RASH: Rheumatoid Arthritis Synthetic Hand Dataset for Joint-Level Inflammation Estimation

Rintaro Chiba[1], Shun Kato[1], Daiki Kitsunai[1], Yasushi Kondo[2], Yoshimitsu Aoki[1], and Mariko Isogawa[1]

[1] Faculty of Science and Technology, Keio University, Yokohama, Japan
[2] School of Medicine, Keio University, Tokyo, Japan
rinrin819@keio.jp, mariko.isogawa@keio.jp

Abstract. Rheumatoid arthritis (RA) is a common autoimmune disease affecting approximately 1% of the population and causing chronic joint pain and swelling. As progressive inflammation can cause irreversible joint damage, early detection and follow-up are essential. Clinical assessment relies on specialist palpation supported by X-ray, MRI, and ultrasonography, which require equipment and expertise, limiting timely evaluation in telemedicine and specialist-scarce settings. RGB image-based screening provides an accessible, low-cost alternative. However, existing methods are limited by scarce, imbalanced joint-level labels, limited joint coverage and inter-joint context, and insufficient modeling of subtle inflammatory cues. This makes them vulnerable to shifts in skin tone, illumination, and acquisition conditions. To address these limitations, we propose a joint-level RA inflammation detection framework. We also construct RASH, a large-scale synthetic image dataset that spans diverse RA appearances across skin tones and disease severities. We integrate adjacency-based inter-joint context across 11 joints, including the wrist, and guide feature learning toward clinically relevant cues while performing domain alignment only within pathology-focused regions. Experiments on a real patient dataset, including cases that are challenging even for expert visual assessment, demonstrate that our method outperforms baseline approaches and approaches expert-level performance under varying imaging conditions.

Keywords: Rheumatoid arthritis (RA) · Synthetic Hand Image Dataset · Joint Inflammation Estimation · Domain Adaptation.

1 Introduction

Rheumatoid arthritis (RA) is a progressive autoimmune disease that can lead to irreversible joint destruction. Thus, early detection of inflammation is crucial, yet early-stage inflammation remains difficult to detect. The clinical gold standard is specialist palpation supported by X-ray [1–3], MRI [4, 5], and ultrasonography [6], which can detect subtle inflammation but require dedicated equipment and expert interpretation [7, 8]. This limits timely assessment in telemedicine and specialist-scarce settings, motivating low-burden screening tools.

RGB images are inexpensive and easy to capture, making them suitable for appearance-based screening and follow-up. However, prior RGB-based methods face three key challenges [9, 10]: scarce labeled data with severe joint-level imbalance, limited joint coverage with insufficient inter-joint context, and lack of explicit modeling of subtle cues such as erythema, gloss, and contour changes [11, 12], leading to vulnerability to domain shifts due to skin tone, illumination, and acquisition conditions. These limitations are most evident in early-stage RA, where visual signs are subtle and expert agreement is limited.

To address these challenges, we propose a robust joint-level inflammation detection framework using RGB images. We mitigate data scarcity and imbalance with a large-scale image dataset from a synthetic hand model reproducing RA-specific manifestations across multiple joints. We steer feature extraction toward clinically relevant cues and learn complementary morphology- and pathology-sensitive representations, integrating adjacency-based inter-joint context with a graph formulation. We further reduce the synthetic-to-real gap by aligning domains only within pathology-focused regions to improve robustness to acquisition variability. The contributions of this paper are summarized as follows:

- We tackle joint-level RA inflammation detection from RGB images under data scarcity, imbalance, and imaging variability, including cases that are difficult even under expert visual assessment.
- To mitigate data scarcity and joint-level imbalance, we construct RASH, the first large-scale synthetic RA image dataset using a controllable 3D hand model that reproduces RA-specific inflammatory manifestations. The dataset will be publicly released upon acceptance.
- To stabilize joint-level predictions under appearance variations, we propose Dual-Path with graph-based inter-joint context modeling to leverage adjacent-joint evidence.
- To model clinically relevant inflammatory cues and mitigate synthetic-to-real domain shift, we propose RGE-LA and CDA, which emphasize pathology-focused regions while suppressing acquisition-dependent nuisance factors.

