Original Article

# Fully automated segmentation of wrist bones on T2-weighted fat-suppressed MR images in early rheumatoid arthritis

**Lun Matthew Wong, Lin Shi, Fan Xiao, James Francis Griffith**

Department of Imaging and Interventional Radiology, The Chinese University of Hong Kong, Hong Kong, China

*Correspondence to:* James Francis Griffith. Department of Imaging and Interventional Radiology, The Chinese University of Hong Kong, Hong Kong, China. Email: griffith@cuhk.edu.hk.

**Background:** Magnetic resonance imaging (MRI) allows accurate determination of soft tissue and bone inflammation in rheumatoid arthritis. Inflammation can be measured semi-quantitatively using the well-established RA-MRI scoring system (RAMRIS), but its application is time consuming in routine clinical practice. To fully realize an automated quantitation of inflammation scoring for clinical use, automatic segmentation of the wrist bones on MR imaging is needed. Most previous studies extracted the wrist bones on T1-weighted (T1W) MR images, and then used registration to segment T2W fat-suppressed images for bone marrow oedema quantification, introducing spatial errors into the process. Relatively little work has tried segmentation directly from T2W fat-suppressed images and no prior study have used convolution neural network (CNN) to segment the wrist bones. The purpose of this study is to develop a CNN-based algorithm for automated segmentation of the wrist bones in early rheumatoid arthritis (ERA) patients on T2W fat-saturated MR images.

**Methods:** As preliminary tests indicated that out-of-the-box segmentation CNN U-net performance was compromised by close apposition of wrist tendons and bone, we separated the volumes prior to segmentation by using classification CNN Inception V3 to group images with similar features. The classified images were then segmented by individually trained U-net. We trained the networks on 40 cases and tested them on 11 cases derived from an MR imaging dataset of 51 patients with varying severity of ERA.

**Results:** We obtained a wrist bone segmentation with an average dice similarity coefficient (DICE) of 0.888±0.014, when compared to a manually drawn label. These results are comparable to existing atlas-based methods.

**Conclusions:** We have developed a fully automatic method to segment the wrist bones in ERA patients of varying severity directly from T2W fat-suppressed MR images. This compares well with manually drawn labels.

**Keywords:** Wrist; rheumatoid arthritis (RA); magnetic resonance imaging (MRI); bone marrow oedema (BME)

Submitted Dec 07, 2018. Accepted for publication Apr 03, 2019.
doi: 10.21037/qims.2019.04.03
**View this article at:** http://dx.doi.org/10.21037/qims.2019.04.03

## Introduction

The wrist is one of the most commonly affected joints in rheumatoid arthritis (RA) and is usually involved at an early stage of the disease. Early rheumatoid arthritis (ERA) is defined as RA with symptom duration of less than 24 months. MRI of the wrist is now commonly used to evaluate the degree of inflammation in patients with ERA (1-3). Determining the level of inflammatory change present has implications for (I) determining treatment need, (II) monitoring treatment response, and (III) predicting disease outcome.

There are two components to wrist inflammation in RA patients. The first is soft tissue inflammation, namely synovitis (inflammation of the joint synovium),

and tenosynovitis (inflammation of the tendon sheath synovium) (2). The second is bony inflammation with osteitis, manifest on MR imaging as bone marrow oedema (BME) (2). Osteitis is a precursor of bone erosion (4).

Currently, the Rheumatoid Arthritis MRI score (RAMRIS) (5,6) system is routinely used to semi-quantitatively quantify inflammation on MRI in RA patients, with the severity of synovitis/tenosynovitis and osteitis being graded visually (7-9). RAMRIS can be applied to both the wrist and metacarpophalangeal joints though incorporating the metacarpophalangeal joints does not strengthen association with patient-related outcomes compared with studying the wrist alone (10). The time taken to perform RAMRIS and its semi-quantitative nature limits uptake into routine clinical practice (10). A fully automated quantitative system would be preferable. One important step in an automated process is accurate segmentation of the wrist bones.

Previous work on MRI wrist bone segmentation involved either (I) atlas- (7,11,12) or (II) seed-based (13-15) algorithms applied to T1-weighted spin echo sequences. Little work has been done on segmenting wrist bones directly from T2-weighted fat-suppressed images, which are the preferred images to depict BME (16). Convolution neural network (CNN) has been used to good effect for medical image segmentation and may be applicable to wrist bone segmentation. To test this possibility, a robust CNN, known as the U-net, designed for medical image segmentation (17), was implemented and tested. Our preliminary findings indicated that U-net yielded an unsatisfactory result for coronal T2W fat-suppressed images with the network consistently trapped in local minima that returned poor labels. The main factor that led to this unsatisfactory result was the conflict between unwanted tendons and wanted wrist bones, both of which appeared as elongated objects of low signal intensity with weak discriminatory contrast on coronal T2-weighted fat-suppressed images.

Although the inability of a trained and converged CNN to handle specific features is usually the result of either inadequate network depth, biased training samples or undesirable initial parameters, this can be mitigated by state-of-the-art hardware, more data, and/or, pertinent strategies that exploit uncommon characteristic features. In this study, we classified and separated images with the candidate features into individual groups, which were trained independently with the same model to yield multiple model parameter sets, allowing performance comparable to a deeper network to be achieved without additional resources. To do this, we proposed a novel strategy of coupling the image classification network Inception V3 (18) and the image segmentation network U-net to, first, classify the wrist region into groups of contiguous 2D images possessing different features and, second, to segment these images into different groups using multiple independently trained U-nets for wrist bone segmentation.

## Methods

### *Data acquisition*

MRI data from a cross-sectional prospective study of treatment naïve ERA patients recruited between October 2012 to January 2016 was utilized. The study protocol was approved by the local Ethics Committee with signed informed consent being obtained from each patient. All 51 patients (mean age: 53±12 years) fulfilled the 2010 American College of Rheumatology (ACR)/European League Against Rheumatism classification (EULAR) criteria for RA (19) with symptom-duration of less than 24 months at the time of recruitment.

