiologists' reliability. $p < 0.05$ indicated statistical significance.

Results In the in-group dataset, the area under the ROC curve (AUC) of VGG16 was 0.92, and the average AUCs of RC-MTL classifiers 1–4 were 0.99, 0.98, 0.97, and 0.97, respectively. The average AUC of the automated multitask deep learning system for groups 0–2.4 was 0.98 and 0.97 in the in-group and out-group datasets, respectively. The ICCs of the radiologists were 0.97–0.99. The automated multitask deep learning system outperformed the radiologists in classifying groups 0–2.4 in both the in-group and out-group datasets ($p < 0.001$).

Conclusion The MRI-based automated multitask deep learning system performed well in diagnosing SST injuries and is comparable to experienced radiologists.

Clinical relevance statement Our study established an automated multitask deep learning system to evaluate supraspinatus tendon (SST) injuries and further determine the location of SST tears. The model can potentially improve radiologists' diagnostic efficiency, reduce diagnostic variability, and accurately assess SST injuries.

Key Points

- A detailed classification of supraspinatus tendon tears can help clinical decision-making.
- Deep learning enables the detailed classification of supraspinatus tendon injuries.

[†]Ming Ni, Yuqing Zhao, Huishu Yuan and Chunyan Tian contributed equally to this work.

*Correspondence:

Chunyan Tian
huishuy@bjmu.edu.cn
Huishu Yuan
tcyzhs@163.com

Full list of author information is available at the end of the article

• *The proposed automated multitask deep learning system is comparable to radiologists.*

Keywords Deep learning, Sports injuries, Shoulder, Rotator cuff, Supraspinatus

Introduction

Supraspinatus tendon (SST) injuries are the most common shoulder joint sports injury [1]. SST injuries have various causes, most commonly long-term repeated and excessive stress, which causes the rate of SST injury to exceed the healing rate, resulting in tendon self-repair failure [2]. In cases of asymptomatic SST injuries, the injury may worsen without timely and accurate diagnosis and intervention [3]. Therefore, accurate assessment of an SST injury can help the clinical development of an individualized treatment strategy, determine the optimal timing for treatment [4], and improve the patient's quality of life.

MRI is the preferred imaging examination method for the preoperative assessment of SST injuries [5–8]. In a previous study, the aggregate sensitivity of MRI in diagnosing partial and complete tears was 0.70 and 0.81, respectively, and the aggregate specificity was 0.95 [9]. CT and MR arthrography are commonly used imaging methods for diagnosing SST injuries, but they are less sensitive to bursal-sided and intratendinous tears and have difficulty in diagnosing SST degeneration [10]. Because related studies recommend MRI as the primary diagnostic method for SST injuries [6], MRI was adopted in the current study. However, there are significant differences in the MRI-based diagnoses of SST injuries among radiologists [11, 12], as they usually do not evaluate the location of tears in detail. Therefore, establishing a more accurate, efficient, and consistent method for effectively evaluating SST injury is important for clinical application.

Deep learning enables the autonomous understanding of the task content and representation of data, automatically extracting valuable feature information without prior feature definitions [13–15]. Currently, the use of deep learning in shoulder joint SST injuries is relatively limited [16, 17], and there are no studies that include SST degeneration as a possible classification, nor have there been studies that provide a detailed classification of tears. With the increasing number of tasks that must be performed simultaneously in medicine, conventional deep learning models are inefficient and have high maintenance costs [18]. However, the introduction and development of multitask learning have provided an opportunity to solve such problems [19].

This study aims to establish an automated multitask deep learning system based on MRI to classify SST injuries in detail and compare model performance with that

of radiologists of different experience levels to help radiologists and clinicians make an accurate, detailed assessment of SST injuries preoperatively.

Materials and methods

The institutional review board and medical science ethics committee of our hospital and another centre approved this retrospective study, and the requirement for informed consent was waived.

Patients

The clinical and imaging data of patients undergoing shoulder surgery in our hospital from January 2013 to October 2022 were collected retrospectively. The inclusion criteria were as follows: MRI examination performed within the two weeks before arthroscopy, and available complete records of arthroscopic surgery. The exclusion criteria were as follows: severe motion artefacts (evaluated by Ming Ni and Yuqing Zhao), calcified tendinitis, shoulder joint tumours, infectious diseases, and previous rotator cuff surgery.

A total of 3087 patients were eventually included in the in-group dataset. The patient enrolment process is shown in Fig. 1. In addition, a separate dataset consisting of 573 patients was treated as an independent out-group dataset. The out-group dataset was subjected to the same inclusion and exclusion criteria as the in-group dataset, and the data came from multiple examination devices at another centre from January 2017 to December 2022.

Arthroscopy

According to the arthroscopy results, the SST injuries were divided into normal, degenerative, and tear groups (groups 0–2, respectively). Group 2 was further divided into bursal-side, articular-side, intratendinous, and full-thickness tear groups (groups 2.1–2.4, respectively). In this study, patients with a normal SST or mild injuries were enrolled if they underwent arthroscopy when other shoulder joint structures were injured. When degeneration and tearing occurred in the SST of a patient at the same time, he or she was included in the tear group. The arthroscopy results were used as labels for model training and ground truth labels for manual diagnosis. Detailed grouping and basic clinical information of the patients in the in-group and out-group datasets are shown in Table 1.

