Hindawi
Complexity
Volume 2022, Article ID 5485606, 9 pages
https://doi.org/10.1155/2022/5485606

Research Article

An Efficient CNN for Hand X-Ray Overall Scoring of Rheumatoid Arthritis

Zijian Wang[1,2] **Jian Liu**[2] **Zongyun Gu**[1,3] **and Chuanfu Li**[1,2,3]

[1]*College of Medical Information Engineering, Anhui University of Traditional Chinese Medicine, Hefei 230012, China*
[2]*First Clinical Medical College, Anhui University of Traditional Chinese Medicine, Hefei 230038, China*
[3]*Artificial Intelligence Research Institute of Hefei Comprehensive National Science Center (Anhui Artificial Intelligence Laboratory), Hefei 230012, China*

Correspondence should be addressed to Zongyun Gu; gzy@ahtcm.edu.cn and Chuanfu Li; licf@ahtcm.edu.cn

Received 20 August 2021; Accepted 19 January 2022; Published 23 February 2022

Academic Editor: Atila Bueno

Rheumatoid arthritis (RA) is a progressive systemic autoimmune disease characterized by inflammation of the joints and surrounding tissues, which seriously affects the life of patients. The Sharp/van der Heijde method has been widely used in clinical evaluation for the RA disease. However, this manual method is time-consuming and laborious. Even if two radiologists evaluate a specific location, their subjective evaluation may lead to low inter-rater reliability. Here, we developed an efficient model powered by deep convolutional neural networks to solve these problems and automated the overall scoring on hand X-rays. The depthwise separable (Dwise) convolution technique is used based on ResNet-50 due to the high resolution of hand X-rays. An inverted residual block is introduced to devise a ResNet-Dwise50 model to enhance the efficiency of the model. The model was trained and tested using bilateral posteroanterior (two-handed, side by side) images of 3818 patients. The experiment results show the ResNet-Dwise50 model achieved an MAE of 14.90 and RMSE of 22.01 while ensuring high efficiency. There was no statistical difference between the average scores given by two experienced radiologists and predicted scores from our model.

1. Introduction

Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disease that primarily strikes the hands and feet joints, causing swelling, pain, and stiffness in the joints [1] and can lead to joint damage in the form of joint space narrowing and joint erosion [2]. Early diagnosis of RA is necessary because early treatment may prevent joints from worsening or slowing the process. The methods of diagnosis usually include physical exams, blood tests, and imaging tests. Radiography is currently the primary tool for evaluating RA because of its wide availability, relatively low cost, and high capability for imaging bones and joints [3]. The commonly used and well-validated method for evaluating RA in hand X-rays is the Sharp/van der Heijde (SvH) method [2]. The radiologist assesses joint space narrowing and joint erosion by grading the specific locations in each hand (wrist) and foot using the SvH method. However, this manual approach is time-consuming and highly subjective, resulting in low inter-rater reliability, and even trained experts often disagree on the final score. We therefore decided to devise a neural network that can automate SvH scoring in radiographs.

With the rapid development of artificial intelligence (AI) technology, deep learning has become a significant trend in medical image analysis. Convolutional neural networks have become the dominant machine learning method for medical image classification [4, 5], medical image segmentation [6, 7], medical image registration [8, 9], medical image fusion [10, 11], and medical image report generation [12–14]. Some researchers have devised models to solve the RA scoring challenge by learning to assess joint injuries directly from data. One approach uses various clinical data to determine whether a patient has RA and diagnose rheumatoid arthritis by training a convolutional neural network on a dataset of radiographs [15, 16]. Another mainstream approach uses a two-step approach to detect finger joint destruction [17, 18]. The first step is a joints' detection task, while the second step is an image classification that uses the convolutional neural network to classify the scoring of joint destruction.

Although these models have shown exemplary performance in some results, they are not ideal in clinical applications. The main reason is the scoring system evaluates specific locations, and it is challenging to detect these locations accurately. Some of these locations are so tight that even when one is detected correctly, it is unlikely to rule out interference from others, especially in carpal bones. Moreover, hand joints may be interlaced in a severe patient with RA (Figure 1). The prospect of training a model to predict overall scores is therefore considerable.

We adopted a similar methodology and proposed an efficient overall scoring framework of RA, which we refer to as ResNet-Dwise50. The model uses a convolutional neural network to extract the features of the hand image and then uses the fully connected layer to predict the overall score. It is an end-to-end model that performs regression based on overall scores. Traditional image classification adopts low resolution (most using 224 × 224), while the medical image analysis generally requires high resolution. Training high-resolution images requires tremendous computing resources. We used a depthwise separable convolution [19] to reduce the number of parameters and multiply-adds (MAdds) required to train high-resolution images. The experiment showed that this trick increases training speed but at a loss of accuracy. We therefore introduced an inverted residual block that is an added expansion layer before the depthwise layer resulting in minimal loss of accuracy while maintaining efficiency. In summary, our contributions are as follows:

(i) An efficient CNN model (ResNet-Dwise50) was designed for the overall scoring of RA in hand X-ray datasets by introducing the techniques of depthwise separable convolution block and inverted residual block

(ii) The hand X-rays of 3818 patients with RA were collected and evaluated by experts using the SvH method

(iii) A new loss function was used to solve the imbalance of Sharp scoring

2. Related Works

2.1. Sharp/van der Heijde (SvH) Method. The Sharp/van der Heijde (SvH) method [2] quantifies erosion and joint space narrowing (JSN) assessment. The method assessed the joints in both hands (all MCPs, 1st IP, all PIPs, scaphoradial, lunate radial, distal ulna, and trapeziometacarpal) and feet (all MTPJs and 1st IPJ). Sixteen areas from each hand and six areas in each foot are included in the erosion score range from 0 to 5. JSN is assessed at 15 areas from each hand and six areas from each foot ranging from 0 to 4. Therefore, the total SvH scores range from 0 to 448. This paper only used the SvH method to assess the hands; overall scores thus ranged from 0 to 280.

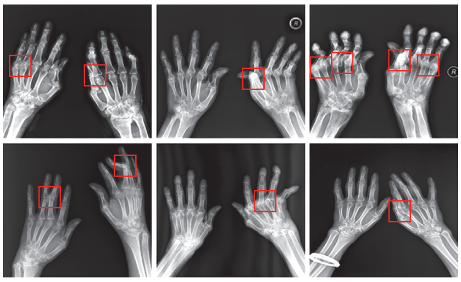


Figure 1: X-rays of some patients with severe RA. Specific positions marked with red ROIs in these images are difficult to detect.

2.2. Convolutional Neural Network. Medical imaging is the process of imaging the interior of a body for clinical analysis and medical interventions. Recent advances in deep learning, especially convolutional neural network (CNN) technology, have brought many breakthroughs to medical image analysis. CNNs can extract image features by their ability to learn features and have been successfully used in medical image classification and detection of lesion areas and medical image segmentation. For example, the VGG, Inception, and ResNet models have been used in medical image classification. The Faster R-CNN and YOLO models are used in the detection of lesion regions. FCN and U-Net models are used in medical image segmentation. The existing radiographic work of rheumatoid arthritis patients mainly focuses on automatic SvH scoring [20], segmentation, and classification [21]. For example, Rohrbach et al. [20] used a CNN to automatically predict bone erosion scores on X-ray images of patients with RA.

2.3. Depthwise Separable Convolution. A CNN can automatically detect important features without human supervision and perform well in image analysis. However, due to the limited hardware of some devices, the computation resources of the neural network still need to be reduced. Google's MobileNet [19] and Xception [22] are optimized by depthwise separable convolution, a technique that factorizes a standard convolution into a depthwise and a pointwise convolution. First, the depthwise convolution applies a single filter to each input channel. The pointwise convolution then uses a 1 × 1 convolution to combine the outputs of the depthwise convolution. In the standard convolution, both filters combine the inputs into a new set of outputs in one step. This factorization has the effect of significantly reducing the computation time and the model size. Depthwise separable convolution has been a key technology in many efficient neural network modules [23–25].

2.4. Transfer Learning. Transfer learning [26] emerged as a technique in developing deep learning models to apply to various domains. In this method, the convolutional neural network model is trained in two stages: (1) pretraining, where the network is generally trained on a large-scale labeled dataset, e.g., the CheXpert dataset [27]; (2) fine-tuning, where the pretrained network is further trained on a specific target dataset, which has fewer labeled samples than the pretraining stage. This technique has become extremely helpful in many settings, particularly in medical imaging. In our experiments, we chose this technique to fine-tune our model. We used the pretrained model parameters in CheXpert to initiate our model and then further trained on our Sharp score dataset.

3. Materials and Methods

3.1. Preprocessing and Dataset Planning

3.1.1. Dataset. To our knowledge, there is currently no RA open-source dataset on SvH scores, so we collected a private RA dataset. The dataset contains a total of 3,818 sets of X-rays. In the past two years, we collected X-rays of RA patients or suspected RA patients who visited the Department of Rheumatology of the First Clinical Medical College of Anhui University of Traditional Chinese Medicine. Two experienced radiologists were invited to score these X-ray images. The scoring rule was that both radiologists read the X-ray images in a hidden chronological order and score the patients' hand JSN and erosion of each group according to the SvH method. Finally, we took the average of the two scores to get the Sharp score of each X-ray image. In this dataset, 2700 images were randomly selected for training, while 760 images were randomly selected from the remaining images for verification, and the last 158 images were used for testing. There was no image overlap between the training set and the test set. Figure 2 shows the frequency distribution of Sharp scores in the RA dataset.

