Interpretable Rheumatoid Arthritis Scoring via Anatomy-aware Multiple Instance Learning

Zhiyan Bo[1(✉)], Laura C. Coates[2], and Bartłomiej W. Papież[1]

[1] Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, UK
zhiyan.bo@reuben.ox.ac.uk & bartlomiej.papiez@bdi.ox.ac.uk
[2] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

Abstract. The Sharp/van der Heijde (SvdH) score has been widely used in clinical trials to quantify radiographic damage in Rheumatoid Arthritis (RA), but its complexity has limited its adoption in routine clinical practice. To address the inefficiency of manual scoring, this work proposes a two-stage pipeline for interpretable image-level SvdH score prediction using dual-hand radiographs. Our approach extracts disease-relevant image regions and integrates them using attention-based multiple instance learning to generate image-level features for prediction. We propose two region extraction schemes: 1) sampling image tiles most likely to contain abnormalities, and 2) cropping patches containing disease-relevant joints. With Scheme 2, our best individual score prediction model achieved a Pearson's correlation coefficient (PCC) of 0.943 and a root mean squared error (RMSE) of 15.73. Ensemble learning further boosted prediction accuracy, yielding a PCC of 0.945 and RMSE of 15.57, achieving state-of-the-art performance that is comparable to that of experienced radiologists (PCC = 0.97, RMSE = 18.75). Finally, our pipeline effectively identified and made decisions based on anatomical structures which clinicians consider relevant to RA progression.

Keywords: Rheumatoid Arthritis · Hand X-ray scoring

1 Introduction

Rheumatoid Arthritis (RA) affects around 18 million people worldwide [29]. This autoimmune disease causes joint inflammation, ultimately leading to structural damage and disability. Commonly manifesting joint erosion and joint space narrowing (JSN) in hands, wrists, and feet, RA is evaluated by plain radiography for diagnosis and monitoring [3]. Several RA quantification schemes exist across imaging modalities, with the van der Heijde modification of the Sharp (SvdH) score being the standard for radiography in clinical trials due to its strong intra- and inter-observer reliability and sensitivity to changes [15,20]. SvdH examines 16 areas and 15 joints in each hand and wrist (see Fig. 1(A)) and 6 areas and 6 joints in each foot [25]. However, its detailed evaluation requires approximately

25 minutes per patient by an experienced radiologist, limiting its practicality in clinical settings [5]. In addition, suboptimal imaging conditions and superimposition contribute to inter-observer variability, particularly in detecting small changes, posing challenges on clinical trials in early-stage patients [15].

Deep learning (DL) provides a possibility to reduce reliance on human scorers for RA quantification while improving sensitivity, consistency, and efficiency [18]. Several automated RA scoring methods have been proposed for ultrasound and X-ray imaging [2,6], typically consisting of a joint localisation stage (using U-Net-based heatmap regression, YOLO model or Mask-RCNN) followed by convolutional neural network (CNN)-based joint-level damage quantification [1,9,11]. Some methods focused on specific anatomical structures, such as fingers [10], making them insufficient for holistic SvdH scoring. The shortage of datasets with joint-level annotations additionally increases the difficulty of method validation and comparison. A few methods developed using in-house datasets [12] lacked image-level or patient-level performance evaluation. Most methods covering both hand and foot joints relied on 674 sets of images from the RA2-DREAM challenge [23], but since participants were not granted access to the final evaluation images, detailed validation was limited [17,24]. External validation was performed only in [26], where the proposed pipeline and the top two solutions of the challenge were tested on a private dataset of 205 patients.