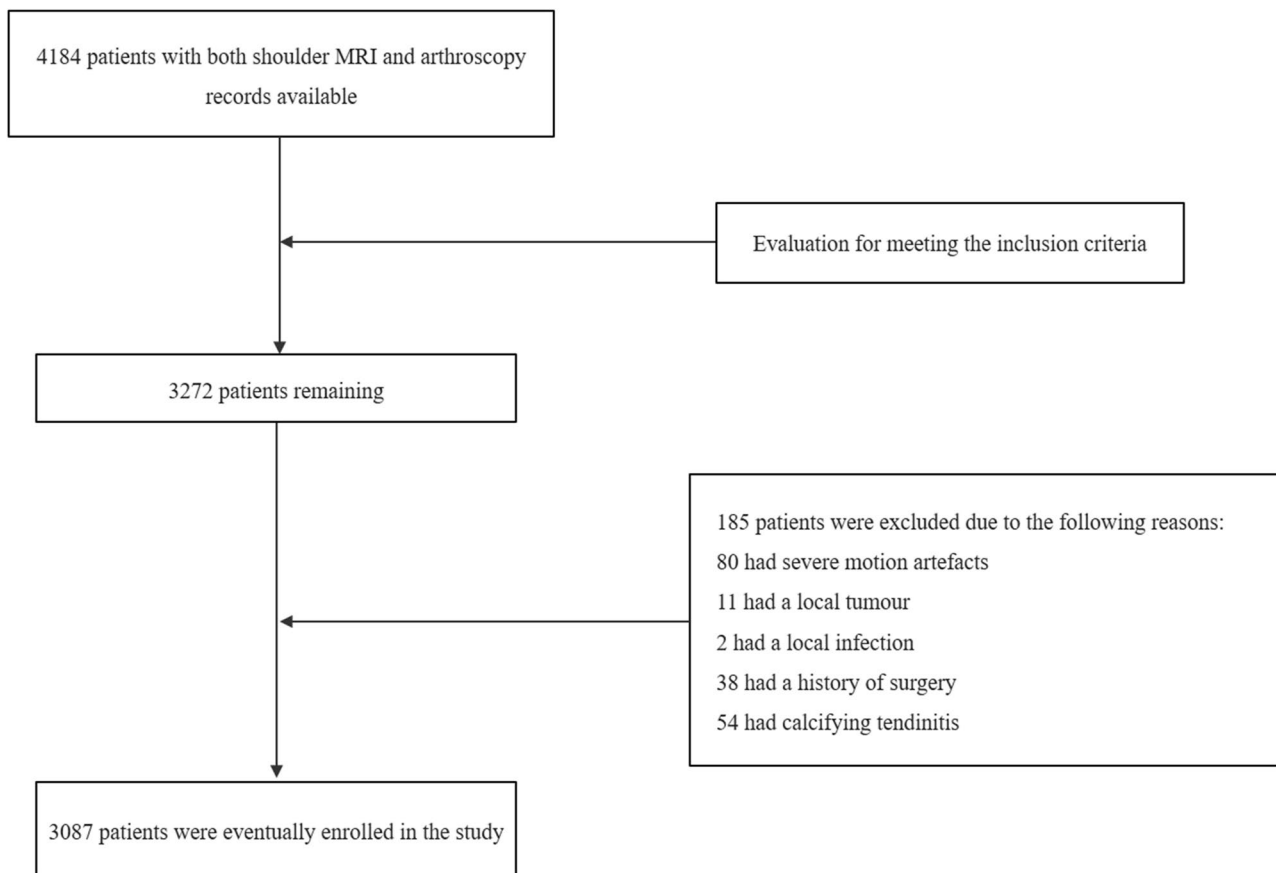


Fig. 1 The detailed process of patient enrolment in this study. Ultimately, 3087 patients were included in the in-group dataset. A total of 573 patients from another centre were included as an out-group dataset according to the same criteria

MRI scanning

In the in-group dataset, MRI was performed with GE Discovery MR750w and MR750w Silent scanners (3.0 T, GE Medical Systems) with a 16-channel flexible coil. During the scan, the patient was positioned supine with a slight external arm rotation of 5–20° to avoid internal rotation. Oblique coronal (OCOR) and oblique sagittal (OSAG) fat saturation proton density-weighted fast spin-echo (PD-FSE-FS) sequences were obtained. The images in the out-group dataset were obtained from a Signa MRexplorer (1.5 T, GE Medical Systems; $n=329$), an Optima MR360 (1.5 T, GE Medical Systems; $n=46$), a Signa HDxt (GE Medical Systems; $n=28$), a UI uMR780i (3.0 T, United imaging; $n=29$), a UI uMR880 (3.0 T, United imaging; $n=43$), a UI Omega (3.0 T, United imaging; $n=85$), and a MAGNETOM Prisma (3.0 T, Siemens Healthineers; $n=13$) scanner. The scanning parameters of the devices are shown in Table 2.

Radiologist evaluation

Four musculoskeletal (MSK) radiologists with different experience levels independently diagnosed the patients in both the in-group and out-group test sets in parallel and compared the diagnostic results with arthroscopic results: radiologist 1, Huishu Yuan, with 32 years of experience; radiologist 2, Wen Chen, with 21 years of experience; radiologist 3, Yuqing Zhao, with 11 years of experience; and radiologist 4, Qizheng Wang, with 4 years of experience. The four radiologists had completed courses on the MRI diagnosis of shoulder joint conditions in our hospital to ensure no diagnostic errors were caused by nonexperience factors.

Regions of interest

The region of interest (ROI) was outlined by two radiologists (Ming Ni and Wen Chen). After Ming Ni created the first outline, Wen Chen corrected the outlined ROI. Using the roLabelImg software (<https://github.com/>

Table 1 Grouping and basic clinical information of the in-group and out-group datasets

Dataset	Group		Subject (n)	Sex (n)	Age (Y, mean ± SD)	Position
In-group dataset	Normal		130	109 M	23.33 ± 7.10	44 left and 65 right
				21 F	30.24 ± 14.33	13 left and 8 right
	Degeneration		326	235 M	30.64 ± 11.40	79 left and 156 right
				91 F	45.07 ± 12.46	39 left and 52 right
	Partial-thickness tear	Bursal-sided tear	736	306 M	49.45 ± 10.71	128 left and 178 right
				430 F	54.26 ± 9.26	170 left and 260 right
		Articular-sided tear	136	57 M	51.74 ± 16.30	23 left and 34 right
				79 F	54.42 ± 13.00	21 left and 58 right
	Full-thickness tear	Intratendinous tear	140	60 M	45.00 ± 11.90	23 left and 37 right
				80 F	52.98 ± 8.44	28 left and 52 right
Out-group dataset	Normal		69	718 M	56.25 ± 9.85	245 left and 473 right
				901 F	60.50 ± 7.69	236 left and 665 right
	Degeneration		97	49 M	26.7 ± 9.65	19 left and 30 right
				20 F	29.78 ± 13.76	7 left and 13 right
	Partial-thickness tear	Bursal-sided tear	96	46 M	32.69 ± 13.73	10 left and 36 right
				51 F	54.8 ± 9.15	23 left and 28 right
		Articular-sided tear	27	49 M	47.53 ± 9.83	30 left and 19 right
				47 F	55.59 ± 8.66	12 left and 35 right
	Full-thickness tear	Intratendinous tear	23	11 M	52.27 ± 14.65	4 left and 7 right
				16 F	56.02 ± 12.70	3 left and 13 right
			280	9 M	49.92 ± 14.53	2 left and 7 left
				14 F	54.56 ± 13.75	5 left and 9 left

M, male; F, female

Table 2 Scanning parameters of the MR devices used in the study. The in-group dataset was obtained from 750 and 750ws scanners, and the out-group dataset was obtained from the remaining devices (from another centre)

Scanner	Sequence	TE (ms)	TR (ms)	FOV (cm)	Slice thickness (mm)	Slice gap (mm)	NEX	Pixel bandwidth (HZ)	Matrix
750w and 750ws	OCOR PD-FSE-FS	50	2021	16	3	0.5	4	198	320 × 224
	OSAG PD-FSE-FS	32	2753	16	3	0.5	4	198	288 × 244
Signa MRexplorer	OCOR PD-FSE-FS	68	2635	18	4	0.5	2	198	288 × 244
	OSAG PD-FSE-FS	37	2474	18	4	0.4	2	198	288 × 244
Optima MR360	OCOR PD-FSE-FS	50	1983	18	4	0.5	2	244	288 × 244
	OSAG PD-FSE-FS	38	2760	18	4.5	0.5	2	198	288 × 244
Signa HDxt	OCOR PD-FSE-FS	50	1983	15	4	0.5	2	244	288 × 244
	OSAG PD-FSE-FS	36	2760	15	4.5	0.5	2	244	320 × 192
UI uMR780i	OCOR PD-FSE-FS	65	3110	16	3	0.3	2	220	320 × 240
	OSAG PD-FSE-FS	65	3110	16	3	0.3	2	200	320 × 240
UI uMR880	OCOR PD-FSE-FS	65	4070	16	3	0.3	1.8	160	336 × 269
	OSAG PD-FSE-FS	55	3100	16	3	0.3	1.8	160	320 × 240
UI Omega	OCOR PD-FSE-FS	65	3550	16	3	0.3	1.5	160	336 × 269
	OSAG PD-FSE-FS	54	3075	16	3	0.3	1.5	160	320 × 240
MAGNETOM Prisma	OCOR PD-FSE-FS	67	2690	16	3	0.3	1	250	320 × 224
	OSAG PD-FSE-FS	43	2900	16	3	0.3	1	215	384 × 269

cgvict/roLabelImg), the bounding box was outlined layer by layer parallel to the long axis of the SST in the OCOR and OSAG PD-FSE-FS images, with the ROI ranging from the greater tuberosity of the humerus to the articular glenoid, maximizing the inclusion of the SST structure and minimizing that of other structures.

