

Deep Learning Models to Automate the Scoring of Hand Radiographs for Rheumatoid Arthritis

Zhiyan Bo¹(✉) , Laura C. Coates² , and Bartłomiej W. Papież¹ 

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 van der Heijde modification of the Sharp (SvdH) score is a widely used radiographic scoring method to quantify damage in Rheumatoid Arthritis (RA) in clinical trials. However, its complexity with a necessity to score each individual joint, and the expertise required limit its application in clinical practice, especially in disease progression measurement. In this work, we addressed this limitation by developing a bespoke, automated pipeline that is capable of predicting the SvdH score and RA severity from hand radiographs without the need to localise the joints first. Using hand radiographs from RA and suspected RA patients, we first investigated the performance of the state-of-the-art architectures in predicting the total SvdH score for hands and wrists and its corresponding severity class. Secondly, we leveraged publicly available data sets to perform transfer learning with different finetuning schemes and ensemble learning, which resulted in substantial improvement in model performance being on par with an experienced human reader. The best model for RA scoring achieved a Pearson’s correlation coefficient (PCC) of 0.925 and root mean squared error (RMSE) of 18.02, while the best model for RA severity classification achieved an accuracy of 0.358 and PCC of 0.859. Our score prediction model attained almost comparable accuracy with experienced radiologists (PCC = 0.97, RMSE = 18.75). Finally, using Grad-CAM, we showed that our models could focus on the anatomical structures in hands and wrists which clinicians deemed as relevant to RA progression in the majority of cases.

Keywords: Rheumatoid Arthritis · Hand X-ray scoring & classification · Deep learning · Transfer learning

1 Introduction

With a prevalence of about 1%, more than 400,000 people in the UK are diagnosed with Rheumatoid Arthritis (RA), an autoimmune disease that causes joint inflammation which could progress to structural damage, pain and disability in patients [16]. Its classic presentation is symmetric arthritis involving multiple joints in both hands, feet and wrists with bone erosions (indicated by blue arrows in Fig. 1(A)) and joint space narrowing (JSN) (indicated by blue boxes in

Fig. 1(A)) [2]. As a widely available tool in clinical settings, plain radiography is the gold standard for RA progression evaluation and is commonly used in its diagnosis, classification and monitoring [2,29].

Several diagnostic and RA quantification criteria have been proposed, with the most commonly used criteria for radiographic scoring being the van der Heijde modification of the Sharp (SvdH) score, which is a standard damage quantification method in clinical trials due to its good intra- and inter-observer reliability, and sensitivity to changes [12,18,19,26]. It includes two scores for bone erosions and JSN in hands, wrists and forefeet. Due to SvdH's complexity, it takes an experienced radiologist or rheumatologist around 4 minutes to score single-hand radiography; moreover imaging artefacts and superimposition can lead to interobserver disagreements especially when measuring small changes [3,25,26]. The above shortcomings limit the method's clinical application and hinder its effectiveness in clinical trials. In routine clinical practice, rheumatologists simply compare images instead of generating and comparing the scores.

Multiple Deep Learning (DL)-based approaches have been developed for RA scoring using ultrasound and X-ray [1,4,11]. Most of those methodologies consist of a joint detection or segmentation stage followed by a severity score prediction stage [7,9,24]. For example, a pipeline proposed adopts a contouring technique to segment the synovial regions from ultrasound images and utilises a convolutional neural network (CNN) to classify the synovial regions into different severity grades [7]. An automatic joint detection (using You Only Look Once v4) and SvdH-based classification system (using EfficientNet-B1) was developed for hand X-ray images [27]. However, these methods assign a severity class to a joint instead of a score, which reduces the resolution of the prediction.

DL-based frameworks that directly predict the SvdH score have been explored as well [1,17]. Hirano et al. focused on the finger joints and developed a pipeline that first detects them by a cascade classifier using Haar-like features and then scores each joint for erosion and JSN by a CNN [8]. In the RA-2-Dialogue for Reverse Engineering Assessment and Methods (RA2-DREAM Challenge), CNNs predicting the overall SvdH score from hands and feet X-rays as well as the joint-by-joint and overall JSN and erosion scores separately were investigated [23]. Among the published solutions, a multi-task DL model adopting U-Net and EfficientNet-B5 architecture that localises the joints and predicts the SvdH scores for JSN and erosion was proposed [13]. Similarly, another proposed pipeline detects the joints using U-Net-based heatmap regression, crops the joint image, and then predicts the amount of damage by EfficientNet-B0 [10].