3.1.2. Data Augmentation. It is essential to augment the training data to improve the model's robustness, especially in insufficient labeled medical image data. We utilize a similar approach in [30] to achieve data augmentation by the following steps:

(1) The original image is resized by its shorter side in the range of scale of 0.9-1.0 for sampling

(2) The 1024×1024 cropping is randomly sampled from the resized image or its horizontal flip

(3) Randomly change the brightness, contrast, and saturation of the cropped image, all of which are set to 0.1

(4) Randomly rotate the image by one of the given angles: [0°, 90°, 180°, 270°]

(5) Normalize a tensor image with mean and standard deviation calculated by the training dataset

The score distribution of our dataset is roughly exponential. As we have tried, using these scores directly for training is hard to convergence. So, it is essential to standardize the score before feeding it into the network. We chose the z-score method to normalize our scores, achieving a promising result.

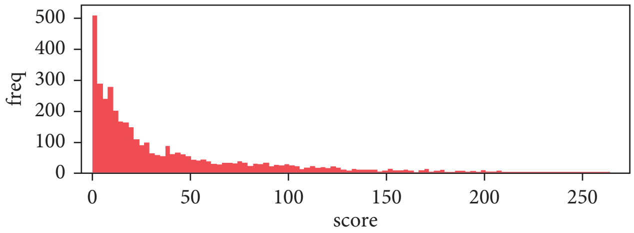


Figure 2: Histogram of frequency distribution of Sharp scores on the RA dataset.

3.2. Deep Learning Model

3.2.1. Overall Pipeline. The pipeline of the overall scoring of RA is shown in Figure 3. We use a customized depthwise separable convolutional neural network to extract a feature matrix, followed by a fully connected layer to predict a score, and then use a loss function to regress the score. In the depthwise separable convolution, standard convolutional layers are divided into two different levels. Depthwise convolution executes convolution in a single depth slice, while pointwise convolution merges the information over the entire depth. The ResNet-Dwise50 model used the inverted residual block with depthwise separable convolution, enhancing the information flowing in the model.

3.2.2. Customized Convolutional Neural Network. Considering the simplicity and high efficiency of residual networks, we use it as our benchmark architecture. He et al. [28] found that full preactivation is better than original activation and better than all other types of activation. He et al. [29] added a 2×2 average pooling layer with a stride of 2 before the 1×1 convolution in the downsampling block. The tweak is called ResNet-D and is illustrated in Figure 4. This method is effective in application and only has a slight influence on calculation cost. We therefore adopted full preactivation and ResNet-D as the benchmark architecture to improve the model performance.

The residual network consists of multiple residual blocks. We formulate the residual block at first. With y_{i-1} as its input, the output of the i^{th} block is recursively defined as the following equation:

$$y_i \equiv f_i(y_{i-1}) + y_{i-1}, \tag{1}$$

where $f_i(x)$ is the residual block's forward propagation, often 2 or 3 stacked convolution layers. Each layer consists of a sequence of convolutions, batch normalizations [24], and rectified linear units (RELUs). For the block of two stacked convolution layers, $f_i(x)$ is defined as the following equation:

$$f_i(x) \equiv W_i * \sigma(BN(W_i * \sigma(BN(x)))), \tag{2}$$

where W_i and W_i' are trainable parameters, $*$ denotes the convolution, $BN(x)$ is the batch normalization, and $\sigma(x) \equiv \max(x, 0)$.

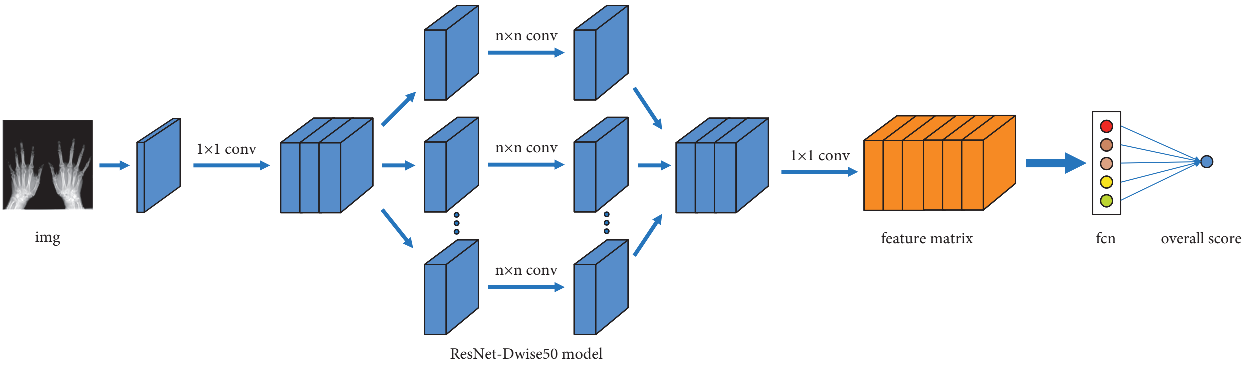


Figure 3: Pipeline of the overall scoring of RA. An inverted residual block with depthwise separable convolution is used to form the ResNet-Dwise50 model to extract the feature matrix followed by a fully connected layer to predict the overall score.