Deep learning workflow

This study established an automated multitask deep learning system based on multitask learning combined with Visual Geometry Group Network 16 (VGG16) and a multistage decision model to classify SST injuries. The in-group dataset was divided into a training set, verification set, and test set at a ratio of 8:1:1. The out-group dataset was used to test the system's generalizability. For all deep learning models, training was implemented with an NVIDIA Tesla V100 graphics card (32 GB video memory) and an Intel(R) Xeon(R) Gold 5215 CPU. The relevant codes used for this study are detailed in the Supplementary Material.

Preprocessing

Random image augmentation was performed on all images by rotating ($\pm 15^\circ$), changing brightness and contrast, and adding Gaussian noise (only for the model training phase) to increase the amount of sample data and reduce the data imbalance between groups. Subsequently, the pixel values in the image were normalized from -1 to 1 to eliminate the effect of dimensionality between different feature data. The formula is as follows:

$$\text{Normalization} = \frac{v}{\max(\|v\|_p, \epsilon)} v$$

(Parameters : $p = 2$; $\epsilon = 1e - 12$)

Finally, all the data were resized (224×224) and shuffled.

Preliminary image screening

MRI scanning of the shoulder joint with various sequences can yield many images (including localization images, OCOR T1-weighted imaging (T1WI), OSAG T1WI, axial, OCOR, and OSAG PD-FSE-FS images). However, due to the large difference in the number of images in the target sequence (OCOR and OSAG PD-FSE-FS sequence) and the nontarget sequence, the rotator cuff multitask learning (RC-MTL) model cannot directly classify images, and selecting the target image from all images is a prerequisite for subsequent analysis. Therefore, in this study, sequence selection was achieved by using VGG16 as a connection model. VGG16 is a classic deep learning model [20] that is widely used in medical research and has achieved excellent results [21, 22].

SST injury classification

Detailed classification of SST injuries was performed using the RC-MTL model. The RC-MTL model is characterized by sharing underlying multilayer network parameters among different tasks, configuring independent task-specific parameters for various tasks at the model output layer, and comprehensively improving the performance of each independent task by simultaneously learning associated tasks [19]. The RC-MTL model involves a relatively small number of repetitive calculations and has good model reasoning speed, high learning efficiency, implicit data enhancement, attention focus, feature data theft, and good generalizability [23].

The RC-MTL model consists of a shared underlying structure and four independent task classifiers. The detailed structure of the model is shown in Fig. 2. The shared underlying structure is ResNet-101, consisting of a feature extraction module and adaptive average pooling. The feature extraction module adds the convolutional block attention module (CBAM), which uses both spatial attention and channel attention mechanisms to improve the representational ability and accuracy of the model [24]. Four task classifiers can output classification results for different tasks simultaneously:

Classifier 1: distinguishes between OSAG and OCOR images;

Classifier 2: distinguishes between images with and without SST;

Classifier 3: distinguishes between images without SST (not in groups 0–2, NIG) and groups 0–2 (GP 0–2);

Classifier 4: distinguishes between NIG, non-SST tears (groups 0 and 1, GP 0–1), and groups 2.1–2.4 (GP 2.1–2.4).

In addition, we added an ROI-based attention module to classifiers 3 and 4, which helps find target points and improve the accuracy of the corresponding classifiers by learning ROIs. The RC-MTL model uses the AdamW optimizer with learning rates of 5×10^{-4} to 1×10^{-6} . Using the cosine annexing strategy, the weights of classifiers 1–4 are 0.3, 0.5, 1.0, and 1.2, respectively. The loss function is as follows: $\text{loss}(g, p) = \sum_{i=0}^n w_{1i} \sum_{j=0}^c w_{2j} g_{ij} \ln p_{ij}$ where g is the ground truth, p is prediction, w_{1i} is weights between classifiers, w_{2j} is weights within groups, n is the total number of classifiers, and c is the number of tasks for the classifier.

Multistage decision model

Although the RC-MTL model can simultaneously output classification results for multiple tasks, the outputs of the

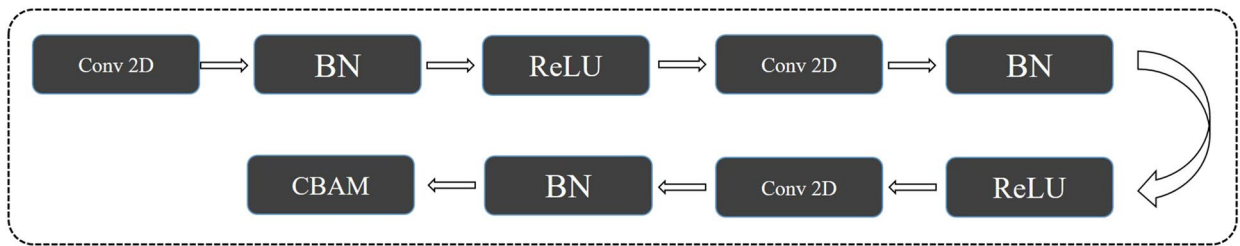


Diagram illustrating the proposed channel and spatial attention mechanism. The process starts with an **Input feature** (represented by a 3D volume). This input is processed by two parallel attention modules: the **Channel Attention Module** and the **Spatial Attention Module**. The outputs of these modules are combined via element-wise multiplication (represented by a circle with an 'X') to produce the final **Refined feature** (represented by a 3D volume).

Fig. 2 The structure of the rotator cuff multitask learning (RC-MTL) model, which consists of a shared underlying structure and four independent task classifiers. The feature extraction module consists of several modules, including the convolutional block attention module (CBAM). The four task classifiers output the classification results of different tasks

multistage decision model weights the output probability values of different classifiers for different tasks to obtain the combined probability values of each category. The classification with the highest probability value is taken as the final result for that image. For the combined distinction between including/not including SST images, the weights of classifiers 2, 3, and 4 are 0.8, 0.1, and 0.1,

