Patterns

RATING: Medical knowledge-guided rheumatoid arthritis assessment from multimodal ultrasound images via deep learning

Highlights

- RATING is a medical knowledge-guided deep learning system for RA assessment
- It leverages diagnostic paradigm and experience to enhance the robustness
- Self-supervised pretraining ensures reliability even with limited training data
- A clinical reader study demonstrates its effectiveness in assisting RA assessment

Authors

Zhanping Zhou, Chenyang Zhao, Hui Qiao, ..., Feng Xu, Qionghai Dai, Meng Yang

Correspondence

qiaohui@mail.tsinghua.edu.cn (H.Q.), wangqian_pumch@126.com (Q.W.), feng-xu@tsinghua.edu.cn (F.X.), daiqh@tsinghua.edu.cn (Q.D.), yangmeng_pumch@126.com (M.Y.)

In brief

The assessment of rheumatoid arthritis activity with ultrasound images suffers from lower intra-observer and inter-observer agreement as well as considerable time and expense to train experienced radiologists. Taking advantage of rheumatoid arthritis knowledge and clinical experience, we propose that the RATING system provides robust and interpretable predictions to assist in radiologists' decision- making. The generalizability and effectiveness of our system have been validated in both internal prospective and external test datasets, respectively.

Zhou et al., 2022, Patterns *3*, 100592
October 14, 2022
https://doi.org/10.1016/j.patter.2022.100592

Patterns

CellPress
OPEN ACCESS

Article

RATING: Medical knowledge-guided rheumatoid arthritis assessment from multimodal ultrasound images via deep learning

Zhanping Zhou,[1,2,6] Chenyang Zhao,[3,6] Hui Qiao,[2,4,*] Ming Wang,[3] Yuchen Guo,[2] Qian Wang,[3,*] Rui Zhang,[3] Huaiyu Wu,[5] Fajin Dong,[5] Zhenhong Qi,[3] Jianchu Li,[3] Xinping Tian,[3] Xiaofeng Zeng,[3] Yuxin Jiang,[3] Feng Xu,[1,2,7,*] Qionghai Dai,[2,4,*] and Meng Yang[3,*]

[1]School of Software, Tsinghua University, Beijing 100084, China
[2]Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing 100084, China
[3]Department of Ultrasound, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China
[4]Department of Automation, Tsinghua University, Beijing 100084, China
[5]Department of Ultrasound, Second Clinical College of Jinan University, First Affiliated Hospital of Southern University of Science and Technology, Shenzhen People's Hospital, Shenzhen 518020, China
[6]These authors contributed equally
[7]Lead contact
*Correspondence: qiaohui@mail.tsinghua.edu.cn (H.Q.), wangqian_pumch@126.com (Q.W.), feng-xu@tsinghua.edu.cn (F.X.), daiqh@tsinghua.edu.cn (Q.D.), yangmeng_pumch@126.com (M.Y.)
https://doi.org/10.1016/j.patter.2022.100592

THE BIGGER PICTURE Rheumatoid arthritis (RA) has detrimental outcomes, including increased disability and mortality. To enhance the clinical assessment of RA, we propose a rheumatoid arthritis knowledge guided (RATING) system for scoring RA activity from multimodal ultrasound images. It combines the knowledge of clinical diagnosis with deep learning, serving as an example of designing deep learning systems for handling real clinical problems. We further integrated the system into the clinical decision-making process via human-machine collaboration and demonstrated significant improvements in assessment performance. We expect that our research will illuminate the road to human-machine collaboration and help transform clinical diagnostics and precision medicine in a wider range of biomedical research.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Multimodal ultrasound has demonstrated its power in the clinical assessment of rheumatoid arthritis (RA). However, for radiologists, it requires strong experience. In this paper, we propose a rheumatoid arthritis knowledge guided (RATING) system that automatically scores the RA activity and generates interpretable features to assist radiologists' decision-making based on deep learning. RATING leverages the complementary advantages of multimodal ultrasound images and solves the limited training data problem with self-supervised pretraining. RATING outperforms all of the existing methods, achieving an accuracy of 86.1% on a prospective test dataset and 85.0% on an external test dataset. A reader study demonstrates that the RATING system improves the average accuracy of 10 radiologists from 41.4% to 64.0%. As an assistive tool, not only can RATING indicate the possible lesions and enhance the diagnostic performance with multimodal ultrasound but it can also enlighten the road to human-machine collaboration in healthcare.

INTRODUCTION

Rheumatoid arthritis (RA), a systemic and chronic inflammation that mainly affects small joints, has detrimental outcomes on both individuals and society, including increased disability and mortality.[1] According to a treatment-to-target strategy, quantitative assessment of RA activity has been deemed as the key for alleviating the disease burden.[2] Due to its

radiation-free, non-invasive, and cost-effective characteristics, ultrasound (US) examination has been commonly used for the RA assessment in clinical practice,[3] which usually comprises two principal modes: grayscale US (GSUS) and Doppler US (either color [CDUS] or power Doppler [PDUS]). GSUS images are examined to investigate the morphological changes of synovial hypertrophy (SH), while Doppler US images are used to detect the synovial hypervascularity.[4] However, there was a long period with no standardized consensus among radiologists on how to evaluate the GSUS and Doppler US measurements until the European League Against Rheumatisms-Outcomes Measures in Rheumatology Synovitis Scoring (EOSS) system.[5] The EOSS system emphasizes the importance of analyzing both the two-mode images and encourages the use of the 0–3 combined score for evaluating the synovitis in RA. Despite the establishment of the EOSS system, the quantitative assessment of disease activity is still hampered by the lower intra-observer and inter-observer agreement in US diagnostics.[6,7] In addition, it often requires considerable time and expense to train experienced radiologists for RA assessment, and integrating both US modes based on the EOSS system further aggravates this problem.[1]

As an emerging alternative solution, deep learning (DL)[8] has demonstrated great potential in various radiology tasks, including breast cancer prediction,[9,10] thyroid cancer diagnosis,[11] lung cancer screening,[12] cardiac function assessment,[13,14] and musculoskeletal image analysis.[15–17] For RA assessment, several DL models have been applied to X-ray[18] and MRI image interpretation.[19] Recently, Andersen et al.[20] and Christensen et al.[21] developed deep neural networks to predict the synovitis combined score using Doppler US images. Meanwhile, Wu et al.[22] proposed a DL technique to determine the combined score with GSUS images.

Although the feasibility of DL methods has been demonstrated, the clinical applicability of DL-assisted US RA assessment has yet to materialize owing to three major limitations. First, previous studies did not consider the multimodal data integration problem, while the EOSS system handled this by human intelligence and expertise. This leads to the potential weakness in diagnostic accuracy and clinical acceptance. Second, the difficulties in the accumulation and annotation of clinical cases remain a barrier to constructing a large-scale training dataset. Therefore, it is necessary to design a data-efficient DL method for clinical deployment. Third, previous studies have not validated to what extent their systems may actually improve clinical diagnostics. Different from the current model-versus-human comparison, it is more promising that humans could collaborate with DL models in real clinical studies.[23]

To overcome the hurdle of RA assessment in clinical practice, we propose a rheumatoid arthritis knowledge guided (RATING) DL system for scoring the RA activity. First, in the RATING system, we explore the MULTI-Task mUltimoDal Ensemble scheme called MULTITUDE, which fully exploits knowledge in clinical diagnosis. It leverages the complementary advantages of dual-mode US images to follow the evaluation paradigm of the EOSS system and combines the diverse perspectives of multitask models in light of clinical expert consultation. As a consequence, MULTITUDE is able to offer a robust and interpretable score determination. Second, we develop a self-supervised pretraining method to enhance the RATING reliability even under the limited training data condition. The pretraining procedure has been designed to promote the morphological feature understanding of human joints; thus, the RATING system can achieve significant accuracy improvement and generalize well to various imaging settings. Third, we devise DL-assisted RA assessment software and conduct clinical trials with it. This software provides the synovitis score predictions and explainable features to facilitate the clinical decision-making process of radiologists. Experiments have demonstrated that our system can not only greatly improve the scoring accuracy but also illuminate the road to human-machine collaboration in healthcare.

We have constructed and used three datasets in this study. The training dataset contains 752 pairs of US images from 104 patients at Peking Union Medical College Hospital (PUMCH). The prospective test dataset contains 274 pairs of US images from 28 patients at PUMCH. The external test dataset contains 293 pairs of US images from 42 patients at Shenzhen People's Hospital (SZPH).

RESULTS

Build of the RATING system on the training dataset

The overall pipeline of building the RATING system is illustrated in Figure 1. To build the model, we retrospectively collected a training dataset, which consisted of 752 pairs of GSUS and CDUS images from 104 patients. For each pair of US images, three experienced radiologists from PUMCH reviewed the images, annotated the region of interest (ROI) for each US image, and decided the synovial hypertrophy score, the vascularity score, and the combined score. The workflow followed the EOSS guidelines (Table S1) and is shown in Figure S1. Detailed patient demographics and EOSS scoring characteristics are summarized in Table S2.

The RATING system consists of five scoring models that share the same architecture. The training dataset was partitioned into five complementary subsets of an equivalent number of samples, and every four of the five subsets were used to train one of the five scoring models in which the remaining subset was used to validate it. Each of the five scoring models separately predicts a synovial hypertrophy score and a vascularity score, and the combined score is predicted by combining all of the predictions using the MULTITUDE scheme.

To further enhance the robustness, every scoring model predicts the synovial hypertrophy and vascularity scores from multiple binary classification outputs using error-correcting output codes (ECOC)[24] rather than a single multiclass classification. By encoding the synovial hypertrophy and vascularity scores in an error-correcting code format in which each bit corresponds to a separate binary classification, the scoring model may be able to recover from the misclassifications. In this study, we trained two networks for each of the three binary classification tasks, whether the score was greater than 0, 1, and 2, so that each time, ECOC integrates six binary classification outputs.