Although the above methods have demonstrated promising performance, their suitability for clinical applications is still uncertain as it could be challenging to localise individual joints accurately. Also, late-stage RA patients may find it painful to straighten their fingers with deformities during an X-ray and algorithms might have difficulty delineating severely damaged bones and joints. Therefore methodologies to predict overall scores might be practically useful and ease the procedure of disease severity quantification. Instead of focusing on each joint, a model ResNet-Dwise50 combining ResNet-50 and MobileNetV2

which predicts an overall SvdH score for hand X-ray was developed [28]. They concluded that using depth-wise separable convolutions and replacing the residual blocks (RBs) with inverted residual blocks (IRBs) resulted in the observed performance improvement compared to baseline ResNet-50.

In this work, we focused on using whole bilateral posteroanterior hand radiographs as input for RA damage quantification and proposed two ensemble models that combine three CNNs to predict 1) the SvdH score and 2) the corresponding RA severity class. Our contributions can be summarised as follows. We first investigated the performance of the state-of-the-art CNNs: MobileNetV2 [21] and ResNet [6] in a publicly available RA hand X-ray dataset. Secondly, we meticulously developed efficient approaches to improve the performance of our models when compared with the state-of-the-art method [28] using transfer learning. Then, ensemble technique was applied, further improving the accuracy of RA scoring to be on par with human readers. Finally, we addressed the limited interpretability of the DL methods using entire radiographs as input by performing Gradient-weighted Class Activation Mapping (Grad-CAM) [22]. The heatmaps generated were qualitatively assessed to evaluate the clinical potential of our models.

2 Data and Methods

2.1 Data

This project utilised a dataset which includes 3,818 hand X-ray images (see exemplar image in Fig. 1(A)) collected from RA patients or suspected RA patients (see details [28]). The average of the SvdH scores reported by two experienced radiologists was provided, alongside patients' age and gender. Most of the samples were older than 30 and 82.6% of them were females. The data publisher also evaluated the inter-rater differences, which demonstrated a Pearson's correlation coefficient (PCC; Eq.3) of 0.97, mean absolute error (MAE; Eq.4) of 12.24, and root mean squared error (RMSE; Eq.5) of 18.75. The dataset was split into training, validation and test sets, each including 2700, 760 and 358 images.

2.2 Pre-processing

The SvdH score looks at 16 areas (yellow in Fig. 1(A)) and 15 joints (orange in Fig. 1(A)) in each hand and wrist for erosion and JSN assessment. The erosion score ranges from 0 to 5 while the JSN score ranges from 0 to 4, so the total score accounting for hands and wrists ranges from 0 to 280 though scores > 140 are less commonly seen [26]. More details are provided in the supplementary material. For direct prediction, the scores of the images were standardised via the z-score method to have a mean of 0 and a standard deviation (SD) of 1. In addition to score prediction, we investigated damage quantification as a classification problem. As in clinical settings, a large proportion of cases were diagnosed or maintained at relatively early stages, the scores were split into 10 classes (Fig. 1(B)), with smaller ranges at the lower end of the score range.

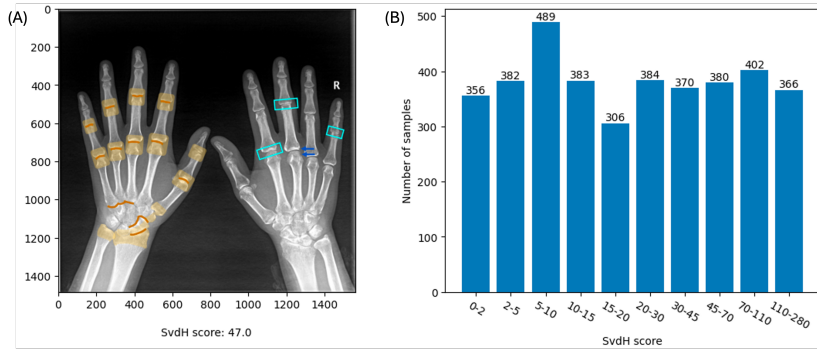


Fig. 1. (A) An example of the radiographs from Wang et al. [28] with an average SvdH score of 47. The bones and joints assessed in the left hand and wrist are highlighted in yellow and orange for reference. Examples of bone erosions and JSN are indicated by blue arrows and boxes. (B) The SvdH class distribution of samples used in this paper.

To increase the sample size and diversity, the training images were first resized to 1024×1024 and augmented with random horizontal flip, random intensity changes with a scale in the range of 0.9 - 1.1, and random rotation by one of the given angles $[0^\circ, 90^\circ, 180^\circ, 270^\circ]$ to avoid exclusion of assessed anatomical structures. Finally image pixel intensities were normalised with the mean and SD of the training set. For the validation and test sets, only image resizing and normalisation were performed.

2.3 Models

In [28] ResNet-50 and MobileNetV2 were used as the baseline models, thus for comparison, the two models were selected for our study. The ResNet architecture adopts skip connections that enable a model layer to learn residual functions regarding the layer inputs to address the issue of vanishing gradient [6]. To investigate the effect of model complexity on performance, ResNet-34, a version of ResNet with fewer layers, was also used. MobileNetV2 uses IRBs with depth-wise separable convolution and linear bottleneck. With fewer model parameters, it demonstrated good efficiency and performance in many image analysis tasks [21]. In this work separate sets of models were trained for the regression and classification tasks, with the output size of the models' last fully connected layer changed to one for regression and 10 for classification.