MRI of the most symptomatic wrist was performed in all patients. Wrists were scanned in the prone position on a 3.0T system (Achieva TX, Philips Healthcare, Best) with a dedicated wrist coil to optimize signal reception. T2-weighted fat-suppressed coronal images were chosen for segmentation as each bone margin is clearly seen on coronal images and BME is only clearly seen on this sequence (6). The sequence used has a TE, TR, flip angle and field-of-view (FOV) of 70 ms, 3,121 ms, 90° and 80 mm × 80 mm respectively. Each coronal image had a 448×448 reconstruction matrix yielding a uniform pixel size of 0.178 mm × 0.178 mm with 1.65 mm inter-slice spacing.

Manual segmentation was undertaken by tracing the margin of each carpal bone on continuous images using the software ITK-SNAP, an open source medical image segmentation tool (20). The distal portion of the radius and ulna, as well as the proximal portions of the five metacarpal bones were also segmented.

### *Image classification*

In early testing, we found U-net unhelpful for bone segmentation on coronal T2-weighted fat-suppressed images due to a conflict between unwanted hypointense tendons and wanted wrist bones, which are anatomically quite closely apposed. To overcome this issue, we first

 

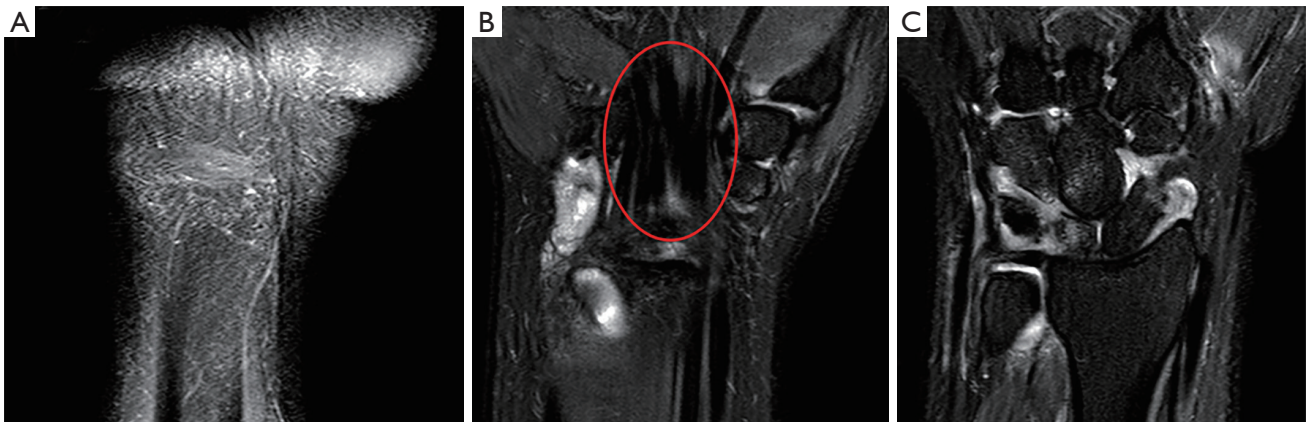


**Figure 1** Classified coronal images. Typical examples of coronal T2W fat-suppressed MR images classified into (A) type A, (B) type B and (C) type C images. The red circle highlights cluster of pixels representing tendons, which are similar in computational features to the wrist bones. This hinders performance and increases training difficulty for single U-net architecture.

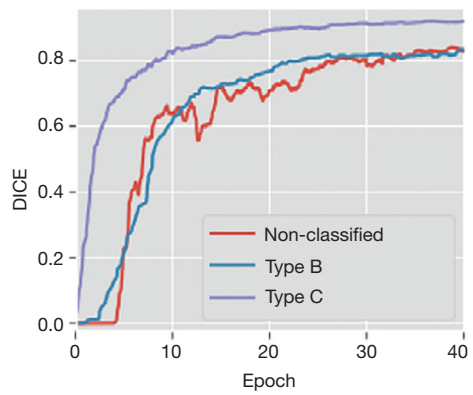


**Figure 2** Dice similarity coefficient (DICE) evolution relative to number of training epochs. Three U-nets are trained against non-classified, type B and type C images respectively under identical conditions for a total of 40 epochs. DICE converge was quickest for type C images, and slightly slower for type B and non-classified images. Non-classified groups had an unstable progression. The final convergences of the curves suggest the combined performance for type B and type C segmentation is better than that of non-classified segmentation.

systematically classified MR images and trained multiple U-nets for wrist bone segmentation. The T2-weighted fat-suppressed coronal images in the training dataset were first classified into three groups as follows:

(I) Type A: superficial images;
(II) Type B: images with tendons;
(III) Type C: other images.

Type A is defined as images reconstructed at superficial locations that do not include any bony structures. Type B refers to images that include both tendons and bone areas, both of which are represented as hypointense objects on T2W fat-suppressed images. Any other images were classified as type C, which contained mainly bone areas and no tendons. The rationale behind this classification was to separate the source of conflicting features, which originates from the co-existence of elongated soft tissue and bony structures. A typical example of the classification is shown in *Figure 1*.

To illustrate this, DICE evolution during training of networks under identical conditions for 50 epochs with and without classification is shown (*Figure 2*). An epoch refers to a single training cycle in which the complete training dataset is passed once through the neural network. DICE provides an objective reference to the networks' performance. Training with classification achieved overall better results than training without classification. The curves also support our impression that the major factor limiting in U-net's performance was the coexistence of tendons and elongated bony structures, present in type B images.

Classification was done automatically by training and utilizing the Inception V3 network (18). The Inception network is a well-established image classification network that was originally adopted for classifying images of nature, but soon became used to classify medical images for detection and diagnosis (21-23). The high-level feature classification capability of the trained network is comparable to human classification under controlled conditions (24). In our case, we utilized the network to classify images into types A–C.