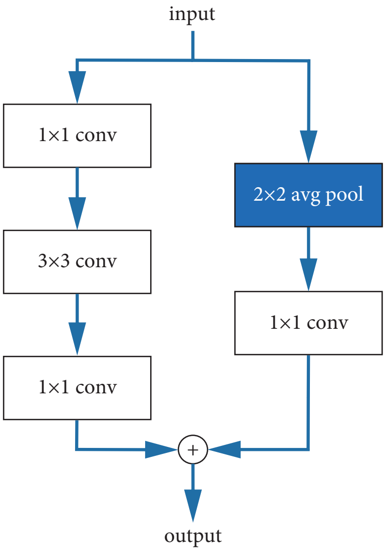


Figure 4: Downsampling block tweak of ResNet-D.

3.2.3. ResNet-Dwise50 Model. Although the standard residual network performs well, due to the high resolution of medical images, the model training consumes a large amount of memory and reduces the training efficiency. Depthwise separable convolution is used instead of standard convolution in this paper to reduce the parameters and training time of the neural network model. We also introduced inverted residual blocks [23] shown in Figure 5 to improve the model's efficiency that are based on the depthwise separable convolution.

Suppose a standard convolutional layer takes a $D_F \times D_F \times M$ input feature map F and produces a $D_G \times D_G \times N$ feature map G where D_F is the width and height of a square input feature map, D_G is the width and height of the square output feature map, M is the number of input channels, and N is the number of output channels.

In the standard convolutional layer, the convolution kernel K is parameterized by a $D_K \times D_K \times M \times N$ matrix, where D_k is the spatial dimension of the kernel and M and N are the previously defined input and output channel number, respectively. The output feature map of the standard convolution (stride is 1 with the same type) is calculated as the following equation:

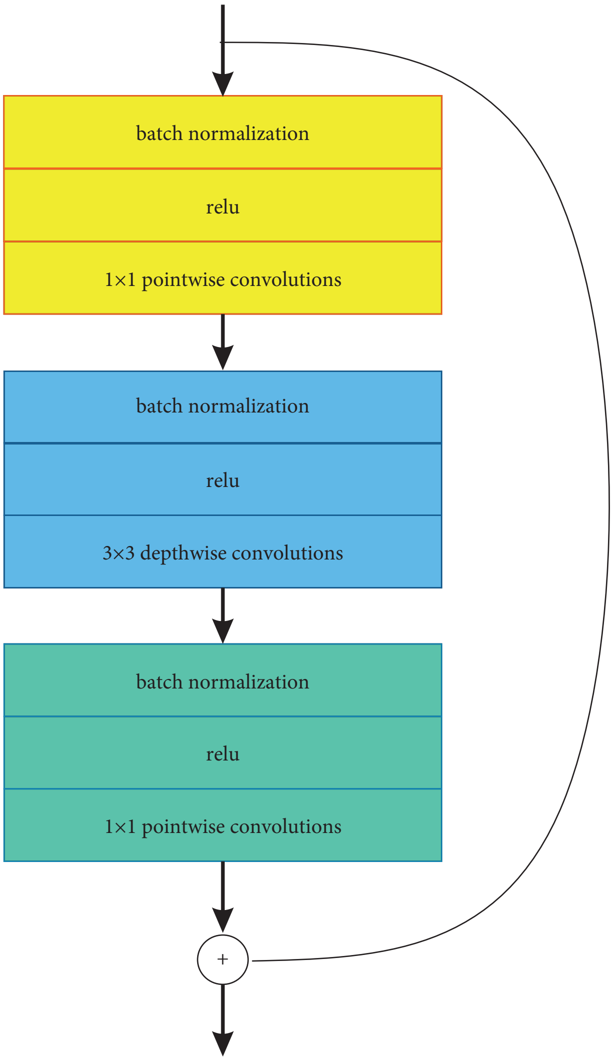

batch normalization
relu
1×1 pointwise convolutions
batch normalization
relu
3×3 depthwise convolutions
batch normalization
relu
1×1 pointwise convolutions
+

Figure 5: Structure of the inverted residual block [20]. The 1 × 1 convolutions were added before depthwise convolutions to expand the dimensions of the input matrix. In this structure, we used full preactivation.

$$\mathbf{G}_{k,l,n} = \sum_{i,j,m} \mathbf{K}_{i,j,m,n} \cdot \mathbf{F}_{k+i-1,l+j-1,m}. \tag{3}$$

So, the calculation cost of the standard convolution is shown in the following equation:

$$D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F. \tag{4}$$

The depthwise separable convolution consists of two layers: depthwise convolutions and pointwise convolutions. The depthwise convolutions apply one filter to each input channel, and the pointwise convolutions use 1 × 1 convolution to create a linear aggregate of the output of the depthwise layer. The output feature map of the depthwise convolution can be written as the following equation:

$$\mathbf{G}_{k,l,m} = \sum_{i,j} \mathbf{K}_{i,j,m} \cdot \mathbf{F}_{k+i-1,l+j-1,m}, \tag{5}$$

where $\mathbf{K}$ is the depthwise convolution kernel of $D_K \times D_K \times M$ and $\mathbf{G}$ is the output feature map. The m^{th} filter in the kernel $\mathbf{K}$ is used to the m^{th} channel in input matrix F to produce the m^{th} channel of the output feature map $\mathbf{G}$.