Loss function. For training classifiers, the cross-entropy loss was employed, while two distinct loss functions were chosen for training the regression models. The first loss was the mean squared error (MSE) loss. The second one was the smooth loss (SL) proposed in [28], which combines MSE and MAE loss defined as follows.

$$SL = \sqrt{\frac{1}{n} \sum_{i=1}^n smooth(y_i - \hat{y}_i)} \quad (1)$$

in which *smooth* is defined as:

$$smooth(x) = \begin{cases} ax^2 & \text{if } |x| < c \\ |x| - b & \text{otherwise} \end{cases} \quad (2)$$

where $a=0.6$, $b=0$, and $c=1.0$. These parameters gave the best performance in [28], so we adopted the same values. In exploratory testing, we obtained more accurate models with MSE loss so it was used for further model training and transfer learning.

2.4 Transfer learning

Transfer learning has been widely adopted in medical imaging to improve the performance of models [20]. In this work, transfer learning was performed using the RSNA Pediatric Bone Age Challenge dataset [5], which includes 14,036 single-hand X-ray images with skeletal age in months. 12,611 (90%) images were used for training and the remaining 1,425 (10%) for validation. Z-score standardisation of bone age, image augmentation, MSE loss and a batch size of 4 were adopted. ResNets and MobileNetV2 were trained for 50 epochs for bone age prediction, and then the weights were transferred for SvdH scoring and RA severity classification. Two types of model-tuning methods were experimented with: 1) Finetuning, where all model layers are trainable; 2) Freezing the first few layers during training and only tuning the higher-level parameters.

Two layer-freezing schemes were picked for each model to find a balance between training accuracy and generalisability. For ResNets, we tuned a set of models while freezing layers up to and including the first stack of residual blocks (RBs; RBs-1) and a second set while freezing layers up to and including the second stack of RBs (RBs-2). For MobileNetV2, two models were tuned separately while freezing up to and including the second stack of inverted residual blocks (IRBs; IRBs-2) and while freezing up to and including the third stack of IRBs (IRBs-3). With RBs-1 and IRBs-2, we investigated the effect of preserving the low-level features that could be common for both the new and the original tasks as similar types of images were used. Bone age assessment and SvdH scoring look at some similar anatomical features, such as joint space, so we also experimented with preserving the middle-level features as well with RBs-2 and IRBs-3.

2.5 Ensemble learning

As a widely adopted ensemble technique, stacking was shown to bring substantial performance gains in medical image classification tasks [15]. It assembles the outputs of independent models and feeds them into an extra algorithm for prediction. In this work, we applied a linear transformation to the outputs from the

best-performing ResNet-50, ResNet-34 and MobileNetV2 to predict the SvdH score or probability of each severity class. As classifiers each produce 10 outputs for 10 classes, two stacking methods were attempted: 1) All classes: using all outputs of three models to make predictions for 10 classes to account for inter-class associations; 2) Single class: using class-specific outputs only for prediction.

2.6 Model explainability

Since the SvdH scoring scheme looks at specific joints and bones in the hand X-ray, we visually assessed the activation maps, which infer the contribution of pixels in an image to the final prediction. Examples of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) were selected. For regression, TN and TP were defined as healthy ($SvdH < 5$) and severe ($SvdH > 200$) cases with an absolute prediction error smaller than 10. FP and FN were defined as cases with an absolute prediction error greater than 50. Grad-CAM was applied to the activation map computed before the fully connected layer. Finally, we invited an experienced rheumatologist to comment on whether the highlighted anatomical structures demonstrate representative pathology.

2.7 Evaluation

For regression, models were evaluated in the test set by measuring the degree of agreement between the predicted scores and the averaged true scores. Metrics used were PCC, MAE and RMSE defined as follows.

$$PCC = \frac{cov(X, Y)}{\sqrt{var(X)}\sqrt{var(Y)}} \quad (3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - y_i| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (5)$$

For classification, models were evaluated using confusion matrix accuracy, balanced accuracy (BA), and PCC, MAE and RMSE to account for the ordinal relationships between classes. The performance of the best regression model in severity classification was evaluated by assigning samples to different predicted classes according to their predicted scores.

2.8 Implementation details

Stochastic gradient descent optimiser with an initial learning rate of 0.001, weight decay of 0.001, and momentum of 0.9 was used. Models with the best validation performance during training were selected. All regression models were

trained on an NVIDIA Tesla V100 SXM2 GPU with 32GB of memory to compare their training time. For loss function comparison, models were trained for 100 epochs with a batch size of 4. With MSE loss, we then investigated the effect of using a larger batch size (16). For transfer learning, models were trained with a batch size of 4. Classification models were trained on different NVIDIA GPUs as we did not plan to measure the training time. All baseline models were trained for 100 epochs with a batch size of 4 or 16. For transfer learning, batch size = 4 was adopted.

