Artificial Intelligence Evolution
http://ojs.wiserpub.com/index.php/AIE/

UNIVERSAL WISER PUBLISHER

Research Article

Rheumatoid Arthritis: Automated Scoring of Radiographic Joint Damage

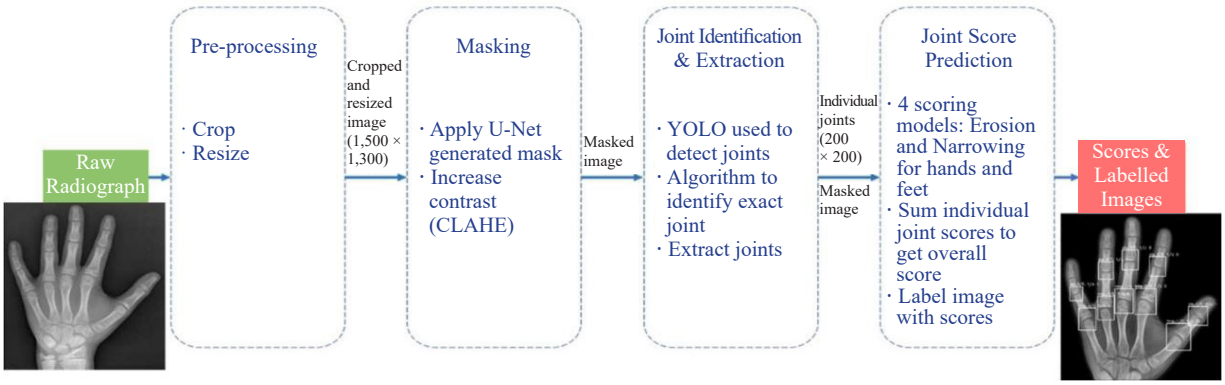
Yan Ming Tan[1], Raphael Quek Hao Chong[2], Carol Anne Hargreaves[1*]

[1]Department of Statistics and Data Science, National University of Singapore, Singapore
[2]Department of Electrical & Computer Engineering, National University of Singapore, Singapore
E-mail: carol.hargreaves@nus.edu.sg

Received: 14 April 2023; **Revised:** 17 July 2023; **Accepted:** 18 July 2023

Abstract: Rheumatoid arthritis is an autoimmune disease that causes joint damage due to inflammation in the soft tissue lining the joints known as the synovium. It is vital to identify joint damage as soon as possible to provide necessary treatment early and prevent further damage to the bone structures. Radiographs are often used to assess the extent of the joint damage. Currently, the scoring of joint damage from the radiograph takes expertise, effort, and time. Joint damage associated with rheumatoid arthritis is also not quantitated in clinical practice and subjective descriptors are used. In this work, a description of a pipeline of deep learning models to automatically identify and score rheumatoid arthritic joint damage from a radiographic image is provided. An automatic tool was built to produce scores with extremely high balanced accuracy within a couple of minutes and utilizing this would remove the subjectivity of the scores between human reviewers. Using a joint segmentation approach and training joint score prediction models with ordinal class encoding, under-sampling, and transfer learning, the joint wised ±1 balanced accuracies ranging from 91.51% to 97.30% were achieved. The ±1 balanced accuracy of the 4 models showed great potential in achieving industry-standard reliability.

Graphical abstract:



DOI: https://doi.org/10.37256/aie.4220232845

***Keywords*:** rheumatoid arthritis, radiograph, convolutional neural network, deep learning, segmentation, object detection

1. Introduction

Rheumatoid Arthritis (RA) is an autoimmune disease where the immune system mistakenly attacks the body's own tissues. This causes inflammation in the synovium, which eventually leads to joint damage. External symptoms can include red and swollen joints accompanied by pain. About 0.5-1% of the global population are affected by RA [1]. Inflammation of the joint will slowly cause cartilage, the layer of tissue that covers the ends of the bones, to erode. As the amount of cartilage decreases, the joint space also narrows. Long-term inflammation can also cause an increase in osteoclasts, cells that break down the tissue in bones, resulting in bone erosion. The degrees of narrowing and erosion observed in radiographs for RA are used in the Sharp/van der Heijde (SvH) method [2] to measure joint damage. This method looks at specific joints in the hands and feet, usually linked to inflammation caused by RA. Radiographs can provide a fair representation of joint damage but are presently not used to their full potential because there is no fast way to measure the damage quantitatively [1]. Currently, the scoring of the degree of joint damage in RA patients is done by manually reviewing their radiographs. This is generally expensive as it takes effort and time. Additionally, joint damage associated with RA is not quantitated in clinical practice, but instead, subjective descriptors such as "mild, moderate, or severe" are used in official reports [1]. Thus, it is desirable to have a method that can quickly and objectively classify joints to allow for more consistent and accurate scoring in clinical and research settings without the need for much medical expertise.