To leverage the complementary advantages of GSUS images and Doppler US images, we proposed the GS-Doppler feature fusion network. It extracts feature vectors from the GSUS and

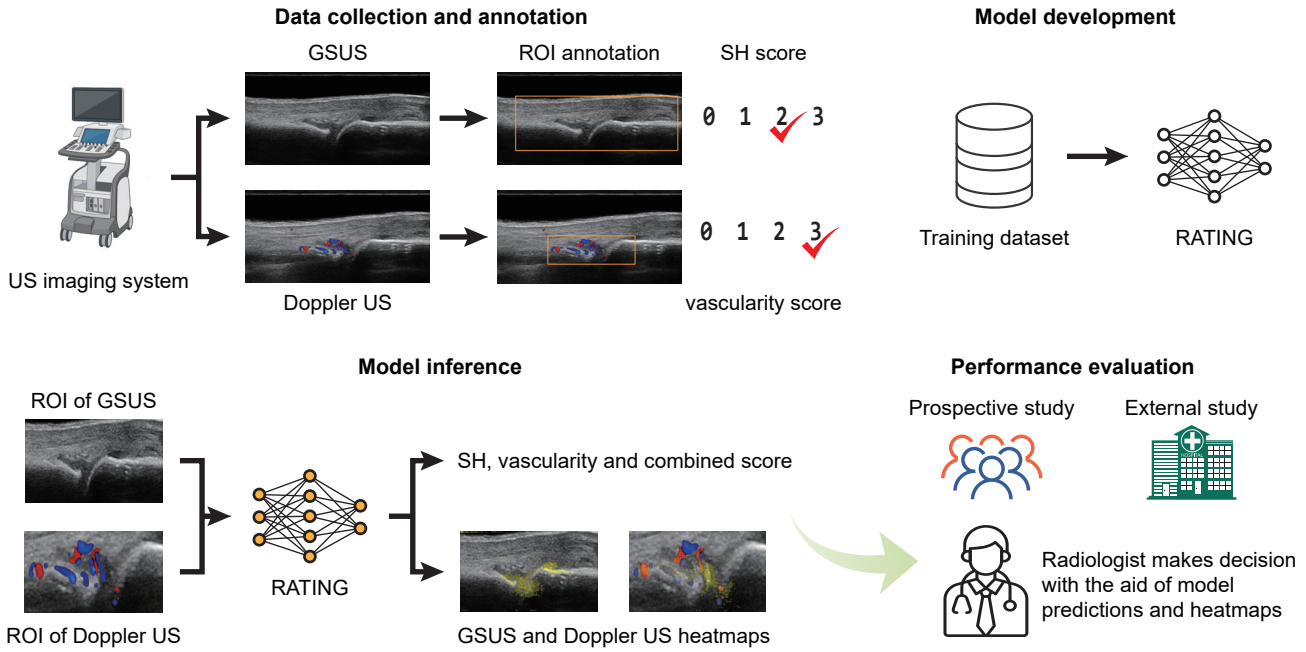


Figure 1. Build of the RATING system for RA scoring

Paired GSUS and Doppler US images were collected for the training dataset, the prospective test dataset, and the external test dataset. Then, ROIs were annotated and scored according to the EOSS system. During model development, the models of RATING system were trained based on the ROIs of US images and the corresponding labels. During model inference, for each pair of GSUS and Doppler US image, the RATING system predicts the synovial hypertrophy score, the vascularity score, and the combined score, and the heatmaps of US images are generated. Performance evaluations were performed on the prospective test dataset and the external test dataset. When used as an assistance tool, the score predictions and heatmaps of the US images are presented to the radiologist.

Doppler US image, concatenates them into a fusion feature vector, and feeds it into a classification network. We used the GS-Doppler feature fusion network to predict the synovial hypertrophy score.

We evaluated the performance of the scoring models on the training dataset by calculating the area under the curve (AUC) of the receiver operating characteristic (ROC) curve for the binary classification tasks. On three synovial hypertrophy score binary classification tasks, the ROC curves on the training dataset are shown in Figure S3 and AUCs were 0.896 (95% confidence interval [CI] = 0.883–0.909), 0.945 (95% CI = 0.935–0.956), and 0.948 (95% CI = 0.932–0.964), respectively. Positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity are shown in Table S3. On three vascularity score binary classification tasks, the ROC curves are shown in Figure S4, and AUCs were 0.980 (95% CI = 0.976–0.984), 0.992 (95% CI = 0.988–0.996), and 0.991 (95% CI = 0.986–0.995), respectively. PPV, NPV, sensitivity, and specificity are shown in Table S4.

Performance of the RATING system on the prospective test dataset

To evaluate the performance of the RATING system in the clinical trial, 28 patients with RA were prospectively recruited from April 20, 2021 to October 2021 and received US examination at PUMCH. After radiologists reviewed and scored the US images, 274 pairs of GSUS and CDUS images were included in the prospective test dataset. There was no patient overlap between the prospective test dataset and training dataset.

We evaluated the performance of the RATING system in two ways. First, we calculated the AUC of the ROC curves for binary classification tasks, including three synovial hypertrophy score binary classification tasks and three vascularity score binary classification tasks. Second, we calculated the four-class accuracy and linearly weighted κ[25] for the synovial hypertrophy score, the vascularity score, and the combined score classification.

On three synovial hypertrophy score binary classification tasks, the ROC curves on the prospective test dataset are shown in Figure S5 and AUCs were 0.930 (95% CI = 0.919–0.941), 0.933 (95% CI = 0.930–0.936), and 0.979 (95% CI = 0.973–0.985), respectively. PPV, NPV, sensitivity, and specificity are shown in Table S5. On three vascularity score binary classification tasks, the ROC curves are shown in Figure S6 and AUCs were 0.986 (95% CI = 0.985–0.987), 0.990 (95% CI = 0.986–0.995), and 0.995 (95% CI = 0.991–0.998), respectively. PPV, NPV, sensitivity, and specificity are shown in Table S6.

The RATING system achieved an accuracy score of 86.1% (95% CI = 82.5%–90.1%) for the combined score prediction and a linearly weighted κ score of 0.853 (95% CI = 0.806–0.900). The confusion matrix is shown in Table S10. For joints of combined scores 0 and 4, the accuracy scores are higher than 90%. It seems to be most difficult for the RATING system to predict the joints of combined score 1, which are frequently predicted to be 0 and 2. As shown in Table S7, the accuracy score of the synovial hypertrophy score and the vascularity score were 79.6% (95% CI = 74.8%–84.3%) and 94.5% (95% CI = 91.6%–97.1%), and the linearly weighted κ scores were 0.757 (95% CI = 0.699–0.885) and 0.919 (95% CI = 0.876–0.966).

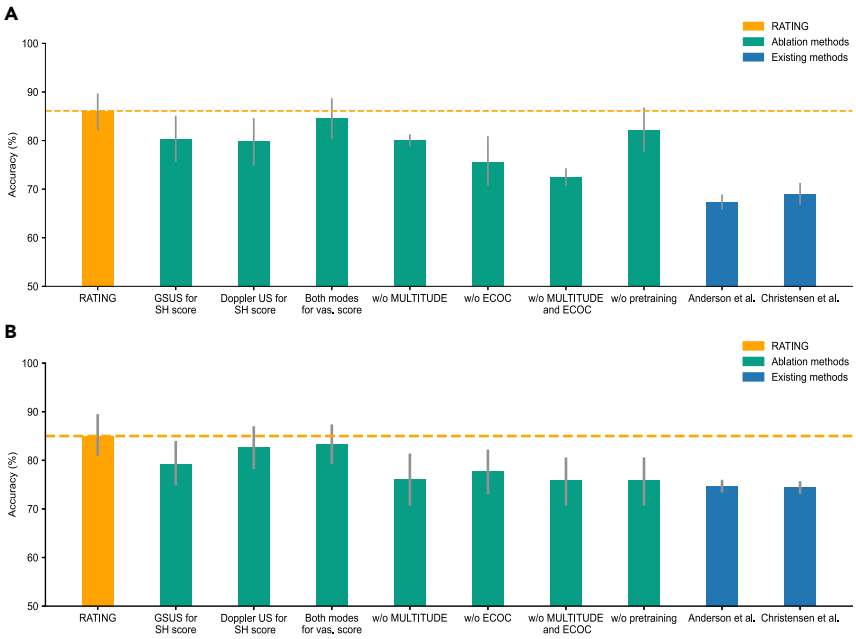


Figure 2. The performance of the RATING system in the classification of the combined score

(A) The RATING system achieved accuracy = 86.1% (95% CI = 82.5%–90.1%) on the prospective test dataset, higher than the ablation methods and the existing methods.

(B) The RATING system achieved accuracy = 85.0% (95% CI = 80.5%–89.1%) on the external test dataset, higher than ablation methods and existing methods. Error bars indicate 95% confidence intervals.

Confusion matrices of the synovial hypertrophy score classification and the vascularity score classification are shown in Tables S8 and S9, respectively.

Performance of the RATING system on the external test dataset

To further demonstrate the generalizability of the RATING system, an external test dataset was collected from SZPH from March 2021 to December 2021. Different from the training dataset and the prospective test dataset, the Doppler US images in the external test dataset are PDUS images rather than CDUS images. After radiologists reviewed and scored the US images, 315 pairs of GSUS and PDUS images from 42 patients were included in the dataset. There was no patient overlap between the prospective test dataset and training dataset.