While joint-level scoring provides detailed assessments, accurately localising individual joints can be challenging in late-stage RA patients, who may struggle to straighten deformed fingers during scans. Also, the intrinsic ambiguity between adjacent scores increases the task difficulty, potentially leading to large radiograph-level error [1]. To address this, image-level scoring methods have been explored such as an overall SvdH score prediction for dual-hand radiographs [4,28]. In [19], a four-stage method with image reorientation, hand segmentation, joint identification, and image-level score prediction was developed. Despite promising performance, these methods lack interpretability. While joint patches were used as inputs [19], the pipeline could not pinpoint damage sites. In our previous work [4], Gradient-weighted Class Activation Mapping (Grad-CAM) was performed to provide explanations for the model's predictions, but our results revealed model confusion by irrelevant regions, such as finger bone diaphyses and arms.

To address these challenges, we propose an anatomy-aware multiple instance learning (MIL) neural network that intelligently extracts RA-relevant patches from dual-hand radiographs and effectively integrates them for total SvdH score prediction. To enable the extraction of regions assessed by SvdH, we also introduce a novel dual-hand joint localisation model. Finally, our MIL framework incorporates an attention mechanism, enhancing both interpretability and predictive performance. Our model achieved state-of-the-art (SOTA) results on a publicly available dataset and performed comparably to experienced radiologists. Additionally, clinician evaluation highlighted the improved clinical relevance of our method in identifying RA-damaged regions.

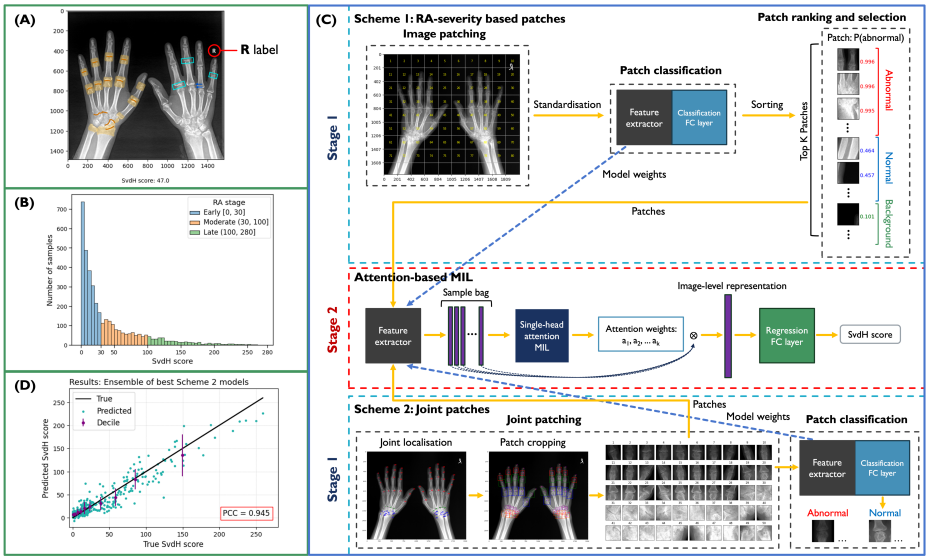

Fig. 1. (A) Example radiograph from [28] highlighting bones (yellow) and joints (orange) assessed by SvdH in the left hand and wrist with examples of bone erosion (blue arrows) and JSN (cyan boxes) in the right hand. (B) SvdH score distribution. (C) Overview of the proposed pipelines. In Scheme 2, 37 landmarks are labelled in each hand, which are then used to crop 25 patches. (D) Predicted vs. true SvdH scores for our best score prediction model.

2 Methods

The proposed framework has two main stages: 1) RA-relevant area sampling and 2) score prediction (see Fig. 1(C)). For Stage 1, we propose two strategies: Scheme 1: in which the most "abnormal" tiles in an image are sampled based on RA relevance, and Scheme 2: in which joint patches are cropped to exclude non-relevant anatomical information. In Stage 2, features extracted from the sampled patches are fed into a single-head Attention-Based MIL (ABMIL) model for total SvdH score prediction.