So, the calculation cost of the depthwise separable convolution is shown in the following equation:

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F, \tag{6}$$

which is the depthwise convolution and the pointwise convolution.

Dividing the standard convolution process into depthwise convolutions and pointwise convolutions separately can reduce the computation:

$$\begin{aligned} &\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} \\ &= \frac{1}{N} + \frac{1}{D_K^2}. \end{aligned} \tag{7}$$

We formulate the total number of multiply-adds of the inverted residual block as shown in the following equation:

$$t \cdot M \cdot D_F \cdot D_F \left(D_K \cdot D_K + M + N\right), \tag{8}$$

where t is the expansion factor. This expression has an additional term compared to equation (6) because we have an extra 1 × 1 convolution. However, the nature of this technique allows us to fully utilize minor input and output dimensions. When a bottleneck layer inputs an M channel tensor and outputs an N channel tensor, the expansion layer is $64t$ channels.

We modified the residual block in the ResNet to the inverted residual block and used full preactivation in the depthwise separable convolution block. All layers are preceded by batchnorm and ReLU nonlinearity except for the final fully connected layer. The ResNet-D tweak was used in the downsampling block to enhance the information transferring. The final average pooling reduces the spatial resolution to 1 before the fully connected layer. The ResNet-Dwise50 model is defined in Table 1. For all experiments using reverse residual blocks, we chose an expansion factor of 6. Counting depthwise as separate layers, ResNet-Dwise50 has 50 layers.

3.3. Loss Function. Class imbalance is ubiquitous in medical data. When the data categories in the training set are highly unbalanced, the efficiency of the prediction model will be significantly affected, resulting in more errors on the classes with fewer samples. Figure 2 shows the frequency distribution histogram of scoring, from which we can see that there is also a severe imbalance in our dataset. To mitigate the effect of this imbalance on the model, we devised a new loss function to treat the error of each sample equally regardless of being a member of majority or minority. We combine L1 and L2 loss advantages to get our smooth loss function and use the smooth loss to focus on complicated, false prediction examples.

For the overall prediction of Sharp scoring, we use the loss as the following equation:

$$L = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \text{smooth}\left(y_i - \hat{y}_i\right)}, \tag{9}$$

in which smooth is shown in the following equation:

$$\text{smooth}(x) = \begin{cases} ax^2 & \text{if} |x| < c, \\ x| - b, & \text{otherwise,} \end{cases} \tag{10}$$

where a, b, and c are factors that are defined manually.

Intuitively, the modulating factor a can reduce the loss contribution from easy examples, and the receiving factor c can extend the range in which an example receives low loss. The bias b for L1 loss should not be too large and plays the role of receiving factor. There are two properties of the smooth loss that might be noticed. (1) When the predicted score of a sample is close to the true score, the L2 loss function is used to reduce the loss for well-predicted examples. The c value defines how close it will be and the a value $(0 < a \leq 1)$ that can regulate the loss reduction. (2) When the predicted score is far from the true score, the L1 loss function is used for hard examples. We found that $a = 0.6$, $b = 0$, and $c = 1.0$ work best in our experiment.

4. Results and Discussion

4.1. Implementation Details. The assessment of Sharp scoring requires high resolution; we used 1024 × 1024 resolution in this experiment. The proposed module is implemented based on PyTorch 1.7. We initialize the weights as in [31] and train ResNet-Dwise50 from scratch on the hand X-rays with the SvH score dataset. We use stochastic gradient descent with a minibatch size of 4, and the learning rate starts from 0.001 and weight decay of 0.001 with a momentum of 0.9. We do not use dropout.

The model was trained on a workstation using the Ubuntu 16.04 operating system. The workstation had an Intel(R) Xeon(R) Silver 4114 CPU @ 2.20 GHz, 64 GB RAM, and two NVIDIA Tesla P4 GPUs.

Table 1: Architecture of the ResNet-Dwise50 model.

Layer name	Output size	ResNet-Dwise50
conv1	112×112	7×7 conv, 64, stride 2
residual_block1	56×56	3×3 max pool, stride 2 $\begin{bmatrix} 1 \times 1 \text{ conv}, 64 \\ 3 \times 3 \text{ dwise} \\ 1 \times 1 \text{ conv}, 128 \end{bmatrix} \times 3$
residual_block2	28×28	$\begin{bmatrix} 1 \times 1 \text{ conv}, 128 \\ 3 \times 3 \text{ dwise} \\ 1 \times 1 \text{ conv}, 256 \end{bmatrix} \times 4$
residual_block3	14×14	$\begin{bmatrix} 1 \times 1 \text{ conv}, 256 \\ 3 \times 3 \text{ dwise} \\ 1 \times 1 \text{ conv}, 512 \end{bmatrix} \times 6$
residual_block4	7×7	$\begin{bmatrix} 1 \times 1 \text{ conv}, 512 \\ 3 \times 3 \text{ dwise} \\ 1 \times 1 \text{ conv}, 1024 \end{bmatrix} \times 3$
Fc	1×1	Average pool, 1-d fc

