



Universiteit
Leiden
The Netherlands

A deep learning-based comparative MRI model to detect inflammatory changes in rheumatoid arthritis

Hassanzadeh, T.; Shamonin, D.P.; Li, Y.L.; Krijbolder, D.I.; Reijnierse, M.; Helm-van Mil, A.H.M.V. van der; Stoel, B.C.

Citation

Hassanzadeh, T., Shamonin, D. P., Li, Y. L., Krijbolder, D. I., Reijnierse, M., Helm-van Mil, A. H. M. V. van der, & Stoel, B. C. (2024). A deep learning-based comparative MRI model to detect inflammatory changes in rheumatoid arthritis. *Biomedical Signal Processing And Control*, 88. doi:10.1016/j.bspc.2023.105612

Version: Publisher's Version

License: [Licensed under Article 25fa Copyright Act/Law \(Amendment Taverne\)](#)

Downloaded from: <https://hdl.handle.net/1887/3720636>

Note: To cite this publication please use the final published version (if applicable).



A deep learning-based comparative MRI model to detect inflammatory changes in rheumatoid arthritis

Tahereh Hassanzadeh^a, Denis P. Shamonin^a, Yanli Li^a, Doortje I. Krijbolder^b,
Monique Reijnierse^c, Annette H.M. van der Helm-van Mil^b, Berend C. Stoel^{a,*}

^a Department of Radiology, Division of Image Processing, Leiden University Medical Center, Leiden, The Netherlands

^b Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands

^c Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands

ARTICLE INFO

Keywords:

Rheumatoid arthritis
Pixel-by-pixel change detection
Deep learning
Wrist MRI

ABSTRACT

Rheumatoid Arthritis (RA) is an autoimmune disease that mainly affects joints in the wrist and hands. It typically results in inflamed and painful joints. MRI is one of the most common imaging modalities to detect and monitor possible inflamed RA-related areas, enabling rheumatologists to treat patients more timely and efficiently. Despite the importance of finding and tracking inflamed areas associated with RA in MRI, there is no previously published work on finding pixel-by-pixel changes related to RA between baseline and follow-up MRIs. Therefore, this paper proposes a hypothesis-free deep learning-based model to discover changes in wrist MRIs on a pixel level to detect changes in inflamed areas related to RA without using prior anatomical information. To do this, a combination of a U-Net-based network and image thresholding was utilised to find pixel-level non-trivial changes between baseline and follow-up MRI images. A wrist MRI dataset including 99 individual pairs of MRI images (each pair constructed of baseline and follow-up images) was used to evaluate the proposed model. Data were collected from patients with clinically suspected arthralgia (CSA), defined as patients at risk of developing RA according to their rheumatologist and already had subclinical inflammation on MRI but could not be diagnosed with RA (yet) since they had not developed clinically detectable arthritis. The obtained results were evaluated using an observer study. The evaluation showed that our proposed model is a promising first step toward developing an automatic model to find RA-related inflammatory changes.

1. Introduction

Rheumatoid Arthritis (RA) is a chronic inflammatory autoimmune disorder predominantly affecting the joints in the hands and feet [1].

RA can be diagnosed by conducting X-rays and lab tests. In research, magnetic resonance imaging (MRI) is one of the common modalities to investigate inflammatory progression in hands and feet joints. MRI scans are currently scored visually, using the validated RAMRIS (Rheumatoid Arthritis Magnetic Resonance Imaging Scoring) system [2]. RAMRIS uses semi-quantitative scores that range between 0 and 3 to score inflammation and damage. RAMRIS has also been modified in order to include more relevant RA features and exclude less relevant features to improve the accuracy of scoring [3,4]. All of the mentioned models use anatomical information to find a specific part of e.g. the wrist to detect and score possible inflammation in these areas, i.e. inflammation in the synovium (synovitis), in the tenosynovium (tenosynovitis) and in bone marrow (bone marrow edema, BME).

However, these mentioned models have an intrinsic limitation, where images are first quantified into a global measure, after which progression is determined by basically subtracting the follow-up from the baseline measurement. This leads to a lack of sensitivity to subtle changes, as it may not show all progression within a bone, synovium or tendon, because areas classified as inflamed may progress or resolve without altering its total area, keeping the visual score unchanged. Similarly, areas labelled as unaffected may still show a subtle intensity increase that has not yet reached the threshold. This can be even more prominent in very early disease stages such as studied here. Another drawback is that these models presume relevant inflammatory patterns to occur only in predefined anatomical regions.

One of the methods to address the above drawbacks is to find inflamed areas related to RA without considering anatomical information and then scoring the detected areas. In this case, it is possible to find and track inflamed areas and find out which regions are exactly affected by RA and how it changed over time.