2 Method

This paper estimates joint-level RA-related inflammation under variability in acquisition conditions and skin appearance, with severe class imbalance (Fig. 1). Given an RGB hand image X_j, we crop joint patches $x_{j,k}$ and predict inflammation probabilities $p_{j,k}$, where j and k denote the image and joint indices.

Our framework comprises data generation and network architecture. We use a 3-Dimensional Computer Graphics (3DCG) hand model to synthesize RA-like images X_j^{S} and extract patches $x_{j,k}^{\mathrm{S}}$ to construct RASH for pre-training. The network has three components: (i) Redness–Gloss–Edge Lesion Attention (RGE-LA) highlights clinically relevant cues to localize candidate inflammatory regions; (ii) Morphology–Pathology Dual-Path Network (Dual-Path) learns complementary representations, producing $\boldsymbol{h}_{j,k}^{\mathrm{M}}$ and $\boldsymbol{h}_{j,k}^{\mathrm{P}}$, fusing them into $\boldsymbol{z}_{j,k}^{(0)}$, and refining

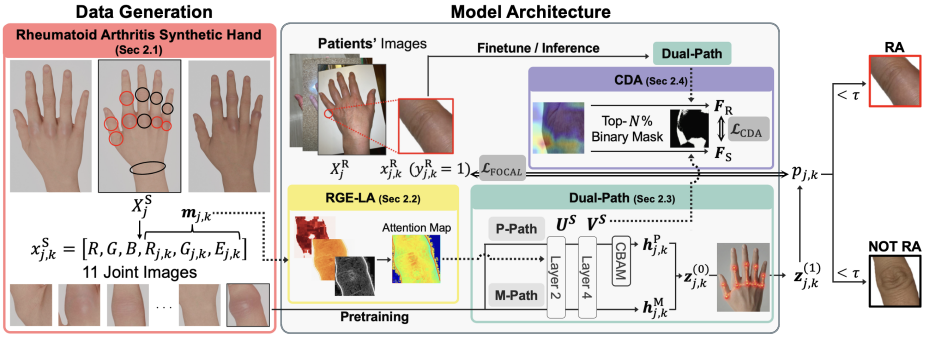

Fig. 1. Overview of the proposed method.

them with inter-joint context via a Graph Neural Network (GNN) to obtain $\boldsymbol{z}_{j,k}^{(1)}$ for stable prediction [13]; (iii) CAM-guided Domain Alignment (CDA) reduces the synthetic-to-real gap by minimizing $\mathcal{L}_{\mathrm{CDA}}$, which aligns $\boldsymbol{F}_{\mathrm{S}}$ and $\boldsymbol{F}_{\mathrm{R}}$ only within regions derived from Class Activation Mapping (CAM) [14].

We pre-train on synthetic patches $x_{j,k}^{\mathrm{S}}$ using Focal Loss [15], and fine-tune on real patches $x_{j,k}^{\mathrm{R}}$ with a weighted sum of $\mathcal{L}_{\mathrm{Focal}}$ and $\mathcal{L}_{\mathrm{CDA}}$, controlled by λ_{CDA}. At inference, $p_{j,k}$ is thresholded by τ.

2.1 RASH: Rheumatoid Arthritis Synthetic Hand Dataset

Convolutional Neural Network (CNN)-based classifiers require large-scale training data, but joint-level labels for RGB-based RA screening are scarce, particularly for subtle early-stage inflammation, with severe joint-wise class imbalance.

To mitigate data scarcity and joint-level imbalance, we extend a 3DCG hand model in Blender [16], since generic models do not support joint-wise, severity-controllable RA-like manifestations. The model independently controls erythema and swelling at each joint with associated cues such as gloss, wrinkles, and contour changes, generating RASH without real patient data. To improve robustness to acquisition variability, we perturb the camera viewpoint within $\pm\theta$ in pitch and yaw while keeping all joints visible. We render synthetic bilateral full-hand images X_j^{S} and sample inflammation at 14 joints per hand. Under specialist supervision, we design joint-wise patterns based on clinically reported prevalence rates [17]. Any joint exhibiting non-zero inflammation is labeled positive. Next, we detect 21 hand landmarks using MediaPipe Hands [18], estimate joint centers from 3D coordinates, and crop patches $x_{j,k}^{\mathrm{S}}$ of size $W \times H$ pixels for the 11 joints used downstream, excluding DIP joints. RASH captures diverse RA lesion appearances across severity levels, skin tones, and acquisition conditions, enabling controlled, reproducible synthesis at scale.