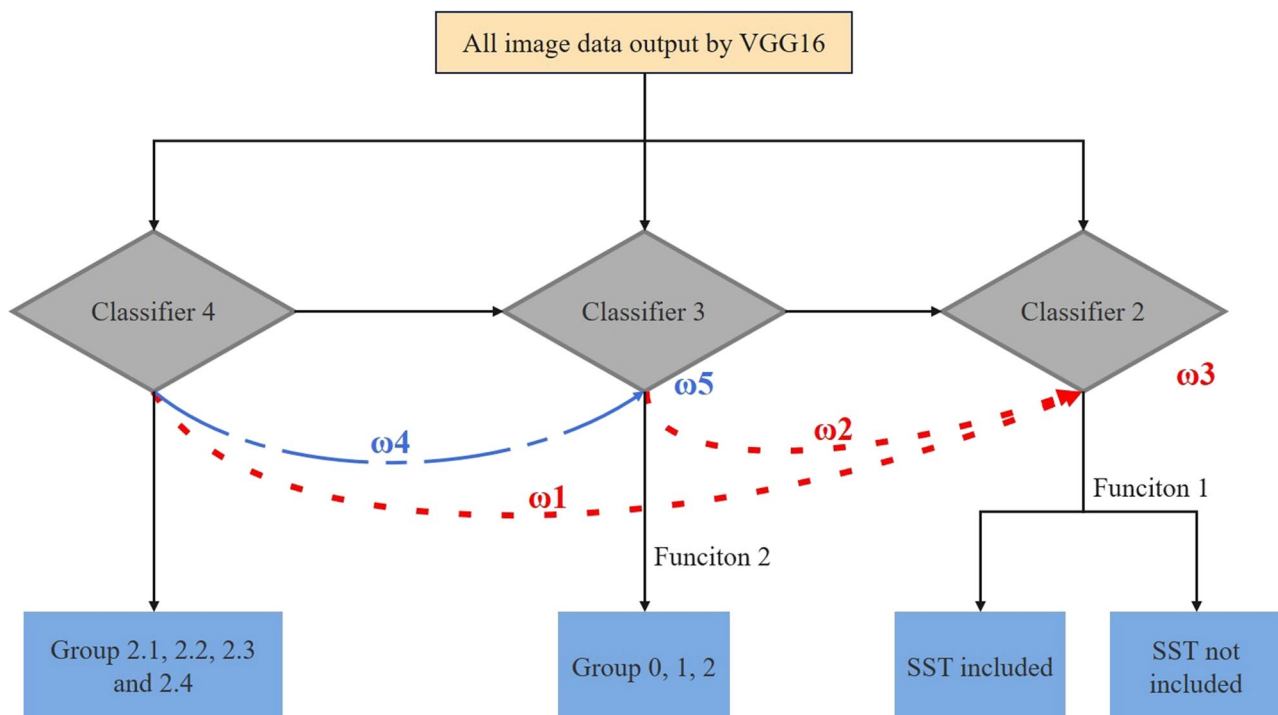


Fig. 3 Calculation method of the multistage decision model. The red dashed line represents the process of combining multiple classifiers to determine the existence of SST, and $\omega 1-3$ represent the weights of classifiers 4, 3, and 2, respectively, where the weight of classifier 2 is the largest. The blue dashed lines represent the combined process of classifying groups 0 and 1, and $\omega 4-5$ represent the weights of classifiers 4 and 3, of which the weight of classifier 3 is the largest

Table 3 Classification results of the four different classifiers of the rotator cuff multitask learning (RC-MTL) model in the in-group dataset

Classifier	Group	AUC (95% CI)	Accuracy	Sensitivity	Specificity	PPV	NPV	Youden Index	F1 score
1	OCOR	0.99 (0.999–1.000)	0.99 (12,458/12,482)	0.99 (6042/6055)	0.99 (6416/6427)	0.99 (6042/6053)	0.99 (6416/6429)	0.99	0.99
	OSAG	0.99 (0.999–1.000)	0.99 (12,458/12,482)	0.99 (6416/6427)	1.00 (6042/6055)	0.99 (6416/6429)	0.99 (6042/6053)	0.99	
2	SST	0.98 (0.974–0.984)	0.95 (11,825/12,482)	0.96 (10,754/11,263)	0.93 (4258/4588)	0.96 (7567/7897)	0.93 (4258/4585)	0.89	0.95
	No-SST	0.98 (0.979–0.986)	0.95 (11,825/12,482)	0.94 (6401/6812)	0.96 (7567/7894)	0.93 (4258/4585)	0.96 (7567/7897)	0.89	
3	NIG	0.98 (0.973–0.980)	0.93 (11,565/12,482)	0.90 (7112/7894)	0.97 (4453/4588)	0.98 (7112/7247)	0.85 (4453/5235)	0.87	0.92
	Group 0	0.99 (0.995–0.997)	0.99 (12,392/12,482)	0.89 (210/236)	0.99 (12,182/12,246)	0.77 (210/274)	0.99 (12,182/12,208)	0.88	
	Group 1	0.97 (0.965–0.974)	0.97 (12,316/12,482)	0.81 (380/467)	0.99 (11,936/12,015)	0.83 (380/459)	0.99 (11,936/12,023)	0.81	
4	Group 2	0.94 (0.932–0.975)	0.93 (11,575/12,482)	0.96 (3740/3885)	0.91 (7835/8597)	0.83 (3740/4502)	0.98 (7835/7980)	0.87	
	NIG	0.97 (0.969–0.978)	0.94 (11,776/12,482)	0.94 (7385/7894)	0.96 (4391/4588)	0.97 (7385/7582)	0.90 (4391/4900)	0.89	0.91
	Groups 0–1	0.97 (0.969–0.985)	0.98 (12,278/12,482)	0.84 (587/703)	0.99 (11,691/11,779)	0.87 (587/675)	0.99 (11,691/11,807)	0.83	
	Group 2.1	0.98 (0.973–0.985)	0.97 (12,057/12,482)	0.88 (923/1049)	0.97 (11,134/11,433)	0.76 (923/1222)	0.99 (11,134/11,260)	0.85	
	Group 2.2	0.96 (0.945–0.996)	0.99 (12,454/12,482)	0.88 (112/127)	0.99 (12,342/12,355)	0.90 (112/125)	0.99 (12,342/12,357)	0.88	
	Group 2.3	0.97 (0.966–0.977)	0.99 (12,451/12,482)	0.97 (128/132)	0.99 (12,323/12,350)	0.83 (128/155)	0.99 (12,323/12,327)	0.97	
	Group 2.4	0.99 (0.991–1.000)	0.94 (11,720/12,482)	0.88 (2269/2577)	0.95 (9451/9905)	0.83 (2269/2723)	0.97 (9451/9759)	0.84	

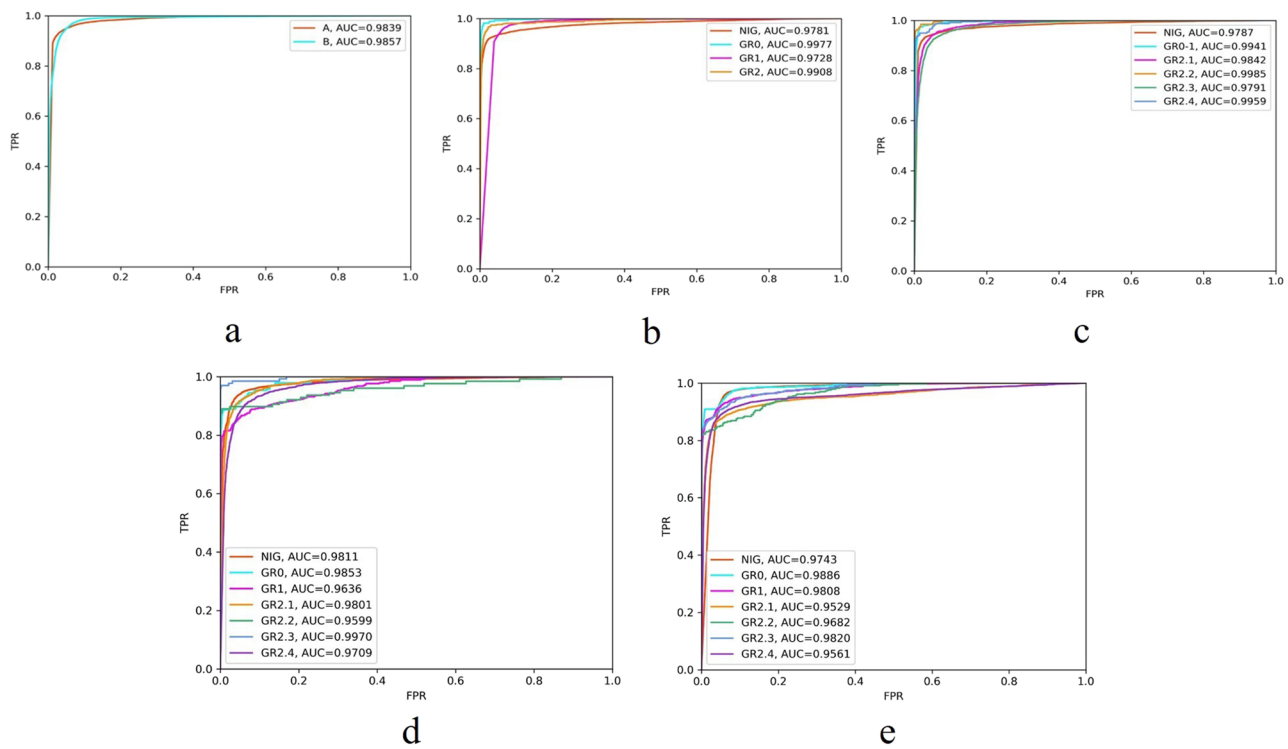


Fig. 4 ROC curves of the RC-MTL model and the automated deep learning system with the in-group and out-group datasets. Panels **a–c** show the ROC curves of the RC-MTL model with the in-group dataset for classifiers 2–4. Categories A and B in panel **a** represent images with and without SST structures, respectively. Panel **d** shows the ROC curves of the automated deep learning system's combined judgement of groups NIG and 0–2.4 with the in-group dataset. Panel **e** shows the ROC curves of the automated deep learning system's combined judgement of groups NIG and 0–2.4 in the out-group dataset