* Corresponding author.

E-mail address: b.c.stoel@lumc.nl (B.C. Stoel).

Since defining a pixel-level change map between two MRI scans with minimum artefacts is a complicated task, this research topic has not been investigated yet. The aim of this paper is therefore to create a pixel-level change map between baseline and follow-up images without using anatomical information, to detect subtle inflammatory changes. As a first step, however, the aim is to find any changes between two MRI scans with a minimum amount of artefacts.

In this paper, we used a combination of deep learning and classical image processing techniques to detect changes between two MRI scans. In the proposed model, we are not only looking for pixel-level changes, but we would also like to find in which areas the inflammation progressed or resolved over time. Therefore, changes are detected forward and backward in time and highlighted separately.

To do this, we propose a U-Net-based [5] network that is able to reconstruct follow-up images from baseline images and visa versa (i.e. forward and backward in time, respectively). After pre-training the network, a joint U-Net-based model is used to improve the accuracy of the image reconstruction. Two copies of the pre-trained networks are used jointly, utilising a joint loss to learn image reconstruction in both forward and backward direction. Since changes in early disease stages with subtle inflammation are rare, the network would not learn these changes, therefore the follow-up reconstructed images would not contain these changes. In the next step, differences are calculated between follow-up and reconstructed baseline image to find progressed areas, and between baseline and reconstructed follow-up image to detect resolved area (where inflammation has decreased over time). Furthermore, to detect non-trivial changes and remove artefacts from the change map Otsu [6] thresholding is utilised. The logic behind our work is, that there are areas with changes that repeatedly occur in different images and their transformation pattern can be learned by our model, and consequently can appear in the reconstructed image. Therefore, less differences will appear in these areas when we compare baseline and reconstructed follow-up images. However, unique changes cannot be learned, therefore they would not emerge in the reconstructed image and will be highlighted in the difference map. Finally, by thresholding the difference map, artefacts will be excluded from the final change map, and unique changes will emerge in the final change map.

To evaluate the proposed model, a wrist MRI dataset including 99 pairs of baseline and follow-up images were used [7]. The images are from 99 individual patients that were considered prone to develop RA (with clinically suspect joint pain). To validate the obtained results, two experts evaluated the calculated change maps. The experimental results showed that our proposed model obtained promising results as a first model to calculate pixel-level change maps as a preliminary stage in detecting RA-related inflamed areas in MRI scans of the wrist.

1.1. Research contribution

The contribution of this paper can be summarised as follows.

- An unsupervised deep learning model is proposed to detect inflammatory changes on a pixel-level.
- The proposed model can detect non-trivial changes forward and backward in time.
- Our model can be considered a significant step in developing an RA-related inflammation detection technique without needing anatomical information.

2. Literature review

Change detection is one of the important computer-based applications in many fields such as remote sensing (satellite imaging), video surveillance, and medical imaging [8]. Below a summary is provided of the various techniques that have been applied for change detection in medical imaging.

2.1. Change detection in medical imaging

Change detection is an active area of research in brain MRI imaging to detect and track changes in brain MRI over time [9–11]. For example, Patriarche et al. [12] proposed an automatic change detection model to find changes between two MRI scans (reference and target) in white-matter, due to brain tumours. In their proposed model, anatomical and lesions' intensity information was used. Subsequently, a combination of feature detection and image processing techniques was utilised to find pixel-wise changes. Furthermore, Seo et al. [13] developed a non-parametric method to detect subtle changes without using prior knowledge. To do this, a local kernel was computed from a reference image, which calculates the similarity of a pixel and its surroundings pixels. Then, it is used to compare against similar features from the target image. Finally, a dissimilarity map was made, indicating the local statistical likelihood of dissimilarity between a baseline and follow-up scan.

Nika et al. [14] proposed an automated change detection model by considering three-dimensional volumetric brain MRI images. Firstly, to align reference and target volumes, a cubic spline interpolation was utilised. In the change detection step, an optimisation model was proposed, where a dictionary composed of reference volumes with a high level of redundancy was used. Further, a Principal Component Analysis (PCA) method was utilised to reduce the dimensionality of the dictionary and consequently the computational time, which can keep more significant features in the dictionary. Subsequently, a statistical model named 3D EigenBlockCD-2 was proposed to compute the background of the reference MRI volume, and foreground blocks containing the significant changes. Finally, the output was thresholded to eliminate noise and false positives. Output intensities greater than the threshold were considered clinically relevant changes between the two MRI volumes of a particular patient, acquired at two different time points.