2.1 Scheme 1: RA-severity based patches

Scheme 1 was inspired by the intelligent sampling method proposed by [22]. Here, images are divided into non-overlapping square patches, each with a side length of $\frac{1}{10}$ of the image's longer dimension. Then, a weakly supervised classifier categorises patches as normal, abnormal, or background, sorts them by abnormality probability, and samples the top K patches in the order (abnormal, normal, background) prioritising RA-relevant regions. Since patch-level labels were unavailable, we used image-level labels as proxies with images having a

total SvdH score < 5 considered normal, and those with a score ≥ 70 considered abnormal. To identify background patches, a nnU-Net [14] was trained using high-quality segmentation masks created by morphological operations (blurring, erosion, dilation, thresholding, and removal of small bright/dark areas). Then segmentation masks were generated for all images involved in patch classifier (PC) development, with small bright areas removed during post-processing. Patches with $\leq 2\%$ of foreground were labelled as background. Following [4,28], MobileNetV2 [21], ResNet-34, and ResNet-50 [8] were selected as the backbone architectures for PC, with the final fully connected (FC) layer modified to output 3 classes. The PCs were later truncated before the FC layer forming feature extractors (FEs) which output 1280, 512, and 2048 features respectively.

2.2 Scheme 2: Joint patches

Joint localisation is a key step of hand radiographic analysis, typically involving identifying joints as landmarks or drawing bounding boxes. Here we defined 37 joint landmarks per hand and wrist (see Fig. 1(C)) and cropped patches accordingly to accommodate variations in imaging protocols. To the best of our knowledge, no existing joint detector is designed for dual-hand radiographs, and severe hand deformities in RA patients prevent straightforward separation of images into left and right hands. To address this, we labelled a subset of the dataset [28] for all landmarks and adapted the multiresolution Hybrid Transformer-CNN (HTC), a top-performing landmark localisation model originally designed for single-hand radiographs of children [27]. The number of output channels was modified to 74 to reflect dual-hand radiographs. Using the predicted landmarks, images were aligned to a standard position (fingers pointing upwards) and 25 patches covering most hand joints (see Fig. 1(C)) were extracted per side. MobileNetV2, ResNet-34, and ResNet-50 were trained to classify joint patches as normal or abnormal using the same PC-development subset as Scheme 1 and then truncated before the FC layer to form Scheme 2 FEs.

2.3 Interpretable score prediction: Attention-based MIL

ABMIL [13] has been adopted for several medical imaging tasks, such as screening of COVID-19 from chest computed tomography [7]. MIL's capacity to account for within-image heterogeneity makes it well-suited for our task as RA-affected hands contain both healthy joints and joints with varying levels of damage, while the attention mechanism enables the model to selectively focus on disease-relevant regions. Our model uses the gated attention [13], which adopts hyperbolic tangent activation to avoid gradient explosion and sigmoid activation to introduce non-linearity. The patch features extracted by FEs are fed into the ABMIL model, scaled by attention weights, and summed to compute the image-level representation, which is subsequently passed through a regression FC layer for score prediction.

3 Experiments and Results

3.1 Dataset, Implementation, and Evaluation

Dataset. This study used a public dataset of 3,818 hand radiographs collected from diagnosed and suspected RA patients (see an example in Fig. 1(A)) [28]. The average SvdH scores, reported by two experienced radiologists, were provided with most samples representing early-stage RA (see Fig. 1(B)). The dataset was split into training, validation, and test sets, containing 2700, 760, and 358 images respectively. Of these, 143, 44, and 24 images were selected for nnU-Net foreground segmentation model development, and 1019, 334, and 153 images were selected for PC development. Using BoneFinder [16], we annotated 351 images for joint localisation model development, consisting of 245 training, 56 validation, and 50 test images.