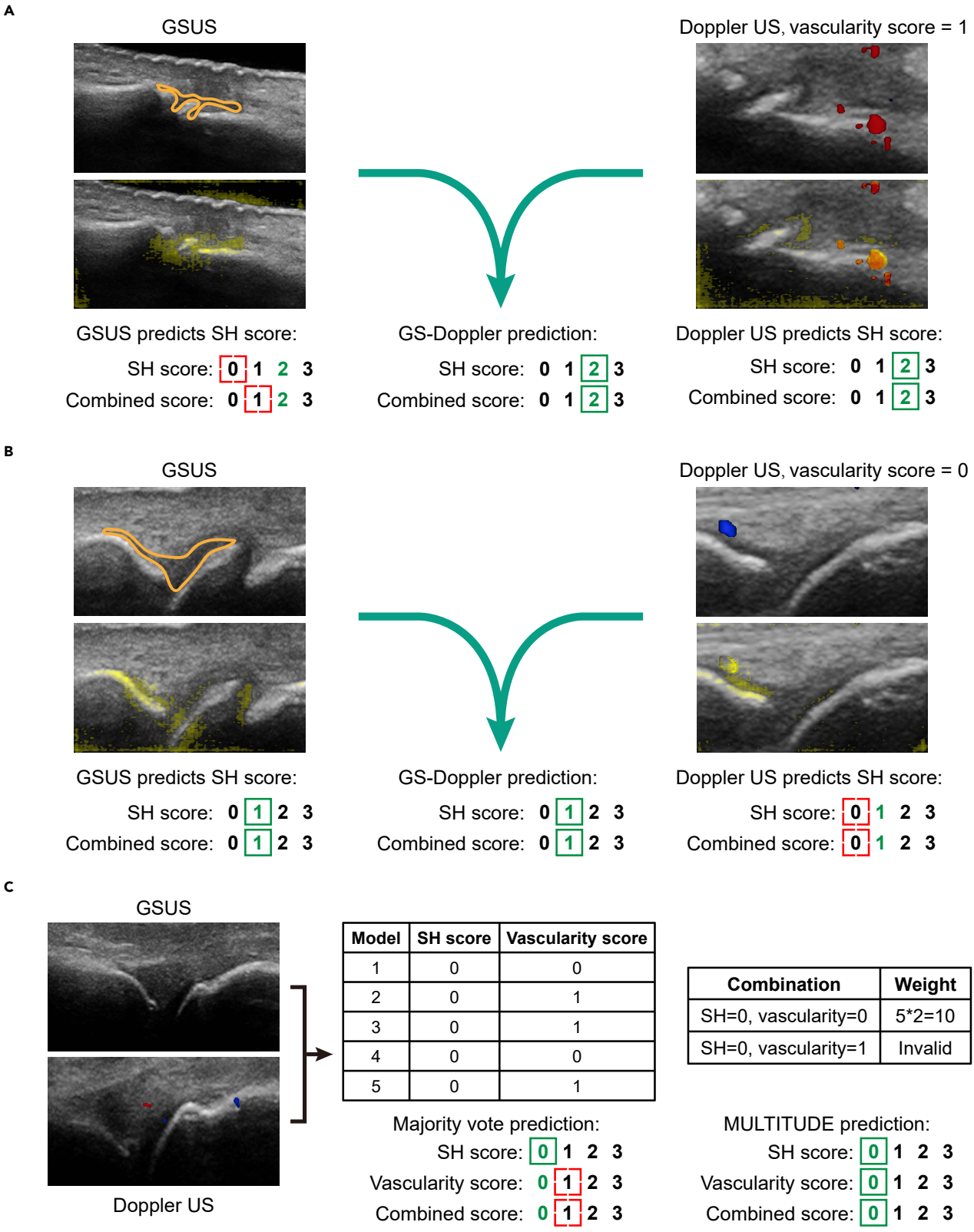
On three synovial hypertrophy score binary classification tasks, the ROC curves on the external test dataset are shown in Figure S7 and AUCs were 0.940 (95% CI = 0.920–0.960), 0.985 (95% CI = 0.983–0.988), and 0.979 (95% CI = 0.974–0.984), respectively. PPV, NPV, sensitivity, and specificity are shown in Table S11. On three vascularity score binary classification tasks, the ROC curves are shown in Figure S8 and AUCs were 0.998 (95% CI = 0.995–1.000), 0.996 (95% CI = 0.994–0.998), and 0.988 (95% CI = 0.974–1.000), respectively. PPV, NPV, sensitivity, and specificity are shown in Table S12.

The RATING system achieved an accuracy score of 85.0% (95% CI = 80.5%–89.1%) for combined score prediction and a linearly weighted κ score of 0.857 (95% CI = 0.817–0.897). The confusion matrix is shown in Table S16. For joints of combined scores 0 and 4, the accuracy scores were higher than 90%. Similar to the result on the prospective test dataset, it seems to be more difficult for the RATING system to predict joints of combined scores 1 and 2. As shown in Table S13, the accuracy score of the synovial hypertrophy score and the vascularity score were 82.9% (95% CI = 78.5%–87.0%) and 96.2% (95% CI = 93.9%–98.3%), and the linearly weighted κ scores were 0.832 (95% CI = 0.789–0.919) and 0.957 (95% CI = 0.932–0.953). Confusion matrices of the synovial hypertrophy score classification and the vascularity score classification are shown in Tables S14 and S15, respectively. The experiments demonstrate that the RATING system generalizes well to different US operators and to PDUS images.

Comparative studies of ablation methods and existing methods

To assess the effectiveness of our method, we conducted ablation studies on the prospective test dataset and external test dataset. For the GS-Doppler feature fusion network, we evaluated three ablation methods: (1) synovial hypertrophy score predicted using only GSUS images, (2) synovial hypertrophy score predicted using only Doppler US images, and (3) vascularity score predicted using both GSUS and Doppler US images. Moreover, we conducted the ablation study for the MULTITUDE scheme, the ECOC method, MULTITUDE and ECOC together, and our self-supervised pretraining strategy. On both the prospective test dataset (Figure 2A) and the external test dataset (Figure 2B), our methods achieved significantly higher accuracy score and linearly weighted κ score than the six ablation methods ($p < 0.001$). The experimental results demonstrate the effectiveness of the MULTITUDE scheme, the self-supervised pretraining strategy, and the GS-Doppler feature fusion network. Detailed results are shown in Tables S7 and S13.

In addition, we compared the performance of the RATING system with the two existing methods proposed by Andersen et al.[20] and Christensen et al.,[21] respectively. For each method, we implemented the model according to the original paper, trained it on our training dataset, and tested it on our prospective test dataset and external test dataset, respectively. To make the experimental results more convincing, we replicated each experiment five times. The method of Andersen et al. directly predicted the combined score using features extracted from Doppler US images, obtaining an accuracy score of 67.4% (95% CI = 65.9%–71.2%) on the prospective test dataset and an accuracy score of 74.7% (95% CI = 73.4%–76.0%) on the external test dataset. Christensen et al. proposed a cascade method that sequentially predicted three binary classification tasks using



Model	SH score	Vascularity score
1	0	0
2	0	1
3	0	1
4	0	0
5	0	1

Combination	Weight
SH=0, vascularity=0	5*2=10
SH=0, vascularity=1	Invalid

Figure 3. Superiority of the GS-Doppler feature fusion network and MULTITUDE

In each GSUS image, the boundary of the synovial hypertrophy area is annotated in orange. The numbers in green are ground truth scores. The green rectangles in the solid line stand for correct predictions, while red rectangles in the dashed line stand for incorrect predictions.

(legend continued on next page)

features extracted from Doppler US images. On the two test datasets, the method of Christensen et al. obtained accuracy scores of 69.0% (95% CI = 66.7%–71.2%) and 74.4% (95% CI = 73.1%–75.7%), respectively. Both the accuracy scores and κ scores of the two existing methods were significantly lower than our RATING system ($p < 0.001$).

The superiority of the GS-Doppler feature fusion network comes from the complementary advantages of the two US modalities. On the one hand, Doppler US images provide information about synovial hypervascularity, which indicates the existence and position of the synovial hypertrophy (Figure 3A). On the other hand, GSUS images do not contain Doppler signals and thus are more suitable for models to focus on morphological characteristics of the synovial hypertrophy (Figure 3B). The superiority of the MULTITUDE scheme comes from a joint analysis of synovial hypertrophy and vascularity score predictions from multiple models. Different from custom model ensemble strategies such as majority voting ensemble that brings together multiple models separately for each classification tasks, MULTITUDE considers the relationship between the tasks (Figure 3C).

Explainability of the RATING system

Explainability of the RATING system is important to understanding how it makes prediction easier and better assists radiologists. For each GSUS image or Doppler US image, the RATING system generates a heatmap that highlights the important areas for deciding the synovial hypertrophy score. The heatmaps of GSUS images may highlight the potential synovial hypertrophy area, and the heatmaps of Doppler US images may highlight the potential synovial hypertrophy area and the potential synovial hypervascularity area. Each heatmap is colorized in yellow and overlaid on the original US image to obtain the heatmap overlay image. The generated heatmap overlay images on the prospective test dataset indicate that the RATING system has learned the features of synovial hypertrophy and blood flow (Figure 4).

Assistance of the RATING system to radiologists

To better assist radiologists for scoring RA, we developed a graphical user interface (Figure 5), which displays the original US images, heatmap overlay images, and the prediction of the RATING system.

To evaluate the usefulness of the RATING system in assisting clinical decision-making, we conducted a reader study and a DL-assisted reader study on the prospective test dataset. We recruited 10 radiologists from PUMCH whose US experience ranged from 4 to 15 years. Detailed experience conditions are shown in Table S17. We first conducted the reader study that each radiologist independently scored the pairs of GSUS and Doppler US images in the prospective test dataset. We evaluated the performance of the RATING system and radiologists in two ways. First, we calculated the four-class accuracy and linearly weighted κ for combined score classification. Second, we calculated the Youden index[26] in three combined score binary classification settings (i.e., 0 versus 1, 2, and 3; 0 and 1 versus 2 and 3; and 0, 1, and 2 versus 3). Besides the 10 human readers, an average reader who achieved the average performance of the 10 real readers was also compared with the RATING system. The RATING system achieved significantly higher combined score accuracy than the 10 radiologists and the average reader ($p < 0.001$; Figure 6A). In all three combined score binary classification settings, the RATING system achieved a significantly higher Youden index than the 10 radiologists and the average reader ($p < 0.001$, Figures 6B–6D).