For image classification, Deep Learning (DL) models have been outperforming classical Machine Learning (ML) models since the 2012 ImageNet challenge [3]. The DL model, AlexNet, which contained 5 layers of Convolutional Neural Network (CNN) had achieved 15.2% Top-5 classification error [4]. Thereafter, subsequent years of ImageNet challenges have been dominated by CNN DL models. CNN DL models can achieve such remarkable performance due to their ability to extract essential features during training [5]. Raw images can be used directly as inputs without the need for prior feature extraction. The accuracies of these CNN DL models have been constantly increasing in plenty of diverse applications in computation vision medical tasks such as disease classification, brain cancer classification, organ segmentation, haemorrhage detection, and tumour detection [6]. Much work has also been done on applying CNN models to classifying X-ray images [7, 8], and how to enhance the image contrast [9]. An automated, accurate method is needed for unbiased assessment quantifying accrual of joint space narrowing and erosions on radiographic images of the hands and wrists, and feet for clinical trials, monitoring of joint damage over time, assisting rheumatologists with treatment decisions. Such a method has the potential to be directly integrated into electronic health records. An international crowdsourcing competition was designed and implemented [10] to catalyse the development of machine learning methods and quantify radiographic damage in Rheumatoid Arthritis (RA). A deep CNN architecture to estimate SvH scores for RA damage was designed which simultaneously performed joint localization, joint erosion assessment, and joint narrowing assessment [11]. Further, a specialized implementation of the Orthogonalizing Expectation Maximization (OEM) algorithm for cross-validation [12] was proposed as it dramatically reduced the computing time for penalized regression and cross-validation. For the joint classification, the average accuracy was 0.88, and the accuracy of severe, mild, and healthy reached 0.91, 0.79, and 0.9, respectively [13]. Existing work done on the automatic scoring of erosion due to RA produced a model based on Visual Geometry Group 16-layer model (VGG16), a CNN architecture [14], that is as accurate as human scorers [15]. Instead of SvH, the Ratingen erosion scoring was used [16]. The segmentation of the joints was not part of the work [16] as pre-extracted joint images were used.

This study built upon the approach [14-16] and achieved an even higher accuracy. As a deviation, this study focused on the problem as a classification of ordinal classes by utilizing ordinal class encoding. To deal with the imbalanced data, an under-sampling approach was attempted. Additionally, the automatic segmentation and extraction of joints were included. State-of-the-art DL models for computer vision were applied to remove image noise, and accurately segment the joints, which ensured the quality of the training samples. A lightweight U-Net architecture which was constructed for bone segmentation [17] was used for the purpose of removing background noise. As for robust joint detection in X-ray images, the You Only Look Once version 3 (YOLOv3) model [18] was implemented. This paper presents the results of our CNN model trained on these joint images for the automatic measurement of joint damage

according to the SvH scoring method. In addition, a Penalized-Regression was applied to predictions, to produce an ensemble prediction. Further, a shrinkage parameter (lambda) was used to minimize the cross-validation error.

2. Materials and methods

2.1 *Dataset*

The X-ray datasets used for the analyses described in this paper were contributed by the University of Alabama at Birmingham and were compiled from two sources - CLEAR (Consortium for the Longitudinal Evaluation of African Americans with Rheumatoid Arthritis [19]), and TETRAD (Treatment Efficacy and Toxicity in RA Database) and Repository: A study of RA patients starting biologic drugs [20].

A total of 367 sets of 4 radiographs each per patient were provided as JPG files. Each patient had a radiograph of their Left Hand (LH), Right Hand (RH), Left Foot (LF), and Right Foot (RF). Corresponding SvH scores were given in CSV format. It includes each patient's overall total damage scores, total erosion scores, total narrowing scores, and the narrowing and erosion scores of each joint. In addition, it was noticed that the size of the X-ray images varies considerably. Thus, this called for the need to resize them before the images could be used for training inputs. In addition, the dataset consists of a large number and percentage of score = 0 for both narrowing and erosion (Figure 1).

