RheumaVIT: transformer-based model for Automated Scoring of Hand Joints in Rheumatoid Arthritis

Alexander Stolpovsky *[1], Elizaveta Dakhova * *[1], Polina Druzhinina * †[1,2], Polina Postnikova[4], Daniil Kudinsky[4], Alexander Smirnov[4], Anastasia Sukhinina[4], Alexander Lila[4], and Anvar Kurmukov[1,3]

[1]Artificial Intelligence Research Institute (AIRI)
[2]Skolkovo Institute of Science and Technology
[3]Institute for Information Transmission Problems
[4]V.A. Nasonova Research Institute of Rheumatology

Abstract

Rheumatoid arthritis (RA) is an autoimmune disease that causes chronic inflammation, joint destruction, and extra-articular manifestations. Radiography is the standard imaging modality for diagnosing and monitoring joint damage in RA. However, the commonly used Sharp method and its variants, which evaluate radiographic progression, are time-consuming and subjective. Automated joint evaluation using deep neural networks can address these challenges. This study introduces RheumaVIT, a novel vision transformer-based pipeline for automatically scoring hand joints affected by RA. The method consists of two stages: a regression model for joint localization and a transformer-based architecture for assessing erosion and joint space narrowing (JSN). Our approach demonstrates superior accuracy (up to 12% higher for erosion and 2% higher for JSN) compared to existing state-of-the-art methods. Moreover, it has a promising ability to detect common patterns of erosion and JSN through roll-out interpretation. To promote further research, we are open-sourcing our clinical collection since there is no annotated dataset on RA available in the public domain. Our paper contributes to the progress of automated joint assessment in rheumatoid arthritis, offering potential applications in both clinical practice and research.

1. Introduction

Rheumatoid arthritis (RA) is an autoimmune disease that causes chronic inflammation, joint destruction, and extra-articular manifestations in multiple organs. It primarily affects the small joints of the hands and feet, leading to irreversible deformities and loss of physical function [17, 22]. The disease can impair an individual's capacity to work, significantly reduce quality of life, and even shorten lifespan. However, modern treatment advances can suppress joint inflammation, prevent joint damage, and improve prognosis [20]. Radiography is the standard imaging modality for diagnosing and monitoring the progression of joint changes in RA [2], with erosion and joint space narrowing (JSN) being the most reliable parameters [4, 41]. The Sharp method [10, 12, 13, 14, 18, 19, 28, 31, 32, 38, 39, 40] and its variants are the most commonly used scoring systems to quantify the severity and rate of joint damage progression in clinical trials and scientific studies [29]. They individually assess erosion by the depth of damage from 0 to 5 points and JSN from 0 to 4 points in the joints of the hands and feet. However, the Sharp-like methods have a significant drawback: the evaluation requires a considerable amount of time as experts must manually appraise and assign scores to each joint. Additionally, this method's consistency is not perfect due to the subjectivity of the evaluation criteria, as shown in Table 1.

Transitioning to automated joint evaluations based on deep neural networks can overcome these limitations. The processing time of the neural network model is significantly shorter than that required for manual image processing, taking just a few seconds. Moreover, because the model uses aggregated evaluations from multiple specialists, its predictions are theoretically more representative than the evaluation of a single physician [4].

With the advancement of machine learning algorithms, the use of neural networks for automatic joint assessment has garnered considerable attention due to their ability to

*dahova.eyu@phystech.edu
†pvdruzhinina@gmail.com

process visual information. A wide array of previous studies have explored models ranging from shallow to deep, including basic multi-layer convolutional neural networks [7, 37] and advanced architectures like VGG [7], U-Net [35, 25], YOLO [42], as well as various data pre-processing techniques such as soft tissue removal, the Sobel filter, and CLAHE [27].

While many current studies boast results comparable to expert annotations, several limiting factors persist. First, all these studies showcase results on different private datasets, which are inaccessible [27, 30, 33, 37, 42]. Furthermore, many lack rigorous evaluations of the proposed methods' performance, often resorting to imbalanced metrics in their analyses [7, 15, 27, 37, 43]. These challenges significantly impede the comparison of approaches and the identification of a definitive state-of-the-art solution.

Transformers are currently gaining popularity in image analysis, presenting new opportunities in the machine learning domain. It was once believed that to achieve quality comparable to convolutional neural networks (CNNs), large datasets and models with numerous parameters were necessary [9]. However, recent studies have introduced compact transformer models such as TinyViT [44], SWIN [24], and MobileViT [26]. Moreover, there have been efforts to train these models on relatively small datasets, still achieving commendable performance [11, 21, 36].

In this paper, we introduce the first vision transformer-based approach for joint assessment, named RheumaVit. This method refines existing techniques, offering a superior and more adaptable solution. The primary contributions of our work include:

- We developed RheumaVIT, an automated system that assigns scores to hand joints affected by RA. This system is two-stage: a regression model identifies joint locations, while a transformer-based architecture assesses the erosion and JSN scores of each joint.
- To leverage the transformer architecture and prevent overfitting, we propose a strategy to augment input data samples in accordance with the number of joints, instead of hand images. We also implemented expanded bounding boxes, transfer learning strategies, and advanced training techniques.
- Our proposed method displayed superior performance (up to 12% increased accuracy for erosion and 2% for JSN) and faster convergence when compared with recent state-of-the-art methods [33]. We also showcased the model's efficacy in precisely identifying patterns related to erosion and JSN in a clinical context through the implementation of roll-out interpretation.
- Lastly, we are releasing an annotated dataset consisting of 330 hand radiographs. This dataset features box annotations for 42 hand joints (21 boxes per hand) accompanied by joint-specific erosion and JSN labels, curated by three radiologists.

The remainder of the paper is structured as follows: Section 2 reviews related work pertinent to our approach. Section 3 delves into the specifics of the proposed pipeline and details the experimental setup. Sections 4 and 5 then present the results, followed by a discussion and conclusions, respectively.