Implementation details. The SvdH scores were standardised to have a mean of 0 and a standard deviation (SD) of 1. For both schemes, random flip, intensity scaling (0.9 - 1.1), and rotation by 0°, 90°, 180°, or 270° were applied to the training images for augmentation. Random affine transformations (small-angle rotation, translation, and scaling only for joint localisation) were additionally applied when training Scheme 2 models, as the joint patching step could exclude the augmentation-induced artefacts. Random Gaussian noise, based on absolute radial distance, was added to the landmark predictions when cropping patches in the training set, using landmark-wise mean radial error (MRE) as SD. Whole images and patches were resized to 1024×1024 and 224×224 respectively, and normalised with the mean and SD of the training set. The PCs were initialised with ImageNet-pretrained model parameters. For Scheme 1 PCs, patches were cropped first and then augmented. For Scheme 2 PCs, images were augmented first, followed by joint patching. The dropout rate was set to 0.1 for ABMIL models. nnU-Net was trained with default configurations [14]. Joint localisation models were trained using the setup in [27] for 300, 450, and 600 epochs with a batch size of 4 or 16. The pixel spacing in each image was estimated by assuming an average wrist width (distance between landmarks at the endpoints of a wrist) of 50 mm. Other models were trained using the stochastic gradient descent optimiser with a learning rate of 0.001, weight decay of 0.001, and momentum of 0.9. We chose mean squared error loss for score prediction and cross-entropy loss for patch classification. For Scheme 1 ABMIL models, we experimented with an input patch number K of 30, 40, or 50. PCs and ABMIL models were initially trained for 100 epochs with a batch size of 4 or 16. An extra 50 or 100 epochs were added for models with slower convergence. The code and newly created dual-hand landmark annotations will be available at: https://github.com/ZhiyanBo/RA-AWMIL.

Evaluation metrics. Dice score was used to evaluate foreground segmentation models and classification accuracy to evaluate PCs. For joint localisation, MRE

in mm and successful detection rate (SDR, %) under 2, 3, 4, and 10 mm conditions were used. For score prediction, Pearson's correlation coefficient (PCC), mean absolute error (MAE), and root mean squared error (RMSE) were used.

3.2 Results and Discussion

Scheme 1: Patch classification. The nnU-Net achieved an average Dice score of 0.986 in 24 test images. Morphological operation-based post-processing further improved the quality of masks for noisy images, though in some cases fingertips were excluded. The MobileNetV2-based, ResNet-34-based, and ResNet-50-based PCs achieved 89.7%, 89.0%, and 90.0% accuracy in the weakly supervised task. They successfully classified over 98% of background patches while the misclassification rate was notably higher between normal and abnormal patches, as expected, since abnormal images may contain normal patches (even in severe RA, not all joints could be affected) and vice versa.

Scheme 2: Joint localisation and patch classification. We selected multiresolution HTC models with best testing performance under two cases: 1) JLval: among models with peak performance in the validation set during training, and 2) JLnoVal: among models extracted at the end of training. Table 1 shows the results in two cases: 1) all images used for testing, and 2) the test set after excluding images with poor-quality patches. Despite achieving over 94% SDR within 4 mm, both models have high MREs with large SDs. We visually examined the images with high MRE and found that in images that are horizontally flipped (i.e., right hand on the left side), models were confused by the **R** label, which indicates the true **right** side. Although we labelled the landmarks based on their positions in the image, the models learned to differentiate between the true left and right hands. While this confusion caused severe MRE, these mistakes did not affect patch quality. However, because the model averages the predictions from two heatmaps of different resolutions, some predictions in flipped images were placed between the hands as one heatmap predicted based on the hands' relative position in the image, while one predicted based on the hands' true side. These errors led to wrong regions being cropped in 2 or 3 out of 50 images in the test set. Our results demonstrate the negative effect of anatomical symmetry on the landmark detection model's performance. To estimate the amount of error our joint patching scheme could tolerate, the landmark-wise MRE was calculated

Table 1. Joint localisation performance of JLval and JLnoVal models in 1) all images involved in testing and 2) the test set after excluding images with poor-quality patches.