3 Results

3.1 Regression task: SvdH score prediction

Baseline. Table 1 displays the results for baseline regression models without pretraining. MSE loss achieved better performances in terms of all three metrics, with fewer fluctuations in the models’ validation performance during training. The average time taken by one epoch was 11:28 min, 12:28 min and 13:45 min for MobileNetV2, ResNet-34 and ResNet-50 when trained with batch size = 4. Increasing the batch size to 16 did not substantially change the performance of ResNet-34 and ResNet-50. Their PCC increased while their MAE and RMSE deteriorated. On the other hand, improvements in two of the three metrics were observed for MobileNetV2. Its PCC increased from 0.877 to 0.903, RMSE decreased from 22.72 to 20.48, and MAE increased slightly from 14.84 to 14.99. Among all models without transfer learning, the MobileNetV2 trained with a batch size of 16 achieved the best performance. However, using a larger batch size increased the speed of convergence but worsened the stability of model performance during training, especially after reaching the plateau in validation loss. For result consistency and model training stability, batch size = 4 was adopted for transfer learning of all three models presented below.

Transfer learning. Transfer learning improved the performance of all three models irrespective of the tuning method (Table 1). Among all regression models, ResNet-50:RBs-1 achieved the best overall accuracy, with a PCC of 0.918, MAE of 13.58 and RMSE of 18.88. For ResNet-34, finetuning increased the PCC by 3.6% and decreased the MAE and RMSE by 1.27 and 2.92. MobileNetV2 yielded worse results than ResNets, and IRBs-3 only raised the PCC by 0.4% and reduced the MAE and RMSE by 1.4 and 0.55. As shown in Fig. 2, models’ accuracy was better among cases with low SvdH scores. Their predictive error increased with the true score and several severe cases were underscored.

Ensemble learning. Ensemble regressors combining ResNet-50:RBs-1, ResNet-34:finetuned and MobileNetV2:IRBs-3 were trained with different batch sizes. As shown in Table 1 and Fig. 2, with batch size = 16, it outperformed all independent models, attaining a PCC of 0.925, MAE of 12.57 and RMSE of 18.02.

Table 1. Model performance in SvdH score prediction. The top two values of a metric are made bold and underlined. Among baseline models, MobileNetV2 demonstrated the highest accuracy. With transfer learning, ResNet-50:RBs-1 achieved the best performance among all independent models. Ensemble learning yielded further performance improvement and the best regression model was obtained with batch size = 16.

Baseline					
Model	Loss function	Batch size	PCC	MAE	RMSE
ResNet-50	Smooth	4	0.838	16.87	26.02
	MSE	4	0.885	15.33	22.72
		16	0.895	17.65	23.26
ResNet-34	Smooth	4	0.828	16.9	27.14
	MSE	4	0.876	15.11	22.7
		16	0.885	17.1	22.89
MobileNetV2	Smooth	4	0.877	14.96	23.11
	MSE	4	0.877	14.84	22.72
		16	0.903	14.99	20.48
Transfer learning (finetuned with MSE loss)					
Model	Layer freezing	Batch size	PCC	MAE	RMSE
ResNet-50	No	4	0.907	12.97	20.14
	RBs-1	4	0.918	13.58	18.88
	RBs-2	4	0.908	13.53	19.86
ResNet-34	No	4	0.912	13.84	19.78
	RBs-1	4	0.902	14.74	20.57
	RBs-2	4	0.9	13.68	20.72
MobileNetV2	No	4	0.898	14.02	20.95
	IRBs-2	4	0.898	13.97	20.81
	IRBs-3	4	0.907	13.59	19.93
Ensemble learning (finetuned with MSE loss)					
Model combination		Batch size	PCC	MAE	RMSE
ResNet-50:RBs-1 & ResNet-34:finetuned &		4	0.925	12.63	18.17
MobileNetV2:IRBs-3		16	0.925	12.57	18.02

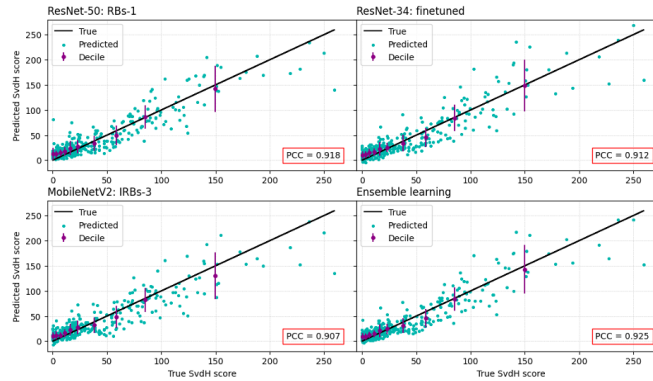


Fig. 2. Predicted vs. true SvdH scores for the best tuned regression models with transfer learning and the ensemble learning model that combines them. The mean and SD of the deciles are plotted for reference. The models yielded better performance in predicting early-stage cases. Ensemble learning reduced prediction errors on average compared to independent models.