2. Related Works

Research on the automated assessment of hand joints in RA varies significantly in aspects from problem formulation to the methods employed and their evaluation. First, some studies concentrate on specific joints, such as the metacarpophalangeal and proximal interphalangeal joints [6, 16]. In contrast, others classify the carpometacarpal (CMC) and intercarpal joints [30]. Intercarpal joints (wrist) are frequently overlooked in studies due to their intricate nature and proximity. Although their inclusion can hamper model performance [16], from medical perspective, they are as crucial as other hand joints.

When considering the number of stages in a solution, all methods for automatic joint assessment fall into two primary categories: single-stage and two-stage methods. Single-stage methods evaluate all joints simultaneously, whereas two-stage methods first identify the joints before conducting an individual assessment [7, 37, 27, 35, 6, 43]. Early attempts to create an automated RA assessment utilized straightforward architectures, such as multi-layer neural networks, complemented by a range of data preprocessing techniques [7, 37]. Later research incorporated more complex models, including pre-trained VGG and CaffeNet, as well as advanced image preprocessing techniques like soft tissue removal, the Sobel filter, MSGVF snakes for phalanx segmentation, and CLAHE [27]. A distinct approach is presented in [35], where the authors employ a lightweight U-Net architecture for bone segmentation, followed by a YOLOv3 model for joint detection, and finally, a VGG for erosion and JSN assessment. The authors of [6] utilize RetinaNet for finger detection and then an EfficientNet with attention mechanisms to score the JSN and erosion of joints. Lastly, [43] treats joint localization as a regression problem.

Several studies have explored single-stage methods, bypassing the intermediate joint localization step in either an end-to-end scoring scenario or by predicting both joint boxes and scores simultaneously [30, 42, 25]. Radke's study showcased the capabilities of a dual-headed RetinaNet network with adaptively changing Intersection over Union (IoU) values for effective recognition of smaller entities like finger and wrist joints [30]. In [42], the authors employ YOLOv4, enhancing it with adjustments to error

functions, aspect ratios, and distinct handling of hand and finger joints. Conversely, [25] treats the scoring task as a segmentation challenge, using a U-shaped architecture to predict a joint class (localization) for each pixel and assigning erosion and JSN scores to each pixel representing a non-background joint.

While single-stage methods might seem more efficient in terms of speed and memory usage, their practical performance largely hinges on the specific models employed and data processing techniques. Such models often possess a global receptive field, facilitating the inclusion of additional context during training. However, this can also introduce potential confounding elements like technical markup variations [30, 42, 25]. In contrast, two-stage methods focus on localizing each individual joint, yielding a higher number of training samples and considerably mitigating the influence of various confounding factors, albeit at the cost of the benefits of a global receptive field.

Previous studies have shown promising results regarding the accuracy and reliability of AI systems compared to manual assessment by experienced clinicians or radiologists. Nonetheless, further research is necessary to address potential issues, such as overfitting or a lack of robustness against varying conditions across different datasets, before these systems can be reliably implemented in clinical practice.

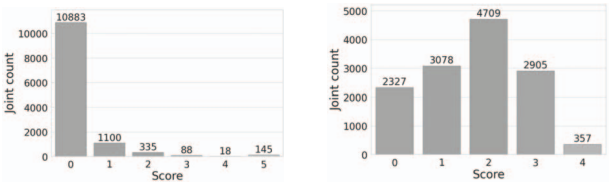
3. Methods and Materials

In this section, we introduce a novel deep learning-based pipeline for automated joint assessment following the Sharp methodology, termed RheumaVIT. Our proposed solution consists of two phases: the localization of joints and followed by scoring each one as depicted in Figure 2.

While the task of joint localization is relatively straightforward and can be efficiently tackled using a regression model, the assessment phase demands greater scrutiny and research. To estimate the erosion and JSN scores, we explore modern deep architectures based on both convolutional neural networks and vision transformers. We consider models such as VGG, EfficientNet, SWIN, MobileViT, and TinyViT [34, 9, 44, 24, 26]. In essence, our study offers the inaugural empirical examination of vision transformer networks' efficacy in scoring hand joints afflicted by RA.

3.1. Study design

To the best of our knowledge, no open radiographic dataset exists that classifies joint space narrowing (JSN) and joint erosion according to the Sharp method. Addressing this gap, the V.A. Nasonova Research Institute of Rheumatology curated a clinical dataset comprising 330 bilateral



(a) Erosion (b) JSN

Figure 1: Distribution of erosion (a) and JSN (b) scores over the entire dataset.

hand radiographs [1]. These images were annotated by three radiologists with respective experiences of 5, 7, and 30 years. Each patient is represented by a single radiograph, with an average age of 48 ± 14 and the male/female ratio is $54 : 298$. These radiographs were collected between 2019 and 2022. 42 patients are classified as normal, while the remaining have a confirmed diagnosis of rheumatoid arthritis.

A modified version of the Sharp method was employed for the annotation process. This adaptation encompasses all hand joints and consolidates closely spaced and overlapping wrist joints into one collective group. The suggested technique identifies 17 erosion sites per hand: interphalangeal, 4 distal interphalangeal, 4 proximal interphalangeal, 5 metacarpophalangeal, 1st metacarpal base, wrist, ulnar, and radiocarpal. For JSN, this method considers 19 areas per hand: interphalangeal, 4 distal interphalangeal, 4 proximal interphalangeal, 5 metacarpophalangeal, carpometacarpal 3 to 5, wrist, and radiocarpal. In line with the standard Sharp methodology, erosion scores span from 0 to 5, and JSN scores from 0 to 4; a higher score indicates more extensive damage [29]. To determine the erosion score for the wrist, we consider the maximum score among the multangular, navicular, lunate, and radius erosion scores. Similarly, the highest score between multangular-navicular and capitate-navicular-lunate is taken as the overall wrist JSN score.