 

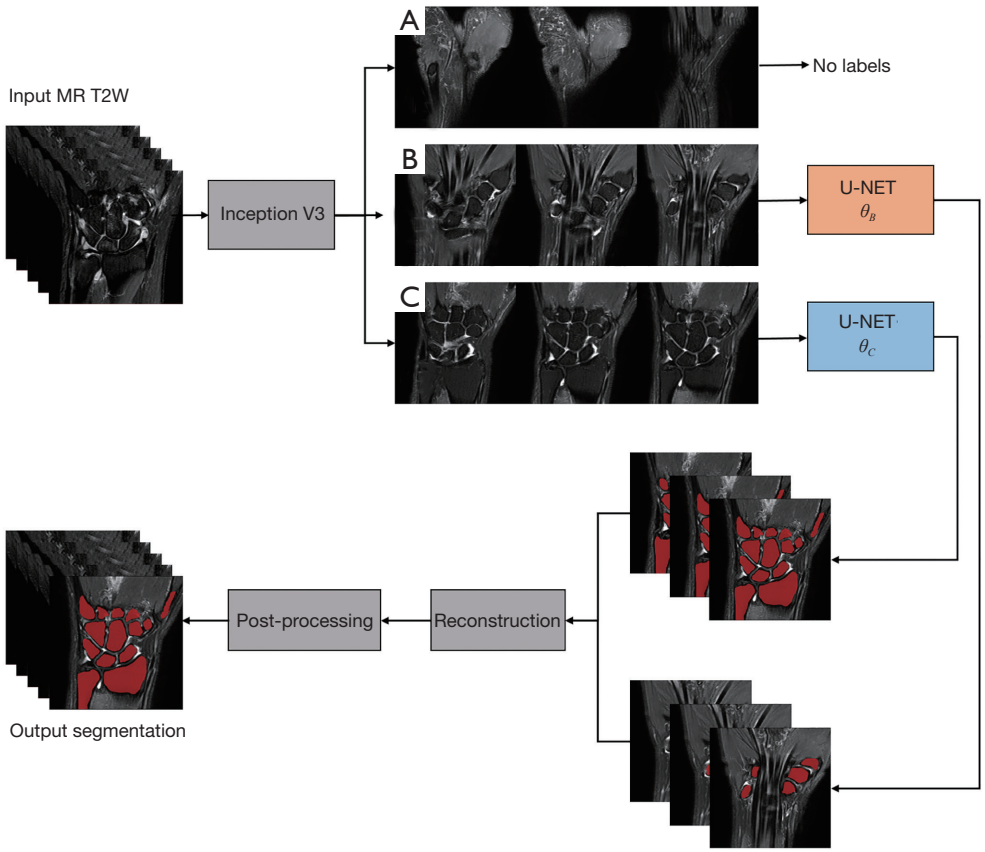


**Figure 3** Segmentation process flow chart. The input volumes were first decomposed into stacks of 2D images. These images were classified automatically into three types through the trained Inception V3 network. By definition, type A images do not contain any useful labels and were not further processed. The remaining type B and type C images were fed into two individually trained U-nets. The labels resulted from the network calculations were then reconstructed into label volumes. Finally, the label volumes underwent post-processing to refine the results.

### *Segmentation*

U-net was employed for segmentation with a minor modification to the original U-net design being made to best suit this particular application. Batch norm kernels were added as the first layer and at the beginning of each downward transition convolutional layer of U-net. The extra batch norm layers helped avoid vanishing or exploding gradients in the training process (25). Upwards transition merging was replaced by a plain bilinear interpolation instead of a concatenation.

Type B and type C classified images were fed separately into two individually trained U-nets and the results reconstructed to give pixel-wise structural-likelihood which was used for bone segmentation.

Bone segmentation results from network output then underwent post-processing: two binary image filters, namely a hole-filling filter and a median-edge-smoothing filter, were applied sequentially to each of the slices in the network output. We used a constant radius of 3×3×1 $px^3$ for the median-edge-smoothing filter. Connected components with a volume <15 $mm^3$ were considered as noise and removed (*Figure 3*).

### *Testing and training*

The image dataset from 51 patients was randomly divided into testing and training subgroups comprising 11 and 40 patients respectively, yielding a total of 222 and 818 coronal slices respectively. Each wrist MR examination was graded by RAMRIS based on the degree of synovitis/tenosynovitis, bone erosion and BME. The mean RAMRIS scores of the two groups were similar at 14.3±15.3 and 16.2±13.2 (P=0.677). Each coronal image was first assigned with a

 

**Table 1** Training key parameters

| Training parameters | Network | |
|---|---|---|
| | Inception V3 | Modified U-Net |
| Initial learning rate | $1\times10^{-5}$ | $1\times10^{-4}$ |
| Initial momentum | 0.9 | 0.2 |
| Training batch size | 80 | 6 |
| Learning rate decay $\tau$ | 0.005 | 0.1 |
| Momentum decay $\beta$ | $6.67\times10^{-4}$ | 0.02 |
| Total epoch ran | 1,000 | 300 |

The parameters used during training of the Inception V3 and UNETs are listed. The networks were trained for a large number of epochs to ensure convergence. The decay constants $\tau$ and $\beta$ adjusted the learning rate and momentum according to equation 2 and 3 respectively.

unique image number and then manually classified into type A, B or C to serve as the classification ground-truth. Image type data was provided with a unique image ID that was referenced to reconstruct the segmented images back into a volume.