The first deep learning-based change detection model in medical imaging was proposed by Dupont et al. [15]. An unsupervised joint auto-encoder model was proposed to detect changes in Age-Related Macular Degeneration (ARMD). In the proposed model, an auto-encoder model constructed from four convolutional layers is used to reconstruct follow-up images. In the first step, the network has been trained to reconstruct images in both directions, and then in the final training stage, a joint auto-encoder network was suggested to learn image reconstruction in each direction (forward and backward in time) separately. Then differences between baseline image and reconstructed follow-up image and follow-up image versus reconstructed baseline image were calculated using Mean Square Error (MSE) as loss function. In the last stage, the average of both reconstruction errors was computed and to select only non-trivial changes, Otsu thresholding was applied to generate the final pixel-by-pixel change map.

As mentioned above, there is only a limited amount of previous works in the area of medical image pixel-by-pixel change detection, of which only one had applied deep learning to do this.

3. Proposed model

The aim of this paper is, therefore, to introduce a model that finds changes related to RA between two MRI images. Inspired by the work of Kalinicheva et al. [16], an unsupervised change detection model is proposed for RA-related change detection between reference and follow-up MRI scans. An overview of the proposed model is given in Fig. 1, where the model is constructed from four stages: data preparation, pre-training, final training and change detection. In the data preparation stage, super-resolution reconstruction was performed to create an isotropic image, images were registered to align the baseline and follow-up images. Subsequently, noise was removed from the areas around the wrist, by truncating the gray values in regions containing air. In the pre-training and training stages, the aim was to create and

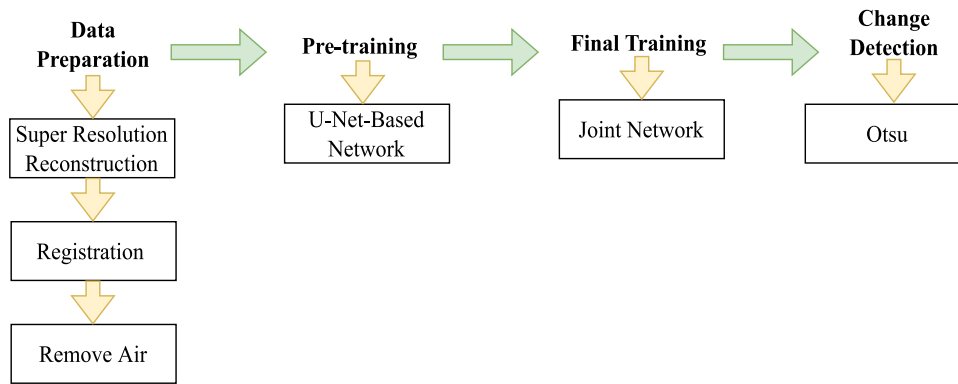


Fig. 1. Overview of the proposed model.

train convolutional neural networks to reconstruct follow-up images using baseline and vice versa. Thus, the purpose was to train a network to learn patterns that repeatedly appear in baseline and follow-up images. After final training, the proposed network can reconstruct follow-up (or baseline) images without unique changes that are assumed to be related to RA. Therefore, by subtracting the original from the reconstructed images, the difference images will contain the RA-related changes. In the final stage of the proposed model, Otsu thresholding was applied to produce the final change maps. A detailed explanation of each stage is provided below

3.1. Data preparation

Contrast-enhanced T1-weighted fat-suppressed MRI scans of the wrist of 236 patients at four time points (baseline, with 4, 12, and 24 months follow-up) were collected from the TREAT-EARLIER trial [7]. At the start of the study, patients were randomly split into a placebo and treatment arm. Treatment started from the beginning of the study and continued up to one year. To evaluate preventive effects of treatment, follow-up continued for one more year without treatment. When conducting this research, wrist MRIs at all time points were available for 107 patients. However, eight low-intensity cases were excluded from the final dataset during the pre-processing stage. Therefore, we have used 99 pairs of MRI scans from baseline and corresponding fourth follow-up, regardless of which group they belonged to. Originally, all MRI images sizes was $20 \times 512 \times 512$. Each image contained 20 slices, and the height and width of the images were 512 pixels. Per visit, two MRI scans were made: one in which the axial plane has the highest resolution (axial scan), and one in which the coronal plane contained most details (coronal scan). In the axial scan, spacing was $0.27 \times 0.27 \times 3.29$ mm and in coronal scan the image spacing was $0.19 \times 0.19 \times 2.19$ mm. To combine these two scans, Super-Resolution Reconstruction (SRR) [17] was applied to obtain one 3D image with isotropic voxels (see Fig. 2). As can be seen from Fig. 2, the axial and coronal scans were combined for each time point by SRR. First, Elastix image registration [18,19] was applied to align the coronal scan to the axial scan. Secondly, intensity matching was applied to match the intensity of the two images, and finally, SRR was applied (see Table 1). After SRR, voxel size was made isotropic to $0.195 \times 0.195 \times 0.195$ mm and consequently the number of slices in each MRI scan increased to 323. An example of a coronal and axial scan in three different cross-sectional planes (coronal, axial, and sagittal) and the corresponding SRR image are provided in Table 1.