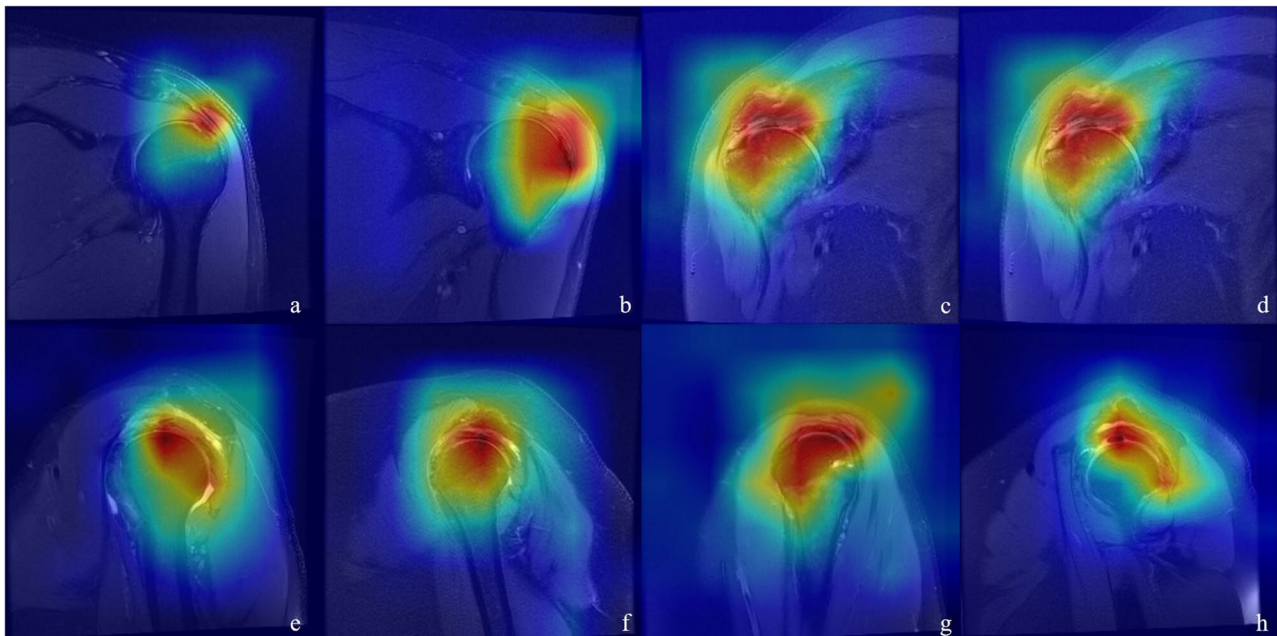


Fig. 5 The Grad-CAM diagram of the last layer of the shared underlying structure in RC-MTL. Redder areas are those to which the model paid more attention, and bluer areas are those to which the model paid less attention. The shared underlying structure can extract common features and correctly map the lesion area instead of perfectly mapping the training data (overfitting). Panels **a–d** show OCOR images, and panels **e–h** show OSAG images

Table 4 Classification results of the automated deep learning system in the in-group and out-group datasets with respect to the arthroscopy results

Dataset	Group	AUC (95% CI)	Accuracy	Sensitivity	Specificity	PPV	NPV	Youden	F1 score
In-group dataset	NIG	0.98 (0.977–0.984)	0.94 (11,705/12,482)	0.92 (7283/7894)	0.96 (4422/4588)	0.98 (7283/7449)	0.88 (4422/5033)	0.89	0.92
	Group 0	0.99 (0.982–0.997)	0.99 (12,391/12,482)	0.87 (205/236)	0.99 (12,186/12,246)	0.77 (205/265)	0.99 (12,186/12,217)	0.86	
	Group 1	0.96 (0.958–0.984)	0.99 (12,301/12,482)	0.80 (374/467)	0.99 (11,927/12,015)	0.81 (374/462)	0.99 (11,927/12,020)	0.79	
	Group 2.1	0.98 (0.971–0.986)	0.96 (12,014/12,482)	0.87 (915/1049)	0.97 (11,099/11,433)	0.73 (915/1249)	0.99 (11,099/11,233)	0.84	
	Group 2.2	0.96 (0.938–0.992)	0.99 (12,443/12,482)	0.87 (111/127)	0.99 (12,332/12,355)	0.83 (111/134)	0.99 (12,332/12,348)	0.87	
	Group 2.3	0.99 (0.990–1.000)	0.99 (12,441/12,482)	0.97 (128/132)	0.99 (12,313/12,350)	0.78 (128/165)	0.99 (12,313/12,317)	0.97	
	Group 2.4	0.97 (0.966–0.975)	0.93 (11,665/12,482)	0.88 (2259/2577)	0.95 (9406/9905)	0.82 (2259/2758)	0.97 (9406/9724)	0.83	
Out-group dataset	NIG	0.97 (0.969–0.977)	0.94 (27,778/29,264)	0.96 (16,008/16,755)	0.94 (11,770/12,509)	0.96 (16,008/16,747)	0.94 (11,770/12,517)	0.90	
	Group 0	0.98 (0.978–0.993)	0.99 (29,004/29,264)	0.87 (609/704)	0.99 (28,395/28,560)	0.79 (609/774)	0.99 (28,395/28,490)	0.86	0.91
	Group 1	0.98 (0.978–0.992)	0.99 (28,865/29,264)	0.84 (975/1161)	0.99 (27,890/28,103)	0.82 (975/1188)	0.99 (27,890/28,076)	0.83	
	Group 2.1	0.95 (0.941–0.960)	0.96 (28,128/29,264)	0.82 (2416/2935)	0.98 (25,712/26,329)	0.80 (2416/3033)	0.98 (25,712/26,231)	0.80	
	Group 2.2	0.97 (0.955–0.991)	0.99 (29,165/29,264)	0.82 (275/336)	0.99 (28,890/28,928)	0.88 (275/313)	0.99 (28,890/28,951)	0.82	
	Group 2.3	0.98 (0.972–0.990)	0.99 (29,130/29,264)	0.84 (327/391)	0.99 (28,803/28,873)	0.82 (327/397)	0.99 (28,803/28,867)	0.83	
	Group 2.4	0.96 (0.951–0.967)	0.94 (27,524/29,264)	0.86 (6027/6982)	0.96 (21,497/22,282)	0.88 (6027/6812)	0.96 (21,497/22,452)	0.83	

respectively; for the combined distinction between group 0 and group 1, the weights of classifiers 3 and 4 are 0.8 and 0.2, respectively. When the discriminations of different classifiers are contradictory, such as when classifier 3 determines group 0 and classifier 4 determines group 2.1, the image is directly judged as NIG.

Since the RC-MTL model judges all images of the patient individually, the output result is the category with the highest judgement probability (excluding the NIG category). In the rare case when two categories have the same judgement probability, the number of images already assigned to both categories is compared. The category with more images is the final output category.