To achieve a more objective assessment, the joints were evaluated by three radiologists. However, due to the inherent subjectivity of this approach, the consistency of their results varied. This underscores the importance of combining assessments from multiple physicians to enhance the representativeness of a particular study. While the majority vote is the most straightforward aggregation method, it proves inadequate in situations with missing scores or ambiguous decisions. Consequently, we employed the David-Skene method [8] across our dataset to derive the final JSN and erosion scores. Originally proposed for medical data aggregation, this method facilitates adjustments for expert discrepancies and produces an outcome closely aligned with

[1]The dataset is published under a license Creative commons BY-NC-SA 3.0 at `https://airi.net/upload/dataset/AIRINIIReumHands.zip`

the observed data.

The distribution of erosion and JSN scores is depicted in Figure 1. Most narrowing categories appear consistently represented, except for scores two and four. The erosion distribution exhibits an exponential trend, with a marked decline in joint count as the erosion score increases.

We also annotated joint locations in 268 radiographs, capturing all 21 standard joint locations used for scoring. The wrist region covers various minor wrist joints, including the multangular, navicular, lunate, multangular-navicular, and capitate-navicular-lunate.

For neural network pre-training, we leveraged an open dataset from [43] comprising data from 3818 patients. In our experiments, we only used radiographs, as the accompanying evaluations lacked joint scores for erosions and JSN.

3.2. Data preprocessing

Each radiograph in the dataset shows the patient's left and right hands. The joints of both hands are symmetrical, which allows us to split the radiograph into two individual images. This approach halves the quantity of joints for which predictions are made, while doubling the amount of data available for training the localization model. To separate the hands in the image, the Minimum Energy Seam algorithm [3] was utilized. Fundamentally, it ascertains a linked sequence of pixels across the entire image with the smallest gradient. We implement morphological opening to reduce the background noise and CLAHE [27] to increase the gradients at the boundary of the hands and the background.

A similar data partitioning approach is used for classification modeling. Based on the predictions of the hand image localisation model, 21 joint regions are identified for which distinctive predictions are then made. This enables the volume of training data to be increased by a factor of 21, and the output size of the model to be reduced by a factor of 21. However, this technique has a limitation - the visibility of the model is confined to a small region around the joint. To overcome this, we randomly increase the area around the joint by up to 4 times during model training. During evaluation, we make use of the average size of the observable area to the model during the training process, which is 2.5 times the size of the original area.

3.3. Proposed method

Joint localization task involves determining the coordinates of each joint's position in an image. This is usually done using bounding boxes, which are defined by the coordinates of two opposite corners. Due to human anatomy, the number of joints remains constant for most patients. In other words, models only need to predict a fixed number of coordinates for the entire image. Therefore, we employ a regression model as the localization task. The backbone of our model is EfficientNet-V2-L [34]. Subsequently, average pooling with an 8×8 kernel is applied to the $1280 \times 16 \times 8$ feature map, followed by flattening into a vector. The head of the model consists of two fully connected layers with ReLU activation and a hidden layer size of 512. To enhance the accuracy of the model, we emploed the CoordConv technique [23] and concatenate the input image with two additional channels, corresponding to x and y pixel coordinates. Additionally, during training, MSE loss was initially minimized in the early epochs, followed by the main part of training using DIoU loss to improve localization metrics, as described in [42].

$$L_{loc} = \begin{cases} \text{MSE}(\mathbf{b}, \widehat{\mathbf{b}}), & \text{if epoch} \leq 50 \\ \text{DIoU}(\mathbf{b}, \widehat{\mathbf{b}}), & \text{otherwise} \end{cases} \tag{1}$$

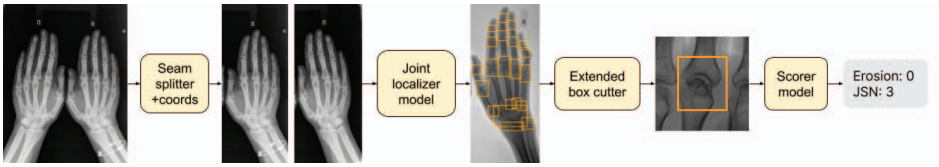
$$\text{DIoU}(\mathbf{b}, \widehat{\mathbf{b}}) = 1 - \text{IoU}(\mathbf{b}, \widehat{\mathbf{b}}) + \frac{\rho^2(\mathbf{b}, \widehat{\mathbf{b}})}{c^2}, \tag{2}$$

where $\mathbf{b}$ denotes the bounding box coordinate vector $(x_{min}, y_{min}, x_{max}, y_{max})^T$ with corresponding prediction $\widehat{\mathbf{b}}$, $\rho(.)$ is the Euclidean distance and c is the diagonal length of the smallest enclosing box covering the two boxes.