Dataset	Implementation details	Epoch number	MRE (mm, SD)	SDR (%)			
				2 mm	3 mm	4 mm	10 mm
All images (in test set/test & validation sets)	JLval	300	9.10 (40.14)	88.89	92.89	94.16	95.27
	JLnoVal	600	8.11 (38.04)	88.37	92.96	94.72	95.93
Images with high-quality patches (in test set)	JLval	300	0.88 (2.88)	93.60	97.71	98.91	99.97
	JLnoVal	600	0.85 (1.63)	93.63	97.75	99.04	99.97

Table 2. Pipeline performance in SvdH score prediction and the published inter-rater differences [28]. The top two values of a metric are made **bold** and underlined.

Method	Implementation details	Feature extractor	PCC	MAE	RMSE
Scheme 1: RA-severity based patches	50 patches as inputs & Fine-tuned FE	MobileNetV2	0.924	12.31	18.04
		ResNet-34	**0.932**	**11.95**	**17.16**
		ResNet-50	0.925	12.68	17.88
Scheme 2: Joint patches	JLval	MobileNetV2	0.938	11.55	16.50
		ResNet-34	0.938	11.46	16.37
		ResNet-50	**0.943**	11.33	**15.73**
	JLnoVal	MobileNetV2	0.941	11.61	16.23
		ResNet-34	0.938	**11.29**	16.53
		ResNet-50	0.938	11.88	16.37
	Ensemble of best models		**0.945**	**11.22**	**15.57**
Published methodologies	ResNet-Dwise50 [28]		0.97	14.90	22.01
	Ensemble of CNNs [4]		0.925	12.57	18.02
	Custom ViT [19]		×	21.14	44.28
Published inter-rater differences [28]			0.97	12.24	18.75

using the test set while excluding 4 or 5 images with symmetry-induced errors described above. With joint patches as inputs, the MobileNetV2-based, ResNet-34-based, and ResNet-50-based PCs achieved 92.3%, 93.0%, and 92.1% accuracy with JLval and 91.8%, 93.1%, and 92.4% accuracy with JLnoVal, suggesting they have successfully learned useful features of RA-related damage.

SvdH score prediction. The results of SvdH score prediction are provided in Table 2. With Scheme 1, the best performance was obtained when sampling 50 patches as inputs and fine-tuning the FE during training. The model with ResNet-34-based FE achieved a PCC of 0.932, MAE of 11.95, and RMSE of 17.16. Consistent performance improvement was observed when switching to Scheme 2 patches, potentially owing to further exclusion of anatomical structures and information irrelevant to SvdH scoring. With JLval joint localisation model, the best model JLval-ResNet50 used a ResNet-50-based FE, while with JLnoVal joint localisation model, the best model JLnoVal-MobileNetV2 used a MobileNetV2-based FE. Balancing all metrics, JLval-ResNet50 was selected as the best individual SvdH scoring model, achieving a PCC of 0.943, MAE of 11.33, and RMSE of 15.73. Ensembling JLval-ResNet50 and JLnoVal-MobileNetV2 by averaging their predictions yielded further performance improvement, reaching a PCC of 0.945, MAE of 11.22, and RMSE of 15.57. Both Scheme 1 and Scheme 2 models outperformed other published methods on dual-hand image-level SvdH score prediction in terms of MAE and RMSE [4,19,28]. Scheme 2 models performed better than Scheme 1 models in scoring moderate and late-stage cases, though their accuracy in early and moderate-stage cases could be further improved, as displayed in Fig. 1(D). Compared with experienced radiologists, JLval-ResNet50 demonstrated as good or even better average prediction accuracy, as shown by its lower MAE and RMSE. However, its PCC was 2.7% lower, suggesting scope for pipeline optimisation.