After 1 week, we conducted the DL-assisted reader study, in which the same group of radiologists scored the same set of images with the assistance of the RATING system. The accuracy of the average reader was significantly improved from 41.4% (95% CI = 35.8%–47.2%) to 64.0% (95% CI = 58.7%–69.5%), and all 10 radiologists achieved significantly higher combined score accuracy than independent assessment with $p < 0.001$ (Figure 6A). Detailed accuracy results of the synovial hypertrophy, vascularity, and combined score are shown in Table S18. As illustrated in Figures 6B–6D, the Youden index of the average reader was significantly improved from 0.226 to 0.520 on the classification of combined score in 0 versus 1, 2, and 3 ($p < 0.001$); from 0.520 to 0.668 on the classification of combined score in 0 versus 1, 2, and 3 ($p < 0.001$); and from 0.492 to 0.660 on the classification of combined score in 0 versus 1, 2, and 3 ($p < 0.001$). Detailed Youden index metrics of all 10 radiologists and the average reader are shown in Table S19.

DISCUSSION

With this work, we developed the RATING system to automatically assess US images for RA scoring. In practice, the US quantitative assessment of disease activity is hampered by low intra-observer and inter-observer agreement during US examination. Moreover, it takes considerable time and expense to train experienced radiologists in US diagnostics. The RATING system is designed to assist radiologists in clinical practice, which not only provides predictions of all three scores rather than only the combined score, but also generates heatmaps to indicate the areas that the DL models focus on. We developed a graphical user interface for radiologists, and conducted a DL-assisted reader study to evaluate the effectiveness of its assistance to the radiologists in clinical trial settings. Experiments demonstrate that the RATING system has great potential in improving the intra-observer and inter-observer agreement for radiologists of various US examination experience.

Previous studies have not evaluated the performance of DL models in the prospective setting; thus, they are not able to fully prove the effectiveness of DL models in clinical practice. Therefore, we collected the prospective test dataset to evaluate the

(A) A sample of combined score grade 2 was underestimated as grade 1 using only the GSUS image. With the aid of synovial hypervascularity information in the Doppler US image, the RATING system made the correct prediction.
(B) A sample of combined score grade 1 was underestimated as grade 0 using only the Doppler US image. With the overall morphological changes of synovial hypertrophy in the GSUS image, the RATING system made the correct prediction.
(C) A sample of combined score grade 0 was incorrectly predicted as grade 1 by custom majority voting ensemble. MULTITUDE excluded the invalid score combination and led to the correct prediction.

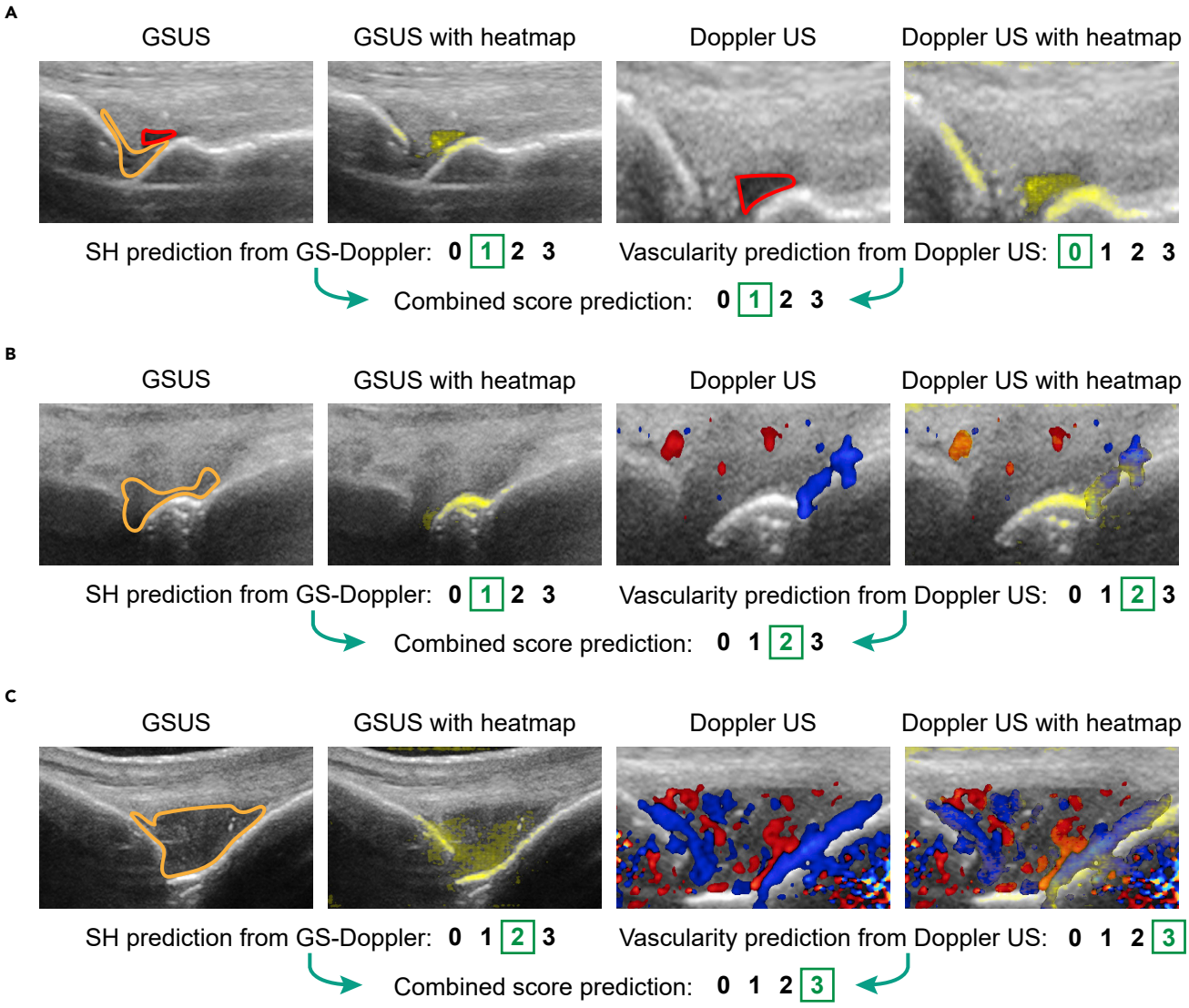


Figure 4. Examples of heatmap visualization

In each GSUS image, the boundary of the synovial hypertrophy area is annotated in orange, and the boundary of the joint effusion area is annotated in red. In each heatmap overlay image, the heatmap is colorized in yellow and overlaid on the original US image.

(A) A sample whose synovial hypertrophy score is 1, vascularity score is 0, and combined score is 1. The joint effusion areas are highlighted in the heatmaps of both GSUS and Doppler US images.

(B) A sample whose synovial hypertrophy score is 1, vascularity score is 2, and combined score is 2. The synovial hypertrophy area near the bone surface is highlighted in the GSUS image and the Doppler US image, and the blood flow areas are highlighted in the Doppler US image.

(C) A sample whose synovial hypertrophy score is 2, vascularity score is 3, and combined score is 3. The synovial hypertrophy area is highlighted in the GSUS image, and the blood flow areas are highlighted in the Doppler US image. The heatmap shows what the RATING system pays attention to, which helps human radiologists understand the predictions of RATING.

model performance and conduct the reader study. The RATING system achieved high accuracy and linearly weighted κ and significantly outperformed the ablation methods, the existing methods, and the 10 radiologists, demonstrating its clinical effectiveness.

To demonstrate the generalizability to different US operators and PDUS images, the RATING system was further evaluated on paired GSUS and PDUS images collected from SZPH. The RATING system achieved a comparable accuracy score of 85.0% (95% CI = 80.5%–89.1%) compared to the accuracy score of 86.1% (95% CI = 82.5%–90.1%) on the prospective test dataset, and achieved a comparable linearly weighted κ score of 0.853 (95% CI = 0.806–0.900) compared to the linearly weighted κ score of 0.857 (95% CI = 0.817–0.897) on the prospective test dataset. The results demonstrate that the RATING system generalizes well to different US operators and to both CDUS and PDUS images.

Although the GS-Doppler feature fusion network achieved the best performance in predicting the synovial hypertrophy score, it did not show superiority in predicting the vascularity score. This is because GSUS images contain no information

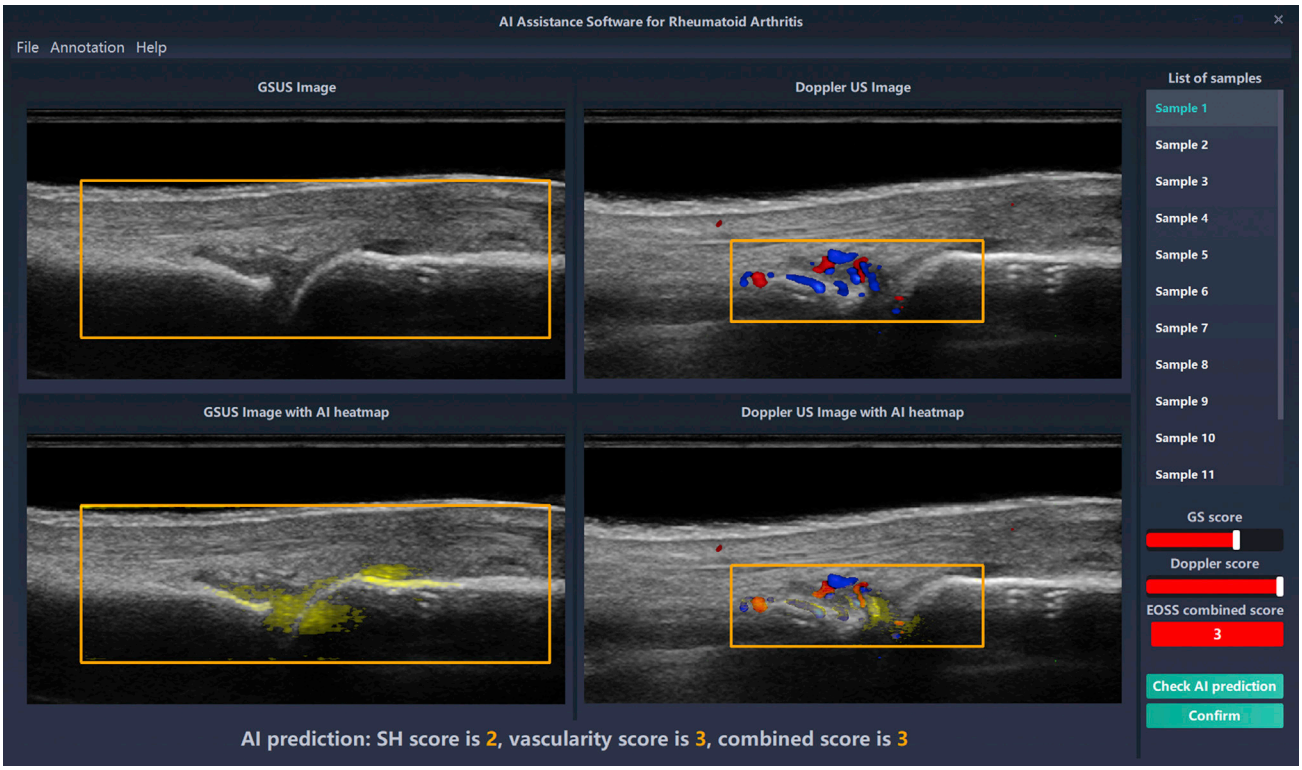


Figure 5. The graphical user interface of the RATING system to assist radiologists for scoring RA