As discussed above, the MRI images have been taken at various time points. The position and orientation of the wrist during MRI scanning can change over time. Since the purpose of our work is to find a pixel-level change map, image registration was needed to align pairs of baseline and follow-up images. To do this, the baseline SRR image was considered the fixed image and the follow-up image as a moving

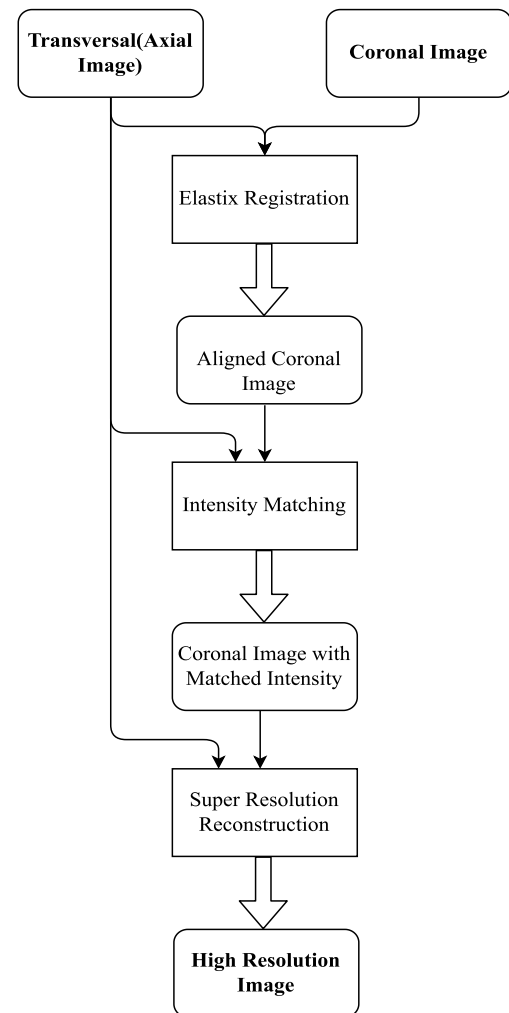



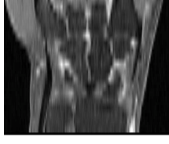
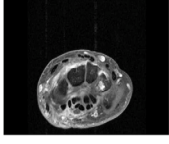
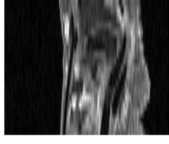

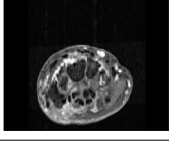
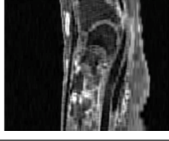


Fig. 2. The overview of super resolution reconstruction method.

image. In this paper, the Elastix registration tool [18,19] was used for registration. First, affine registration was applied; subsequently, to have a more precise and pixel-wise alignment, B-spline registration was utilised after affine registration. An example of a registered SRR image is provided in Table 2.

As can be seen from the first row, the fixed image stayed the same, but the follow-up image has been aligned to the baseline image. It needs to be noted that, during registration, missing pixels may occur especially in the first and last slices of an MRI image. This is caused by

Table 1
An example image of super-resolution reconstruction.

	Coronal Plane	Axial Plane	Sagittal Plane
Coronal Image			
Axial Image			
SRR			

the fact that not always the exact same volume of interest is scanned over time, leading to pixels that appear in one image but are absent in the other. To differentiate missing pixel pairs from complete pixel pairs, we set the intensity values of missing pixels to -500 during the registration.

In the final pre-processing stage, we applied background removal [20] to delete pixels from MRI scan containing only air. Moreover, all pixel values (except for missing pixels) between the 2.5 and 97.5 percentile were normalised to an interval of $[0,1]$.