### *Training*

The networks were implemented and trained with the PyTorch API (26) on a machine with an NVIDIA TITAN Xp graphic processing unit (GPU). For both networks, we optimized the parameters by using stochastic gradient descent (SGD) to minimize the negative-log-likelihood loss $L$ which was defined as:

$$L[P(X),Y;\theta]=-\sum_{i=1}^{N}\sum_{c=1}^{c}\ln[P(x_i=c|y_i=c)] \quad [1]$$

where $P(X)$ is the network output, $x_i$, $y_i$ are the $i$-th component of predicted label $X$ and ground-truth $Y$, $C$ is the number of classes, $N$ is the total number of components in $X$ or $Y$, each element of $P(X)$ is a length $C$ vector $\{P_C(x_i=c);c\in[1,C]\}$ which stored the probabilities predicted by the model with parameters while each component of Y should only contain one truth value. The networks were iterated until convergences were observed. The learning rates $r_k$ and momentum $m_k$ were decayed before the ($k$+1)-th epoch by following the equation:

$$r_{k+1}=r_k e^{-k\tau} \quad [2]$$

$$m_{k+1}=\max\{0.1,m_k e^{-k\beta}\} \quad [3]$$

We ran the network for a large number of epochs to ensure convergence. Some key training parameters were tabulated in *Table 1*.

### *Performance evaluation*

The classification accuracies were evaluated by simple accuracy. A confusion matrix was plotted from the test result. Columns represent the predicted values by the network and rows the truth values.

For the segmentation, fitness was evaluated by DICE (27) and Jaccard similarity coefficient (JAC) (9), which are defined as follows for binary labels:

$$\text{DICE}=\frac{2TP}{2TP+FP+FN} \quad [4]$$

$$\text{JAC}=\frac{\text{DICE}}{2-\text{DICE}} \quad [5]$$

The error was evaluated with global consistency error (GCE) (28) and the distance of volume between output and ground-truth evaluated by volumetric distance (VD) (29) normalized to the range 0–1. Definitions are listed as follow:

$$\text{VD}=\frac{|FN-FP|}{2TP+FP+FN} \quad [6]$$

$$\text{GCE}=\frac{1}{n}\min\left[\frac{FP(FP+2TN)}{TN+FP}+\frac{FN(FN+2TP)}{TP+FN},\ \frac{FP(FP+2TP)}{TP+FP}+\frac{FN(FN+2TN)}{TN+FN}\right] \quad [7]$$

for the upper equations, $TP$, $FP$, $TN$ and $FN$ represent true-positive, false-positive, true-negative and false-negative statistics respectively, and $n$ is the number of voxels.

We also compared our results with recent work that involves atlas-based segmentation against coronal T1-weight fat-saturated images (11). The recall rate (RR), also known as the percentage match or positive predictive value, was defined according to (11):

 

$$RR=\frac{TP}{TP+FN} \quad [8]$$

To provide quantitative references for segmentation performance from both technical and clinical perspectives, the results are evaluated by-image and by-case respectively. This is because our technique processes data image-by-image independent of inter-image information while clinical relevance relates to the complete wrist volume.

**Table 2** Confusion matrix of classification results

| Guess/truth | A | B | C | Total |
|---|---|---|---|---|
| A | 24 | 9 | 0 | 33 |
| B | 6 | 100 | 15 | 121 |
| C | 0 | 10 | 58 | 68 |
| Total | 30 | 119 | 73 | 222 |

Columns represent the predicted values by the network and rows the truth values. The trained Inception V3 network attained an 82% accuracy with a 2.7% rate of mis-classifying type B or type C images to type A.

## Results

A total of 222 test images were classified. Accuracy was 82.0%, with 6 (2.7%) of the 222 images misclassified as type A. The confusion matrix listing the hit/miss frequency is shown in *Table 2*.

Trained models were used to process the testing data. Similarity and distance metrics were evaluated image by image for different severity of BME and RAMRIS. A single U-net was also trained with identical parameters and tested on non-classified data to show improvements after employing the classification. In the image by image comparison, images with either empty predicted label or empty ground truth were excluded as comparison metrics are not well defined in these cases (*Table 2*). Exclusion was not necessary for case-by case comparison as none of the involved label images were empty. Results are shown in *Tables 3-5*.

All statistical analyses were performed using SPSS (30). ANOVA and least significant distance were used to evaluate the mean difference and performance comparisons by-image while Mann-Whitney test was used for comparisons by-case.

**Table 3** Quantitative analysis of segmentation performance

| Tested group | N | Similarity | | Distance | |
|---|---|---|---|---|---|
| | | DICE | JAC | VD | GCE |
| By image | | | | | |
| Non-classified | 188 | 0.81±0.15 | 0.70±0.17 | 0.08±0.12 | 0.06±0.03 |
| Type B | 109 | 0.83±0.10 | 0.72±0.12 | 0.07±0.08 | 0.04±0.03* |
| Type C | 73 | 0.90±0.04* | 0.83±0.06* | 0.02±0.03* | 0.07±0.02* |
| Type B + type C | 182 | 0.86±0.09* | 0.76±0.12* | 0.05±0.07* | 0.05±0.03 |
| Type B (manual) | 120 | 0.83±0.10 | 0.72±0.12 | 0.07±0.08 | 0.04±0.03* |
| Type C (manual) | 68 | 0.91±0.03* | 0.84±0.05* | 0.02±0.02* | 0.07±0.02* |
| Type B + type C (manual) | 188 | 0.86±0.09* | 0.76±0.12* | 0.05±0.07* | 0.05±0.03 |
| By case | | | | | |
| Non-classified | 11 | 0.87±0.01 | 0.78±0.02 | 0.02±0.02 | 0.05±0.01 |
| With classification | 11 | 0.89±0.01* | 0.80±0.02* | 0.02±0.01 | 0.04±0.01 |
| Manual classification | 11 | 0.89±0.01* | 0.80±0.02* | 0.02±0.01 | 0.04±0.01 |

The experiment was repeated for automatic and manual image classification, using the non-classified group as a control. The evaluation was done image-by-image and case-by-case to provide insight into both technical and clinical perspectives. The overall performance of the segmentation attained DICE 0.86 when compared by image and 0.89 when compared by case. Statistically significant improvement ($P<0.05$) to mean DICE and JAC ($P<0.05$) were observed by introducing the classification step. There are no significant differences between the manual and automatic classification groups. *, a significant mean difference ($P<0.05$) between "Tested" group and "Not-classified" group. DICE, dice similarity coefficient; JAC, Jaccard similarity coefficient; VD, volumetric distance; GCE, global consistency error.