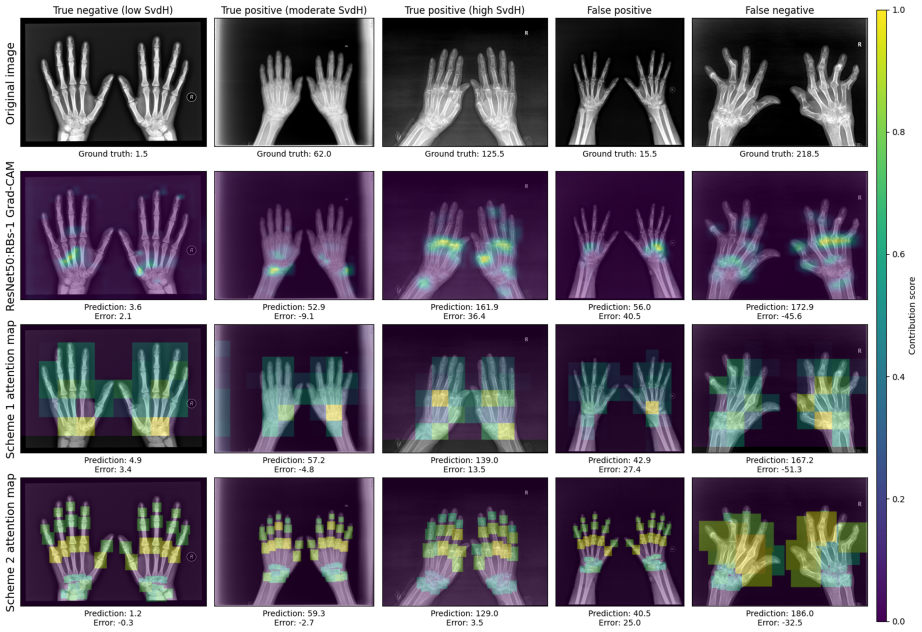

Fig. 2. Grad-CAM and attention maps of example images from the best individual whole-image model (ResNet-50:RBs-1) [4], and our Scheme 1 and Scheme 2 models.

Attention maps for interpretable RA scoring. Fig. 2 displays the Grad-CAM heatmaps of ResNet-50:RBs-1, the SOTA individual single-stage model from our previous work [4] that takes a whole image as input, and the attention maps of our models for images with varying RA severity and prediction accuracy. An experienced rheumatologist was invited to review the visual explanations, who noted that ResNet-50:RBs-1 could incorrectly focus on RA-irrelevant structures. Capable of identifying some but not all damage in each example, it wrongly highlighted diaphyses and arm bones in true negative and true positives (TPs). Scheme 1 performed better in detecting wrist damage, but failed to highlight finger damage in TPs. Also, some joints were split into multiple patches, causing information loss. On the other hand, Scheme 2 offered better interpretability than Scheme 1 as its patches were more specific and included larger proportions of RA-relevant structures. However, joint localisation accuracy may deteriorate in images with curved fingers, leading to suboptimal patches as in the false negative example and subsequently large prediction errors. Best at detecting finger damage, Scheme 2 could over-focus on finger joints, especially the metacarpophalangeal joints, while not highlighting wrist damage enough, as in false positive and TPs. For clarification, this clinician evaluation is based on a small number of images, which may not cover the full spectrum of model behaviours.

4 Conclusion

This work presents an interpretable anatomy-aware ABMIL framework for image-level RA quantification by SvdH scoring, achieving SOTA performance and accuracy comparable to experienced radiologists. Our two RA-relevant region sampling strategies effectively capture damaged features, with joint patches acquired using a multiresolution HTC further enhancing model performance and interpretability. Our work shows that by incorporating prior knowledge of the disease, ABMIL could generate better image-level representations, further boosting prediction accuracy. Future work will focus on improving joint localisation robustness through noise removal and enhancing prediction performance, particularly in early-stage cases. A structured clinician evaluation study would also provide insights into the perceived accuracy and usefulness of the pipeline.
Prospect of Application: Our SvdH scoring model has the potential to serve as an additional rater in clinical trials, offering more consistent trial endpoints by reducing inter-rater variability. Furthermore, by providing anatomy-aware visual explanations, it may enhance clinician trust and enable objective quantification of joint damage for the diagnosis and monitoring of RA.