3.2. Pre-training stage

In the proposed model, a U-Net (see Fig. 3) was first used for image reconstruction, where the network learns to reconstruct follow-up images from the baseline images and vice versa. At this stage, the network will learn to reconstruct images in both directions. It means, that the input can be a baseline to reconstruct follow-up or the other way around. During pre-training, the network learns to detect common patterns, textures and small intensity changes. However, unique changes such as inflamed areas that occurs rarely and irregularly would not be learned by network. The fact that these latter patterns have not been learned, can be used to detect these uncommon changes by image subtraction in the next phase. As shown in Fig. 3, the proposed network was constructed from six blocks for downsampling, six blocks for upsampling, and a bridging block to connect the two parts of the network. Each block had two convolutional layers with a filter size of 3×3 , along with a Batch Normalisation (BN) layer. To prevent overfitting, dropout between two convolutional layers was added to the network, with a probability of 0.1. In the downsampling part, there was a Maxpooling layer after each block, to reduce the size of the feature maps. In the upsampling section, a deconvolution layer was utilised to increase the size of feature maps. Further, to concatenate the block's input feature maps, a shortcut connection to its output feature maps was utilised for each block. To increase the quality of the up-sampled feature maps, long connections were used to transfer and concatenate down-sampled features to the up-sampled features. Finally, a convolutional layer with a filter size of 1×1 was used to generate the final output of the network.

3.3. Final training

After pre-training, two copies of the trained network is created and used as a joint network utilising a joint loss function (see Fig. 4). In

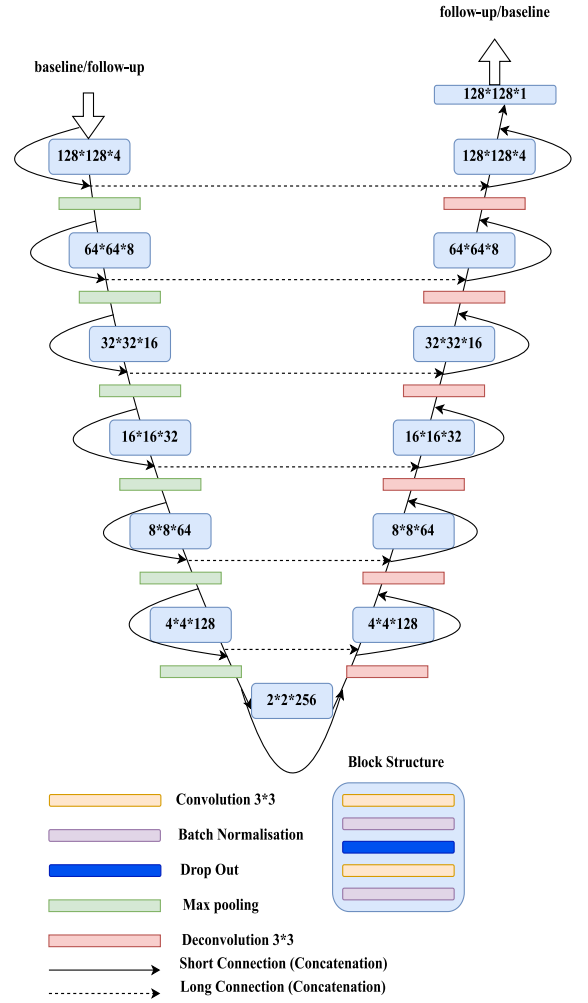


Fig. 3. The U-Net structure for image reconstruction.

the final training stage, the first U-Net is trained just using baseline images to learn to reconstruct follow-up images, and the second U-Net is utilised to reconstruct baseline images using follow-up images (see Fig. 4, original images with orange border and reconstructed images with pink border). In the proposed model, the Mean Square Error (MSE) is used as a loss function to calculate the differences between the predicted follow-up image versus the original follow-up image. First, the MSE for the predicted reconstructed image is obtained in each branch, and then the average of achieved losses is calculated as the joint loss:

$$JointMSE = \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2 + \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_i)^2}{2}, \quad (1)$$

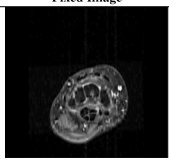
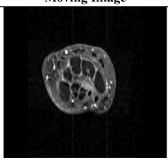
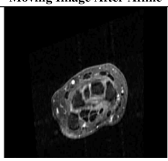
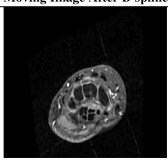
where y_i represents the true follow-up image, \bar{y}_i is the predicted follow-up image, x_i indicates the true baseline image, \bar{x}_i is the predicted baseline image, and n represents the number of pixels.

After final training, the jointly trained network is used to generate the predicted images in the test stage.

3.4. Change detection

Subsequently, to find differences between baseline and reconstructed follow-up images, these images were subtracted. Similarly, differences were calculated between follow-up and reconstructed baseline images (see Fig. 4, difference images with red border). To determine significant changes in the forward versus backward direction, Otsu

Table 2
An example image of registration.