The paired GSUS and Doppler US images are shown in the first row, and the ROI of each image is illustrated by an orange rectangle. When the radiologist clicks the button to check the predictions of the RATING system, the heatmap overlay images are presented in the second row, and the predictions appear at the bottom.

about synovial hypervascularity, and radiologists only examine the Doppler US images to assess the synovial hypervascularity condition and decide the vascularity score in clinical practice. Furthermore, the pointless use of GSUS images also increases the overfitting risk of the DL models. Experiments also demonstrate that using the GS-Doppler feature fusion network obtains lower accuracy and linearly weighted κ in classifying the vascularity score.

It should be noted that the methods used in the RATING system can easily be extended to other medical examination tasks. For multimodal US evaluations such as assessment of breast cancer and thyroid cancer, a feature extraction network can be built for each imaging mode, and their features can be fused in a way similar to that of the GS-Doppler feature fusion network. The MULTITUDE scheme is also appropriate for other medical diagnosis in which the radiological decision consists of multiple related components, such as the TNM staging system for malignant tumor classification, which includes primary tumor (T), regional lymph node (N), and distant metastasis (M).[27]

To better understand when and why the RATING system may make incorrect predictions, we analyzed the incorrect cases in the prospective test dataset and external test dataset. Based on the confusion matrices, we analyzed the four most common types of incorrect predictions and illustrate one typical example for each type (Figure 7). Incorrect predictions may occur when the situation is on the borderline between adjacent grades. When this happens, human experts should carefully analyze the US images according to the EOSS guidelines.

One limitation of our study regards the test data. Because we collected data from only two hospitals in China, the generalizability of the RATING system to other population has not been demonstrated. Moreover, all of the US images are acquired using Mindray US machines; therefore, the generalizability of the RATING system to US machines from other manufacturers has not been demonstrated. Future evaluation of the RATING system should include data from other population and other US machine manufacturers.

Another potential limitation of our study is that the RATING system consists of 60 convolutional neural networks, which may consume considerable inference time and result in applicability issues. Since these networks are independent of one another, one feasible solution is to infer the input US images in parallel using multiprocessing technique. Lastly, we have not compared the RATING system with the method of Wu et al.,[22] which requires additional segmentation annotations.

In conclusion, we propose the RATING system to evaluate the RA activity from GSUS and Doppler US images. We demonstrate that the RATING system outperforms all existing methods and can be implemented well in clinical trial settings. Moreover, we demonstrate that RATING generalizes well to different US operators and both CDUS and PDUS images. The RATING system is

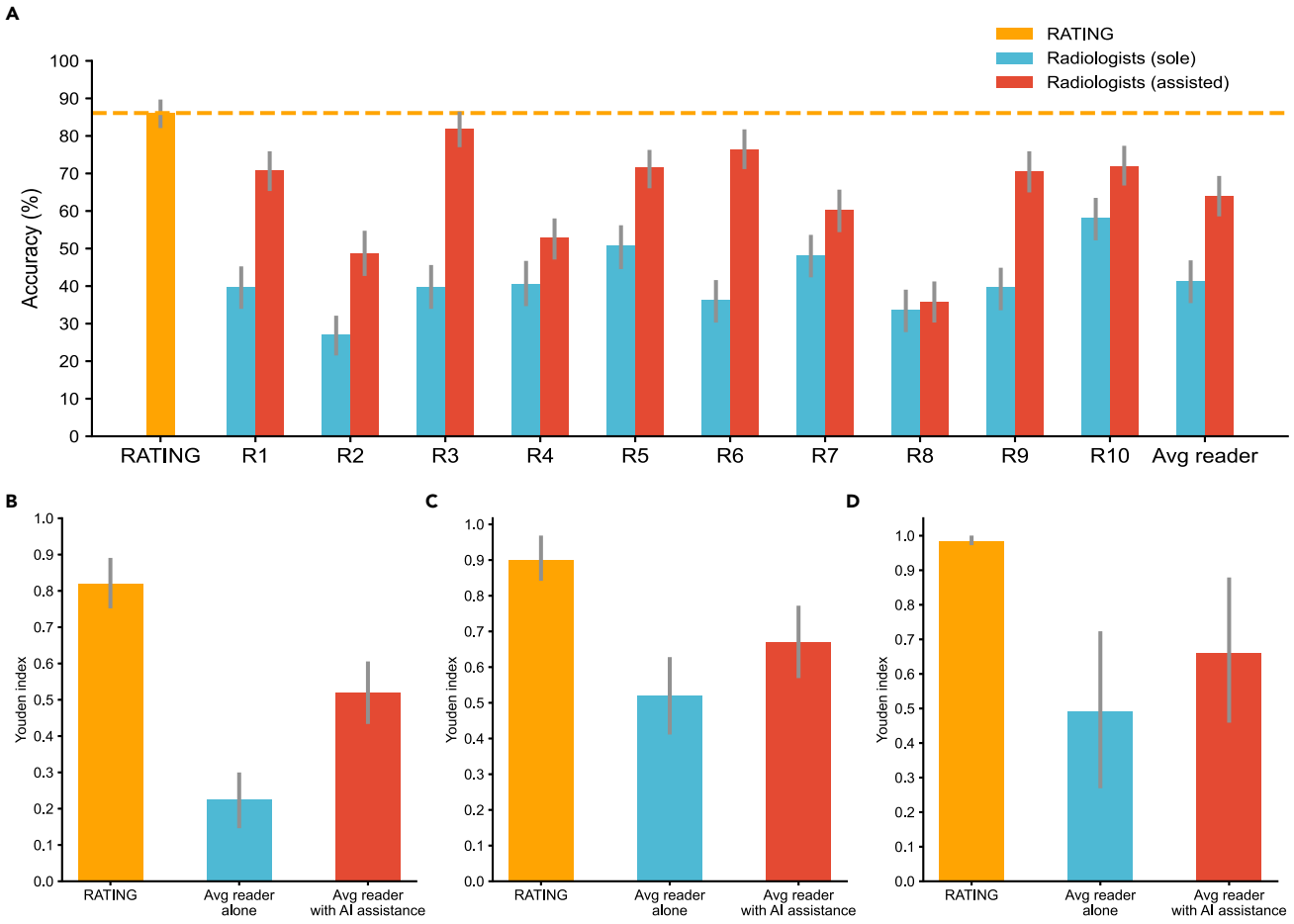


Figure 6. Performance comparison of the RATING system, radiologists alone, and with the assistance of the RATING system

(A) With the assistance of the RATING system, radiologists (R1–R10) and the average reader achieved higher accuracy in the classification of combined score. (B–D) The Youden index of radiologists' combined score binary classification without and with the assistance of the RATING system: 0 versus 1, 2, and 3 (B); 0 and 1 versus 2 and 3 (C); and 0, 1, and 2 versus 3 (D). Error bars indicate 95% confidence intervals.

interpretable and has showed great potential in assisting radiologists in clinical RA assessment. The future looks promising for incorporating the RATING system into US machines, displaying heatmaps and model predictions on the screen in real time. In addition to the clinical assistance, the RATING system may also be used to train junior radiologists.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Feng Xu, feng-xu@tsinghua.edu.cn.

Materials availability

This study did not generate new unique reagents.

Data and code availability

The original data reported in this study cannot be deposited in a public repository because biogenetic information involving humans collected during the project, including imaging data of patients, should be kept confidential until the end of the projects that are supported by the National Natural Science Foundation of China. All of the original code has been deposited at Zenodo under the https://doi.org/10.5281/zenodo.7005383 and is publicly available as of the date of publication. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Methods

Ethical approval

The study was registered at clinicaltrials.gov (NCT04297475) and approved by the institutional review boards of PUMCH (approval no. JS-1923). The prospective study was an observational one and did not involve interventional methods. The recruited patients of the retrospective and prospective parts were well informed of the study and provided signed informed consent. Patients or the public were not involved in the design, recruitment, and conduct of the study.

Patients and data collection

To build the training dataset, we retrospectively collected GSUS and Doppler US images from patients who were diagnosed with RA according to the 2010 American College of Rheumatology/European League Against Rheumatism (ACR/EULAR) classification criteria between March 2019 and April 10, 2021 at the PUMCH. To evaluate the performance of the RATING system, we prospectively recruited patients with RA between April 20, 2021 and October 2021 at PUMCH. To further evaluate the generalizability of the RATING system to different US operators and to PDUS images, we recruited patients between March 2021 and December 2021 at SZPH. In the retrospective, prospective, and external workflows, the US images of the patients who had comorbid inflammatory joint diseases were excluded. Details of data collection workflow are illustrated in Figure S1.