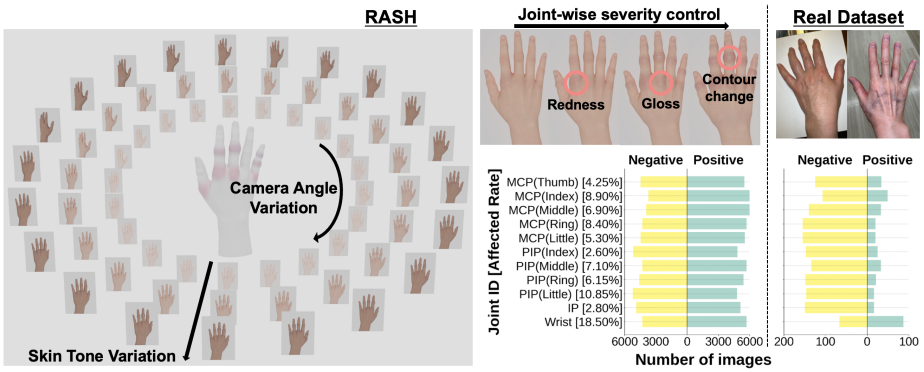

Fig. 2. Left: RASH. **Right**: Real Dataset. Top: representative samples. Bottom: joint-wise class distributions.

2.2 Redness–Gloss–Edge Lesion Attention

Early inflammatory findings in RGB images often appear as subtle erythema, swelling-related contour changes, and specular gloss, yet prior work does not explicitly model these cues. This can bias predictions toward acquisition-dependent appearance and reduce sensitivity.

RGE-LA localizes candidate inflammatory regions using erythema, swelling-related gloss, and contour changes. It takes synthetic patches during pre-training and real patches during fine-tuning and inference. For each patch, we compute three cue maps $R_{j,k}$, $G_{j,k}$, and $E_{j,k}$. Let x_R, x_G, x_B be the pixel-wise RGB values of the joint patch and $x_R^{\mathrm{d}}, x_G^{\mathrm{d}}, x_B^{\mathrm{d}}$ those of a reference dorsum patch x^{d} (cropped via 21 MediaPipe landmarks with a fixed margin). The redness map $R_{j,k}$ uses the dorsum as a normal-skin reference and is computed as

$$R_{j,k} = \max\Big(0,\ \big[x_R - \max(x_G, x_B)\big] - \big[x_R^{\mathrm{d}} - \max(x_G^{\mathrm{d}}, x_B^{\mathrm{d}})\big]\Big) \tag{1}$$

The gloss map $G_{j,k}$ captures reflection on swollen surfaces and is computed in HSV as $V(1-S)$. The edge map $E_{j,k}$ captures swelling-induced shape changes via the Sobel gradient magnitude on a grayscale-converted patch.

We form the cue tensor as $\boldsymbol{m}_{j,k} = [R_{j,k}, G_{j,k}, E_{j,k}]$ and concatenate it with the RGB patch to form a 6-channel input to a lightweight CNN with three parallel cue-specific branches, each producing one attention map. Each branch comprises two convolutional layers with 3×3 kernels and ReLU activation, followed by bilinear upsampling to match the target feature map resolution.

The three attention maps undergo sigmoid activation, contrast enhancement, and normalization to $[0, 1]$. Their weighted sum yields $\boldsymbol{a}_{j,k}$. We apply this attention map to modulate an intermediate feature map $\boldsymbol{U}$, yielding the attention-modulated feature map $\boldsymbol{U}'$:

$$\boldsymbol{U}' = \boldsymbol{U} \odot \{1 + \alpha\,(\tilde{\boldsymbol{a}}_{j,k} - 0.5)\} \tag{2}$$

where $\tilde{\boldsymbol{a}}_{j,k}$ is the normalized attention map and α controls the modulation strength. This increases sensitivity to subtle cues while reducing background-driven responses.