4.2. Evaluation Metrics. We have used different metrics to evaluate the performance of the proposed method, including Pearson's correlation coefficient (ρ), mean absolute error (MAE), and root-mean-square error (RMSE). These metrics are defined in equations (11)–(13):

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right]}{\sigma_X \sigma_Y}, \tag{11}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left| y_i - x_i \right|}{n}, \tag{12}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(x_i - y_i\right)^2}. \tag{13}$$

We also calculated the t-test on predicted scores and true scores from the test dataset. We used the two-sided t-test for the null hypothesis that two related samples have identical average values—the test measures whether the average score differs significantly across samples. If we observe a significant p value, for example, greater than 0.05, then we cannot reject the null hypothesis of identical average scores. If the p value is smaller than the threshold, then we reject the null hypothesis of equal averages. We first evaluated differences in SvH scores given by two different radiologists using the above metrics. These differences are shown in Table 2. The results show that even based on the consistent cognition in the diagnosis of RA, there is still a slight inconsistency between given scores.

As described in Section 4.1, we used the mean of two SvH scores evaluated by two different readers as the true score. This study aims to measure the degree of agreement between the model's prediction as one rater and the provided true score by two different readers as another rater.

Table 2: Differences in Sharp scores given by two radiologists.

Metric	Value
ρ	0.97
MAE	12.24
RMSE	18.75
p value	>0.05
Mean of scores from radiologist 1	39.76
Mean of scores from radiologist 2	39.85
SD of scores from radiologist 1	40.90
SD of scores from radiologist 2	38.78

4.3. Ablation Studies. We used the ResNet-50 model as the baseline (B) with a bottleneck building block. Using too many parameters can introduce much variability into the model, leading to overfitting. Considering our dataset is not big enough to settle the overfitting, we reduced the number of parameters. We used the depthwise separable convolution block (DSC) to replace the bottleneck building block for acceleration. The inverted residual block (IRB) added an expansion layer in front of the depthwise separable convolution block, and the results show that this modification improves generalization capability and makes the predicted Sharp scoring closer to the actual values. Table 3 shows that when we used DSC based on ResNet-50, the performance is somewhat reduced, but the training parameters are reduced to 1/8 of the original, and the training time is reduced by half. On this basis, we continued to replace the standard residual block with IRB, and the performance of the model is greatly improved while maintaining high efficiency. The ablation studies prove that each component in our architecture plays an important role and helps to improve performance.

4.4. Quantitative Comparisons. Table 4 summarizes the performance of different models in predicting overall scores and their criteria on the RA dataset. Our ResNet-Dwise50 provided the best performance with a good compromise

Table 3: Ablation analysis for the proposed architecture. ↑ denotes larger is better, and ↓ denotes smaller is better. The best results with different backbones are highlighted in red.

Model	ρ↑	MAE↓	RMSE↓	params↓ (M)	Training time (each epoch)↓	p value
Baseline (B)	0.95	18.85	28.95	25	5 min 59 s	>0.05
B + DSC	0.93	18.96	29.48	3.7	2 min 49 s	>0.05
B + DSC + IRB	0.95	17.76	26.34	4.3	2 min 59 s	>0.05

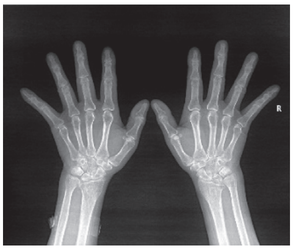
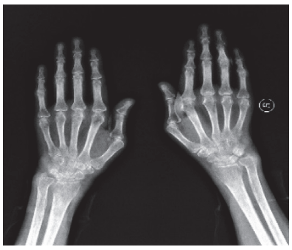
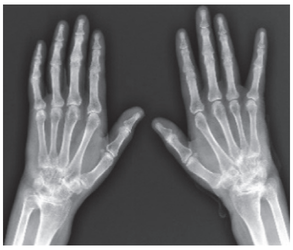
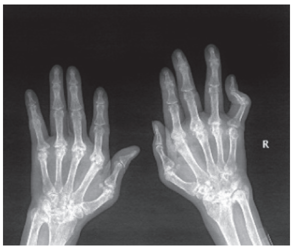
Table 4: Performance on our Sharp scoring dataset, comparison of different networks.

Model	ρ↑	MAE↓	RMSE↓	params↓ (M)	Training time (each epoch)↓	p value
ResNet-50 [30]	0.95	18.85	28.95	25	5 min 59 s	>0.05
MobileNetV2 [23]	0.93	18.94	29.85	2.2	2 min 42 s	>0.05
Ours	0.95	17.76	26.34	4.3	2 min 59 s	>0.05

Table 5: Results of models with transfer learning on our dataset.