	Fixed Image	Moving Image	Moving Image After Affine	Moving Image After B-spline
Forward Direction				

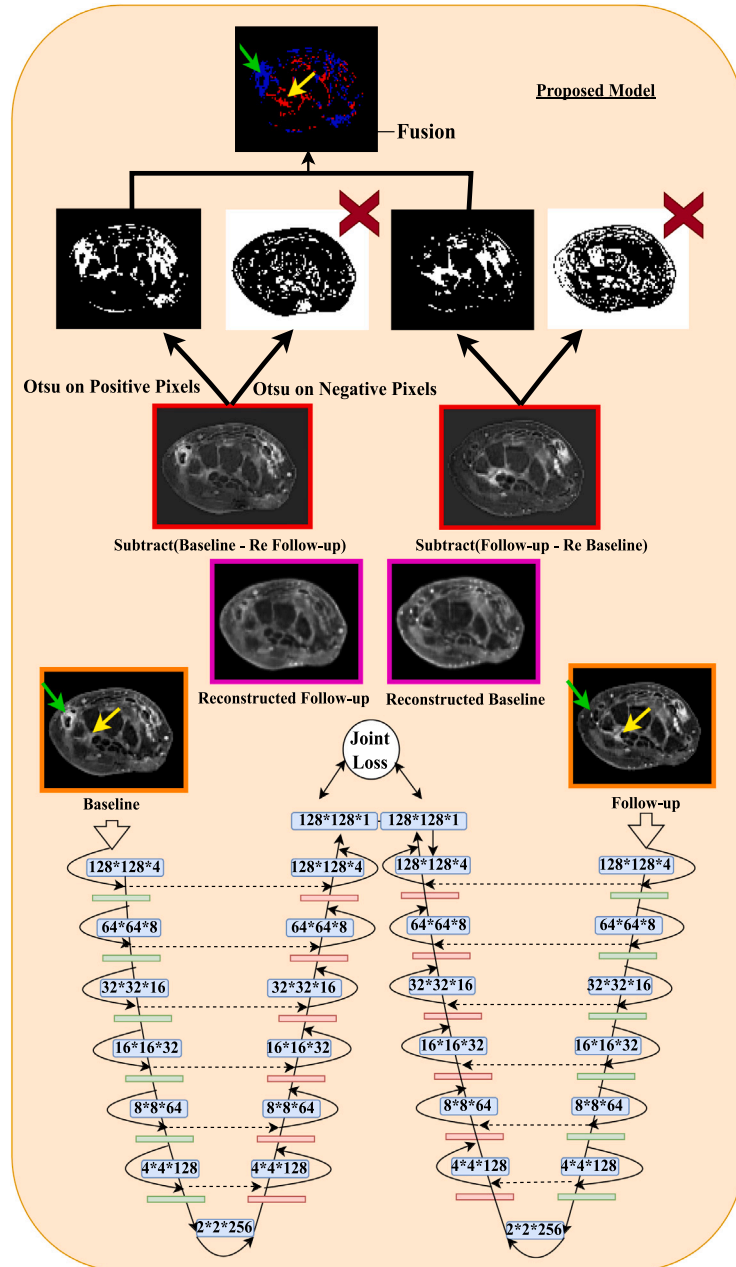


Fig. 4. Overview of the proposed model for change detection. Orange border: original baseline and follow-up images, pink border: reconstructed baseline and follow-up images, red border: difference images, Green arrows: inflammation remission, and yellow arrows: inflammation progression.

thresholding was applied to each of the difference maps. For each of the difference maps, Otsu thresholding was applied on positive and negative pixels separately. As a result, two change maps were generated. As can be seen from Fig. 4, the RA-related changes are among the positive pixels, therefore we excluded the results obtained from negative pixels from the final change map.

To calculate the final change map, the two obtained change maps were merged, such that pixels appearing in the forward change map (left branch) but not in backward one (right branch) represent inflammation remission (blue pixels). Similarly, pixels that are in backward change map but not in the forward one, show inflammation progression (red pixels). To clarify how blue and red pixels are obtained, we added green and yellow arrows to indicate corresponding areas.

As shown in Fig. 4, the proposed model was able to extract changes between two MRI scans, using the combination of convolution neural network and conventional image processing techniques.

4. Experiments

4.1. Dataset

For evaluation of the proposed model, a wrist (right and left) MRI dataset was used [7]. The dataset has MRI scans of 236 patients from four time points (baseline, with 4, 12, and 24 months follow-up). However, when this research conducted, MRI scans for all time points were available for 99 cases. Since, our purpose was finding the changes related to RA between first and last time point (follow-up after 24 months) MRI scans, we just used scans from the first and last follow-up of these 99 patient. Therefore our training (69 cases), testing (15 patients), and validation (15 cases) sets were from patients with four-time point data (first and last time points pairs).