2.3 Morphology–Pathology Dual-Path Network

Single-stream CNNs can entangle morphology variations such as skeletal shape and skin texture with lesion cues such as erythema and swelling, destabilizing predictions across acquisition conditions and skin appearance. Treating joints discards intra-hand context, although RA often affects adjacent joints.

To disentangle these two factors, Dual-Path employs two parallel ResNet18 branches with different supervision signals and feature conditioning. The *Morpho -logy-Path* is trained without lesion attention conditioning, encouraging it to encode acquisition-invariant structural features of the joint, namely skeletal geometry and skin texture that are stable across patients. In contrast, the *Pathology-Path* receives lesion attention modulation from RGE-LA at its intermediate layers and applies the Convolutional Block Attention Module (CBAM) [20] at its final layer, explicitly steering its feature space toward inflammation-discriminative cues such as erythema patterns and swelling-induced appearance changes. This complementary design is inspired by clinical practice, where a specialist simultaneously assesses joint shape and pathological skin changes as independent but synergistic signals.

Concretely, Dual-Path applies the same pipeline to $x^{\mathrm{S}}_{j,k}$ during pre-training and to $x^{\mathrm{R}}_{j,k}$ during fine-tuning and inference. The Morphology-Path encodes joint morphology using ResNet18 [19], producing a morphology embedding $\boldsymbol{h}^{\mathrm{M}}_{j,k}$. In parallel, the Pathology-Path uses another ResNet18 whose Layer-2 features are modulated by the lesion attention map $\boldsymbol{a}_{j,k}$ and whose final-layer features are refined by CBAM [20], yielding a pathology embedding $\boldsymbol{h}^{\mathrm{P}}_{j,k}$. We concatenate and project these embeddings to obtain the fused representation $\boldsymbol{z}^{(0)}_{j,k}$.

RA inflammation often co-occurs across adjacent joints, so per-joint patches can miss cross-joint evidence. To incorporate intra-hand context, we apply a GNN to node features $\boldsymbol{z}^{(0)}_{j,k}$, obtaining context-refined features $\boldsymbol{z}^{(1)}_{j,k}$ [13]. Under specialist guidance, we fully connect the metacarpophalangeal (MCP) joints and connect each finger's MCP to its proximal interphalangeal (PIP) joint [17]. Each node is updated by aggregating neighboring node features, followed by a learnable linear transformation and non-linear activation. A classifier then predicts the joint-wise probability $p_{j,k}$ from $\boldsymbol{z}^{(1)}_{j,k}$, and the final decision is obtained using a threshold τ.

2.4 CAM-guided Domain Alignment

Applying a synthetic-trained model to real images can suffer from shifts in illumination and background. CORrelation ALignment (CORAL) [21] globally matches feature statistics, but can also align nuisance factors unrelated to inflammation, risking negative transfer.

We introduce CDA to restrict alignment to pathology-relevant regions estimated by CAM. We use the P-Path last-layer feature maps $\boldsymbol{V}^{\mathrm{S}}$ and $\boldsymbol{V}^{\mathrm{R}}$. From the positive-class CAM map, we form a binary mask ρ by setting the Top-N% pixels to 1 and the rest to 0, obtaining lesion-focused features $\boldsymbol{F}_{\mathrm{S}}$ and $\boldsymbol{F}_{\mathrm{R}}$. We then minimize

$$\mathcal{L}_{\mathrm{CDA}} = \frac{1}{4d^2}\|\mathrm{Cov}(\boldsymbol{F}_{\mathrm{S}}) - \mathrm{Cov}(\boldsymbol{F}_{\mathrm{R}})\|_F^2 \tag{3}$$

where $\mathrm{Cov}(\cdot)$ computes the channel-wise covariance using spatial locations as samples, $\|\cdot\|_F$ is the Frobenius norm, and d is the number of channels. This prioritizes transferring inflammation-specific cues while reducing sensitivity to domain-specific variations, improving robustness across acquisition conditions.