Activation mapping. The Grad-CAM heatmaps of TP, TN, FP, and FN examples predicted by the best independent regression model (i.e., ResNet-50:RBs-1) were provided in Fig. 3. Except for TN, joints and bones assessed by the SvdH score were highlighted, suggesting that the model was looking at the desirable regions when making the prediction. However, in FP the highlighted joints did not demonstrate any pathology and the model failed to pick up a lot of sclerosis in the wrists in FN (depicted by circles in the heatmap). In TN, the diaphysis of some bones was highlighted.

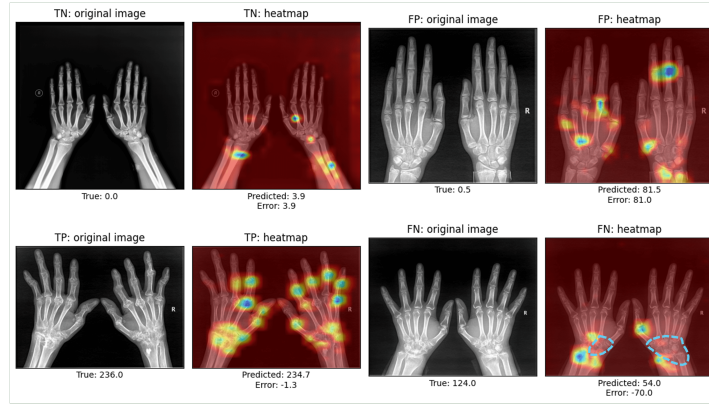


Fig. 3. Grad-CAM heatmaps of ResNet-50:RBs-1 regression model for examples of TN, FP, TP and FN. The true and predicted scores are provided. In FN, the overlooked changes in wrists are circled.

3.2 Classification task: RA severity class prediction

Baseline. As shown in Table 2, in general using a batch size of 4 yielded better results in terms of accuracy and BA. However, when referring to MAE, RMSE and PCC, MobileNetV2 trained with a batch size of 16 performed better. ResNet-50 achieved the highest accuracy (0.341) and BA (0.327). MobileNetV2 achieved the highest PCC (0.833). ResNet-34 achieved the lowest MAE (1.24) and RMSE (1.76). Balancing all four metrics, ResNet-34 demonstrated the best performance. Its confusion matrix is provided in Fig. 4(A). Although the accuracy was low, especially for class 15-20, most mistakes were made between neighbouring classes and the misclassification rate decreased with severity.

Transfer learning. As shown in Table 2, model pretraining improved the performance of MobileNetV2 irrespective of the tuning scheme, with IRBs-3 achieving the highest accuracy (0.349) and BA (0.339) and lowest MAE (1.17) and RMSE (1.67). Transfer learning’s effect on ResNets was less consistent. Though

Table 2. Model performance in RA severity class prediction. The top two values of a metric are made bold and underlined. ResNet-34, ResNet-50:finetuned, and the ensemble classifier that combines class-specific predictions were the best-performing baseline, transfer learning and ensemble learning models. Substantial performance improvement was observed with transfer and ensemble learning. Using the best regression model for severity classification yielded the highest PCC and RMSE but much lower accuracy.

Baseline							
Model	Batch size	Accuracy	BA	PCC	MAE	RMSE	
ResNet-50	4	0.341	0.327	0.809	1.42	2.09	
	16	0.249	0.242	0.75	1.65	2.31	
ResNet-34	4	0.321	0.312	0.829	1.24	1.76	
	16	0.296	0.289	0.784	1.36	1.92	
MobileNet V2	4	0.316	0.31	0.822	1.33	1.86	
	16	0.313	0.301	0.833	1.24	1.77	
Transfer learning							
Model	Layer freezing	Batch size	Accuracy	BA	PCC	MAE	RMSE
ResNet-50	No	4	0.352	0.344	0.842	1.18	1.72
	RBs-1	4	0.335	0.326	0.833	1.23	1.75
	RBs-2	4	0.316	0.302	0.82	1.29	1.84
ResNet-34	No	4	0.318	0.307	0.844	1.2	1.67
	RBs-1	4	0.296	0.288	0.828	1.27	1.81
	RBs-2	4	0.358	0.351	0.834	1.22	1.8
MobileNetV2	No	4	0.349	0.337	0.834	1.23	1.73
	IRBs-2	4	0.341	0.327	0.819	1.24	1.79
	IRBs-3	4	0.349	0.339	0.832	1.17	1.67
Ensemble learning							
Model combination	Stacking principle	Batch size	Accuracy	BA	PCC	MAE	RMSE
ResNet-50:finetuned & ResNet-34:RBs-2 & MobileNetV2:IRBs-3	All classes	4	0.355	0.348	0.854	1.13	1.64
		16	0.355	0.342	0.851	1.19	1.71
	Single class	4	0.352	0.344	0.853	1.13	1.63
		16	0.358	0.349	0.859	1.12	1.62
Regression to classification							
Ensemble regression model		16	0.288	0.29	0.861	1.14	1.54