 

**Table 4** Quantitative analysis of segmentation performance by images grouped by RAMRIS

| Tested group | RAMRIS | N | Similarity | | Distance | |
|---|---|---|---|---|---|---|
| | | | DICE | JAC | VD | GCE |
| Non-classified | 0–14 | 114 | 0.83±0.11 | 0.72±0.14 | 0.07±0.09 | 0.05±0.03 |
| | 15–30 | 35 | 0.81±0.22 | 0.71±0.21 | 0.08±0.15 | 0.05±0.04 |
| | ≥31 | 39 | 0.78±0.16 | 0.66±0.19 | 0.11±0.17 | 0.07±0.04 |
| Type B | 0–14 | 65 | 0.85±0.07 | 0.74±0.10 | 0.05±0.05 | 0.04±0.03 |
| | 15–30 | 21 | 0.85±0.07 | 0.74±0.10 | 0.05±0.05 | 0.05±0.03 |
| | ≥31 | 23 | 0.76±0.15 | 0.63±0.17 | 0.13±0.13 | 0.05±0.03 |
| Type C | 0–14 | 46 | 0.90±0.04 | 0.82±0.06 | 0.03±0.03 | 0.07±0.02 |
| | 15–30 | 12 | 0.92±0.02 | 0.85±0.04 | 0.01±0.01 | 0.06±0.02 |
| | ≥31 | 15 | 0.90±0.04 | 0.82±0.06 | 0.02±0.01 | 0.09±0.03 |
| Type B + type C | 0–14 | 111 | 0.87±0.06 | 0.77±0.10 | 0.04±0.04 | 0.05±0.03 |
| | 15–30 | 33 | 0.88±0.07 | 0.78±0.10 | 0.04±0.05 | 0.05±0.03 |
| | ≥31 | 38 | 0.82±0.13 | 0.71±0.17 | 0.09±0.11 | 0.06±0.04 |

The table shows how various degrees of ERA progression, reflected by RAMRIS score, can affect the segmentation algorithm. It is performance is slightly compromised for high RAMRIS cases (i.e., ≥31), influencing segmentation of type B images most severely. RAMRIS, RA-MRI scoring system; DICE, dice similarity coefficient; JAC, Jaccard similarity coefficient; VD, volumetric distance; GCE, global consistency error.

**Table 5** Quantitative analysis of segmentation performance by images grouped by BME

| Tested group | BME | N | Similarity | | Distance | |
|---|---|---|---|---|---|---|
| | | | DICE | JAC | VD | GCE |
| Non-classified | 0–14 | 149 | 0.82±0.14 | 0.72±0.16 | 0.07±0.11 | 0.05±0.03 |
| | ≥15 | 39 | 0.78±0.16 | 0.66±0.19 | 0.11±0.17 | 0.07±0.04 |
| Type B | 0–14 | 86 | 0.85±0.07 | 0.74±0.10 | 0.05±0.05 | 0.04±0.03 |
| | ≥15 | 23 | 0.76±0.15 | 0.63±0.17 | 0.13±0.13 | 0.05±0.03 |
| Type C | 0–14 | 58 | 0.90±0.04 | 0.83±0.06 | 0.02±0.03 | 0.07±0.02 |
| | ≥15 | 15 | 0.90±0.04 | 0.82±0.06 | 0.02±0.02 | 0.09±0.03 |
| Type B + type C | 0–14 | 144 | 0.87±0.06 | 0.77±0.09 | 0.04±0.04 | 0.05±0.03 |
| | ≥15 | 38 | 0.82±0.13 | 0.71±0.17 | 0.09±0.11 | 0.06±0.04 |

This table shows how various degrees of BME, manifest as high intensity voxels within low intensity bony structures, can affect the segmentation algorithm. Performance was compromised at high BME, mostly affecting segmentation of type B images. BME, bone marrow oedema; DICE, dice similarity coefficient; JAC, Jaccard similarity coefficient; VD, volumetric distance; GCE, global consistency error.

Coupling U-net with classification achieved 0.86 DICE coefficient compared with 0.81 DICE without classification for by-image comparison and 0.89 DICE coefficient compared with 0.87 DICE without classification for by-case comparison (*Table 3*). The smaller difference in the by-case comparisons was mainly due to the inclusion of empty images. Type B images of patients with severe ERA (RAMRIS score ≥30, or BME score ≥15) most strongly influenced segmentation accuracy (*Tables 4,5*).

Graphical results showed that the algorithm delivered

 