Acknowledgments. This work was supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1). We would like to thank the Oxford Biomedical Research Computing Facility for providing the computational resources.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bird, A., Oakden-Rayner, L., Chakradeo, K., Thomas, R., Gupta, D., Jain, S., Jacob, R., Ray, S., Wechalekar, M.D., Proudman, S., Palmer, L.J.: AI automated radiographic scoring in rheumatoid arthritis: Shedding light on barriers to implementation through comprehensive evaluation. Seminars in Arthritis and Rheumatism **74** (2025). https://doi.org/10.1016/j.semarthrit.2025.152761
2. Bird, A., Oakden-Rayner, L., McMaster, C., Smith, L.A., Zeng, M., Wechalekar, M.D., Ray, S., Proudman, S., Palmer, L.J.: Artificial intelligence and the future of radiographic scoring in rheumatoid arthritis: a viewpoint. Arthritis Research and Therapy **24**, 1–10 (2022). https://doi.org/10.1186/S13075-022-02972-X
3. BMJ Best Practice: Rheumatoid arthritis - symptoms, diagnosis and treatment (4 2024), https://bestpractice.bmj.com/topics/en-gb/105
4. Bo, Z., Coates, L.C., Papież, B.W.: Deep learning models to automate the scoring of hand radiographs for rheumatoid arthritis. In: Medical Image Understanding and Analysis. pp. 398–413 (2024)
5. Boini, S., Guillemin, F.: Radiographic scoring methods as outcome measures in rheumatoid arthritis: properties and advantages. Annals of the Rheumatic Diseases **60**, 817 (2001). https://doi.org/10.1016/s0003-4967(24)43379-7
6. Gilvaz, V.J., Reginato, A.M.: Artificial intelligence in rheumatoid arthritis: potential applications and future implications. Frontiers in Medicine **10** (2023). https://doi.org/10.3389/FMED.2023.1280312

7. Han, Z., Wei, B., Hong, Y., Li, T., Cong, J., Zhu, X., Wei, H., Zhang, W.: Accurate Screening of COVID-19 Using Attention-Based Deep 3D Multiple Instance Learning. IEEE Transactions on Medical Imaging **39**, 2584–2594 (2020). https://doi.org/10.1109/TMI.2020.2996256
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition. pp. 770–778 (2016)
9. Hemalatha, R.J., Vijaybaskar, V., Thamizhvani, T.R.: Automatic localization of anatomical regions in medical ultrasound images of rheumatoid arthritis using deep learning. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine **233**, 657–667 (2019). https://doi.org/10.1177/0954411919845747
10. Hirano, T., Nishide, M., Nonaka, N., Seita, J., Ebina, K., Sakurada, K., Kumanogoh, A.: Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. Rheumatology Advances in Practice **3** (2019). https://doi.org/10.1093/RAP/RKZ047
11. Ho, S., Elamvazuthi, I., Lu, C.: Classification of rheumatoid arthritis using machine learning algorithms. In: 2018 IEEE 4th International Symposium in Robotics and Manufacturing Automation (ROMA) (2018)
12. Honda, S., Yano, K., Tanaka, E., Ikari, K., Harigai, M.: Development of a scoring model for the Sharp/van der Heijde score using convolutional neural networks and its clinical application. Rheumatology **62**, 2272–2283 (2023). https://doi.org/10.1093/RHEUMATOLOGY/KEAC586
13. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning. pp. 2127–2136. PMLR (2018)
14. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 2020 18:2 **18**, 203–211 (12 2020). https://doi.org/10.1038/s41592-020-01008-z
15. Landewé, R.B., Connell, C.A., Bradley, J.D., Wilkinson, B., Gruben, D., Strengholt, S., van der Heijde, D.: Is radiographic progression in modern rheumatoid arthritis trials still a robust outcome? Experience from tofacitinib clinical trials. Arthritis Research and Therapy **18** (2016). https://doi.org/10.1186/s13075-016-1106-y
16. Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F.: Robust and accurate shape model matching using random forest regression-voting. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**, 1862–1874 (2015). https://doi.org/10.1109/TPAMI.2014.2382106
17. Maziarz, K., Krason, A., Wojna, Z.: Deep learning for rheumatoid arthritis: Joint detection and damage scoring in x-rays (2022), https://arxiv.org/abs/2104.13915
18. Momtazmanesh, S., Nowroozi, A., Rezaei, N.: Artificial intelligence in rheumatoid arthritis: Current status and future perspectives: A state-of-the-art review. Rheumatology and Therapy **9**, 1249 (2022)
19. Moradmand, H., Ren, L.: Multistage deep learning methods for automating radiographic sharp score prediction in rheumatoid arthritis. Scientific Reports 2025 15:1 **15**, 1–14 (2025). https://doi.org/10.1038/s41598-025-86073-0
20. Salaffi, F., Carotti, M., Beci, G., Carlo, M.D., Giovagnoni, A.: Radiographic scoring methods in rheumatoid arthritis and psoriatic arthritis. Radiologia Medica **124**, 1071–1086 (2019). https://doi.org/10.1007/S11547-019-01001-3
21. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR. pp. 4510–4520 (2018)