Statistical analysis

Statistical and deep learning analyses were performed using Python (version 3.6.0) and R (version 4.0) languages, and data processing was carried out with the PyTorch (version 1.1.0) library based on dataflow programming. Model performance was evaluated using the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC), 95% confidence intervals (95% CIs, calculated with the cross-validation method),

accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, and Youden index. The intraclass correlation coefficient (ICC) was used to assess diagnostic consistency among radiologists. McNemar’s test was used to compare the differences between the diagnoses of model and radiologists. $p < 0.05$ indicated a statistically significant result.

Results

In the in-group dataset, the AUC of VGG16 for preliminary image screening was 0.92 (95% CI: 0.906–0.924). The AUCs for classifier 1 were all 0.99; those for classifier 2 were 0.98; those for classifier 3 were 0.98, 0.99, 0.97, and 0.94; and those for classifier 4 were 0.97, 0.97, and 0.98. After combining the RC-MTL model with VGG16 and the multistage decision model (automated multitask deep learning system), the AUCs for groups NIG and 0–2.4 were 0.98, 0.99, 0.96, 0.98, 0.96, 0.99, and 0.97, respectively. The detailed results are shown in Table 3, the corresponding ROC curves are shown in Fig. 4, and the Grad-Cam output plots of the automated multitask deep learning system (the last layer of the shared underlying structure) for the in-group dataset are shown in Fig. 5.

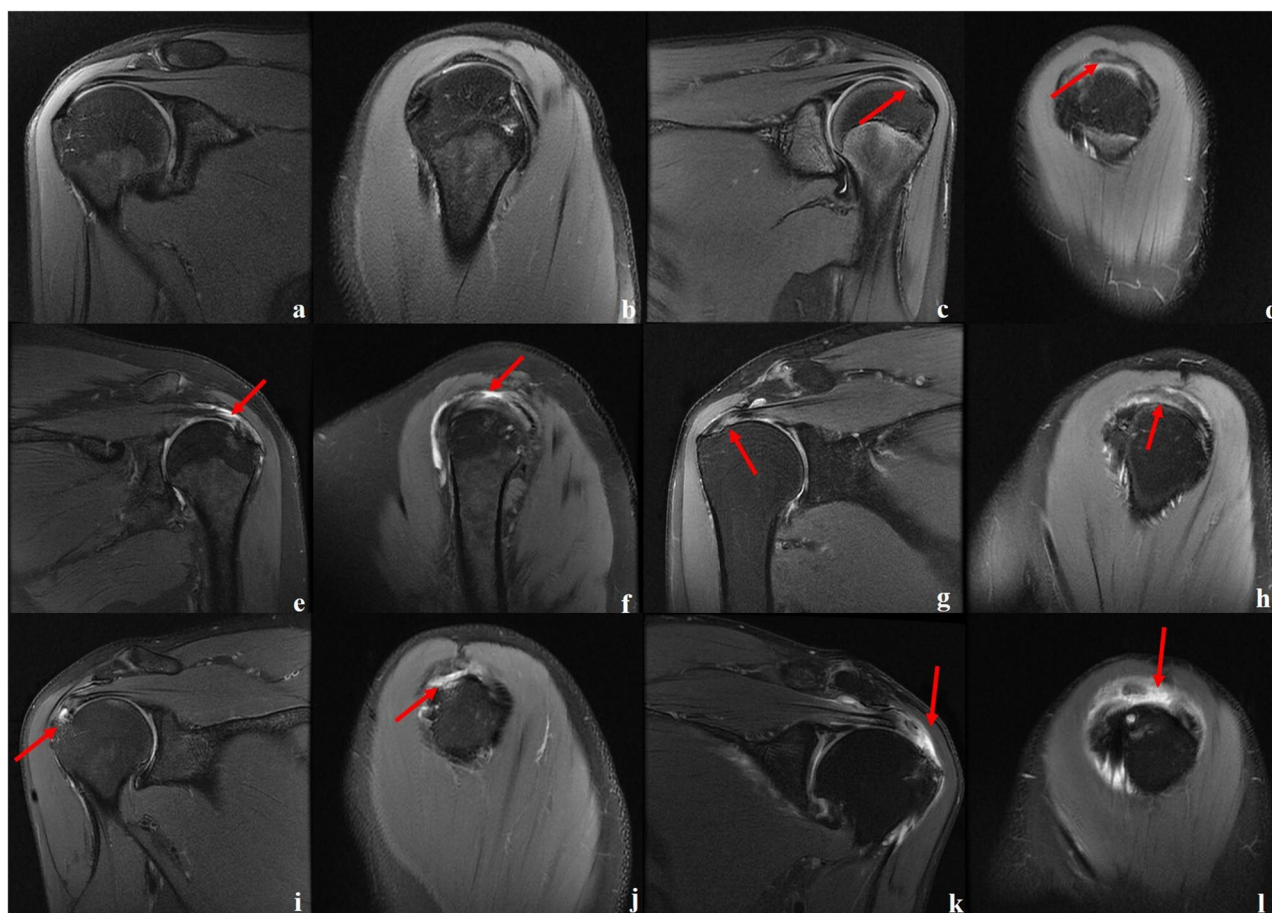


Fig. 6 Normal SST and different types of SST injuries. The red arrow represents the location of the lesion. Panels **a** and **b** represent a normal SST (group 0); panels **c** and **d** represent SST degeneration (group 1), and a slightly high signal can be seen in the SST; panels **e** and **f** represent SST bursal-sided tears (group 2.1), showing that the upper surface of the SST is filled with a liquid signal, which is more obvious on OSAG sequence images; panels **g** and **h** represent SST articular-sided tears (group 2.2), showing a flaky PD-FS high signal on the underside of the SST, accompanied by degenerative changes; panels **i** and **j** represent intratendinous tears (group 2.3), and the internal fluid signal filling of the SST attachment can be seen; panels **k** and **l** show that the SST is completely torn (group 2.4), the SST attachment is completely filled with liquid signals, and the SST stump is slightly retracted

In the out-group dataset, the AUCs of the automated multitask deep learning system for groups NIG and 0–2.4 were 0.97, 0.98, 0.98, 0.95, 0.97, 0.98, and 0.96, respectively. The detailed results are shown in Table 4, and the ROC curve is shown in Fig. 4.

For groups 0–2, four radiologists diagnosed the in-group test set with an ICC of 0.999 (95% CI: 0.997–1.000) and the out-group test set with an ICC of 0.996 (95% CI: 0.990–0.999); the in-group and out-group ICCs for groups 2.1–2.4 were 0.972 (95% CI: 0.954–0.984) and 0.978 (95% CI: 0.965–0.988), respectively. Examples of MR images of different types of SST injury are shown in Fig. 6. In the in-group and out-group test sets, the automated multitask deep learning system showed a better performance than the four radiologists with different

experience levels. The detailed results are shown in Table 5.

Discussion

In this study, we developed an automated multitask deep learning system based on MRI to assess SST injuries. We also compared the model with MSK radiologists with different experience levels. The results show that the system developed in this study has a good diagnostic performance and generalizability and is comparable to experienced radiologists.