As mentioned before, each MRI has 323 slices after SRR. However, after registration, we found some missing pixels in the first and last slices of each patients. Therefore, 230 middle slices from each patient were used for final training and evaluation of the proposed model. Accordingly, 3450 MRI slices were in the testing and validation sets, and 15 870 MRI slices in the training set. Also, to increase the number of samples in the training set, horizontal and vertical translation was utilised. It needs to be noted that, all image slices were resized to 128×128 for our proposed model.

4.2. Implementation

The proposed model was implemented using the Keras Python package [21]. All experiments were carried out on a GeForce GTX 1080 Ti (3584/12 Gb) or Quadro RTX 6000 (4608/24 Gb) GPU. The training of the proposed model was conducted in two stages. In the first stage (pre-training), the U-Net-based network was trained with data in both directions. After pre-training, two copies of the networks were used along with a joint loss function for the final training. One of the networks was specifically trained to reconstruct images in the forward direction and the second one was to reconstruct images in the backward direction. Detailed information regarding the network's pre- and final training is provided in Table 3.

In both training stages, Mean Square Error (MSE) is used as a loss function to calculate differences between constructed output and original output. However, we exclude pixels that come from out of the picture (because of registration) from our calculation (pixels with -500 as the pixel's value).

4.3. Experimental results

Since no ground-truth is available for change maps, we evaluated our results by post-hoc subjective evaluation and an observer study.

Table 3

The general parameters and their corresponding values for training the models.

Training parameters	Range
Number of epochs	50
Steps per epochs	100 000
Batch size	32
Early Stopping	30
Optimiser	Adam
Learning rate	0.00001
Loss function	MSE

4.3.1. Subjective evaluation

A few examples of obtained change maps are provided in Fig. 5, where each row represents a different patient: the first column shows baseline images; the second column contains follow-up images; the third gives the change maps obtained from simple subtraction between original baseline and original follow up images (Baseline model), where we just calculated a difference map and then applied Otsu thresholding; and the last column contains the obtained results of the proposed model (subtraction between original baseline and reconstructed follow up-images). The pixel-by-pixel change maps are displayed in two different colours. Blue pixels indicate bright areas converted to dark over time, that can indicate resolving inflammation (MRI intensities have decreased). Red colours illustrate dark areas that turned bright over time (MRI intensities have increased), which can indicate newly inflamed areas emerging in the follow-up MRI scan. As can be seen, our proposed change map found intensity changes with less noise and less artefacts, as compared to the baseline model. If the reconstruction for a particular patient went well, the obtained change map generally obtained a higher accuracy in all MRI slices, as compared to the baseline model.

4.3.2. Observer study

To evaluate our proposed model, we performed an observer study. To do this, we designed a viewer application to read and evaluate the obtained change maps. In this section, firstly, an overview of viewer is provided, then the obtained results are discussed in the following subsection.

4.3.2.1. Viewer. For evaluating the change maps by observers, we developed a viewer in C# (see Fig. 6), containing four main sections. In the first section, the baseline image, follow-up image and the change map are presented. In the second section, tools to load and scroll through various MRI studies are provided. There are buttons to load MRI scans of next or previous patients, and also buttons and sliders to scroll through next or previous slices. In addition, there is a possibility to overlay the change map onto the image slice and scroll through different slices with overlay. Although our purpose was to score axial scans, the coronal plane was also provided to help readers to have a better understanding of the overall changes between baseline and follow-up images.

The main part of the viewer is on scoring images visually in four different domains. We asked the readers to check and evaluate changes in the synovium, tenosynovium, bone marrow and the remainder of the image (entire image except areas of synovium, tenosynovium and bone marrow). For each section, we separately evaluated intensity increase and decrease as follows:

- False Positive (FP) $[-10, -1]$: If there is no change between the baseline and follow-up image, but the change map did show changes. Dependent on the size or severity of the presented erroneous changes, the score can be between -1 to -10 (small/insignificant to extensive/severe errors).
- False Negative (FN) (0): If there is a change between baseline and follow-up, but it is not shown in the change map, the score is 0.