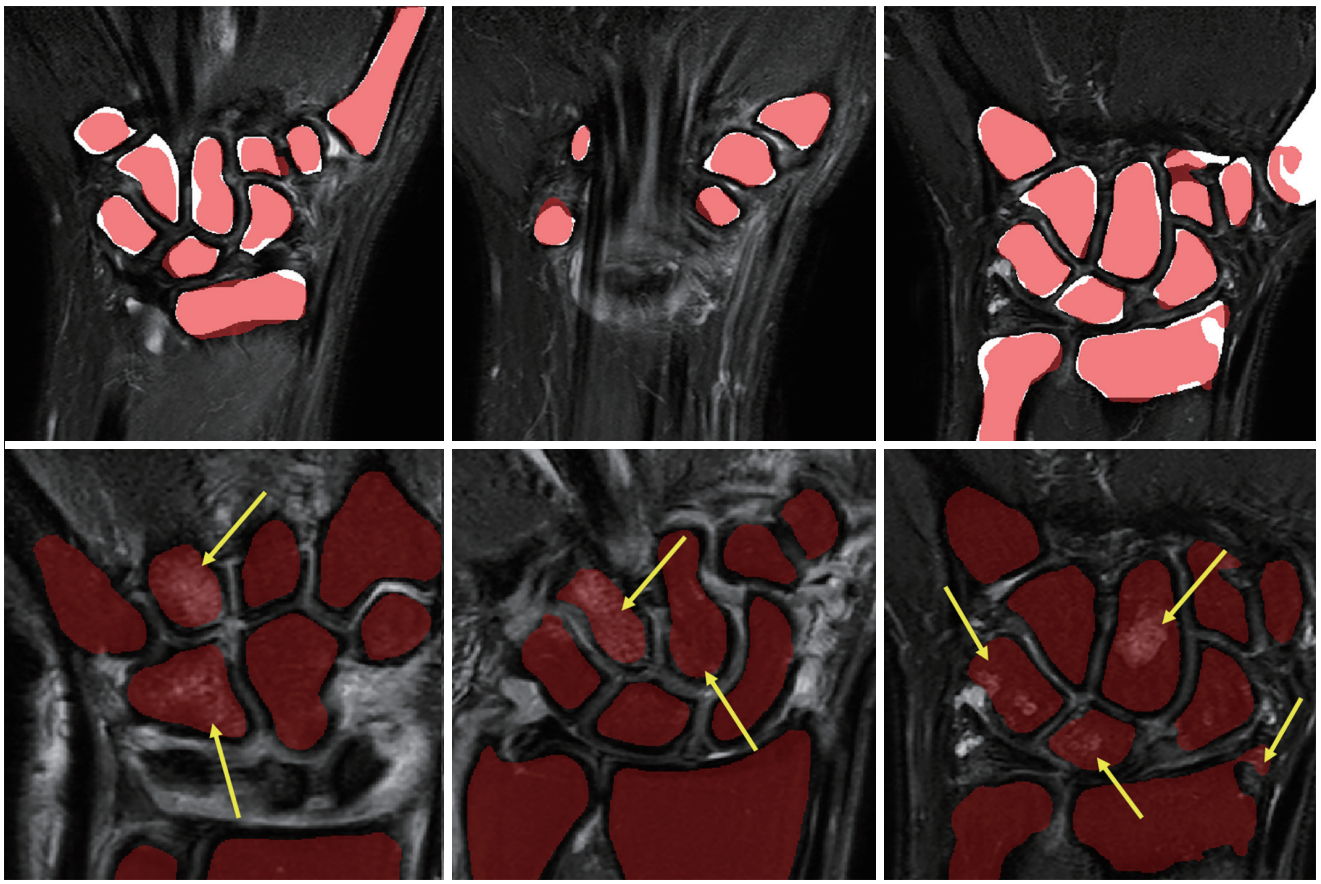

**Figure 4** Automated bone segmentation versus manual segmentation. The three images on the upper row are manually segmented T2W fat-suppressed MR images with the ground-truth displayed as opaque white labels and model predicted label as transparent red labels. The three images on the bottom row show selected slices with different degrees of bone marrow oedema and model predicted label with the yellow arrow pointing to oedematous regions. Images in the right-most column were captured at an identical location in the patient with the most severe RA-MRI scoring system (RAMRIS) score (=40) in the testing set. The algorithm delivers satisfactory accuracy even with quite severe oedema being present.

good accuracy for T2-fat-suppressed segmentation with oedematous tissues (*Figure 4*). This can be seen visually in *Figure 5*, where meshes were rendered on a single testing case. The calculated Hausdorff distance, which provides a measure of mesh differences (31) between the predicted and manually drawn label, was small (*Figure 5*).

Recently, a feasibility study addressing fully automated evaluation of BME was conducted, using an atlas-based method to segment carpals from T1-weighted fat-saturated images (11). The reported RR was 0.58–0.82 for different carpal bones. In that study, although the predicted label mostly lies within the manually labelled region, it tends to miss out a relatively high portion of the carpal volume. In comparison, our method achieved a mean RR of 0.88±0.02 based on segmentation direct from T2W fat-saturated images.

## Discussion

We have proposed a multi-step and automated CNN-based process to segment wrist bones from T2W fat-saturated images of ERA patients. Regarding the classification step, although the accuracy of Inception V3 was good, the overall accuracy of classification was not as high as expected. Misclassification into type A occurred at a rate of 2.7%, strongly affecting segmentation accuracy as these type A slices were not processed for segmentation. This inaccuracy is also likely resulted from network overfitting as the

 