A total of four US operators at PUMCH and SZPH performed US scanning. Two US operators from PUMCH both have 5 years of experience in musculoskeletal US (MSK-US), with approximately 1,000 MSK-US scanning cases per year. The other two US operators from SZPH have 4 and

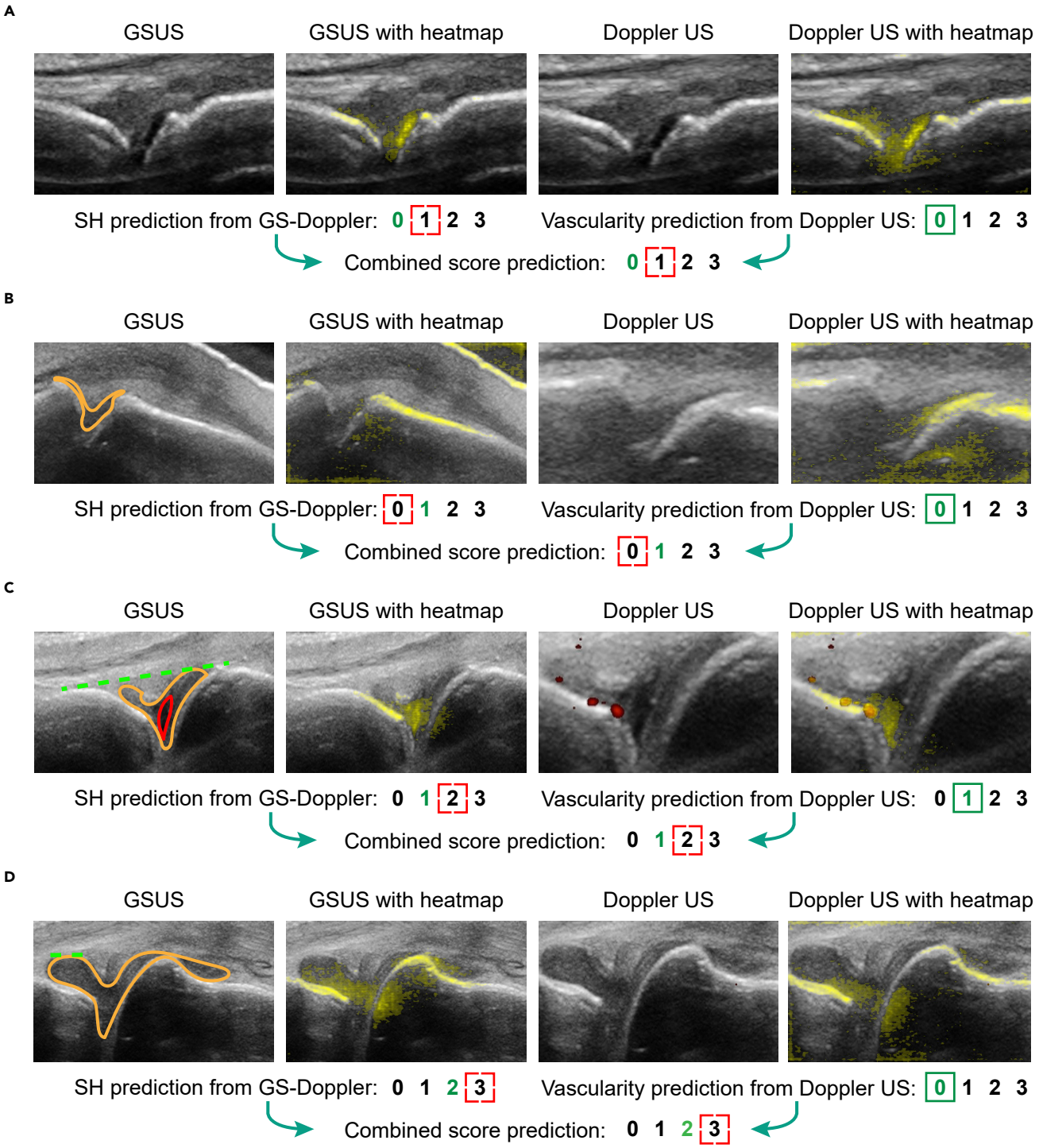


Figure 7. Typical examples of incorrect predictions

In each GSUS image, the boundary of the synovial hypertrophy area is annotated in orange, and the joint effusion area is annotated in red. In each heatmap overlay image, the heatmap is colorized in yellow and overlaid on the original US image. The numbers in green are ground truth scores. The green rectangle in the solid line stands for correct predictions, while the red rectangle in the dashed line stands for incorrect predictions.

(A) The sample of grade 0 was incorrectly predicted as grade 1. The model correctly identified the mild synovial hypertrophy in both GSUS and Doppler US images, but overestimated it and predicted the combined score as 1.

(B) The model underestimated the mild synovial hypertrophy and incorrectly predicted the sample of grade 1 as grade 0.

(C) The sample of grade 1 was incorrectly predicted as grade 2. Although there is obvious synovial hypertrophy and effusion, they do not exceed the joint line across the left and right bones illustrated by the green dashed line.

(legend continued on next page)

5 years of experience in MSK-US, with approximately 700 MSK-US scanning cases per year. All four US operators received training in the standard scanning protocol for small joints. In both sites, the recruited patients received US scanning of metacarpophalangeal (MCP) joints and proximal interphalangeal (PIP) joints using the same commercial US system (Resona 7, Mindray Bio-Medical Electronics) and probe (L23-15 MHz, central frequency of 20 MHz, Mindray Bio-Medical Electronics). For each patient, GSUS and Doppler US images were performed consecutively at a depth of 1.5–2 cm. The CDUS settings included pulse repetition frequency (PRF) of 1,000 Hz, wall filter of 80 Hz, maximum gain of 50, and scale of 3 cm/s. The PDUS settings included PRF of 700 Hz, wall filter of 37 Hz, scale of 3 cm/s, and maximum gain of 50. Both Doppler US settings use a rectangle sampling box with no angulation. During the examination, patients placed their hands on the white surface of examining tables with a bubble-free gel pad put on the dorsal side of the hands. The operator positioned the probe longitudinally on the dorsal surface of the patient's fingers. The static GSUS image showing the bone surfaces of the long bones at both ends and synovial area clearly was saved, and subsequently the corresponding Doppler US image on the same section was also saved. The GSUS and Doppler US images were exported from US machines in either JPEG or TIFF format.

After image collection, three experienced radiologists at PUMCH selected and scored the GSUS and Doppler images. The three radiologists have 10, 7, and 6 years of experience in US and 5, 5, and 4 years of experience in MSK-US, reading more than 1,000 sets of MSK-US images per year. Before the study, all of them participated in a 6-month training program about the EOSS system. The exclusion criterion for the images included (1) images with significant artifacts (GSUS: blurred images, anisotropic artifacts; Doppler US: aliasing, motion artifacts) and (2) images not clearly showing bone surfaces and synovium. Then, the radiologists scored the selected GSUS and Doppler US images according to the EOSS system.[5] After scoring all of the images, the three radiologists discussed the images, with inconsistent scores, to reach consensus. When disagreement still existed, another professional radiologist at PUMCH with 15 years of experience in US and 10 years of experience in MSK-US re-evaluated the images and made the final decision. All of the radiologists and US operators participating were blinded to the patients' clinical information.

Image preprocessing

To eliminate irrelevant information, for each GSUS image and Doppler US image, the ROI was annotated as a rectangular area by radiologists using a custom annotation software. Then, the ROIs of the US images were cropped and resized to 224 × 224 pixel size. To build a model that generalizes to both CDUS and PDUS images, each Doppler US image was segmented to obtain a binary mask of the Doppler signals. The binary mask was the same size as the original Doppler US image and its pixel values were either 0 or 1, where 1 indicated the existence of Doppler signals. Afterward, a transformed Doppler US image was generated for each Doppler US image by transforming the pixel color to red where the corresponding pixel in the binary mask was 1.

Overall pipeline of the RATING system

The RATING system (Figure S2) is composed of five scoring models that separately predict the synovial hypertrophy score and the vascularity score. The combined score is predicted by comprehensively considering five scoring models' predictions using our proposed MULTITUDE scheme.

Each of the five scoring models is composed of a synovial hypertrophy scoring module that predicts the synovial hypertrophy score, as well as a vascularity scoring module that predicts the vascularity score. Instead of directly predicting the scores using multiclassification networks, we adopted a technique called ECOC that uses a series of binary classification models.[24] Specifically, synovial hypertrophy scoring modules and vascularity scoring modules trained models for three binary classification tasks (i.e., whether the score was greater than 0, 1, and 2. Theoretically, at least three binary classification models are needed in a four-class classification task. To improve accuracy and robustness, for each binary classification task, we trained two models with the same training settings except for randomization seeds, resulting in a total of six classification networks.

For the binary classification networks of the synovial hypertrophy scoring modules, we propose the GS-Doppler feature fusion network, which leverages the complementary advantages of the GSUS image and the Doppler US image. For the binary classification networks of the vascularity scoring modules, only the Doppler US image is used to predict the vascularity score, because the Doppler US image is sufficient to decide the vascularity score in clinical practice.

MULTITUDE

Model ensemble is a type of machine learning technique that combines a number of weak learners to achieve better performance than each individual learner. Custom model ensemble methods such as majority voting, stacking,[28] and AdaBoost[29] solve only one prediction task at a time. Recently, the idea of ensemble has been introduced to multitask learning,[30] but each task is predicted separately. Our proposed MULTITUDE scheme develops the problem formulation of custom model ensemble methods from independent tasks to multiple corelated tasks, and naturally combines the medical knowledge into the method design. MULTITUDE has a general formulation and can be used in tasks other than US RA assessment.