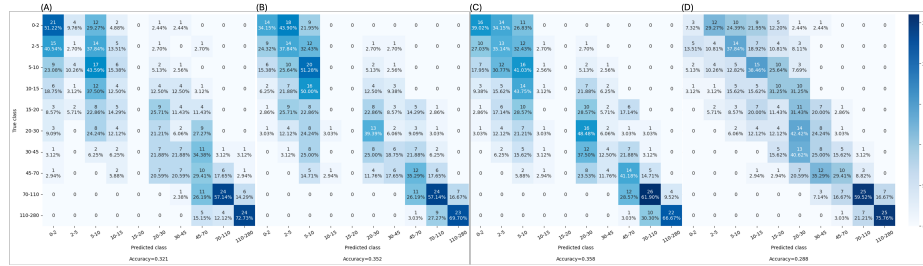


Fig. 4. Confusion matrices of (A) the best baseline classifier (ResNet-34), (B) the best pretrained classifier (ResNet-50:finetuned), (C) the best ensemble classifier, and (D) the ensemble regression model. The numbers and proportions of images from the same class that fell into different predicted classes are provided. The ensemble classifier yielded the smallest misclassifications and the highest overall accuracy.

improvements in the regression metrics were observed for all tuning schemes of ResNet-50, only finetuning improved the classification accuracy (0.352) and BA (0.344), with a PCC of 0.842, MAE of 1.18, and RMSE of 1.72. For ResNet-

34, though finetuning yielded the best PCC, MAE and RMSE, RBs-2 was the only scheme attaining a higher accuracy (0.358) and BA (0.351) than the baseline. Balancing all metrics, we concluded that ResNet-50:finetuned was the best model. As shown in Fig. 4(B), there were fewer misclassifications between far-away classes, which is reflected by the rise in PCC and drop in MAE and RMSE.

Ensemble learning. Stacking class-specific outputs of ResNet-50:finetuned, ResNet-34:RBs-2, and MobileNetV2:IRBs-3 to predict the probability of a single class yielded the best results with batch size = 16 (see Table 2). Achieving a PCC of 0.859, MAE of 1.12 and RMSE of 1.62, it outperformed all independent classifiers.

Activation mapping. Fig. 5 displays the Grad-CAM heatmaps extracted from the best independent classifier (ResNet-50:finetuned) for TP, TN, FP and FN examples. Despite highlighting the joints and bones assessed in SvdH scoring, the heatmaps show relatively high activation at regions of irrelevant anatomical structures, such as the diaphysis of metacarpals and the distal interphalangeal joints in TP. Although the damages in the wrists were picked up in TP, several finger bones and joints in the left hand and the metacarpophalangeal joints in the right hand were overlooked (circled in the heatmap). In TN, normal joints in the wrists were highlighted. The reason for the high intensity around the right thumb is unclear, but this could be an area of confusion for the model because of a high prevalence of osteoarthritis in this area. Despite looking at the correct area (only the wrists as the fingers are normal) for damage, the model underestimated the RA severity of FN. In FP, the identified right thumb base displays some irregularities that do not map to the score, thus might confuse our model.

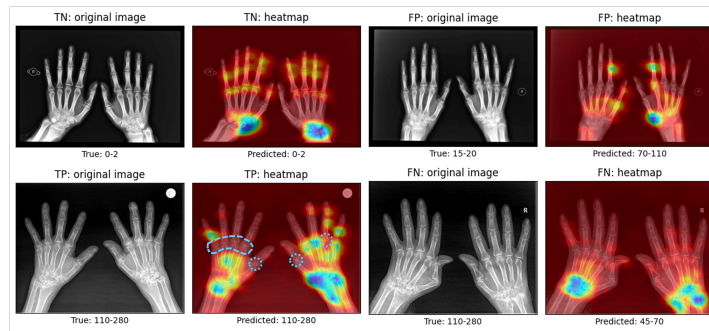


Fig. 5. Grad-CAM heatmaps of ResNet-50:finetuned classifier for examples of TN, FP, TP and FN. In TP, some overlooked changes in fingers are circled.

Regression to classification. As shown in Table 2 and Fig. 4(D), the ensemble regression model achieved higher PCC (0.861) and RMSE (1.54) than the best ensemble classifier in severity classification but much lower accuracy (0.288) and BA (0.29), especially among samples with SvdH < 10. Its performance was better than the classification model in classes 10-15 and 15-20.