Since T1WI is not as sensitive to SST injuries as the PD sequence, PD-FSE-FS sequences were used in this study [25]. Current studies on SST injuries have mainly

Table 5 Comparison of the classification of supraspinatus tendon injuries between four MSK radiologists and the automated deep learning system

Dataset	Reader	Group	Accuracy	Sensitivity	Specificity	F1 score	χ^2*	p value*
In-group test set	Automated multi-task deep learning system	0	0.99	0.87	0.99	0.92		/
		1	0.99	0.8	0.99			
		2.1	0.96	0.87	0.96			
		2.2	0.99	0.87	0.99			
		2.3	0.99	0.97	0.99			
		2.4	0.93	0.88	0.93			
	Radiologist 1	0	0.99	0.85	0.99	0.71	55.84	< 0.001
		1	0.96	0.88	0.96			
		2.1	0.92	0.88	0.94			
		2.2	0.94	0.57	0.95			
		2.3	0.98	0.67	0.99			
		2.4	0.87	0.78	0.96			
	Radiologist 2	0	0.96	0.77	0.97	0.61	71.32	< 0.001
		1	0.93	0.72	0.95			
		2.1	0.91	0.78	0.96			
		2.2	0.96	0.57	0.97			
		2.3	0.94	0.56	0.95			
		2.4	0.81	0.72	0.93			
	Radiologist 3	0	0.95	0.77	0.96	0.56	106.96	< 0.001
		1	0.92	0.64	0.95			
		2.1	0.83	0.6	0.91			
		2.2	0.93	0.57	0.93			
		2.3	0.96	0.44	0.98			
		2.4	0.83	0.78	0.88			
	Radiologist 4	0	0.96	0.69	0.97	0.52	215.57	< 0.001
		1	0.92	0.76	0.94			
		2.1	0.83	0.61	0.9			
		2.2	0.9	0.57	0.91			
		2.3	0.92	0.44	0.94			
		2.4	0.76	0.63	0.91			

focused on classifying normal, partial, and complete tears in SST [16, 17]. However, SST degeneration is prevalent in clinical practice [26]. Ignoring SST degeneration may lead to the limited application of AI models; furthermore, patients with SST degeneration have a higher risk of developing tears [27], which is a disease process that cannot be ignored in SST injuries. Therefore, the system developed in this study treats SST degeneration as an independent category to help identify high-risk patients as early as possible, conduct reasonable interventions, and improve prognosis.

Although a combination of multiple factors can cause SST tears, the common causative factors differ between tear locations [28]. Bursal-sided tears are associated with subacromial impingement syndrome, for which clinicians need to pay more attention to acromion morphology, and articular-sided tears may be associated with

SST degeneration and chronic microtrauma [4]; intra-tendinous tears alone are related to mechanical stress, have a lower incidence, are more challenging to diagnose, and are easily missed [29]. Different types of partial SST tears help determine the arthroscopic approach and are also conducive to developing patient postoperative rehabilitation plans. At the same time, some studies have shown that due to differences in the aetiology of different types of partial tears, there are differences in their treatment options and clinical outcomes [30]. Therefore, the method proposed in this study provides a detailed classification of SST tears based on their location to provide clinicians with more information preoperatively and to assist intraoperative exploration, which facilitates individualized treatment planning.

Our study developed an automated multitask deep learning system that can interface with an MRI

Table 5 (continued)

Dataset	Reader	Group	Accuracy	Sensitivity	Specificity	F1 score	χ^2 *	p value*
Out-group test set	Automated multi-task deep learning system	0	0.99	0.87	0.99	0.91	/	/
		1	0.99	0.84	0.99			
		2.1	0.96	0.82	0.98			
		2.2	0.99	0.82	0.99			
		2.3	0.99	0.84	0.99			
		2.4	0.94	0.86	0.96			
	Radiologist 1	0	0.96	0.83	0.98	0.74	13.35	<0.001
		1	0.93	0.84	0.95			
		2.1	0.88	0.87	0.88			
		2.2	0.95	0.52	0.97			
		2.3	0.98	0.61	0.99			
		2.4	0.88	0.79	0.96			
	Radiologist 2	0	0.96	0.62	1	0.66	23.22	<0.001
		1	0.9	0.8	0.91			
		2.1	0.85	0.8	0.86			
		2.2	0.94	0.59	0.96			
		2.3	0.96	0.56	0.97			
		2.4	0.89	0.76	1			
	Radiologist 3	0	0.95	0.8	0.61	0.61	31.109	<0.001
		1	0.89	0.6	0.95			
		2.1	0.84	0.63	0.88			
		2.2	0.93	0.59	0.94			
		2.3	0.94	0.52	0.96			
		2.4	0.87	0.77	0.96			
	Radiologist 4	0	0.93	0.62	0.97	0.61	43.96	<0.001
		1	0.87	0.74	0.9			
		2.1	0.8	0.62	0.83			
		2.2	0.96	0.59	0.98			
		2.3	0.97	0.46	0.99			
		2.4	0.78	0.66	0.88			

"/" no corresponding value
* Radiologist vs. automated multitask deep learning system in the in-group and out-group test sets

scanning system. The multitask deep learning model increases the number of model parameters and reduces the learning cost of each classifier while outputting multiple classification results. The multistage decision model logically connects and aggregates the multiple results output by the RC-MTL model. Notably, due to the characteristics of the multitask model, each classifier must be able to complete the classification of the input data independently. Therefore, the NIG and group 0–1 categories are inevitably introduced in classifiers 3 and 4. Because of this, we combined the repeated outputs of different classifiers in the multistage decision model to modify the corresponding classification results to more accurately output the final results of the model. Classifier 1 in the RC-MTL model was used to further distinguish the OCOR and OSAG

sequences obtained by VGG16 and accelerate the training of other classifiers using the implicit data enhancement and feature theft aspects of the multitasking deep learning model.

This study has a data bias limitation, and the number of normal patients is small. Although this problem is partially alleviated by weighted loss functions and data augmentation, the impact caused by data bias cannot be eliminated. This problem may cause the system's performance to be overestimated and reduce the performance of distinguishing between normal and abnormal cases. Therefore, the system's performance in real-world cases is still uncertain, and more verification in actual work is needed. An out-group test was conducted to verify the system's performance, which initially proved that the system has the potential to assist in diagnosing SST lesions. For inexperienced, non-MSK radiologists or community physicians, the system could be used as an auxiliary tool to help assess SST injuries and reduce misdiagnosis due to inexperience, fatigue, and equipment differences.

Limitations

First, there was an imbalance in the number of patients included in this study between groups and a discrepancy with the current epidemiologically reported incidence, which may be related to the inclusion of only surgically treated patients in the study. Second, the arthroscopic surgeries were performed by different teams, and differences in outcomes could not be wholly avoided. Finally, no further grading studies of partial and complete tears were performed in this study, but we are currently conducting the relevant research.

Conclusion

The proposed automated multitask deep learning system based on MRI has a good diagnostic performance in assessing SST injuries. The system is comparable to experienced radiologists and has the potential to assist radiologists and clinicians in the detailed preoperative assessment of SST injuries.

Abbreviations

FOV	Field of view
MSK	Musculoskeletal
NEX	Number of excitations
NIG	Not in groups 0–2
OCOR	Oblique coronal
OSAG	Oblique sagittal
PD-FSE-FS	Fat saturation proton density-weighted fast spin-echo
RC-MTL	Rotator cuff multitask learning
ROI	Region of interest
SST	Supraspinatus tendon
TE	Echo time
TR	Repetition time
VGG16	Visual geometry group network 16

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1007/s00330-023-10392-x>.

Below is the link to the electronic supplementary material. Supplementary file1 (PDF 133 kb)

Funding

The National Natural Science Foundation of China, 2171927, Huishu Yuan, Beijing Natural Science Foundation, 7212126, Huishu Yuan, Beijing New Health Industry Development Foundation, XM2020-02-006, Huishu Yuan.

Declarations

Ethical approval

Institutional Review Board approval was obtained (number: IRB00006761-M2020458).

Study subjects or cohorts overlap

None

Methodology

- retrospective
- diagnostic or prognostic study
- multicentre study

Author details

¹Department of Radiology, Peking University Third Hospital, Haidian District, Beijing, People's Republic of China.