Model	ρ↑	MAE↓	RMSE↓	p value
ResNet-50 [30]	0.97	15.32	22.58	>0.05
MobileNetV2 [23]	0.96	15.70	22.89	>0.05
Ours	0.97	14.90	22.01	>0.05

Table 6: Four sample cases of hand X-rays. The score of two radiologists (R1, R2) is shown on the 2nd and 3rd rows, respectively. The 4th row (G) is the average score of two radiologists. The 5th row (P) denotes the predicted score of the ResNet-Dwise50 model. Green: near average score. Red: more than 15 points from the average score.

Hand X-ray				
R1	4	42	74	175
R2	8	58	90	163
G	6	50	82	169
P	15	64	83	186

with the smallest RMSE and computational cost. The ResNet-50 model does not perform as well as the lightweight network, indicating that too many parameters can lead to overfitting on the small dataset. The ResNet-Dwise50 model outperformed MobileNetV2, reducing the RMSE 3.51 scores, which may prove the validity of our introduction of the inverted residual block to the baseline model. It is worth noting that the p values of all the models were greater than 0.05, so there was no statistical difference between the true score and predicted score, suggesting that these models were sufficient to predict the overall score accurately.

The small dataset may lead to overfitting or underfitting, and the accuracy of the neural network model trained from scratch is generally not high. We therefore used our proposed models pretrained on the CheXpert dataset and then reserved the parameters of the feature extractor for migrating to our specific task. The results of these models with transfer learning on our Sharp scoring dataset are shown in Table 5. We can see that transfer learning dramatically improves the training accuracy of all models.

To explain the advantages of our approach, we visualized four sets of sample images along with two radiologists and the predicted score of the ResNet-Dwise50 mode with transfer learning in Table 6. It can be easily seen that our approach can predict the overall score well, with a slight difference from the average score of the two radiologists.

To further visualize the model performance results, a sample of 100 patients was randomly selected from the test dataset, comparing true scores and predicted scores. Figure 6 shows the predicted results of our ResNet-Dwise50 model with transfer learning. It proves that our model has good performance in the overall scoring of RA. It is necessary to acknowledge that the error in predicting the Sharp scoring is different if such prediction is higher or lower than the true score. For example, the difference between

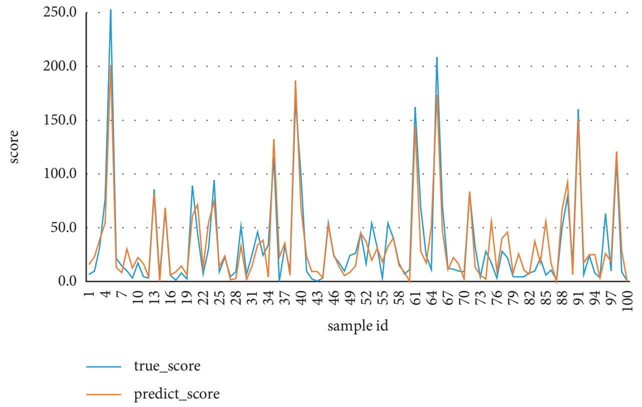


Figure 6: Comparison of prediction errors in Sharp scoring for RA disease. The graph shows the distribution of the score in the test dataset used to compare the true score and predicted score computed by the ResNet-Dwise50 model with transfer learning.

250 and 270 is not significant for a severe patient. However, the difference between 0 and 20 is significant for a milder patient. Future research may aim to find criteria to better predict and adjust the results according to the scoring strategies.

5. Conclusions

It is challenging to utilize deep learning to automatically assign SvH scores: (1) the number of training data is limited; (2) scores are severely unbalanced; (3) high-resolution images are required; (4) data labelling is based on subjective human scores, without real ground truth. We cooperated with the hospital to collect the hand X-ray radiographs of 3818 patients with RA, invited two experienced radiologists using the SvH method to evaluate the radiographs, and then took the average score as the ground truth. We proposed a smooth loss to address the imbalance of our dataset based on scoring distribution. An efficient ResNet-Dwise50 model for predicting the overall scoring of RA is proposed in this paper to improve training efficiency on high-resolution images. We replace the standard convolution with the depthwise separable convolution block to reduce the computation of the model and introduced the inverted residual block to reserve the information integrity in the bottleneck to improve the network's performance. As a result, the ResNet-Dwise50 model outperformed the MobileNetV2 model with a similar speed. In the future work, we will find a better metric for predicting and adjusting the results according to the scoring strategy, continue to carry out automated research on the overall scoring of RA foot X-rays, and apply the research results in the clinic.

Data Availability

The data used to support the results of this study are available from the corresponding authors upon request. One can also download the dataset from https://drive.google.com/drive/folders/1O11ROJypqyBksrfYqTIjdXxdc6MBao27?usp=sharing.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to acknowledge the Department of Rheumatology of the First Clinical Medical College of Anhui University of Traditional Chinese Medicine for the infrastructure and radiologist evaluation support. This work was supported in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2020-015, Natural Science Research Project of Anhui Universities under Grant KJ2020A0392, and Natural Science Research Fund Project of Department of Education of Anhui Province under Grant 2020zrzd16.