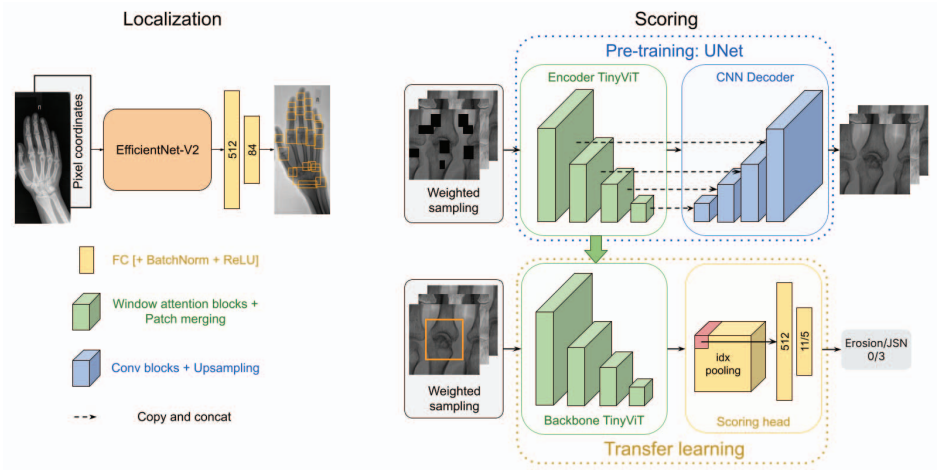
Joint assessment follows joint localization. The goal is to determine the erosion and JSN scores for each joint, depending on the type of joint. These scores have discrete scales ranging from 0 to 5 and 0 to 4, respectively. The higher the score, the greater the damage is to the joint. This makes joint scoring to be treated as a classification task. However, if a degenerate distribution is used as the target during model training, there will be no gradations between classes. To counter this, we specify the following distribution for the targets. Let p be the probability of labeling error. For the true class, we set the probability of the target p_c to $1 - p$ if there are two neighboring classes, and $1 - \frac{p}{2}$ otherwise. For neighboring classes, we set the target probability to $\frac{p}{2}$, and for all other classes - 0. In this way, regardless of the error function, the model is penalized less for deviations from the norm of 1 point than for more significant deviations. We set $p = 0.4$. As a loss function we use the focal loss. Assuming q is the predicted probability of target class, the error function equation is

$$L_{scor} = FL = -\alpha(1 - q(x))^{\gamma} p_c(x) \log q(x) \tag{3}$$

TinyViT was used as a backbone for the erosion and JSN estimation models, providing the feature map. An idx pooling operation was applied to the obtained feature map with dimensions $channels\_cnt \times 7 \times 7$, which selects the top-left vector, i.e., the elements with zero values for the last two coordinates. This approach effectively creates an equivalent of the CLS token for the TinyViT transformer, compensating for its absence. The model's head consists of two fully connected layers with a hidden dimension of 512, BatchNorm



(a) Full pipeline



(b) Pipeline model architectures

Figure 2: (a) RheumaVIT is composed of the following steps. Input X-ray image of hands is split into two images of left and right hands via seam splitting; a localization model then estimates the bounding boxes for each hand's joints. An extended region around the joint is cropped and fed to the input of the scoring model for erosion and JSN assessment. Specifically, (b) The localization stage is formulated as a regression task and solved using the EfficientNet backbone, while the TinyViT-based scoring model handles erosion and JSN classification. Pre-training the scoring model on external data in an unsupervised fashion substantially enhanced the model's quality.

and ReLU activation function between them. The output dimension of the head is 11, comprising 6 logits for erosion and 5 logits for narrowing. During training, weighted sampling was employed to form batches, with the weight set to be inversely proportional to the frequency of the corresponding class.

Transfer learning was used to reduce model overfitting. We pre-trained the backbones (TinyViT and SWIN) on another publicly available dataset of 3818 hand radiographs [5]. The pre-training was performed in an image-to-image self-supervised setup, where the task was to reconstruct the clean image from a corrupted version. Our pre-training used the backbone as a block encoder in the UNet model. The corruption process involved adding Gaussian noise, applying geometric elastic transformations, and cropping small rectangles. The error metric for pre-training was an L1 loss function.

Medical imaging models are more likely to be positively received by physicians if they have good interpretability. Inspired by the attention rollout method [1], we adopted it as the main idea for our interpretation method. However, it was not possible to apply it directly to the used transformers due to the differences from ViT [9]. Firstly, attention is not computed over the whole image, but over windows that are shifted by half in the case of the SWIN transformer. Secondly, downsampling sometimes occurs when four windows are merged into one by combining 2x2 patch blocks into a single block. Lastly, these transformers lack a CLS token, using instead our idx pooling equivalent. Therefore, we modified the rollout procedure to obtain a correct attention matrix as output. In addition, due to the significantly smaller patch size in these transformers compared to ViT, the interpretation results are more detailed.

3.4. Implementaion details

The models were implemented using PyTorch and trained on a server equipped with two NVIDIA V100 GPUs, each with 32 GB of memory, and an Intel Xeon Gold 6278C processor.

The models for erosion and narrowing assessment were also trained in two stages. The first lasted 50 epoch and assumed a frozen backbone with a starting learning rate of

10^{-3}. The next implied 100 epochs with a fully unfrozen backbone and a starting learning rate of 5×10^{-5}. The AdamW optimizer was used with an L2 weight decay parameter of 10^{-2} and gradient clipping. The batch size was set to 32. The focal loss was employed with the standard parameter, indicating that the weight is quadratically proportional to the deviation from the true probability. The choice of the lr-scheduler depended on the backbone architecture: ReduceLROnPlateau was used for regular convolutional networks, and WarmupCosineSchedule for transformer networks. During training the following image augmentations were applied: InvertImg, ColorJitter, HorizontalFlip, RandomRotate90, Rotate, CropAndPad.

3.5. Evaluation Metrics

As quality metrics for the localization task, we used the traditional Intersection over Union (IoU) score, as well as a new metric called localization accuracy. This metric evaluates the relative accuracy of the predicted bounding box location, in relation to the true bounding box location. Specifically, as long as the center of the predicted box is within the target bounding box plus or minus half the width and height of the box, it is considered a true positive. Localization accuracy has been found to be more intuitive and reliable than traditional detection and segmentation metrics, as it accounts for any errors in bounding box annotation.

In order to address the class imbalance associated with the joint scoring task, we employed the weighted metrics. These included multi-class weighted accuracy (5 classes for JSN, 6 classes for erosion), and weighted mean absolute error (MAE).