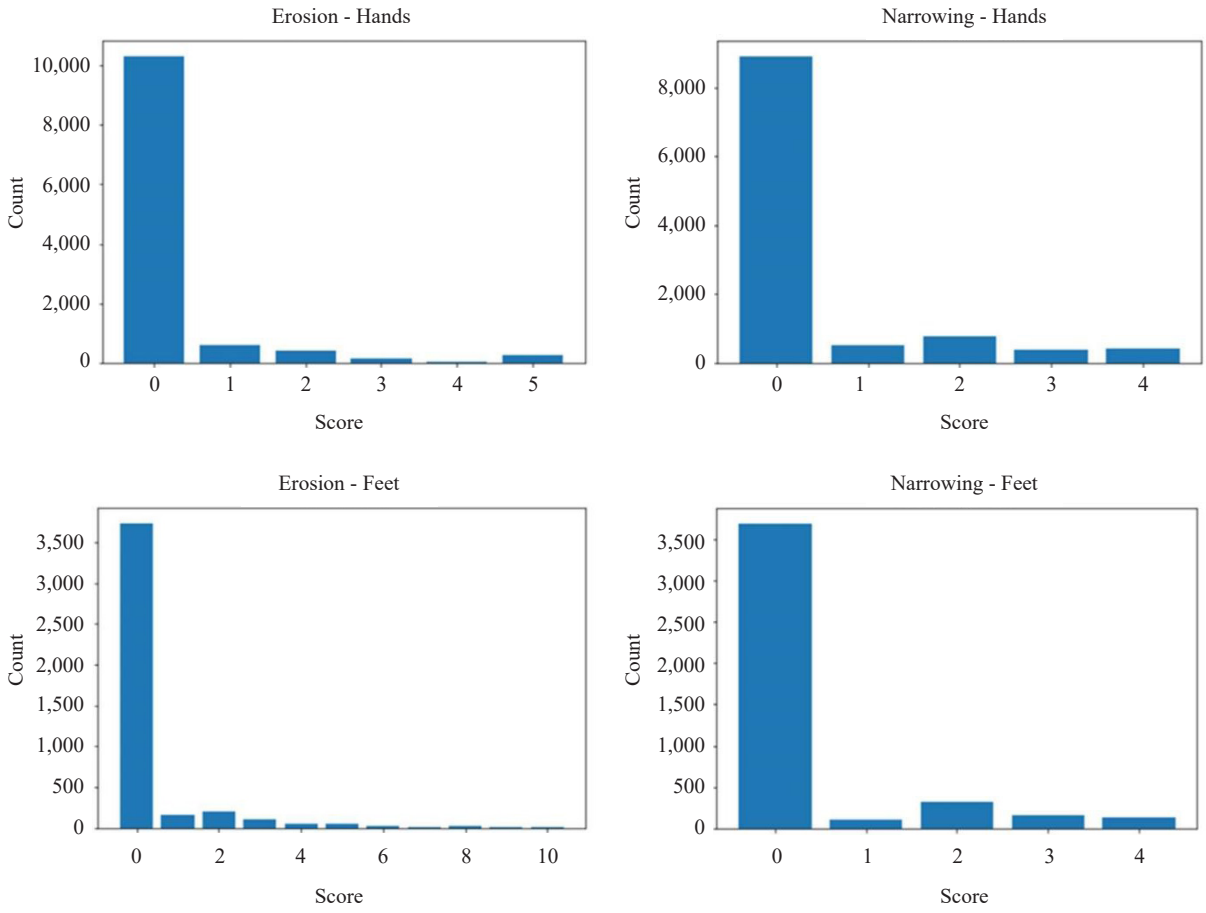


Figure 1. Distributions of scores for erosion and narrowing of each limb

2.2 *Data pre-processing*

2.2.1 *Normalization*

All images were normalized to have pixel values ranging from 0 to 1 by dividing the image arrays by 255 before training as the original pixel values belong to {0, 1, …, 255}.

2.2.2 *Image re-scaling and padding*

All the images were re-sized to 1,500 × 1,200 pixels. The original aspect ratio was retained via padding with black pixels at the borders where necessary to ensure the aspect ratio of the original is maintained.

2.2.3 *Cropping*

The raw images were first cropped to remove unnecessary parts of the limb that do not include the joints. For the hands, the bottom of the image was removed. This mostly removed the beginnings of the ulna and radius bones and part of the wrist. For the feet, the bottom 1/4 of the image was removed. This did not remove any of the joints involved in scoring as the toes are found nearer the top of the feet.

2.2.4 *Noise removal and increasing image contrast.*

A paper on Noise Removal and Contrast Enhancement for X-Ray Images [9] was replicated and the unessential parts of the feet images were additionally cropped out. Figure 2 illustrates the increase in contrast and a significant improvement in the image quality of the joints. The contrast of the images was then increased using Contrast Limited Adaptive Histogram Equalization (CLAHE) with a clip limit of 2 and the default grid size of (8, 8). This was done to address the issue that some images were almost too dark or too light to be seen clearly. Figure 2 shows the importance of improving contrast.

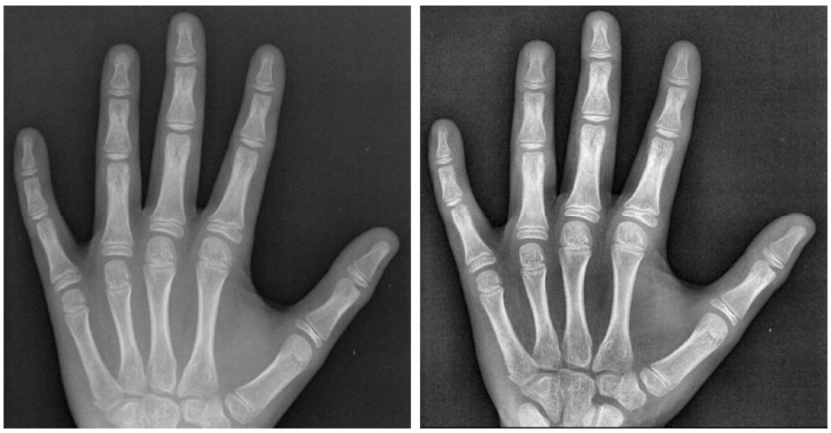


Figure 2. Randomly selected before and after CLAHE processing images for side-by-side comparison

2.3 *Joint segmentation and identification*

Apart from the joints themselves, the rest of the image is unnecessary information. As such, first extracting only the joint images and using them to train a model to predict their scores would save computation and improve accuracy. Hence, a two-step method was used: U-Net background Masking, and YOLOv3.

2.3.1 *Mask extraction algorithm*

Before the U-Net could be trained and used for background removal, an algorithm was generated and applied to

the images first. This included 3 main steps: 1) Getting the entropy of the image, 2) applying Otsu thresholding, and 3) further removal of mask noise.

Entropy is a measure of the amount of randomness in an image [21]. As there is variation in pixel values between the background and the skin and bones of the limb, this textural feature can be obtained as the entropy of the areas across the image. Entropy can be defined as follows:

$$H = -\sum_0^n p(x_i)\log_2 p(x_i) \tag{1}$$

where n is the number of discrete levels of pixel values within a region of 37 × 37 pixels, and $p(x_i)$ is the probability of a pixel belonging to the discrete level i, which is just the proportion of pixels that are in that level.

After obtaining the entropy across the image, Otsu thresholding [22] is applied to differentiate between background and limb. An intensity threshold level is supplied by minimizing the intensity variance within each class. This is defined as minimizing:

$$\sigma_w^2(t) = w_0(t)\sigma_0^2(t) + w_1(t)\sigma_1^2(t) \tag{2}$$

where $w_0(t)$ and $w_1(t)$ are the probabilities of a pixel being in the two classes when separated by the threshold t. σ_0^2 and σ_1^2 are the intensity variances of the two classes.

The mask obtained from applying the threshold mostly have some amount of noise due to the varying and random X-ray photons detected in the background of the raw image. This method is adapted from a paper that only used flood filling [23], but it was found to be unsatisfactory. Hence, additional steps, such as contour identification and filling, and flood filling from multiple origins, were taken to further remove this noise.

Contours, the borders of a region with the same intensity, were first identified and filled with white pixels. Regions with sizes smaller than a threshold of 1% of the total white area were then removed by applying a layer of black pixels along its boundary. This mostly cleared any small regions of noise in the background.