In general, we define a classification task in which a total of t measurements $S_1, S_2, \ldots, S_t$ need to be classified. For any measurement $S_i (i = 1, 2, \ldots, t)$, we denote all of its c_i possible values as $v_1^i, v_2^i, \ldots, v_{c_i}^i$. For the above task, m models $M_1, M_2, \ldots, M_n$ are built to represent m independent experts. For any model $M_j (j = 1, 2, \ldots, m)$, it independently predicts the t measurements as $p_j^1, p_j^2, \ldots, p_j^t$ where $p_j^i \in \{v_k^i\}_{k=1}^{c_i}$. We define q_c^i as the number of models that predicts $S_i (i = 1, 2, \ldots, t)$ as $c \in \{v_j^i\}_{j=1}^{c_i}$. It should satisfy Equation 1:

$$\sum_{j=1}^{c_i} q_{v_j^i}^i = m \qquad \text{(Equation 1)}$$

We use the term *value combination* to refer a possible condition of the t measurements, which is represented as a tuple $(v^1, v^2, \ldots, v^t)$. Theoretically, there are $\prod_{i=1}^{t} c_i$ possible value combinations of the t measurements. However, some value combinations contradict domain knowledge. MULTITUDE figures out all of the valid combinations and calculates a weight for each valid combination. The weight of a valid combination $(v^1, v^2, \ldots, v^t)$ quantifies the level of agreement that the models reach on it, which is defined as $\prod_{i=1}^{t} q_{v^i}^i$. A larger weight of a valid combination indicates that more models agree on it. Finally, the predictions of the t measurements are obtained as the valid combination with the largest weight. If more than one valid combination gains the largest weight, the first one in the alphabetical order is selected.

Different from the custom majority voting ensemble strategy that separately makes predictions for each measurement, MULTITUDE takes advantage of relationships between different measurements. As a result, invalid value combinations are excluded by MULTITUDE, leading to a more reasonable prediction. If all of the value combinations are valid, MULTITUDE can yield the same prediction as the majority voting ensemble strategy.

As for RA assessment, the combined score is decided by the synovial hypertrophy score and the vascularity score, both of which range from 0 to 3, resulting in 16 theoretically possible combinations. According to the EOSS system, three combinations whose synovial hypertrophy score is 0 and vascularity score is >0 are invalid. We used $m = 5$ models in this study.

GS-Doppler feature fusion network

The GS-Doppler feature fusion network (Figure S2C) comprehensively analyzes the GSUS image and Doppler US image of a sample. It is composed of a GSUS feature extraction network F, a Doppler US feature extraction network G, and a fusion classification network H. For a sample of a GSUS ROI x_G and a transformed Doppler US ROI x_D, F extracts a 512-dimension feature vector h_G from x_G, and G extracts a 512-dimension feature vector h_D from x_D. h_G and h_D contains potentially useful information in the GSUS image and the transformed Doppler US image for predicting the synovial hypertrophy score. Subsequently, h_G and h_D are concatenated into a 1,024-dimension feature vector h_{fusion} and it is fed into H to predict the binary classification

(D) The sample of grade 2 was incorrectly predicted as grade 3. The synovial hypertrophy score is determined by expert radiologists as 2 rather than 3 because the surface of the left synovial hypertrophy area is only slightly convex rather than obviously convex, which is just on the borderline between grades 2 and 3. The green dashed line illustrates the synovial hypertrophy surface line, which is approximately horizontal.

result. In this study, we used the ResNet-18 network[31] for F and G, and built a two-layer multilayer perceptron (MLP) as H.

The GS-Doppler feature fusion network is trained in a three-stage manner. In the first stage, a self-supervised pretraining method is adopted to train a ResNet-18 network that learns both a feature mapping of joint parts in US images and their correct spatial arrangement.[32] To be specific, the ROI of each GSUS image is resized to 225 × 225 and split into a 3 × 3 grid. Then, a 64 × 64 tile is randomly cropped from each 75 × 75 grid cell. These 9 tiles are reordered via a randomly chosen permutation from 1,000 predefined permutations and then fed to the ResNet-18 network to obtain 9 feature vectors. Finally, these feature vectors are concatenated and fed into a MLP to predict probabilities that the chosen permutation belongs to 1,000 predefined permutations. In the second stage, the pretrained ResNet-18 networks in the first stage is maintained as initial weights and further trained to extract features from the GSUS and transformed Doppler US images for predicting the synovial hypertrophy score. These networks are kept as F and G. In the third stage, with the parameters in F and G frozen, H is trained for the synovial hypertrophy score binary classification task.

Training details

We implemented the networks using the PyTorch DL framework. Cross-entropy loss was used to optimize classification networks. All of the networks were optimized using an adaptive moment estimation (ADAM) optimizer[33] in a batch size of 64 with an initial learning rate of 0.0003, which then decayed every 15 epochs, with a decay factor of 0.3. To address the class imbalance issue, training images were randomly resampled at the beginning of every epoch so that there was the same amount of training samples in different classes. Data augmentation was also performed so that the training images were augmented by applying random cropping and color jittering. To aid the training of feature extraction networks and the vascularity score classification networks, we adopted the transfer learning strategy. Specifically, we initialized network parameters using the model pretrained on ImageNet[34] and then trained on the target tasks.

Heatmap generation

To ensure trust by human experts and assist radiologists in the clinical setting, heatmaps were calculated from the GS-Doppler feature fusion networks that predict whether the synovial hypertrophy score was >0. The heatmaps are supposed to indicate the potential synovial hypertrophy area.

We adopted the recently proposed integrated gradient (IG) technique, which assigns an importance score of the prediction to each input feature.[35] For a model $\mathcal{M}(\cdot)$ and an input x, we defined the baseline input x' as a zero-filled tensor that has the same shape as x and the integrated gradient as the path integral of the gradients along the straight-line path from x' to x. Specifically, for an input x of n features $\{x_i\}_{i=1}^{n}$, the integrated gradient along x_i is defined using Equation 2:

$$\mathrm{IG}(x,i) = (x_i - x_i') \int_0^1 \frac{\partial \mathcal{M}(x' + t(x - x'))}{\partial x_i} \mathrm{d}t \qquad \text{(Equation 2)}$$

In practice, the integration is approximated via a summation using Equation 3:

$$\mathrm{IG}^{approx}(x,i) = (x_i - x_i') \sum_{k=1}^{m} \frac{1}{m} \frac{\partial \mathcal{M}\left(x' + \frac{k}{m}(x - x')\right)}{\partial x_i} \qquad \text{(Equation 3)}$$

where we used $m = 50$ in all of the experiments. Thus, the integrated gradients of the n features were obtained, and they formed a new image $\mathrm{IG}(x) = \{\mathrm{IG}^{approx}(x,i)\}_{i=1}^{n}$ that was of the same size as the input image x.

To make the calculation of heatmaps more robust, Gaussian noise $\delta \sim \mathcal{N}(0,1)$ is randomly added to the original image x eight times, and eight heatmaps are generated using IG. The final heatmap for model M and input x is generated by averaging the eight heatmaps. To visualize the heatmaps, a yellow mask is overlaid on the original image x. The alpha channel pixel values are decided by the corresponding heatmap pixel values. To remove less important features, pixel values smaller than 0.2 are set to zero.

In the RATING system, each of the five scoring models contains two GS-Doppler feature fusion networks that predict whether the synovial hypertrophy score was >0. Therefore, a total of 10 heatmaps of the GSUS image and 10 heatmaps of the Doppler US image are generated for each pair of US images. Because the 10 networks differ in either training data or randomization seed, they view the same US image in different perspectives and concentrate on different areas. To combine the knowledge of all of the networks, the heatmaps are averaged to generate a GSUS heatmap and a Doppler US heatmap, as shown in Figure S9.

Reader study

A reader study was conducted to compare the performance of the RATING system with the performance of the 10 radiologists, whose US examination experience ranges from 4 to 15 years (Table S17). A total of 274 samples from 28 patients in the prospective test dataset were presented to the radiologists and the RATING system in random order. For each sample, the ROIs of GSUS and Doppler US were provided. Each radiologist reviewed the same set of samples independently and decided the synovial hypertrophy score, the vascularity score, and the combined score according to the EOSS system.

DL-assisted reader study

To evaluate the assistance of the RATING system to the radiologists in guiding clinical decisions, we conducted a DL-assisted reader study. Two weeks after the reader study, the same 274 samples in the prospective test dataset were presented to the same 10 radiologists again in a different order. For each sample, together with the ROIs of the GSUS and Doppler US image, the heatmaps and predictions of the RATING system were also provided to the readers. The heatmaps indicated the potential synovial hypertrophy area and the synovial hypervascularity area, and the scores predicted by the RATING system served as a reference that may assist the readers when they lacked confidence in their own judgment. The radiologists were blinded to the first-time interpretation and to one another.

Statistical analysis

For synovial hypertrophy score binary classification tasks and vascularity score binary classification tasks, in which 10 models are trained in each task, the 95% CI of AUC, PPV, NPV, sensitivity, and specificity are computed as $\pm 1.96\sigma/\sqrt{10}$, where σ is the standard error across the 10 models. For the two ablation methods without MULTITUDE and the two existing methods, the 95% CI are computed as $\pm 1.96\sigma/\sqrt{5}$, where σ is the standard error across the five scoring models. For the remaining experiments, in which only a single accuracy and a single κ value is available, we bootstrapped the estimation for 1,000 iterations and reported the 2.5th and 97.5th percentiles as the 95% CI.

The agreement between DL models and experienced radiologists is graded as follows: poor ($\kappa \leq 0.20$), moderate ($0.20 < \kappa \leq 0.40$), fair ($0.40 < \kappa \leq 0.60$), good ($0.60 < \kappa \leq 0.80$), or very good ($0.80 < \kappa \leq 1.00$). To compare differences between accuracy scores and between linearly weighted κ, we bootstrapped the estimation for 1,000 iterations and compared using the z test. All of the statistical tests are two sided, and $p < 0.001$ indicates statistically significant differences. The analyses are performed using the Python scikit-learn library and the statsmodel library.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100592.