3 Experiments and Results

3.1 Experimental Settings

RASH Dataset We synthesized 10,000 full-hand images using a 3D hand model and cropped joint patches. Based on the raw joint-wise prevalence rates [17], we defined the sampling probability for each joint by adding a constant offset of 0.45 to the raw prevalence rates to obtain a roughly balanced positive rate close to 0.5 for stable pre-training under severe imbalance, and generated samples according to this distribution. RA prevalence is primarily concentrated in the MCP and PIP joints, while DIP joints show substantially lower clinical involvement [17]; we therefore focus on these 11 joints (MCP and PIP across five fingers, plus the wrist), excluding DIP joints, to match the downstream evaluation protocol and ultrasound-annotated real dataset. We constructed RASH with 110,000 joint images across these 11 joints. Fig. 2 shows the distribution across joints. This is substantially larger and more balanced than existing RA datasets [9, 22, 23].

Real Dataset We collected bilateral full-hand RGB images from 75 RA patients at a collaborating hospital. Images were acquired at baseline and 12 weeks later using an iPhone 11 under specialist supervision across diverse acquisition conditions. Ultrasonography-based Gray Scale (GS) and Power Doppler (PD) scores were assigned to 14 finger joints and the wrist, and a joint was defined as positive if it satisfied GS ≥ 1 and PD ≥ 1. We normalized images, detected 21 landmarks using MediaPipe Hands, and cropped 256×256 pixel joint patches. Under rheumatologist supervision, we curated 354 positive and 1,474 negative patches after excluding DIP joints and removing images with poor joint visibility or improper framing, as summarized in Fig. 2. This dataset reflects realistic variations in illumination, framing, and skin appearance.

Implementation Details All methods were evaluated using patient-level 5-fold cross-validation (no patient appears in more than one split per fold) repeated 30 times. In each fold, a validation split was used to tune hyperparameters and the decision threshold τ, and we report mean test performance over 5 folds and 30 trials. Training consisted of synthetic pre-training (batch size 32, up to 20 epochs) followed by up to 5 epochs of fine-tuning on real data (with early stopping given the small real fine-tuning split, $\sim$1,800 patches). We used AdamW [24] with

Table 1. Comparison with baseline methods and rheumatologists (all sharing the same pre-training/fine-tuning protocol, Sec. 3.1). **Bold**/underline: best/second-best excluding Rheumatologists.

Method	Accuracy[%] ↑	Sensitivity[%]↑	Specificity[%]↑	F1↑	Gmean↑
Ours	**66.3**	**58.4**	67.7	**0.392**	**0.628**
Kato et al.	53.6	46.7	55.3	0.281	0.508
Pathak et al.	63.7	43.4	**68.3**	0.320	0.544
MobileNetV3-Large	61.0	43.1	64.8	0.295	0.529
EfficientNet-B0	62.2	44.6	65.9	0.308	0.537
Rheumatologists	66.8	66.4	66.9	0.417	0.667

horizontal flips, rotations, and color jitter. Hyperparameters were either tuned automatically on the validation split via Optuna's [30] TPE sampler (learning rate, weight decay, Balanced MixUp [25] α, Focal Loss γ, λ_{cda}) or fixed a priori (GNN hidden dimension/dropout, camera perturbation θ, RGE-LA strength α, patch size, CAM ratio N); values and ranges are in the supplementary material. The decision threshold τ, tuned on the validation set to maximize G-mean, averaged $\tau = 0.42$ across folds and trials. Experiments were conducted on Ubuntu 22.04 using an NVIDIA RTX 6000 Ada Generation GPU.

Baseline Methods As baselines, we compare four methods. Kato et al. [9] is a dual-encoder model using DINO-pretrained ResNet18 to fuse global and local joint features [27]. Pathak et al. [10] is an Inception-ResNet-v2 classifier [26]. We additionally include MobileNetV3-Large [28] and EfficientNet-B0 [29] as lightweight-backbone baselines. Our method uses ResNet18 with all proposed components. As a clinical reference, we report the mean performance of three rheumatologists who independently assessed the same RGB images.

3.2 Experiments

Comparison with Baseline Methods Table 1 shows that our method outperforms all baselines across most metrics. A paired t-test on trial-wise G-mean values confirms a significant difference vs. Pathak et al., our strongest baseline ($p = 0.038 < 0.05$); the remaining baselines are not yet tested for significance. Outperforming MobileNetV3-Large and EfficientNet-B0 under the same pre-training protocol further suggests that the gains stem from the proposed components rather than backbone capacity alone.