Flood filling was then used to remove any noise found within the limb. The mask image was flood filled from the four corners in case any corner might have some noise remaining which could prevent the flood filling from working. These steps resulted in a mostly full and clean mask of the limb. The mask obtained was then applied to the cropped image to remove all the noise from the background.

Flood filling was then used to remove any noise found within the limb. The mask image was flood filled from the four corners in case any corner might have some noise remaining which could prevent the flood filling from working. These steps resulted in a mostly full and clean mask of the limb.

2.3.2 *U-net masking*

To achieve robustness against the background noise, the effectiveness of the nonuniform background noise removal step had to be improved. It is almost impossible to develop a robust unsupervised algorithm to remove noises that can come in various forms. As such, a deep learning model was used to learn how to generate masks of the limbs to remove noise in the background. The U-Net architecture was selected for this model. The traditional U-Net has been commonly used for semantic image segmentation, especially for biomedical applications. It has been modified from a conventional Fully Convolutional Network (FCN) [24] so that it is able to perform well on medical images [25]. It is implemented like an encoder-decoder network but contains skip connections [25]. These skip connections create a link between layers and others that are deeper in the network. Unlike the classical encoder-decoder network, the output space mapping depends on both the latent space and the input space instead of only the latent space.

The specific architecture chosen was a lightweight U-Net architecture which was constructed for bone segmentation [17], but in this project, it was used for the purpose of background masking. Here, the number of down and up-sampling operations were adjusted to achieve higher performance in radiographic image segmentation [17]. Additionally, a Multi-Scale Block (MSB) structure was used to do feature extraction. The MSB utilizes filters of different kernel sizes to deal with features at various scales [17].

Masks that were used to train the U-Net were first obtained using the Mask Extraction algorithm. From there, good mask outputs which match the limb well and do not contain noise from the background were selected by the eye. A total of 296 hand masks and 238 feet masks were selected for training. These were then used to train 2 separate U-Net models, one for each type of limb. Figure 3 shows examples of good masks that were chosen from the Mask Extraction algorithm.

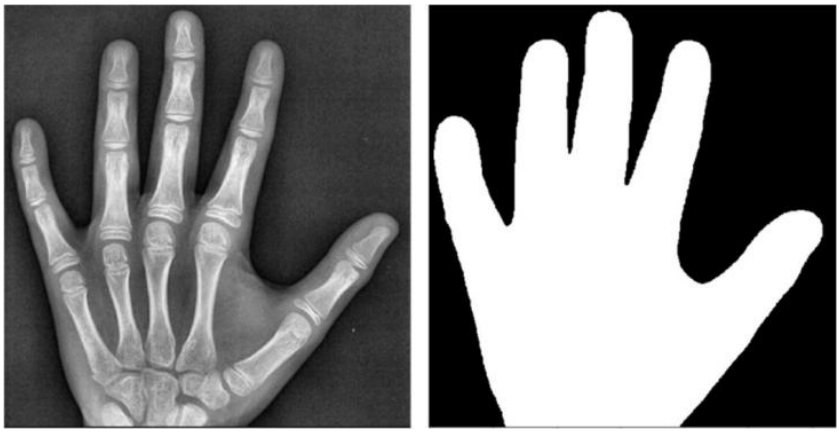


Figure 3. Examples of good masks with their corresponding original images

The optimiser used was the Adam optimizer with a learning rate set as 0.0001. The batch size was set at 16, and training ran for 200 epochs. Early stopping was used to prevent overfitting of the model. The loss function selected was binary cross entropy since a mask pixel can only either be 0 or 1. This is defined mathematically by:

$$BCE\ Loss = -\frac{1}{N}\sum_{i=0}^{N} y_i \times \log \widehat{y_i} + (1 - y_i) \times \log(1 - \widehat{y_i}) \tag{3}$$

where y_i is the value of the i-th pixel from the predicted output, y_i is its corresponding target value, and N is the number of pixels in total.

2.3.3 *YOLOv3 and joint identification*

YOLOv3 is used for fast object detection where multiple bounding boxes would be predicted at the same time together with their class probabilities from one full image directly by a single network [18]. It outperforms Faster R-CNN (FRCNN) in terms of fewer background errors [18]. YOLOv3 is also generalizable as it was able to learn broader representations of images. Hence, it can be applied to medical images such as radiographs in this case.

The overall process of using YOLOv3 for joint detection and identification includes selecting images to be labelled, initial pre-processing, manual labelling of selected images, training of 2 models for both hands and feet separately, followed by an additional step to identify the type of joints identified and their respective coordinates in the image.

All selected images were annotated via the use of an open-source graphical image annotation tool called labelImg [26]. Since all the joints are of varying sizes, all boundary boxes had to be manually drawn individually.

The hands and feet images were trained separately. Since both the hands and feet contain the same number of classes, Proximal Interphalangeal (PIP) and Metacarpophalangeal (MCP) for hands, PIP and Metatarsal-Phalangeal (MTP) for feet, the parameters for training these 2 sets of data were the same. The configurations that were changed from the default settings are as follows:

1. max_batches = 4,000.
2. steps = 3,200, 3,600.

3. classes = 2.

4. filters = 21 for the end of each convolution block - layers 82, 94, and 106.

However, the YOLOv3 models were not trained from scratch. A pre-trained weight on Common Object in Context (COCO) dataset which could predict 80 classes was used [18]. The YOLOv3 trained models could now detect the ROIs and identify the type of joint - whether it is PIP or MCP for hand images, and whether it is PIP or MTP for foot images, by keeping the threshold at 0.5 confidence level even though most of the joints were classified to be around 95-100% confident. If more than 10 or 6 joints were identified for the hands and feet respectively, the top 10 or 6 joints identified based on their confidence levels were picked (Figure 4).