4. Results

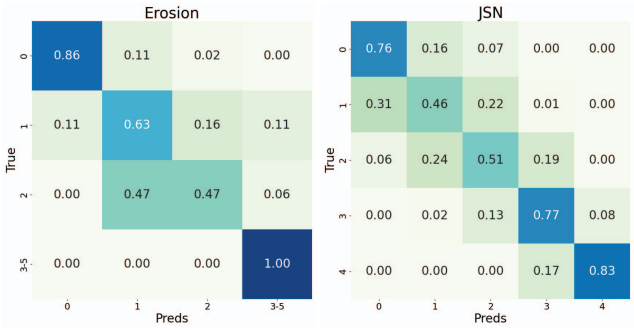
In this section, we demonstrate the efficacy of a transformer-based deep learning approach for automatic joint assessment. We compare the performance of various existing solutions and the latest state-of-the-art models based on convolutional and transformer architectures. We also detail the techniques applied to further promote productive learning and increase the quality of the resulting method RheumaVIT.

Table 1: Erosion and JSN scoring performance. Accuracy and MAE are weighted across the joints. Average experts' variability is measured in a 1 vs 2 regime.

	Erosion		JSN	
	Accuracy ↑	MAE ↓	Accuracy ↑	MAE ↓
RetinaNet [30]	0.43 ± 0.09	1.68 ± 0.24	0.41 ± 0.05	0.58 ± 0.08
Team Shirin	0.62 ± 0.05	0.93 ± 0.17	0.60 ± 0.04	0.43 ± 0.05
Csabaibio	0.56 ± 0.05	0.95 ± 0.16	0.65 ± 0.04	0.36 ± 0.04
Zbigniew Wojna [25]	0.53 ± 0.05	1.22 ± 0.21	0.58 ± 0.04	0.45 ± 0.05
RheumaVIT (ours)	$\mathbf{0.74 \pm 0.05}$	$\mathbf{0.75 \pm 0.18}$	$\mathbf{0.67 \pm 0.04}$	$\mathbf{0.36 \pm 0.05}$
Average expert	0.79 ± 0.03	0.48 ± 0.13	0.76 ± 0.03	0.25 ± 0.04

To evaluate the effectiveness of the proposed pipeline, we compared it against the state-of-the-art solutions. First, we consider the top solutions in three different RA2 Dream Challenge rankings [33]. Owing to the annotation constraints, our analysis covers the top-ranked solution from the Team Shirin [2], Csabaibio [3] and Zbigniew Wojna teams [25] and a recent detection-based method for a more extensive analysis [30]. The methodologies and associated hyperparameters of each competing solution were implemented in accordance with the specifications described in the corresponding original papers or published open-sourced code.

Team Shirin's two-stage approach, which utilizes transfer learning and fine-tuning of high-performance CNN models (ResNet34, DenseNet201), proved stronger than other existing approaches for the erosion task. At the same time, Csabaibio's two-step solution, which applies detection model and ensemble of CNNs, achieves the highest JSN scoring. Based on the Accuracy and MAE metrics displayed in Table 1, our proposed pipeline RheumaVIT outperforms the competing solutions in both tasks. We observe an Accuracy of 0.74 ± 0.05 and MAE of 0.75 ± 0.18 for erosion estimation, and 0.67 ± 0.04 Accuracy and 0.36 ± 0.05 MAE for JSN estimation. The confusion matrices of the best model for erosions and JSN are presented in Figure 3a and Figure 3b, respectively. The model's prediction quality clearly aligns with that of human annotators. Notably, for certain joints, the model's estimation of erosion and JSN scores even surpasses that of some physicians.



(a) Erosion (b) JSN

Figure 3: Confusion matrices for the RheuthmaVIT predictions of erosion (a) and JSN (b) scores.

4.1. Localization

We considered two main approaches to address the localization problem: detection and regression. We implemented RetinaNet and SSD as detection models and observed that

[2] https://doi.org/10.7303/syn21478998
[3] https://github.com/patbaa/RA2_dream

Table 2: Joints localization results.

	IoU	Loc Accuracy
RetinaNet	0.21	0.18
SSD	0.68	0.93
Resnet-50 + MSE loss	0.56	0.90
Resnet-50 + DIoU-loss	0.65	0.98
Resnet-50 + DIoU-loss + coordinate grid	0.68	0.98
EfficientNet-V2-S + DIoU-loss + coordinate grid	0.71	**0.99**
EfficientNet-V2-L + DIoU-loss + coordinate grid	**0.72**	0.98

both detection-based models and the basic variant of the regression model bring comparable results in terms of localization accuracy. For instance, as Table 2 demonstrates, the model of SSD yields an IoU score of around 0.68, which is roughly equivalent to the experiment with ResNet-50 trained via DIoU-loss and a coordinate grid.

Nevertheless, the detection-based model architecture is excessive for joint localization tasks, since it considers the possibility of multiple boxes of the same class, while in current task all hand joints are unique and have the exact amount. Therefore, we prioritized the further exploration of the regression approach. We examined different loss functions (MSE and DIoU) and backbone architectures (ResNet and EfficientNet). Further, we found that utilization of a coordinate grid as an additional input channel and generation of a loss minimization scheme through a gradual transition from MSE loss to DIoU-loss in later epochs resulted in remarkable improvement of both IoU and localization accuracy scores (0.72 and 0.99 respectively).

4.2. Ablation study of erosion and JSN scoring

Table 3: Ablation study to establish most efficient backbone for scoring stage in RheumaVIT. Accuracy and MAE are weighted across joints. Average experts' variability is measured in a 1 vs 2 regime.