4 Discussion

Our results on SvdH score prediction suggested that MSE was a more suitable loss function than the smooth loss proposed in [28]. Without pretraining, MobileNetV2 outperformed the ResNets, which contradicts Wang et al.’s observations. When trained from scratch, the baseline models achieved better results than ResNet-Dwise50 in the paper, which obtained a PCC of 0.95, MAE of 17.76, and RMSE of 26.34 without transfer learning. Our baseline MobileNetV2’s testing MAE and RMSE were 2.77 and 5.86 lower. In this work, ResNet-34 outperformed ResNet-50, suggesting that the improvement might result from a more lightweight model architecture that reduces the risk of overfitting.

Among all baseline classifiers, ResNet-34 achieved the best performance. Considering the high MAE and RMSE between the scores assigned by two radiologists, we expected there would be more errors when predicting milder cases as narrower classes were set. Since cross-entropy loss cannot account for the ordinal relationships between classes, our models with higher classification accuracy did not always demonstrate better performance in terms of regression metrics. Other loss functions which penalise bigger errors could be adopted in future studies.

Our results with transfer learning showed the potential of optimising RA scoring models’ performance by pretraining them in a larger dataset of similar types of images. Although the RSNA challenge images were from children and single-hand, the models still acquired useful features for our tasks. This is additionally supported by the improved performance when tuning models with early layers frozen. Unlike models trained from scratch for score prediction, tuned ResNet-50 achieved the best performance while tuned MobileNetV2 performed the worst. One potential reason might be that ResNet-50 could learn more features from the bigger pretraining dataset that later became useful for SvdH score prediction. Also, it might suffer less from overfitting in the larger dataset. With transfer learning, more complex model architectures could be experimented with. ResNet-Dwise50 pretrained in CheXpert achieved a PCC of 0.97, MAE of 14.9 and RMSE of 22.01. Our best independent regression model (ResNet-50:RBs-1) outperformed it by further reducing the MAE by 1.32 and RMSE by 3.13. Furthermore, our model achieved almost as good RMSE as the radiologists and its MAE was only 1.34 higher. However, the PCC was much lower, suggesting scope for further improvement. As shown by the activation maps, the model could identify and focus on the joints in RA images. However, it might not pick up all the pathologies as shown by FN and might confuse healthy joints with damaged joints as in FP. In future, other medical imaging datasets of similar or different types of anatomical structures such as RadImageNet [14] could be

adopted for transfer learning to acquire more comprehensive feature representations. The performance of the regression models deteriorated in more severe cases, probably because of the reduction in sample size. In future model optimisation, methods to handle data imbalance should be introduced.

For classification, no specific patterns in the results of different model tuning schemes were observed and some set-ups even worsened the prediction accuracy despite improving the other metrics. This again shows the weaknesses of cross-entropy loss and emphasises the need for a more suitable loss for ordinal classification problems. The activation maps suggested that the model also looked at some anatomical structures that are irrelevant to SvdH scoring. Since late-stage patients may fail to straighten their fingers during an X-ray, it is important to investigate if the positioning of fingers influences the prediction of both regression and classification models. Our Grad-CAM results did not suggest this issue, but additional testing is needed to improve the trustworthiness of the models.

Ensemble learning of the best pretrained independent models yielded further performance improvement in both regression and classification tasks. The ensemble regressor even achieved smaller RMSE than radiologists. Stacking has effectively combined the strengths of each model and evened out the failures. Though not directly comparable, regression models outperformed classification models in terms of PCC and provided higher resolution in the predictions. When using the best regression model for classification, the higher prediction resolution resulted in higher PCC and lower RMSE. Since the entire score range was treated equally during regression training, it performed much worse than the best classifier in classifying very early-stage cases.

5 Conclusion

Our results of predicting the SvdH score and SvdH-based severity class of hand radiographs using state-of-the-art CNN models demonstrated the great potential of DL techniques in quantifying RA on a whole image level. Transfer learning using a paediatric single-hand X-ray dataset and ensemble learning improved the prediction accuracy substantially, resulting in models that largely outperformed previously proposed models [28] and achieved performance close to the experienced radiologists. For score prediction, future work should explore ways to deal with data imbalance to reduce large errors in severe cases and improve the correlation between predictions and true scores. For classification, alternative loss functions and model architectures could be considered to improve prediction accuracy and reduce distances of misclassification.

Acknowledgements. Zhiyan Bo is 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.