However, the predictions do not provide information on which PIP, MCP/MTP the joints belonged to. To determine which fingers and toes, such as the index, middle, etc, a simple algorithm was developed to help in the identification. The algorithm was conducted based on the detected joints' positions along the horizontal axis. A backup algorithm was also created to determine joint type and location if the YOLOv3 model predicted the number of each joint type wrongly.

In the case where the model detects fewer than the required number of joints (10 for hands, and 6 for feet), the joint types identified by the model would not be utilized and their locations cannot be determined accurately. As such, this information would be unavailable during the prediction of the scores for these joints. Consequently, the scores cannot be tagged to the exact joint.

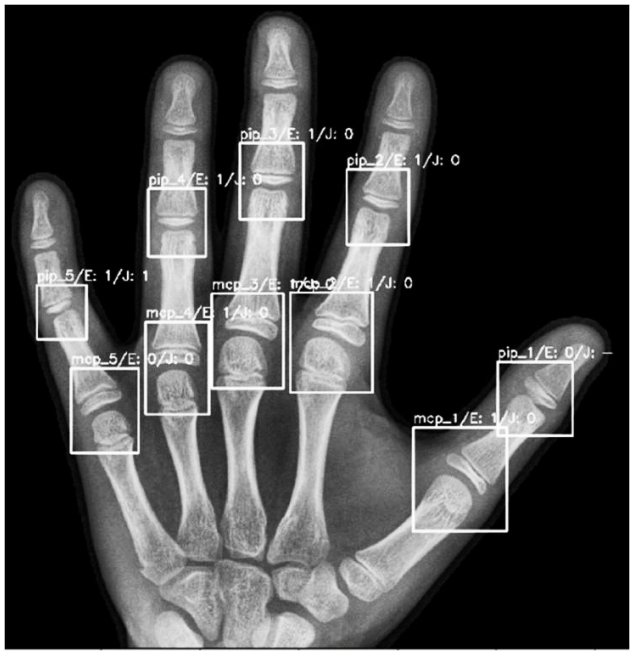


Figure 4. Example of a patient's hand image with their ROI joints identified

The YOLOv3 models performed tremendously well at detecting joints. They were tested on the remaining images that were not used for training and all the joints, except one, were correctly identified with high confidence level. The only joint that was not identified was an MTP joint of the third toe. Upon observation, it was noticed that the joints in that image had a severe amount of overlap which could have caused the error. This implied that the YOLOv3-trained models had a high combined percentage accuracy of 99.991%.

2.4 *Joint scoring*

2.4.1 *Train, validation, and test splits*

All train-test splits were conducted with a test size equal to 10% of the entire data with a random state of 42. Note

that the test sets were not touched during training and were only used for the final prediction and accuracy evaluation. During training, a validation split of 10% of the training sets was used.

2.4.2 *Metric for evaluating the model performance*

The joint scores for both erosion and narrowing take discrete integer values. As such, they can be deemed as separate classes. This means that joint scoring becomes a classification problem as compared to initially using RMSE for regression. However, it should be noted that these classes should be considered ordinal in nature.

The metric used to measure the performance of the models tested thus had to change. A modified accuracy score, the ±1 balanced accuracy, was used. To understand this, balanced accuracy is first defined as:

$$Balanced\ Acc = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{Number\ of\ Correctly\ Predicted\ Class\ i}{Total\ Number\ of\ Class\ i}\right) \tag{4}$$

where N is the total number of classes.

The ±1 balanced accuracy is like balanced accuracy just that samples that were predicted as a neighbouring class to its actual class (off by one class) would be considered as correctly predicted. This helps to account for the ordinality of the classes as well and an off-by-one classification is still medically acceptable [15]. This is also a suitable metric for a test set with an imbalanced distribution of the classes. In this case, a prediction of all samples as class-0 could achieve a high percentage accuracy. Hence, a better measure of performance would be how accurate the model is at predicting samples amongst each class. Using a balanced accuracy achieves this.

2.4.3 *VGG16 with Transfer Learning (TL)*

VGG16 is a vision model with CNN architecture that performs well in image classification [25]. Transfer learning refers to applying previously learned knowledge from other tasks to new but related tasks. Unfortunately, the pre-trained weights for VGG16 were not trained on any radiographic images. As such, the model had to be trained on radiographs using transfer learning.

The original VGG16 architecture was used but the convolution layers were frozen, leaving only the weights on the fully connected dense layers to be trainable. This was to prevent any changes in the pre-trained weights so the model could be tweaked to suit radiographic images. The original architecture consists of convolutional layers that have kernel sizes of (3, 3) and stride 1, and Max Pooling layers with a pool size of (2, 2) and stride 2. The same padding is used throughout. It ends with fully connected layers using Rectified Linear Unit (ReLU) as activation. Sixteen of its layers have weights and have about 138 million parameters.

In the modified model used, each convolution block has 2 convolution layers before ending with Max Pooling. In total, it had 6 convolution blocks unlike the original which only has 5. Here, ReLU activation was still used throughout the model after each Batch Normalization layer that can be found in each block, twice in the convolution blocks, and once in the fully connected blocks. The dropout used was set at 0.5. For the final activation, what was used depending on the approach taken. This could include ‘softmax’, ‘sigmoid’, and ‘linear’.

2.4.4 *Ordinal class encoding*

For further improvement, the use of class ordering was relooked into. Hence, another approach was required to include the ordering into the training. This was done with ordinal class encoding. This approach deviated from the method used by [15] where only Weighted Categorical Cross Entropy (WCCE) loss was used. This was an improvement upon their method as it utilizes class order which is essential information. Ordinal class encoding is a special encoding that is similar to multi-label classification encoding. It is as though the higher numbered classes must have the labels of all the classes lower than them and one more additional label. This suggests that each class is a subset of the classes lower than it. For example, the encoding for classes 0-2 would be as follows:

Class 0: [1, 0, 0].