22. Su, Z., Tavolara, T.E., Carreno-Galeano, G., Lee, S.J., Gurcan, M.N., Niazi, M.K.: Attention2majority: Weak multiple instance learning for regenerative kidney grading on whole slide images. Medical Image Analysis **79** (2022). https://doi.org/10.1016/j.media.2022.102462
23. Sun, D., Nguyen, T.M., Allaway, R.J., Wang, J., Chung, V., Yu, T.V., Mason, M., Dimitrovsky, I., Ericson, L., Li, H., Guan, Y., Israel, A., Olar, A., Pataki, B.A., Stolovitzky, G., Guinney, J., Gulko, P.S., Frazier, M.B., Chen, J.Y., Costello, J.C., Bridges, S. Louis, J., Community, R.D.C.: A Crowdsourcing Approach to Develop Machine Learning Models to Quantify Radiographic Joint Damage in Rheumatoid Arthritis. JAMA Network Open **5**(8), e2227423 (2022)
24. Tan, Y.M., Quek, R., Chong, H., Hargreaves, C.A.: Rheumatoid Arthritis: Automated Scoring of Radiographic Joint Damage (2021), https://arxiv.org/abs/2110.08812
25. van der Heijde, D., Dankert, T., Nieman, F., Rau, R., Boers, M.: Reliability and sensitivity to change of a simplification of the Sharp/van der Heijde radiological assessment in rheumatoid arthritis. Rheumatology **38**, 941–947 (1999). https://doi.org/10.1093/RHEUMATOLOGY/38.10.941
26. Venäläinen, M.S., Biehl, A., Holstila, M., Kuusalo, L., Elo, L.L.: Deep learning enables automatic detection of joint damage progression in rheumatoid arthritis—model development and external validation. Rheumatology **64**(3), 1068–1076 (2024). https://doi.org/10.1093/rheumatology/keae215
27. Viriyasaranon, T., Ma, S., Choi, J.H.: Anatomical Landmark Detection Using a Multiresolution Learning Approach with a Hybrid Transformer-CNN Model. In: MICCAI. pp. 433–443 (2023)
28. Wang, Z., Liu, J., Gu, Z., Li, C.: An Efficient CNN for Hand X-Ray Overall Scoring of Rheumatoid Arthritis. Complexity **2022** (2022). https://doi.org/10.1155/2022/5485606
29. WHO: Rheumatoid arthritis (6 2023), https://www.who.int/news-room/fact-sheets/detail/rheumatoid-arthritis