References

1. 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>
2. BMJ Best Practice: Rheumatoid arthritis - symptoms, diagnosis and treatment (4 2024), <https://bestpractice.bmj.com/topics/en-gb/105>
3. Boini, S., Guillemin, F.: Radiographic scoring methods as outcome measures in rheumatoid arthritis: properties and advantages. *Annals of the Rheumatic Diseases* **60**, 817–827 (2001), <https://ard.bmj.com/content/60/9/817>
4. 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>
5. Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., Mamonov, A.B., Bilbily, A., Cicero, M., Pan, I., Pereira, L.A., Sousa, R.T., Abdala, N., Kitamura, F.C., Thodberg, H.H., Chen, L., Shih, G., Andriole, K., Kohli, M.D., Erickson, B.J., Flanders, A.E.: The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology* **290**, 498–503 (2019). <https://doi.org/10.1148/RADIOL.2018180736>
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2016-December**, 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
7. 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>
8. 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>
9. Ho, S., Elamvazuthi, I., Lu, C.K.: Classification of Rheumatoid Arthritis using Machine Learning Algorithms. 2018 IEEE 4th International Symposium in Robotics and Manufacturing Automation, ROMA 2018 (2018). <https://doi.org/10.1109/ROMA46407.2018.8986700>
10. 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>
11. Imtiaz, M., Shah, S.A.A., ur Rehman, Z.: A review of arthritis diagnosis techniques in artificial intelligence era: Current trends and research challenges. *Neuroscience Informatics* **2**, 100079 (2022). <https://doi.org/10.1016/J.NEURI.2022.100079>
12. Landewé, R.B., Connell, C.A., Bradley, J.D., Wilkinson, B., Gruben, D., Strenghtolt, 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>
13. Maziarz, K., Krason, A., Wojna, Z.: Deep Learning for Rheumatoid Arthritis: Joint Detection and Damage Scoring in X-rays (2022). <https://doi.org/10.48550/arXiv.2104.13915>

14. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z.A., Yang, Y.: RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiology: Artificial Intelligence* **4** (2022). <https://doi.org/10.1148/RYAI.210315>
15. Muller, D., Soto-Rey, I., Kramer, F.: An Analysis on Ensemble Learning Optimized Medical Image Classification with Deep Convolutional Neural Networks. *IEEE Access* **10** (2022). <https://doi.org/10.1109/ACCESS.2022.3182399>
16. NICE: Prevalence and incidence of rheumatoid arthritis (4 2024), <https://cks.nice.org.uk/topics/rheumatoid-arthritis/background-information/prevalence-incidence/>
17. Parashar, A., Rishi, R., Parashar, A., Rida, I.: Medical imaging in rheumatoid arthritis: A review on deep learning approach. *Open Life Sciences* **18** (2023). <https://doi.org/10.1515/BIOL-2022-0611>
18. Radu, A.F., Bungau, S.G.: Management of Rheumatoid Arthritis: An Overview. *Cells* 2021, Vol. 10, Page 2857 **10**, 2857 (2021). <https://doi.org/10.3390/CELLS10112857>
19. 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>
20. Salehi, A.W., Khan, S., Gupta, G., Alabdullah, B.I., Almjally, A., Alsolai, H., Siddiqui, T., Mellit, A.: A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. *Sustainability* 2023, Vol. 15, Page 5930 **15**, 5930 (2023). <https://doi.org/10.3390/SU15075930>
21. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp. 4510–4520 (2018). <https://doi.org/10.1109/CVPR.2018.00474>
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision 2017-October*, 618–626 (2017). <https://doi.org/10.1109/ICCV.2017.74>
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–e2227423 (2022). <https://doi.org/10.1001/jamanetworkopen.2022.27423>
24. Tan, Y.M., Quek, R., Chong, H., Hargreaves, C.A.: Rheumatoid Arthritis: Automated Scoring of Radiographic Joint Damage (2021). <https://doi.org/https://doi.org/10.48550/arXiv.2110.08812>
25. van der Heijde, D., Boonen, A., Boers, M., Kostense, P., van der Linden, S.: Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. *Rheumatology (Oxford, England)* **38**, 1213–1220 (1999). <https://doi.org/10.1093/RHEUMATOLOGY/38.12.1213>
26. 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>

27. Wang, H.J., Su, C.P., Lai, C.C., Chen, W.R., Chen, C., Ho, L.Y., Chu, W.C., Lien, C.Y.: Deep Learning-Based Computer-Aided Diagnosis of Rheumatoid Arthritis with Hand X-ray Images Conforming to Modified Total Sharp/van der Heijde Score. *Biomedicines* 2022, Vol. 10, Page 1355 **10**, 1355 (2022). <https://doi.org/10.3390/BIMEDICINES10061355>
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. Welsing, P.M., Landewé, R.B., van Riel, P.L., Boers, M., van Gestel, A.M., van der Linden, S., Swinkels, H.L., van der Heijde, D.M.: The relationship between disease activity and radiologic progression in patients with rheumatoid arthritis: A longitudinal analysis. *Arthritis and Rheumatism* **50**, 2082–2093 (2004). <https://doi.org/10.1002/ART.20350>