Class 1: [1, 1, 0].

Class 2: [1, 1, 1].

2.4.5 *Loss and activation functions*

When using ordinal class encoding for the training of joints, the loss function needs to be Binary Cross Entropy (BCE) with a 'sigmoid' final activation. The sigmoid function allows for the multi-labels to work. BCE is used here because this classification is done as multiple binary classifications. For example, the first binary classification would be between the 0-class samples and the non-0-class samples. The next binary classification would be between the 1-class samples and the non-1-class samples. This continues with all the other classes. In this way, the multi-labels and BCE take the class ordering into account during training. Ordering the classes was important information to be used. However, by removing the use of loss weights, the imbalance in classes would have a strong effect.

2.4.6 *Under-sampling*

The paper by [15] used WCCE loss to deal with the class imbalance. In our study, the imbalance in class distribution was very large where the number of 0s was 10 times the size of any other class. To deal with the imbalance now that weights could not be added, an under-sampling approach was used.

Under-sampling refers to using fewer samples of a class for training. So, in this case, the number of 0 s was reduced to only 3 times the size of the other classes. Under-sampling refers to using fewer samples of a class for training. In this case, the number of joints with 0 s was reduced to an amount close to the class that had the next highest sample size.

2.4.7 *Training*

A total of 4 models were trained. Two separate models for the erosion and narrowing of the joints in the hands, and 2 separate models for erosion and narrowing of the feet joints. All models were trained for 250 epochs with a learning rate of 0.0001. As the convolution layers were frozen, the number of trainable parameters was reduced to:

- Total parameters: 15,113,049.
- Trainable parameters: 396,815.
- Non-trainable parameters: 14,716,234.

Google Colaboratory was used for all the joint model training. The Google Colaboratory provided a single 25 GB NVIDIA Tesla K80 GPU and 2vCPU @ 2.2 GHz.

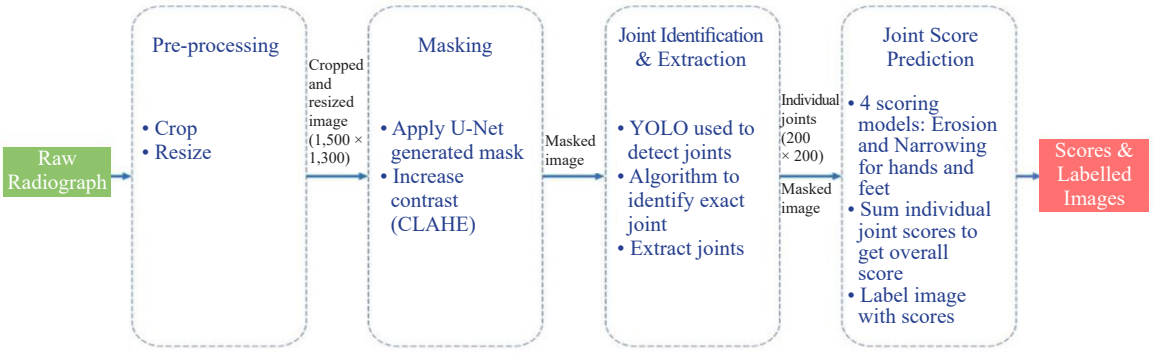


Figure 5. Full pipeline

2.5 *Full pipeline*

These models were then integrated into the end of the pipeline which includes the stages of pre-processing, masking, joint identification, and extraction. They are meant to perform the final step in the pipeline: joint score prediction. The raw radiographs would be passed into this pipeline as input while the individual joint scores and overall

scores would be outputted together with the labelled images. See Figure 5 for the visualization of the full pipeline.

3. Results

A summary of the model results can be found in Table 1 below, which shows the joint-wise ±1 balanced accuracy for each prediction model. Note that the performance on the unseen test dataset outdid the results achieved in the paper by [15] (83%). An extremely high ±1 balanced average accuracy of 94.63% was attained. Table 1 below shows the joint-wise wise ±1 balanced accuracy of each prediction model and Figure 6 below shows the models' confusion matrix.

Table 1. Joint-wise ±1 balanced accuracy of each prediction model

Model	Balanced test accuracy (%)	±1 Balanced accuracy (%)
Narrowing (Hands)	82.07	97.30
Narrowing (Feet)	83.76	91.51
Erosion (Hands)	78.77	94.63
Erosion (Feet)	57.83	92.49

4. Discussion

4.1 *Summary of results*

The ±1 balanced accuracy of the 4 models showed great potential in achieving industry-standard reliability. However, the balanced accuracy without ±1 tolerance is not as promising as it ranges from 57.83% to 83.76%. This suggests that if being exact is important, this approach does not do very well in achieving that. However, since most methods to determine RA severity relies on subjective measures, it shows that such a tool would still be able to help by getting a gauge of the disease progression. It might not be suitable to rely on this method to get an objective measure with high accuracy for standardized use in research. On the other hand, by allowing physicians to save time by automatically predicting joint damage at high accuracy, it can still be used as part of the process of determining RA severity rather than just on its own.

4.2 *Possible impact*

Currently, the method can be used as an automatic tool to get a gauge on the degree of joint damage. This might not be able to serve as a standardization for clinical or research use, but it can still help in determining RA severity. Subjective measures are still used in the clinical setting due to limited time and resources. As such, this method can provide a highly accurate and quick measure of the amount of joint damage which is indicative of RA severity.