Ablation Study As shown in Table 2, removing synthetic pre-training yields the largest drop, confirming that RASH mitigates data scarcity and joint-level imbalance and helps learn RA-like patterns. Among the remaining components, removing Balanced MixUp reduces Sensitivity under imbalance, removing RGE-LA weakens subtle cue modeling, removing Dual-Path or GNN degrades joint-level stability by discarding inter-joint context, and removing CDA reduces robustness to acquisition variability. On challenging real images with low inter-rater agreement (Cohen's Kappa = 0.161), our method achieves performance comparable to expert assessment; a paired t-test on G-mean shows no significant difference ($p = 0.053 > 0.05$).

Table 2. Ablation study. **Bold** denotes the best and underline the second-best within this table.

Method	Accuracy[%] ↑	Sensitivity[%]↑	Specificity[%]↑	F1↑	Gmean↑
Ours	66.3	**58.4**	67.7	**0.392**	**0.628**
Ours w/o Balanced MixUp	**69.8**	51.9	73.1	0.360	0.616
Ours w/o CBAM	68.1	56.0	70.4	0.375	**0.628**
Ours w/o RGE-LA	60.1	57.2	62.9	0.355	0.600
Ours w/o Dual-Path	62.4	**58.4**	63.1	0.350	0.607
Ours w/o GNN	62.9	51.4	**74.4**	0.340	0.619
Ours w/o CDA	60.5	57.9	61.0	0.358	0.594
Ours w/o Pre-training	56.0	45.7	58.4	0.305	0.517
Ours w/o Fine-tuning	63.0	54.6	61.8	0.332	0.581

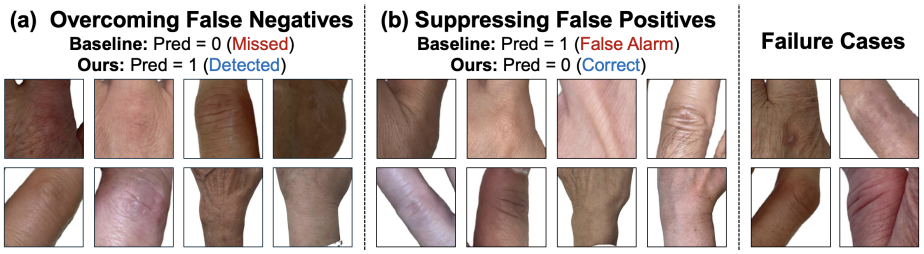

Fig. 3. Qualitative comparison between the proposed method and the baselines of Kato et al. and Pathak et al., illustrating reduced false negatives (a) and false positives (b), with representative failure cases of the proposed method shown on the right.

Qualitative Evaluation Fig. 3 shows representative test-set predictions from our method and the baselines of Kato et al. and Pathak et al. Our method correctly predicts joint-level inflammation where the baselines fail, reducing false negatives and false positives under challenging acquisition conditions.

4 Conclusion and Discussion

We address joint-level binary classification of RA-related inflammation, including early-stage cases, with a framework targeting 11 joints. The approach leverages RASH, a large-scale synthetic dataset reproducing RA manifestations under specialist supervision with control over skin tone, disease severity, camera viewpoints, erythema, and swelling. On a clinically realistic RGB test set with severe class imbalance and visually challenging cases, the proposed components outperform baselines, approach physicians' visual assessment, and remain robust to imaging conditions and skin appearance. This indicates potential for diagnostic support in telemedicine and specialist-scarce settings.

Limitations include subtle cases undetectable from RGB alone, a single-hospital 75-patient cohort, and low inter-rater agreement (Kappa = 0.161). Future work includes multi-stage detection and generative data augmentation.

Acknowledgment This work was partially supported by the KGRI Challenge Grant.