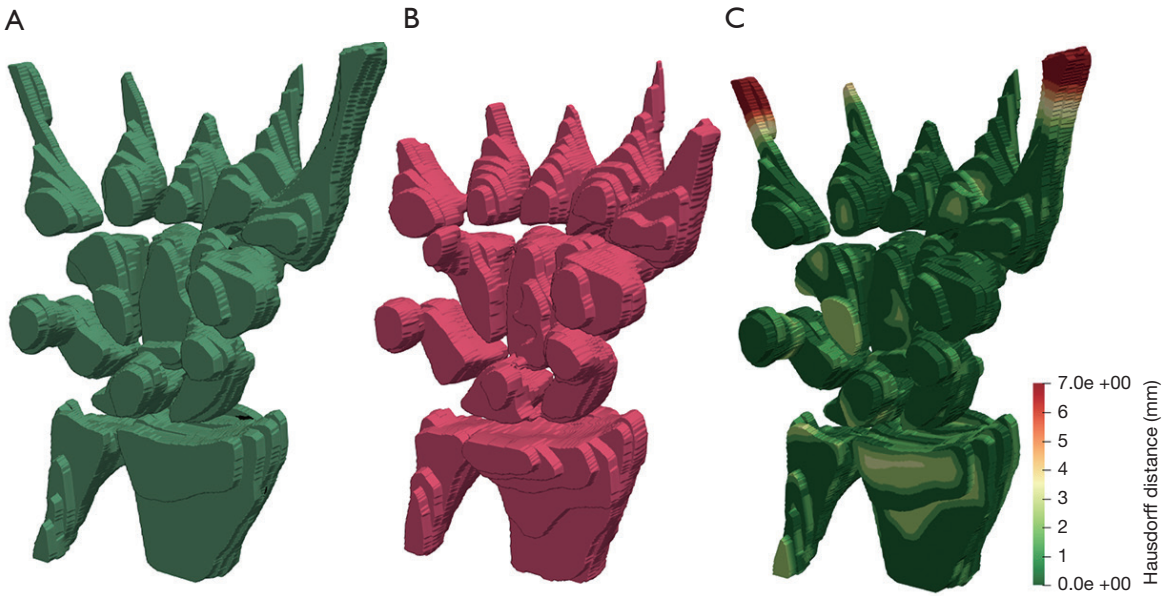


**Figure 5** Surface rendering of segmented wrist bones showing (A) the manually draw ground truth (left), (B) the model output label and (C) the Hausdorff distance sampled on ground truth mesh for a single case in the study. The Hausdorff distance calculates the distance between the sampled point and its closest counter-part in the target mesh. Most margins has an error <1 mm.

Inception network demonstrated a 99.9% accuracy during training. An increased in sample size will help to reduce this inaccuracy.

The algorithm was made more robust by verifying if type A image positions were within type B and C images. Falsely classifying type A images into type B and C images only occurs in about 4% cases.

For segmentation, there was a consistent underestimation of wrist bone volume by U-net (*Figure 4*). The wrist bone labels are slightly smaller than the ground truth in most cases. This is not a significant limiting factor as underestimation of the wrist bone margin is preferred over overestimation since the latter will lead to the inclusion of inflammatory soft tissue surrounding the bone which will affect BME quantification. The average volumetric distance is small at only 2% when compared by-case. Global consistency error values also indicate the algorithm is reasonably consistent with margins drawn manually by a trained expert.

This study has some limitations. First, the segmentation labels do not differentiate between carpal bones. Quantification of BME considers the volume of oedematous tissue relative to the whole label rather than on a bone-by-bone basis. Multi-label segmentation by out-of-the-box U-net was tested without success. This limitation is to be anticipated because 2D features between different carpals are similar, and U-net is weak in defining positional information. We have not tested our data on 3D segmentation networks and thus, cannot make any inferences as to whether 3D segmentation is superior to 2D segmentation by U-net. Nevertheless, we do expect that additional post-processing, such as connected component analysis or surface point cloud k-mean clustering, may be helpful in further refining the U-net segmented binary label into different wrist bones.

Second, the study focused on the wrist bones rather than the metacarpal joints. Focusing on the wrist allows higher resolution data acquisition. Also incorporating the metacarpophalangeal joints does not seem to strengthen association with patient-related outcomes compared with studying the wrist alone (10). We anticipate that the technique could be easily extended to larger field-of view images as the anatomy of the metacarpals and phalanges is not as complex as the carpus.

In conclusion, a novel strategy of coupling Inception V3 image classification network with segmentation network U-net to achieve segmentation of the wrist bones in ERA patients from T2W fat-saturated images was presented. The proposed method is comparable to existing atlas-based methods that utilize T1W sequences. We also showed that adding a classification step can improve training stability,

 

lessen difficulty and improve the performance of wrist bone segmentation by CNN.

## Acknowledgements

None.

## Footnote

*Conflicts of Interest:* The authors have no conflicts of interest to declare.

*Ethical Statement:* The study protocol was approved by the local Ethics Committee with signed informed consent being obtained from each patient.