	Erosion		JSN	
	Accuracy ↑	MAE ↓	Accuracy ↑	MAE ↓
VGG	0.63 ± 0.05	0.79 ± 0.19	0.57 ± 0.04	0.56 ± 0.07
EfficientNet	0.58 ± 0.05	$1.23 \pm 0,28$	0.62 ± 0.04	0.40 ± 0.05
SWIN	0.62 ± 0.05	0.95 ± 0.21	0.64 ± 0.04	0.40 ± 0.05
TinyViT	0.74 ± 0.05	0.75 ± 0.18	0.67 ± 0.04	0.36 ± 0.05
MobileViT	0.61 ± 0.05	1.06 ± 0.23	0.63 ± 0.04	0.41 ± 0.05
Average expert	0.79 ± 0.03	0.48 ± 0.13	0.76 ± 0.03	0.25 ± 0.04

To determine the optimal solution for the feature evaluation task of rheumatoid arthritis, we undertook a comprehensive ablation study. Through our experiments, we identified the best combination of factors and the most effective backbone architecture, which together produced the highest-quality results for evaluating scoring models.

We identified three most important factors that influence the quality of the models: bounding box expansion coefficients, pretraining of backbone on an external dataset, and the use of two independent models or a single shared model to predict erosion and JSN scores. (Appendix B) For erosion estimation the most effective was pre-training the backbone model on an external dataset and then fine-tuning the final model with a combination of extended joint images for both erosion estimation and JSN tasks. 3b Remarkably, the models for JSN scoring showed superb performance when trained independently from the erosion estimation models, with pre-training the backbone model and usage of augmented boxes having negligible influence on the final quality.

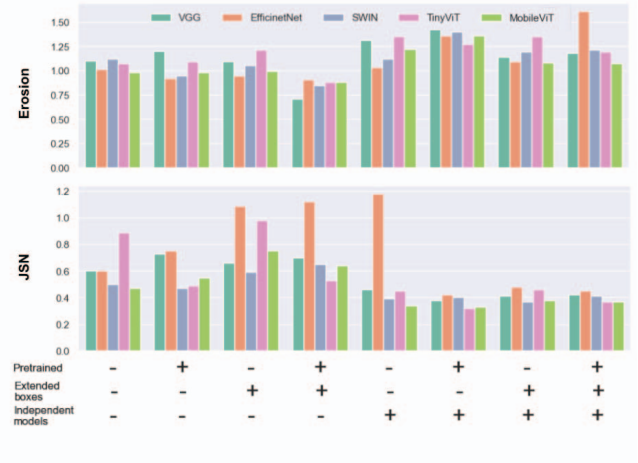
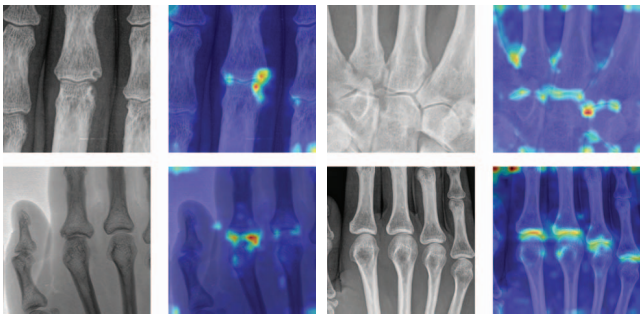


Figure 4: The histograms corresponding to MAE metric of the models for erosion (a) and JSN (b) evaluation

We also considered five different models as backbones: VGG, EfficientNet, Swin, TinyViT, and MobileViT. We carefully chose the optimal hyperparameters for training for each backbone model. Through a thorough search on three key factors and five different backbones, conducted on a subset of the data, we successfully determined the most suitable combination of factors for evaluating erosion and JSN models. VGG, SWIN, and TinyViT models demonstrated the highest performance. However, it can be observed that the TinyViT transformer model excelled as among other competing backbones Table 3 both for each metrics and for each task demonstrating for erosion 0.74 ± 0.05 Accuracy and 0.75 ± 0.18 MAE and 0.67 ± 0.04 Accuracy, 0.36 ± 0.05 MAE for JSN, accordingly.

4.3. Interpretation

To gain insights into our model, we applied feature map visualization using the attention rollout method. This technique provides a comprehensive understanding of how input patterns are processed internally, allowing us to trace the paths of individual neurons through the network's layers and observe the contributions of different elements to the



(a) Erosion (b) JSN

Figure 5: Interpretation maps for erosion (a) and JSN (b), obtained using attention roll-out method [1]. Target joint is in the center of each image.

model's decision-making process. The interpretation examples of the best models are shown in Figure 5. For JSN, the model primarily focuses on the area between the joints, while for erosion, the focus is on the bone borders. This observation is consistent with the definition of these destructive changes. Furthermore, when examining a specific joint, the JSN model tends to consider neighboring joints more frequently and extensively than the erosion model. Based on the results, it can be inferred that the model effectively captures task-specific patterns that are essential for both erosion and JSN tasks.

5. Discussion and conclusion

This paper offers a thorough review of the methodologies currently in use for automated joint assessment and showcases the effectiveness of transformer-based models. Through extensive empirical evidence, we confirm that the proposed vision transformer-based pipeline, RheumaVIT, surpasses existing methods, delivering superior accuracy and precision.

Our study utilized a dataset curated by the V.A. Nasonova Research Institute of Rheumatology, comprising 330 hand radiographs annotated according to the Sharp method. It's worth noting that while current datasets aren't publicly accessible, we will make ours available for future research. Additionally, we furnish a detailed analysis of established approaches and modern models.

Despite the complexity and variability among patient images, the most efficient algorithms consistently achieved a commendable prediction accuracy and reproducibility.

Our investigation led to several significant observations. If the prior research established that JSN task was probably much more straightforward for the model, given that it involved measuring distances, while the identification of erosions necessitated in-depth understanding of bone morphology and destruction. It can be deduced that the experiments based on transformer models yielded a superior performance in the JSN task, potentially due to the advanced components of the architecture that most likely enabled a more exact understanding of the various morphological properties of erosion identification. Besides, considering the characteristic postures present in radiographs, such as bent fingers, the true distance between the bones can be misinterpreted, thus making it more challenging for the model to accurately predict JSN score.