For more extensive use, the model needs to be able to be reliable enough for use in the industry where patients' well-being is involved. Hence, the standards for the model must be extremely high. The approach proposed here and the performance it achieved shows great promise that it can reach this high standard. Once it is achieved, the solution would have a great impact on RA patient care through better disease management. This would be due to saving time and effort from manual scoring and providing an objective measure of the degree of damage for clinical diagnoses. It would also aid research that requires the use of objective joint damage scores.

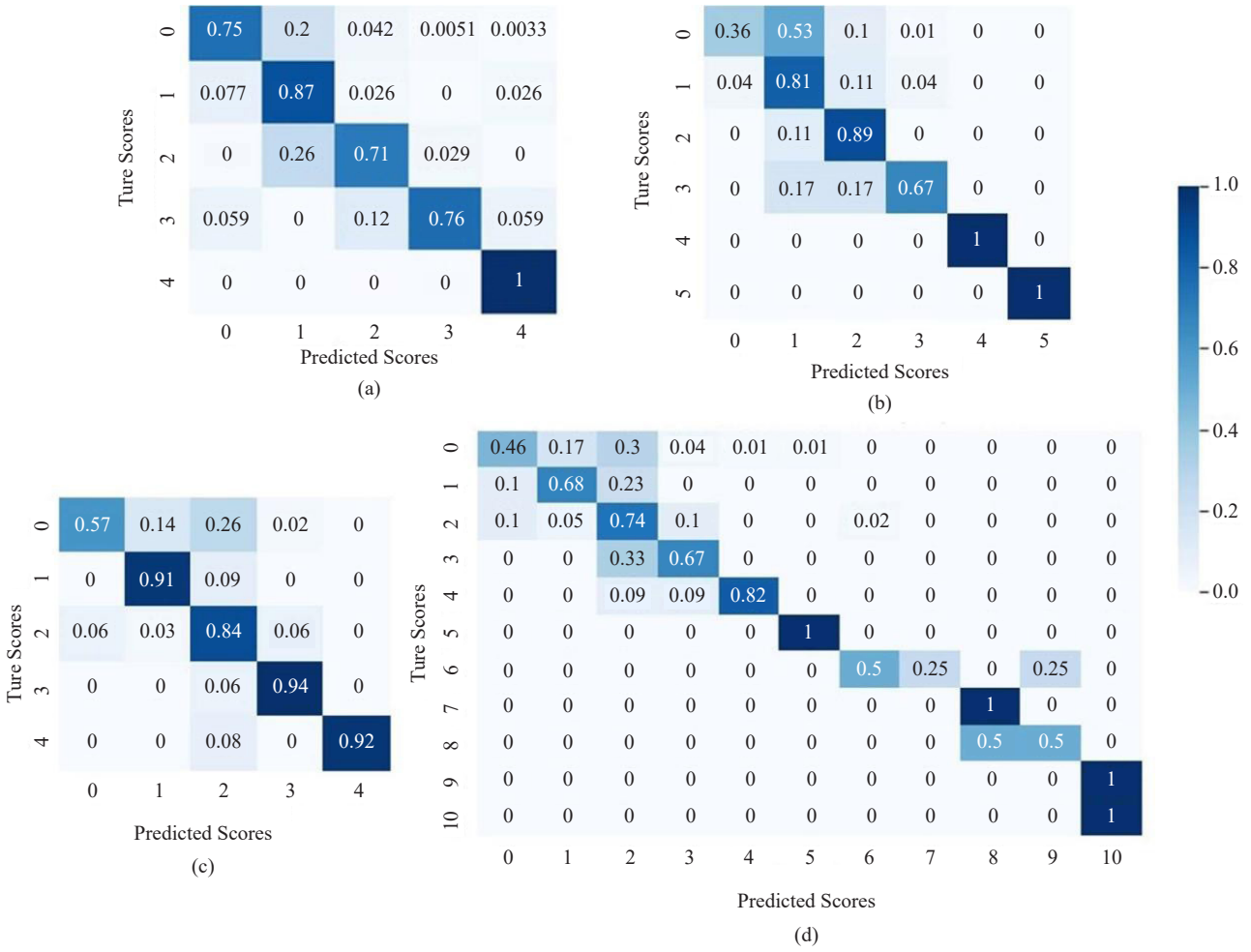


Figure 6. Confusion matrices for the +- balanced accuracies on the test set of (a) joint narrowing (Hands) model; (b) Joint erosion (Hands) model; (c) Joint narrowing (Feet) model; (d) Joint erosion (Feet) model

4.3 *Future work/limitations*

To achieve a standard high enough for use, some possible further improvements to overcome certain limitations identified could be implemented.

Firstly, improvements in the approach could be made. Bone segmentation could be used to remove the skin tissue in the radiograph. This could help improve the joint detection rate and increase prediction accuracy as there is less unnecessary information from the skin. Another improvement could be to do a binary classification of the 0 class and non-0 classes before differentiating between the non-0 classes. This could help with the imbalance in the distribution of classes where there are a lot of 0 class joints. Computer vision could also be used to get the physical characteristics of the joint spaces such as distance measures for narrowing scores or the contours of the joint.

Apart from adjusting the method, the accuracy of the prediction model can also be improved by obtaining more data samples for training. This could be done by getting data from hospitals and laboratories. Hopefully, with these improvements, the model will be able to achieve a high level of accuracy and hence, reliability, so that it can be deployed, and have its benefits realized. Further, it has been found that researchers mainly focus (93%) on medical images as data input for their models, and only one work (7%) used medical patient data in addition to medical images [28]. A valuable and important improvement can be made in integrating the image data with the clinical data that doctors use for patient diagnosis.