ACKNOWLEDGMENTS

This work was supported by the International Science and Technology Cooperation Programme (2015DFA30440), the National Natural Science Foundation of China (81421004, 62071271, 62021002, 61727808, 61971447, and 81301268), the National Key Technology R&D Program of China (2019YFC0840603, 2017YFC0907601, 2017YFC0907604, 2017YFE0104200, and 2018YFA0704000), the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (2020-I2M-C&T-B-035 and 2021-I2M-1–005), the Non-profit Central Research Institute Fund of the Chinese Academy of Medical Sciences (2021-PT320-002), and the Beijing Municipal Natural Science Foundation (JQ18023, JQ19015, and JQ21012). We thank Dr. Xuelan Li, Dr. Cheng Chen, Dr. Sirui Liu, Dr. Yang Gui, Dr. Na Su, Dr. Ruojiao Wang, Dr. Yao Wei, Dr. Yanwen Luo, Dr. Zihan Niu, and Dr. Yuanjing Gao for participating in the reader study and the DL-assisted reader study. We thank the Institute for Brain and Cognitive Science, Tsinghua University (THUIBCS), and the Beijing Laboratory

of Brain and Cognitive Intelligence, Beijing Municipal Education Commission (BLBCI) for their support. Several figures were created with BioRender.com and Vecteezy.

AUTHOR CONTRIBUTIONS

Z.Z. developed the system and analyzed the experiment results. Z.Z. and C.Z. wrote the manuscript. C.Z., M.W., and M.Y. provided clinical expertise. Z.Z., H.Q., and Y.G. discussed the techniques. C.Z., M.W., Q.W., and R.Z. created the datasets and scored the US images. Q.W., R.Z., H.W., F.D., Z.Q., X.T., and X.Z. collected the data. J.L. and Y.J. discussed the medical issues. H.Q., F.X., and M.Y. revised the manuscript. H.Q., F.X., Q.D., and M.Y. supervised the study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 11, 2022
Revised: August 4, 2022
Accepted: August 30, 2022
Published: September 29, 2022

REFERENCES

1. Atchia, I., Brown, A.K., Chitale, S., Ciechomska, A., Estrach, C., Karim, Z., and Wakefield, R.J.; British Society for Rheumatology Ultrasound Special Interest Group BSRUSSIG (2021). British society for Rheumatology ultrasound special interest group (BSRUSSIG) (2021). Recommendations for rheumatology ultrasound training and practice in the UK. Rheumatology *60*, 2647–2652. https://doi.org/10.1093/rheumatology/keaa656.
2. van Vollenhoven, R. (2019). Treat-to-target in rheumatoid arthritis - are we there yet? Nat. Rev. Rheumatol. *15*, 180–186. https://doi.org/10.1038/s41584-019-0170-5.
3. Colebatch, A.N., Edwards, C.J., Østergaard, M., van der Heijde, D., Balint, P.V., D'Agostino, M.A., Forslind, K., Grassi, W., Haavardsholm, E.A., Haugeberg, G., et al. (2013). EULAR recommendations for the use of imaging of the joints in the clinical management of rheumatoid arthritis. Ann. Rheum. Dis. *72*, 804–814. https://doi.org/10.1136/annrheumdis-2012-203158.
4. Avramidis, G.P., Avramidou, M.P., and Papakostas, G.A. (2021). Rheumatoid arthritis diagnosis: deep learning vs. Appl. Sci. *12*, 10. https://doi.org/10.3390/app12010010.
5. Gadeholt, O. (2019). Forward to the past: ultrasound might be necessary in some patients with rheumatoid arthritis. Ann. Rheum. Dis. *78*, e56. https://doi.org/10.1136/annrheumdis-2018-213278.
6. D'Agostino, M.A., Terslev, L., Aegerter, P., Backhaus, M., Balint, P., Bruyn, G.A., Filippucci, E., Grassi, W., Iagnocco, A., Jousse-Joulin, S., et al. (2017). Scoring ultrasound synovitis in rheumatoid arthritis: a EULAR-OMERACT ultrasound task force-Part 1: definition and development of a standardised, consensus-based scoring system. RMD Open *3*, e000428. https://doi.org/10.1136/rmdopen-2016-000428.
7. Ventura-Ríos, L., Hernández-Díaz, C., Ferrusquia-Toríz, D., Cruz-Arenas, E., Rodríguez-Henríquez, P., Alvarez Del Castillo, A.L., Campaña-Parra, A., Canul, E., Guerrero Yeo, G., Mendoza-Ruiz, J.J., et al.; Grupo Mexicano de Ecografía Musculoesquelética AC ECOMER (2017). Reliability of ultrasound grading traditional score and new global OMERACT-EULAR score system (GLOESS): results from an inter and intra-reading exercise by rheumatologists. Clin. Rheumatol. *36*, 2799–2804. https://doi.org/10.1007/s10067-017-3662-1.
8. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.
9. Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., Han, C., Gao, X., Zhang, H., Zheng, W., et al. (2021). Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat. Biomed. Eng. *5*, 522–532. https://doi.org/10.1038/s41551-021-00711-2.
10. Shen, Y., Shamout, F.E., Oliver, J.R., Witowski, J., Kannan, K., Park, J., Wu, N., Huddleston, C., Wolfson, S., Millet, A., et al. (2021). Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat. Commun. *12*, 5645–5713. https://doi.org/10.1038/s41467-021-26023-2.
11. Li, X., Zhang, S., Zhang, Q., Wei, X., Pan, Y., Zhao, J., Xin, X., Qin, C., Wang, X., Li, J., et al. (2019). Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol. *20*, 193–201. https://doi.org/10.1016/S1470-2045(18)30762-9.
12. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. *25*, 954–961. https://doi.org/10.1038/s41591-019-0447-x.
13. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., and Zou, J.Y. (2020). Video-based AI for beat-to-beat assessment of cardiac function. Nature *580*, 252–256. https://doi.org/10.1038/s41586-020-2145-8.
14. Arnaout, R., Curran, L., Zhao, Y., Levine, J.C., Chinn, E., and Moon-Grady, A.J. (2021). An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. Nat. Med. *27*, 882–891. https://doi.org/10.1038/s41591-021-01342-5.
15. Smerilli, G., Cipolletta, E., Sartini, G., Moscioni, E., Di Cosmo, M., Fiorentino, M.C., Moccia, S., Frontoni, E., Grassi, W., and Filippucci, E. (2022). Development of a convolutional neural network for the identification and the measurement of the median nerve on ultrasound images acquired at carpal tunnel level. Arthritis Res. Ther. *24*, 38. https://doi.org/10.1186/s13075-022-02729-6.
16. Cosmo, M.D., Chiara Fiorentino, M., Villani, F.P., Sartini, G., Smerilli, G., Filippucci, E., Frontoni, E., and Moccia, S. (2021). Learning-based median nerve segmentation from ultrasound images for carpal tunnel Syndrome evaluation. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE), pp. 3025–3028. https://doi.org/10.1109/EMBC46164.2021.9631057.
17. Fiorentino, M.C., Cipolletta, E., Filippucci, E., Grassi, W., Frontoni, E., and Moccia, S. (2022). A deep-learning framework for metacarpal-head cartilage-thickness estimation in ultrasound rheumatological images. Comput. Biol. Med. *141*, 105–117. https://doi.org/10.1016/j.compbiomed.2021.105117.
18. Wang, H.J., Su, C.P., Lai, C.C., Chen, W.R., Chen, C., Ho, L.Y., Chu, W.C., and Lien, C.Y. (2022). Deep learning-based Computer-Aided diagnosis of rheumatoid arthritis with hand X-ray images Conforming to Modified total Sharp/van der Heijde score. Biomedicines *10*, 1355. https://doi.org/10.3390/biomedicines10061355.
19. Chang, G.H., Felson, D.T., Qiu, S., Guermazi, A., Capellini, T.D., and Kolachalama, V.B. (2020). Assessment of knee pain from MR imaging using a convolutional Siamese network. Eur. Radiol. *30*, 3538–3548. https://doi.org/10.1007/s00330-020-06658-3.
20. Andersen, J.K.H., Pedersen, J.S., Laursen, M.S., Holtz, K., Grauslund, J., Savarimuthu, T.R., and Just, S.A. (2019). Neural networks for automatic scoring of arthritis disease activity on ultrasound images. RMD Open *5*, e000891. https://doi.org/10.1136/rmdopen-2018-000891.
21. Christensen, A.B.H., Just, S.A., Andersen, J.K.H., and Savarimuthu, T.R. (2020). Applying cascaded convolutional neural network design further enhances automatic scoring of arthritis disease activity on ultrasound images from rheumatoid arthritis patients. Ann. Rheum. Dis. *79*, 1189–1193. https://doi.org/10.1136/annrheumdis-2019-216636.
22. Wu, M., Wu, H., Wu, L., Cui, C., Shi, S., Xu, J., Liu, Y., and Dong, F. (2022). A deep learning classification of metacarpophalangeal joints synovial proliferation in rheumatoid arthritis by ultrasound images. J. Clin. Ultrasound *50*, 296–301. https://doi.org/10.1002/jcu.23143.
23. Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E.J. (2022). AI in health and medicine. Nat. Med. *28*, 31–38. https://doi.org/10.1038/s41591-021-01614-0.

24. Dietterich, T.G., and Bakiri, G. (1994). Solving multiclass learning problems via error-correcting output codes. J. Artif. Intell. Res. *2*, 263–286. https://doi.org/10.5555/1622826.1622834.

25. Cicchetti, D.V., and Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. Am. J. EEG Technol. *11*, 101–110. https://doi.org/10.1080/00029238.1971.11080840.

26. Youden, W.J. (1950). Index for rating diagnostic tests. Cancer *3*, 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3.

27. Sawaki, M., Shien, T., and Iwata, H. (2019). TNM classification of malignant tumors (breast cancer study group). Jpn. J. Clin. Oncol. *49*, 228–231. https://doi.org/10.1093/jjco/hyy182.

28. Wolpert, D.H. (1992). Stacked generalization. Neural Network. *5*, 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.

29. Freund, Y., and Schapire, R.E. (1996). Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, pp. 148–156.

30. Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., and Chi, E.H. (2018). Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1930–1939. https://doi.org/10.1145/3219819.3220007.

31. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

32. Noroozi, M., and Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In Proceedings of the 14th European Conference on Computer Vision, pp. 69–84. https://doi.org/10.1007/978-3-319-46466-4_5.

33. Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. https://arxiv.org/abs/1412.6980.

34. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009). Imagenet: a large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

35. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, pp. 3319–3328. http://proceedings.mlr.press/v70/sundararajan17a.html.