In this study, it was observed that for the assessment of erosion and JSN, a single shared model demonstrated the best quality for erosion, while independent models tended to provide a better performance for the JSN assessment. Clinically, arthritis progression typically entails firstly the appearance of JSN and then the emergence of erosions, hence it was hypothesized that accounting for both JSN and erosion in a shared model might improve the accuracy of the overall assessment of erosion. But an independent model seemed to perform better than a shared model for the assessment of JSN alone, though the exact reason behind this remains inconclusive. Nonetheless, in this particular investigation, we assumed the main reason is the chosen pipeline due to the uneven distribution of erosion weights, and training a single shared model for JSN might have had an adverse effect on its assesment.

The evaluation of the best model with roll-out interpretation revealed its ability to capture the clinical specific patterns typical for erosion and JSN progression. This suggests that the model is capable of performing sophisticated diagnoses regarding erosion and JSN status of joint areas based on the images.

In comparison to convolution models, transformers are typically perceived as requiring ample data to be effectively trained. The development of complex architectures and extensive research on the behavior of vision transformers have made it possible to adapt them to the most existing task. Furthermore, there are a lot of methods that can improve the quality and prevent overfitting, such as pre-training, transfer learning, augmentations techniques, as well as more specific measures such as increasing the size of bounding boxes, weighted sampling, smoothed loss function, and careful selection of hyperparameters. Through a combination of these methods, it is possible to reach new heights of performance when dealing with these tasks.

This study presents certain limitations associated with the number of images available for training. For algorithms to possess a greater level of reliability, large repositories of annotated images taken from extensive observational studies and clinical trials must be used for experiments. Additionally, the models should be subject to constant validation on new data in order to ascertain the accuracy and robustness of the algorithms.

For potential future research, extending the joints assess-

ment to the additional body areas commonly affected by arthritis deserves further exploration. Several studies already include the diagnosis of small joints of the feet. Additionally, there is a particular benefit and practical value in increasing the model's diagnostic capability to recognize the stages of arthritis and even determine the type of arthritis (e.g., rheumatoid, acute, and psoriatic). Also, prediction of disease progression could offer greater insight and assistance in clinical trials and research. This provides an opportunity to leverage the most modern technology, including, for example, modern multimodal approaches as well as may need additional features or modalities to obtain accurate results.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[2] Daniel Aletaha and Josef S Smolen. Diagnosis and management of rheumatoid arthritis: a review. *Jama*, 320(13):1360–1372, 2018.

[3] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. pages 10–es, 2007.

[4] Alix Bird, Lauren Oakden-Rayner, Christopher McMaster, Luke A Smith, Minyan Zeng, Mihir D Wechalekar, Shonket Ray, Susanna Proudman, and Lyle J Palmer. Artificial intelligence and the future of radiographic scoring in rheumatoid arthritis: a viewpoint. *Arthritis Research & Therapy*, 24(1):1–10, 2022.

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 205–218. Springer, 2023.

[6] Neelambuj Chaturvedi. Deepra: predicting joint damage from radiographs using cnn with attention. *arXiv preprint arXiv:2102.06982*, 2021.

[7] Son Do Hai Dang and Leigh Allison. Using deep learning to assign rheumatoid arthritis scores. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 399–402. IEEE, 2020.

[8] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] James F Fries, Daniel A Bloch, John T Sharp, Dennis J McShane, Patricia Spitz, Gilbert B Bluhm, Deborah Forrester, Harry Genant, Philip Gofton, Steven Richman, et al. Assessment of radiologic progression in rheumatoid arthritis. a randomized, controlled trial. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 29(1):1–9, 1986.

[11] Hanan Gani, Muzammal Naseer, and Mohammad Yaqub. How to train vision transformer on small-scale datasets? *arXiv preprint arXiv:2210.07240*, 2022.

[12] Harry K Genant. Methods of assessing radiographic change in rheumatoid arthritis. *The American journal of medicine*, 75(6):35–47, 1983.

[13] Harry K Genant, Yebin Jiang, Charles Peterfy, Ying Lu, Janos Rédei, and Peter J Countryman. Assessment of rheumatoid arthritis using a modified scoring method on digitized and original radiographs. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 41(9):1583–1590, 1998.

[14] Désirée MFM Van Der Heijde, Miek A Van Leeuwen, Piet LCM Van Riel, Anja M Koster, Martin A Van't Hof, Martin H Van Rijswijk, and Levinus BA Van De Putte. Biannual radiographic assessments of hands and feet in a three-year prospective followup of patients with early rheumatoid arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 35(1):26–34, 1992.

[15] Toru Hirano, Masayuki Nishide, Naoki Nonaka, Jun Seita, Kosuke Ebina, Kazuhiro Sakurada, and Atsushi Kumanogoh. Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. *Rheumatology advances in practice*, 3(2):rkz047, 2019.

[16] Yun-Ju Huang, Chang-Fu Kuo, Fakai Wang, Shun Miao, Kang Zheng, and Le Lu. Automatic joint space assessment in hand radiographs with deep learning among patients with rheumatoid arthritis. In *ARTHRITIS & RHEUMATOLOGY*, volume 72. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2020.

[17] Takuji Iwamoto, Hiroshi Okamoto, Yoshiaki Toyama, and Shigeki Momohara. Molecular aspects of rheumatoid arthritis: chemokines in the joints of patients. *The FEBS journal*, 275(18):4448–4455, 2008.

[18] Yebin Jiang, Harry K Genant, Iain Watt, Mark Cobby, Barry Bresnihan, Roger Aitchison, and Dorothy McCabe. A multicenter, double-blind, dose-ranging, randomized, placebo-controlled study of recombinant human interleukin-1 receptor antagonist in patients with rheumatoid arthritis: radiologic progression and correlation of genant and larsen scores. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 43(5):1001–1009, 2000.