References

[1] M. Zamanpoor, "The genetic pathogenesis, diagnosis and therapeutic insight of rheumatoid arthritis," *Clinical Genetics*, vol. 95, no. 5, pp. 547–557, 2019.

[2] D. Van der Heijde, M. Van Leeuwen, P. Van Riel, and L. Van de Putte, "Radiographic progression on radiographs of hands and feet during the first 3 years of rheumatoid arthritis measured according to Sharp's method (van der Heijde modification)," *Journal of Rheumatology*, vol. 22, no. 9, pp. 1792–6, 1995.

[3] K. Kato, K. Sutherland, Y. Tanaka et al., "Fully automatic quantitative software for assessment of minute finger joint space narrowing progression on radiographs: evaluation in rheumatoid arthritis patients with long-term sustained clinical low disease activity," *Japanese Journal of Radiology*, vol. 38, no. 10, pp. 979–986, 2020.

[4] X. Li, Z. Guo, S. Zhang, and X. Guo, "A new deep model based on part pooling for medical image classification," in *IEEE International Conference on Artificial Intelligence and Industrial Design (AIID)*, pp. 145–149, IEEE, Guangzhou, China, May 2021.

[5] A. H. Masquelin, N. Cheney, C. M. Kinsey, and J. H. T. Bates, "Wavelet decomposition facilitates training on small datasets for medical image classification by deep learning," *Histochemistry and Cell Biology*, vol. 155, no. 2, pp. 309–317, 2021.

[6] K.-K. Tseng, R. Zhang, C.-M. Chen, and M. M. Hassan, "DNetUnet: a semi-supervised CNN of medical image segmentation for super-computing AI service," *The Journal of Supercomputing*, vol. 77, no. 4, pp. 3594–3615, 2021.

[7] D. Müller and F. Kramer, "MIScnn: a framework for medical image segmentation with convolutional neural networks and deep learning," *BMC Medical Imaging*, vol. 21, no. 1, pp. 1–11, 2021.

[8] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1, pp. 1–18, 2020.

[9] T. Fechter and D. Baltas, "One-shot learning for deformable medical image registration and periodic motion tracking," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2506–2517, 2020.

[10] Y. Li, J. Zhao, Z. Lv, and J. Li, "Medical image fusion method by deep learning," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 21–29, 2021.

[11] M. Kaur and D. Singh, "Multi-modality medical image fusion technique using multi-objective differential evolution based deep neural networks," *Journal of Ambient*

Intelligence and Humanized Computing, vol. 12, no. 2, pp. 2483–2493, 2021.

[12] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13753–13762, 2021.

[13] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12910–12917, 2020.

[14] M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: a survey," *Artificial Intelligence in Medicine*, vol. 106, Article ID 101878, 2020.

[15] K. Üreten, H. Erbay, and H. H. Maraş, "Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network," *Clinical Rheumatology*, vol. 39, no. 4, pp. 969–974, 2020.

[16] G. S. Mate, A. K. Kureshi, and B. K. Singh, "An efficient CNN for hand X-ray classification of rheumatoid arthritis," *Journal of Healthcare Engineering*, vol. 2021, 2021.

[17] N. Chaturvedi, "DeepRA: predicting joint damage from radiographs using CNN with attention," arXiv preprint arXiv: 2102.06982, 2021.

[18] T. Hirano, M Nishide, N Nonaka et al., "Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis," *Rheumatology advances in practice*, vol. 3, no. 2, p. rkz047, 2019.

[19] A. G. Howard, *Mobilenets: Efficient Convolutional Neural Networks for mobile Vision Applications*, arXiv preprint arXiv: 1704.04861, 2017.

[20] J. Rohrbach, T. Reinhard, B. Sick, and O. Dürr, "Bone erosion scoring for rheumatoid arthritis with deep convolutional neural networks," *Computers & Electrical Engineering*, vol. 78, pp. 472–481, 2019.

[21] U. Snekhalatha and M. Anburajan, "Automated hand radiograph segmentation, feature extraction and classification using feed forward BPN network in assessment of rheumatoid arthritis," in *Proceedings of 2nd International Conference on Intelligent Computing and Applications*, pp. 133–154, Springer, 2017.

[22] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[24] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.

[25] J. Liu, M. Li, Y. Luo, S. Yang, and S. Qiu, "Human body posture recognition using wearable devices," *Artificial Neural Networks and Machine Learning - ICANN 2019: Workshop and Special Sessions*, in *International Conference on Artificial Neural Networks*, Springer, pp. 326–337, 2019.

[26] T. George Karimpanal and R. Bouffanais, "Self-organizing maps for storage and transfer of knowledge in reinforcement learning," *Adaptive Behavior*, vol. 27, no. 2, pp. 111–126, 2019.

[27] J. Irvin, P. Rajpurkar, M. Ko et al., "Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison," *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, pp. 590–597, 2019.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *Computer Vision - ECCV 2016*, in *European Conference on Computer Vision*, Springer, pp. 630–645, 2016.

[29] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 558–567, 2019.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.