5. Conclusion

The YOLOv3 models performed tremendously well at detecting the joints. All the joints, except one, were correctly identified with high confidence levels. The only joint that was not identified was an MTP joint of the third toe. The YOLOv3-trained models had a high combined percentage accuracy of 99.991%. An extremely high ±1 balanced average accuracy of 94.63% was attained, with 97.30% accuracy for Hands Narrowing, 91.50% for Feet Narrowing, 94.63% for Hand Erosion, and 92.49% for Feet Erosion. The automated algorithms may be highly beneficial for physicians as they may save physicians time by automatically predicting joint damage at high accuracy. In addition, automated deep learning algorithms may also be used as part of the process of determining rheumatoid arthritis severity rather than just whether the patient has rheumatoid arthritis or not.

Acknowledgements

The Datasets used for the analyses described in this manuscript (or publication) were contributed by the University of Alabama at Birmingham. They were obtained as part of the RA2-DREAM Challenge: Automated Scoring of Radiographic Damage in Rheumatoid Arthritis through Synapse ID [syn20545111].

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Smolen JS, Aletaha D, Barton A, Burmester GR, Emery P, Firestein GS, et al. Rheumatoid arthritis. *Nature Reviews Diseases Primers*. 2018; 4: 18001.

[2] Van der Heijde DMFM, Van Leeuwen MA, Van Riel PLCM, Van de Putte LBA. 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*. 1995; 22(9): 1792-1796.

[3] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; 521(7553): 436-444.

[4] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 2012; 25: 1097-1105.

[5] Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. *Japanese Journal of Radiology*. 2018; 36(4): 257-272.

[6] Saba L, Biswas M, Kuppili V, Godia EC, Suri HS, Edla DR, et al. The present and future of deep learning in radiology. *European Journal of Radiology*. 2019; 114: 14-24.

[7] Kesim E, Dokur Z, Olmez T. X-ray chest image classification by a SmallSized convolutional neural network. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*. Istanbul, Turkey: IEEE; 2019. p. 1-5.

[8] Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest X-ray classification. *Scientific Reports*. 2019; 9(1): 1-10.

[9] Huang RY, Dung LR, Chu CF, Wu YY. Noise removal and contrast enhancement for x-ray images. *Journal of Biomedical Engineering and Medical Imaging*. 2016; 3(1): 56.

[10] Sun D, Nguyen TM, Allaway RJ, Wang J, Chung V, Yu TV, et al. A crowdsourcing approach to develop machine learning models to quantify radiographic joint damage in rheumatoid arthritis. *JAMA Network Open*. 2022; 5(8): e2227423.

[11] Krzysztof M, Krason A, Wojna Z. *Deep learning for rheumatoid arthritis: Joint detection and damage scoring in x-rays*. arXiv [Preprint]. 2021. Available from: doi: 10.48550/arXiv.2104.13915.

[12] Huling JD, Chien P. Fast penalized regression and cross-validation for tall data with the oem package. *Journal of Statistical Software*. 2022; 104(6): 1-24.

[13] Wang HJ, Su CP, Lai CC, Chen WR, Chen C, Ho LY, et al. 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; 10(6): 1355.
[14] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks. *Advances in Neural Information Processing Systems*. 2015; 28: 2017-2025.
[15] Rohrbach J, Reinhard T, Sick B, Duerr O. Bone erosion scoring for rheumatoid arthritis with deep convolutional neural networks. *Computers & Electrical Engineering. Computers & Electrical Engineering*. 2019; 78: 472-481.
[16] Rau R, Wassenberg S. Scoring methods in rheumatoid arthritis. In Imaging techniques in rheumatology. *Zeitschrift für Rheumatologie*. 2003; 62(6): 555-565. Available from: doi: 10.1007/s00393-003-0516-9.
[17] Ding L, Zhao K, Zhang X, Wang X, Zhang J. A lightweight U-net architecture multi-scale convolutional network for pediatric hand bone segmentation in X-ray image. *IEEE Access*. 2019; 7: 68436-68445.
[18] Redmon J, Farhadi A. *Yolov3: An incremental improvement*. arXiv [Preprint]. 2018. Available from: doi: 10.48550/arXiv.1804.02767.
[19] Bridges Jr SL, Causey ZL, Burgos PI, Huynh BQ, Hughes LB, Danila MI. Radiographic severity of rheumatoid arthritis in African Americans: results from a multicenter observational study. *Arthritis Care & Research*. 2010; 62(5): 624-631.
[20] Danila MI, Laufer VA, Reynolds RJ, Yan Q, Liu N, Gregersen PK, et al. Dense genotyping of immune-related regions identifies loci for rheumatoid arthritis risk and damage in African Americans. *Molecular Medicine*. 2017; 23: 177-187.
[21] Thum C. Measurement of the entropy of an image with application to image focusing. *Optica Acta: International Journal of Optics*. 1984; 31(2): 203-211.
[22] Otsu N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*. 1979; 9(1): 62-66.
[23] Huo Y, Vincken KL, Viergever MA, Lafeber FP. Automatic joint detection in rheumatoid arthritis hand radiographs. In *2013 IEEE 10th International Symposium on Biomedical Imaging*. San Francisco, CA, USA: IEEE; 2013. p. 125-128. Available from: doi: 10.1109/ISBI.2013.6556428.
[24] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE; 2015. p. 3431-3440. Available from: doi: 10.1109/CVPR.2015.7298965.
[25] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015: 18th International Conference*. Munich, Germany: Springer International Publishing; 2015. p. 234-241.
[26] Simonyan K, Zisserman A. *Very deep convolutional networks for large-scale image recognition*. arXiv [Preprint]. 2014. Available from: doi: 10.48550/arXiv.1409.1556.
[27] Zieliński B. Hand radiograph analysis and joint space location improvement for image interpretation. *Schedae Informaticae*. 2009; 17/18: 45-61.
[28] Avramidis GP, Avramidou MP, Papakostas GA. Rheumatoid arthritis diagnosis: Deep learning vs. humane. *Applied Sciences*. 2022; 12(1): 10.