## References

1. Haavardsholm EA, Bøyesen P, Østergaard M, Schildvold A, Kvien TK. Magnetic resonance imaging findings in 84 patients with early rheumatoid arthritis: Bone marrow oedema predicts erosive progression. Ann Rheum Dis 2008;67:794-800.
2. Østergaard M, McQueen F, Wiell C, Bird P, Bøyesen P, Ejbjerg B, Peterfy C, Gandjbakhch F, Duer-Jensen A, Coates L, Haavardsholm EA, Hermann KGA, Lassere M, O'Connor P, Emery P, Genant H, Conaghan PG. The OMERACT Psoriatic Arthritis Magnetic Resonance Imaging Scoring System (PsAMRIS): Definitions of key pathologies, suggested MRI sequences, and preliminary scoring system for PsA hands. J Rheumatol 2009;36:1816-24.
3. Tam LS, Griffith JF, Yu AB, Li TK, Li EK. Rapid improvement in rheumatoid arthritis patients on combination of methotrexate and infliximab: Clinical and magnetic resonance imaging evaluation. Clin Rheumatol 2007;26:941-6.
4. McQueen FM, Stewart N, Crabbe J, Robinson E, Yeoman S, Tan PL, McLean L. Magnetic resonance imaging of the wrist in early rheumatoid arthritis reveals a high prevalence of erosions at four months after symptom onset. Ann Rheum Dis 1998;57:350-6.
5. Østergaard M, Peterfy C, Conaghan P, McQueen F, Bird P, Ejbjerg B, Shnier R, O'Connor P, Klarlund M, Emery P, Genant H, Lassere M, Edmonds J. OMERACT rheumatoid arthritis magnetic resonance imaging studies. Core set of MRI acquisitions, joint pathology definitions, and the OMERACT RA-MRI scoring system. J Rheumatol 2003;30:1385-6.
6. Østergaard M, Peterfy CG, Bird P, Gandjbakhch F, Glinatsi D, Eshed I, Haavardsholm EA, Lillegraven S, Bøyesen P, Ejbjerg B, Foltz V, Emery P, Genant HK, Conaghan PG. The OMERACT rheumatoid arthritis magnetic resonance imaging (MRI) scoring system: Updated recommendations by the OMERACT MRI in arthritis working group. J Rheumatol 2017;44:1706-12.
7. Alphonse E, Roex H. Early Detection of Rheumatoid Arthritis using extremity MRI: Quantification of Bone Marrow Edema in the Carpal bones. Available online: http://resolver.tudelft.nl/uuid:7145d7a6-25bb-42a4-ba48-240d70a68792
8. Chand AS, McHaffie A, Clarke AW, Reeves Q, Tan YM, Dalbeth N, Williams M, McQueen F. Quantifying synovitis in rheumatoid arthritis using computer-assisted manual segmentation with 3 tesla MRI scanning. J Magn Reson Imaging 2011;33:1106-13.
9. Crum WR, Camara O, Hill DLG. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans Med Imaging 2006;25:1451-61.
10. Glinatsi D, Baker JF, Hetland ML, Hørslev-Petersen K, Ejbjerg BJ, Stengaard-Pedersen K, Junker P, Ellingsen T, Lindegaard HM, Hansen I, Lottenburger T, Møller JM, Ørnbjerg L, Vestergaard A, Jurik AG, Thomsen HS, Torfing T, Møller-Bisgaard S, Axelsen MB, Østergaard M. Magnetic resonance imaging assessed inflammation in the wrist is associated with patient-reported physical impairment, global assessment of disease activity and pain in early rheumatoid arthritis: longitudinal results from two randomised controlled trials. Ann Rheum Dis 2017;76:1707-15.
11. Aizenberg E, Roex EAH, Nieuwenhuis WP, Mangnus L, van der Helm-van Mil AHM, Reijnierse M, Bloem JL, Lelieveldt BPF, Stoel BC. Automatic quantification of bone marrow edema on MRI of the wrist in patients with early arthritis: A feasibility study. Magn Reson Med 2018;79:1127-34.
12. Gemme L, Nardotto S, Dellepiane SG. Automatic MPST-cut for segmentation of carpal bones from MR volumes. Comput Biol Med 2017;87:335-46.
13. Conte D, Foggia P, Tufano F, Vento M. An Enhanced Level Set Algorithm for Wrist Bone Segmentation. Image Segmentation, Image Segmentation, Pei-Gee Ho, IntechOpen. (April 19th 2011). Available online: https://www.intechopen.com/books/image-segmentation/an-enhanced-level-set-algorithm-for-wrist-bone-segmentation
14. Włodarczyk J, Czaplicka K, Tabor Z, Wojciechowski W,

 

Urbanik A. Segmentation of bones in magnetic resonance images of the wrist. Int J Comput Assist Radiol Surg 2015;10:419-31.
15. Włodarczyk J, Wojciechowski W, Czaplicka K, Urbanik A, Tabor Z. Fast automated segmentation of wrist bones in magnetic resonance images. Comput Biol Med 2015;65:44-53.
16. Manara M, Varenna M. A clinical overview of bone marrow edema. Reumatismo 2014;66:184.
17. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A. editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer, 2015.
18. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016:2818-26.
19. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, Birnbaum NS, Burmester GR, Bykerk VP, Cohen MD, Combe B, Costenbader KH, Dougados M, Emery P, Ferraccioli G, Hazes JMW, Hobbs K, Huizinga TWJ, Kavanaugh A, Kay J, Kvien TK, Laing T, Mease P, Ménard HA, Moreland LW, Naden RL, Pincus T, Smolen JS, Stanislawska-Biernat E, Symmons D, Tak PP, Upchurch KS, Vencovský J, Wolfe F, Hawker G. 2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62:2569-81.
20. Yushkevich PA, Piven J, Hazlett C, Smith G, Ho S, Gee JC, Gerig G. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 2006;31:1116-28.
21. Chang J, Yu J, Han T, Chang HJ, Park E. A method for classifying medical images using transfer learning: a pilot study on histopathology of breast cancer. In: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom). Piscataway, NJ: IEEE; 2017:1-4.
22. Midya A, Chakraborty J, Pak LM, Zheng J, Jarnagin WR, Do RKG, Simpson AL. Deep convolutional neural network for the classification of hepatocellular carcinoma and intrahepatic cholangiocarcinoma. Proc. SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis, 1057528 (27 February 2018). Available online: https://doi.org/10.1117/12.2293683
23. Torres Figueroa F, Salinas Miranda E, Bravo Sarmiento MA, Triana G, Arbeláez Escalante PA. Bone age detection via carpogram analysis using convolutional neural networks. 13th Int Conf Med Inf Process Anal 2017;1057217:45.
24. Tschandl P, Kittler H, Argenziano G. A pretrained neural network shows similar diagnostic accuracy to medical students in categorizing dermatoscopic images after comparable training conditions. Br J Dermatol 2017;177:867-9.
25. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Available online: http://proceedings.mlr.press/v37/ioffe15.pdf
26. Paszke A, Chanan G, Lin Z, Gross S, Yang E, Antiga L, Devito Z. Automatic differentiation in PyTorch. In: The NIPS workshop on the future of gradient-based machine learning software & techniques, 2017.
27. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 1945;26:297–302.
28. Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to\nevaluating segmentation algorithms and measuring ecological statistics. Proc Eighth IEEE Int Conf Comput Vision ICCV 2001;2:416-23.
29. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging 2015;15:29.
30. IBM SPSS Inc. SPSS Statistics for Windows. IBM Corp Released 2012 2012;Version 20:1-8.
31. Aspert N, Santa-cruz D, Ebrahimi T. MESH: measuring errors between surfaces using the Hausdorff distance. In: IEEE International Conference on Multimedia and Expo. Lausanne, Switzerland: IEEE, 2002:705-8.

**Cite this article as:** Wong LM, Shi L, Xiao F, Griffith JF. Fully automated segmentation of wrist bones on T2-weighted fat-suppressed MR images in early rheumatoid arthritis. Quant Imaging Med Surg 2019;9(4):579-589. doi: 10.21037/qims.2019.04.03

 