Received: 9 June 2023 Revised: 1 August 2023

Accepted: 8 September 2023 Published online: 15 November 2023

References

1. Zhao J, Luo M, Liang G, et al (2021) Risk factors for supraspinatus tears: a meta-analysis of observational studies. *Orthop J Sports Med*, 9(10): 23259671211042826. <https://doi.org/10.1177/23259671211042826>
2. Griffith K M, Hammer L C, Iannuzzi N P, et al (2022) Review of human supraspinatus tendon mechanics. Part I: fatigue damage accumulation and failure. *J Shoulder Elb Surg*, 31(12): 2671–7. <https://doi.org/10.1016/j.jse.2022.06.017>
3. Lawrence R L, Moutzouros V, Bey M J (2019) Asymptomatic rotator cuff tears. *JBJS reviews*, 7(6): e9. <https://doi.org/10.2106/jbjs.Rvw.18.00149>
4. Plancher K D, Shanmugam J, Briggs K, Petterson S C (2021) Diagnosis and management of partial thickness rotator cuff tears: a comprehensive review. *J Am Acad Orthop Surg*, 29(24): 1031–43. <https://doi.org/10.5435/jaaos-d-20-01092>
5. Morag Y, Jacobson J A, Miller B, De Maeseneer M, Girish G, Jamadar D (2006) MR imaging of rotator cuff injury: what the clinician needs to know. *Radiographics* 26(4): 1045–65. <https://doi.org/10.1148/rg.264055087>
6. Zoga A C, Kamel S I, Hynes J P, Kavanagh E C, O'Connor P J, Forster B B (2021) The evolving roles of MRI and ultrasound in first-line imaging of rotator cuff injuries. *AJR Am J Roentgenol*, 217(6): 1390–400. <https://doi.org/10.2214/ajr.21.25606>
7. Roy J S, Bräen C, Leblond J, et al (2015) Diagnostic accuracy of ultrasonography, MRI and MR arthrography in the characterisation of rotator cuff disorders: a systematic review and meta-analysis. *Br J Sports Med*, 49(20): 1316–28. <https://doi.org/10.1136/bjsports-2014-094148>
8. Pierce J, Anderson M (2023) Update on diagnostic imaging of the rotator cuff. *Clin Sports Med*, 42(1): 25–52. <https://doi.org/10.1016/j.csm.2022.08.009>

9. Liu F, Cheng X, Dong J, Zhou D, Han S, Yang Y (2020) Comparison of MRI and MRA for the diagnosis of rotator cuff tears: a meta-analysis. *Medicine* (Baltimore) 99(12): e19579. <https://doi.org/10.1097/md.00000000000019579>
10. Omoumi P, Bafort A C, Dubuc J E, Malghem J, Vande Berg B C, Lecouvet F E (2012) Evaluation of rotator cuff tendon tears: comparison of multi-detector CT arthrography and 1.5-T MR arthrography. *Radiology*, 264(3): 812–22. <https://doi.org/10.1148/radiol.12112062>
11. Bauer S, Wang A, Butler R, et al (2014) Reliability of a 3 T MRI protocol for objective grading of supraspinatus tendinosis and partial thickness tears. *J Orthop Surg Res*, 9(128). <https://doi.org/10.1186/s13018-014-0128-x>
12. Pow R E, Bokor D, Deady L, D'Souza M, Ansari S (2022) Grading the severity of the rotator cuff tendinosis on MRI: assessment of inter-observer agreement and evaluation of a novel objective assessment tool. *J Med Imaging Radiat Oncol*, 66(3): 357–61. <https://doi.org/10.1111/1754-9485.13306>
13. D'Angelo T, Caudo D, Blandino A, et al (2022) Artificial intelligence, machine learning and deep learning in musculoskeletal imaging: current applications. *J Clin Ultrasound*, 50(9): 1414–31. <https://doi.org/10.1002/jcu.23321>
14. Kijowski R, Liu F, Caliva F, Pedoia V (2020) Deep learning for lesion detection, progression, and prediction of musculoskeletal disease. *J Magn Reson Imaging*, 52(6): 1607–19. <https://doi.org/10.1002/jmri.27001>
15. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature*, 521(7553): 436–44. <https://doi.org/10.1038/nature14539>
16. Yao J, Chepelev L, Nisha Y, Sathiadoss P, Rybicki F J, Sheikh A M (2022) Evaluation of a deep learning method for the automated detection of supraspinatus tears on MRI. *Skeletal Radiol*, 51(9): 1765–75. <https://doi.org/10.1007/s00256-022-04008-6>
17. Lin D J, Schwier M, Geiger B, et al (2023) Deep learning diagnosis and classification of rotator cuff tears on shoulder MRI. *Invest Radiol*. <http://https://doi.org/10.1097/rli.0000000000000951>
18. Kim Y J, Kim K G (2022) [Understanding and application of multi-task learning in medical artificial intelligence]. *J Korean Soc Radiol*, 83(6): 1208–18. <https://doi.org/10.3348/jksr.2022.0155>
19. Vandenhende S, Georgoulis S, Proesmans M, Dai D, Gool L V (2020) Revisiting multi-task learning in the deep learning era. *arXiv:2004.13379*. <https://doi.org/10.48550/arXiv.2004.13379>
20. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*. <https://doi.org/10.48550/arXiv.1409.1556>
21. Li C, Yan Y, Xu H, et al (2022) Comparison of transfer learning models in pelvic tilt and rotation measurement in pediatric anteroposterior pelvic radiographs. *J Digit Imaging*, 35(6): 1506–13. <https://doi.org/10.1007/s10278-022-00672-1>
22. Klontzas M E, Vassalou E E, Kakkos G A, et al (2022) Differentiation between subchondral insufficiency fractures and advanced osteoarthritis of the knee using transfer learning and an ensemble of convolutional neural networks. *Injury*, 53(6): 2035–40. <https://doi.org/10.1016/j.injury.2022.03.008>
23. Liu T, Tao D, Song M, Maybank S J (2017) Algorithm-dependent generalization bounds for multi-task learning. *IEEE Trans Pattern Anal Mach Intell*, 39(2): 227–41. <https://doi.org/10.1109/tpami.2016.2544314>
24. Woo S, Park J, Lee J-Y, Kweon I S J a e-p. CBAM: convolutional block attention module 2018, *arXiv:1807.06521*. <https://ui.adsabs.harvard.edu/abs/2018arXiv180706521W>.
25. McCrum E (2020) MR imaging of the rotator cuff. *Magn Reson Imaging Clin N Am*, 28(2): 165–79. <https://doi.org/10.1016/j.mric.2019.12.002>
26. Teunis T, Lubberts B, Reilly B T, Ring D (2014) A systematic review and pooled analysis of the prevalence of rotator cuff disease with increasing age. *J Shoulder Elb Surg*, 23(12): 1913–21. <https://doi.org/10.1016/j.jse.2014.08.001>
27. Miller R M, Thunes J, Maiti S, Musahl V, Debski R E (2019) Effects of tendon degeneration on predictions of supraspinatus tear propagation. *Ann Biomed Eng*, 47(1): 154–61. <https://doi.org/10.1007/s10439-018-02132-w>
28. Sambandam S N, Khanna V, Gul A, Mounasamy V (2015) Rotator cuff tears: an evidence based approach [J]. *World J Orthop*, 6(11): 902–18. <https://doi.org/10.5312/wjo.v6.i11.902>
29. Clavert P, Le Coniat Y, Kempf J F, Walch G (2016) Intratendinous rupture of the supraspinatus: anatomical and functional results of 24 operative cases. *Eur J Orthop Surg Traumatol*, 26(2): 133–8. <https://doi.org/10.1007/s00590-015-1716-0>
30. Thangarajah T, Lo I K (2022) Optimal management of partial thickness rotator cuff tears: clinical considerations and practical management. *Orthop Res Rev*, 14(59–70). <https://doi.org/10.2147/orr.S348726>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.