References

1. Mate, G.S., Kureshi, A.K., Singh, B.K.: An efficient CNN for hand X-ray classification of rheumatoid arthritis. Journal of Healthcare Engineering **2021**, Article ID 6712785 (2021). https://doi.org/10.1155/2021/6712785
2. Perez-Siguas, R., Matta-Solis, E., Remuzgo-Artezano, A., Matta-Solis, H., Matta-Perez, H., Perez-Siguas, L.: Hand X-ray image processing system for the early detection of rheumatoid arthritis. In: 2023 3rd International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1–4. IEEE (2023)
3. Huo, Y., Vincken, K.L., Viergever, M.A., Lafeber, F.P.: Automatic joint detection in rheumatoid arthritis hand radiographs. In: 2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI), pp. 125–128. IEEE (2013). https://doi.org/10.1109/ISBI.2013.6556428
4. Ostendorf, B., Peters, R., Dann, P., Becker, A., Scherer, A., Wedekind, F., Schneider, M.: Magnetic resonance imaging and miniarthroscopy of metacarpophalangeal joints: sensitive detection of morphologic changes in rheumatoid arthritis. Arthritis Rheum. **44**(11), 2492–2502 (2001)
5. Rubin, D.A.: MRI and ultrasound of the hands and wrists in rheumatoid arthritis. I. Imaging findings. Skeletal Radiol. **48**(5), 677–695 (2019)
6. Tang, J., Jin, Z., Zhou, X., Zhang, W., Wu, M., Shen, Q., Yuan, J.: Enhancing convolutional neural network scheme for rheumatoid arthritis grading with limited clinical data. Chinese Physics B **28**(3), 038701 (2019)
7. Kondo, Y., Kaneko, Y., Takeuchi, T.: Differential diagnosis of inflammatory arthritis from musculoskeletal ultrasound view. Rheumatol. Immunol. Res. **3**(2), 54–60 (2022)
8. Quintana-López, G., Maldonado-Cañón, K., Flórez-Suárez, J.B., Méndez-Patarroyo, P., Coral-Alvarado, P., Calvo, E.: Correlation and Agreement between Physical and Ultrasound Examination after a Training Session Dedicated to the Standardization of Synovitis Assessment in Rheumatoid Arthritis Patients. Advances in Rheumatology **61**(1), 68 (2021)
9. Kato, S., Kondo, Y., Saito, S., Aoki, Y., Isogawa, M.: Image-based joint-level detection for inflammation in rheumatoid arthritis from small and imbalanced data. arXiv preprint arXiv:2602.14365 (2026)
10. Pathak, S.: Deep learning approach to early detection of rheumatoid arthritis. In: IEEE Integrated STEM Education Conference (ISEC), pp. 1–2 (2024)
11. Sayah, A., English III, J.C.: Rheumatoid arthritis: A review of the cutaneous manifestations. J. Am. Acad. Dermatol. **53**(2), 191–209 (2005). https://doi.org/10.1016/j.jaad.2004.07.023
12. Habibullah, T., Habibullah, A., Simsim, R.: Skin Manifestations of Rheumatological Diseases. In: Skills in Rheumatology. Springer, Cham (2021)
13. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Transactions on Neural Networks **20**(1), 61–80 (2009)
14. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929 (2016)
15. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2980–2988 (2017)
16. Blender Foundation: Blender, https://www.blender.jp/, last accessed 2025/02/25

17. Terao, C., et al.: Three groups in the 28 joints for rheumatoid arthritis synovitis–analysis using more than 17,000 assessments in the KURAMA database. PLoS ONE **8**(3), e59341 (2013)
18. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., Grundmann, M.: MediaPipe Hands: On-device real-time hand tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
20. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
21. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, no. 1 (2016)
22. Chaturvedi, N.: DeepRA: Predicting joint damage from radiographs using CNN with attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
23. Lien, C.-Y., Wang, H.-J., Lu, C.-K., Hsu, T.-H., Chu, W.-C., Lai, C.-C.: Deep learning with an attention mechanism for enhancing automated modified total sharp/van der Heijde scoring of hand X-ray images in rheumatoid arthritis. Journal of Medical and Biological Engineering (2025)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
25. Galdran, A., Carneiro, G., González Ballester, M.A.: Balanced-mixup for highly imbalanced medical image classification. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS, pp. 323–333. Springer, Cham (2021)
26. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016)
27. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 9650–9660 (2021)
28. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for MobileNetV3. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1314–1324 (2019)
29. Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 6105–6114 (2019)
30. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), pp. 2623–2631 (2019)