[19] JEREMY J Kaye, E PAUL Nance Jr, LEIGH F Callahan, FRANK E Carroll, ALAN C Winfield, WEBB J Earthman, KEITH A Phillips, HOWARD A Fuchs, and THEODORE Pincus. Observer variation in quantitative assessment of rheumatoid arthritis part ii. a simplified scoring system. *Investigative radiology*, 22(1):41–46, 1987.

[20] Birgit M Köhler, Janine Günther, Dorothee Kaudewitz, and Hanns-Martin Lorenz. Current therapeutic options in the treatment of rheumatoid arthritis. *Journal of clinical medicine*, 8(7):938, 2019.

[21] Seunghoon Lee, Seunghyun Lee, and Byung Cheol Song. Improving vision transformers to learn small-size dataset from scratch. *IEEE Access*, 10:123212–123224, 2022.

[22] Ning Li, Jun C Wang, Toong H Liang, Ming H Zhu, Jia Y Wang, Xue L Fu, Jie R Zhou, Song G Zheng, Paul Chan, and Jie Han. Pathologic finding of increased expression of interleukin-17 in the synovial tissue of rheumatoid arthritis patients. *International journal of clinical and experimental pathology*, 6(7):1375, 2013.

[23] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31, 2018.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. pages 10012–10022, 2021.

[25] Krzysztof Maziarz, Anna Krason, and Zbigniew Wojna. Deep learning for rheumatoid arthritis: Joint detection and damage scoring in x-rays. *arXiv preprint arXiv:2104.13915*, 2021.

[26] Sachin Mehta and Mohammad Rastegari. Mobilevit: lightweight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[27] Seiichi Murakami, Kazuhiro Hatano, JooKooi Tan, Hyoungseop Kim, and Takatoshi Aoki. Automatic identification of bone erosions in rheumatoid arthritis from hand radiographs based on deep convolutional neural network. *Multimedia Tools and Applications*, 77:10921–10937, 2018.

[28] E PAUL Nance Jr, JEREMY J Kaye, LEIGH F Callahan, FRANK E Carroll, ALAN C Winfield, WEBB J Earthman, KEITH A Phillips, HOWARD A Fuchs, and THEODORE Pincus. Observer variation in quantitative assessment of rheumatoid arthritis. part i. scoring erosions and joint space narrowing. *Investigative radiology*, 21(12):922–927, 1986.

[29] Yune-Jung Park, Ana Maria Gherghe, and Desirée van der Heijde. Radiographic progression in clinical trials in rheumatoid arthritis: a systemic literature review of trials performed by industry. *RMD open*, 6(2):e001277, 2020.

[30] Karl Ludger Radke, Matthias Kors, Anja Müller-Lutz, Miriam Frenken, Lena Marie Wilms, Xenofon Baraliakos, Hans-Jörg Wittsack, Jörg HW Distler, Daniel B Abrar, Gerald Antoch, et al. Adaptive iou thresholding for improving small object detection: A proof-of-concept study of hand erosions classification of patients with rheumatic arthritis on x-ray images. *Diagnostics*, 13(1):104, 2022.

[31] John T Sharp, Frederick Wolfe, Donald M Mitchell, and Daniel A Bloch. The progression of erosion and joint space narrowing scores in rheumatoid arthritis during the first twenty-five years of disease. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 34(6):660–668, 1991.

[32] John T Sharp, Donald Y Young, Gilbert B Bluhm, Andrew Brook, Anne C Brower, Mary Corbett, John L Decker, Harry K Genant, J Philip Gofton, Neal Goodman, et al. How many joints in the hands and wrists should be included in a score of radiologic abnormalities used to assess rheumatoid arthritis? *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 28(12):1326–1335, 1985.

[33] Dongmei Sun, Thanh M Nguyen, Robert J Allaway, Jelai Wang, Verena Chung, V Yu Thomas, Michael Mason, Isaac Dimitrovsky, Lars Ericson, Hongyang Li, et al. A crowdsourcing approach to develop machine learning models to quantify radiographic joint damage in rheumatoid arthritis. *JAMA network open*, 5(8):e2227423–e2227423, 2022.

[34] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[35] Yan Ming Tan, Raphael Quek Hao Chong, and Carol Anne Hargreaves. Rheumatoid arthritis: Automated scoring of radiographic joint damage. *arXiv preprint arXiv:2110.08812*, 2021.

[36] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[37] Kemal Üreten, Hasan Erbay, and Hadi Hakan Maraş. Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network. *Clinical rheumatology*, 39:969–974, 2020.

[38] DMFM Van der Heijde. How to read radiographs according to the sharp/van der heijde method. *The Journal of rheumatology*, 26(3):743–745, 1999.

[39] DMFM Van der Heijde, T Dankert, F Nieman, R Rau, and M Boers. Reliability and sensitivity to change of a simplification of the sharp/van der heijde radiological assessment in rheumatoid arthritis. *Rheumatology*, 38(10):941–947, 1999.

[40] DésiréeM Van Der Heijde, PietL Van Riel, FrankW Gribnau, IkeH Nuver-Zwart, and LevinusB Van De Putte. Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis. *The Lancet*, 333(8646):1036–1038, 1989.

[41] Desiree MFM Van der Heijde. Plain x-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. *Bailliere's clinical rheumatology*, 10(3):435–453, 1996.

[42] Hao-Jan Wang, Chi-Ping Su, Chien-Chih Lai, Wun-Rong Chen, Chi Chen, Liang-Ying Ho, Woei-Chyn Chu, and Chung-Yueh Lien. 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(6):1355, 2022.

[43] Zijian Wang, Jian Liu, Zongyun Gu, and Chuanfu Li. An efficient cnn for hand x-ray overall scoring of rheumatoid arthritis. *Complexity*, 2022, 2022.

[44] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision*, pages 68–85. Springer, 2022.