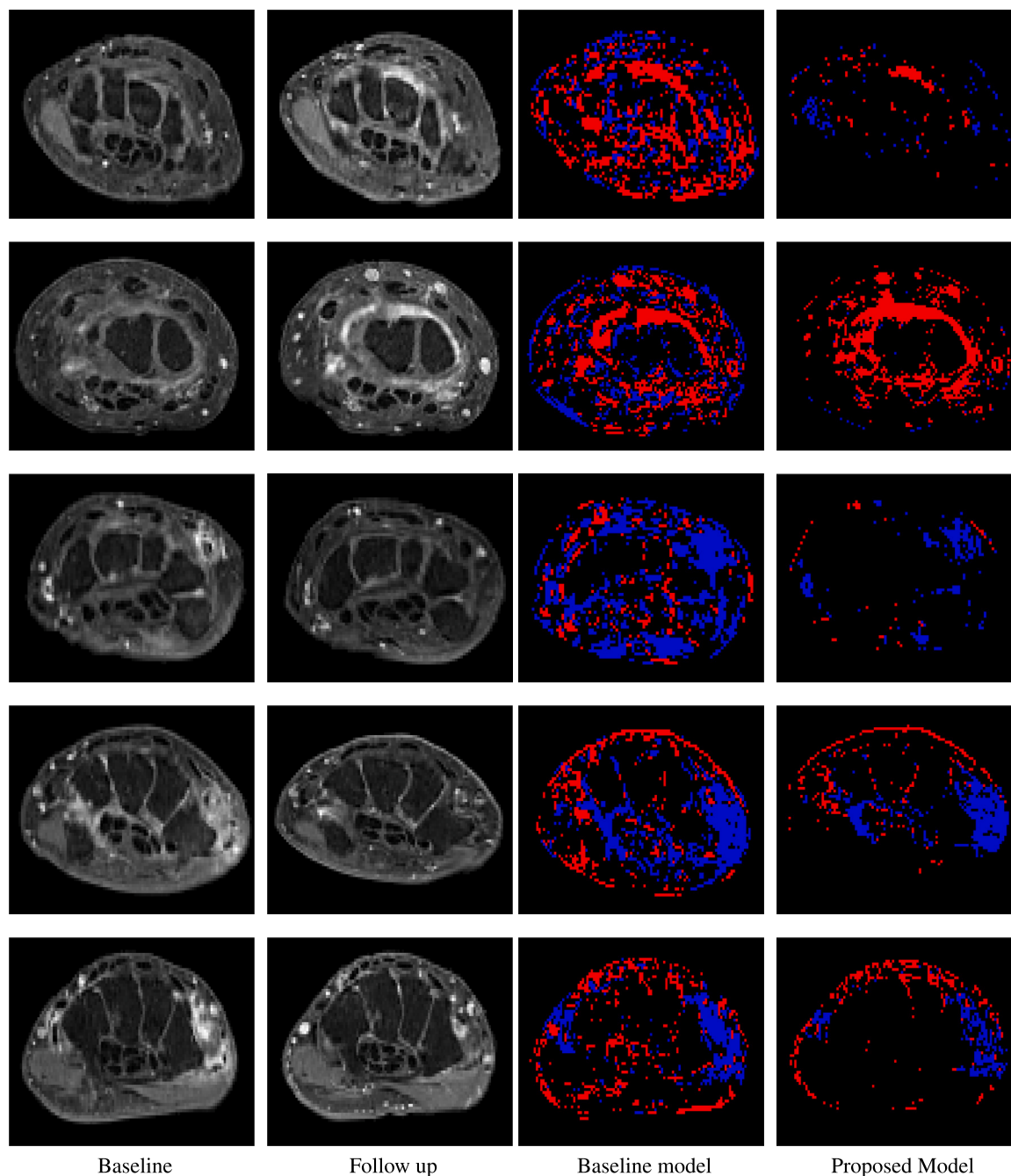


Fig. 5. Five sample of obtained change maps.

- True Positive (TP) [1, 10]: If there is a change between baseline and follow-up, and it is presented in the change map. According to the accuracy of the presented changes, the score can be between 1 and 10.
- True Negative (TN): There is no change between baseline and follow-up, and the change map also does not show any changes. This is indicated separately.

Since each SRR image consists of hundreds of slices and scoring all slices would therefore be a too time-consuming process, we decided to only score the central slice for each patient. As can be seen from section four in Fig. 6, there is a table that shows the anonymised study ID of the patient, central slice and saving status. Therefore, readers can choose the central slice and save this slice number for each patient, then score that specific slice and save it to file.

4.3.2.2. Observer study results. To evaluate the obtained change maps, we invited a radiologist and rheumatologist in training to analyse the results. One of the challenges to evaluate the change maps was the large number of slices. To do this, we tried to select a slice on a location where all features could be assessed properly. Therefore, slices with the radio-ulnar joint were selected as the central slice, since this is the best place to also score for tenosynovitis [22]. Firstly, we asked Reader 1 to select the central slices for each MRI scan and score that specific slice. The obtained scores from Reader 1 are shown in Table 4. Subsequently, Reader 2 scored the same slices blinded for the scores of Reader 1, and these scores are provided in Table 5.

According to Reader 1's point of view, there was a considerable amount of TN in the synovium, the changes in which were correctly recognised by our model, as shown in Table 4. Also, there were five True Positive cases (1, 9, 9, 10, 9), where in four of them our proposed

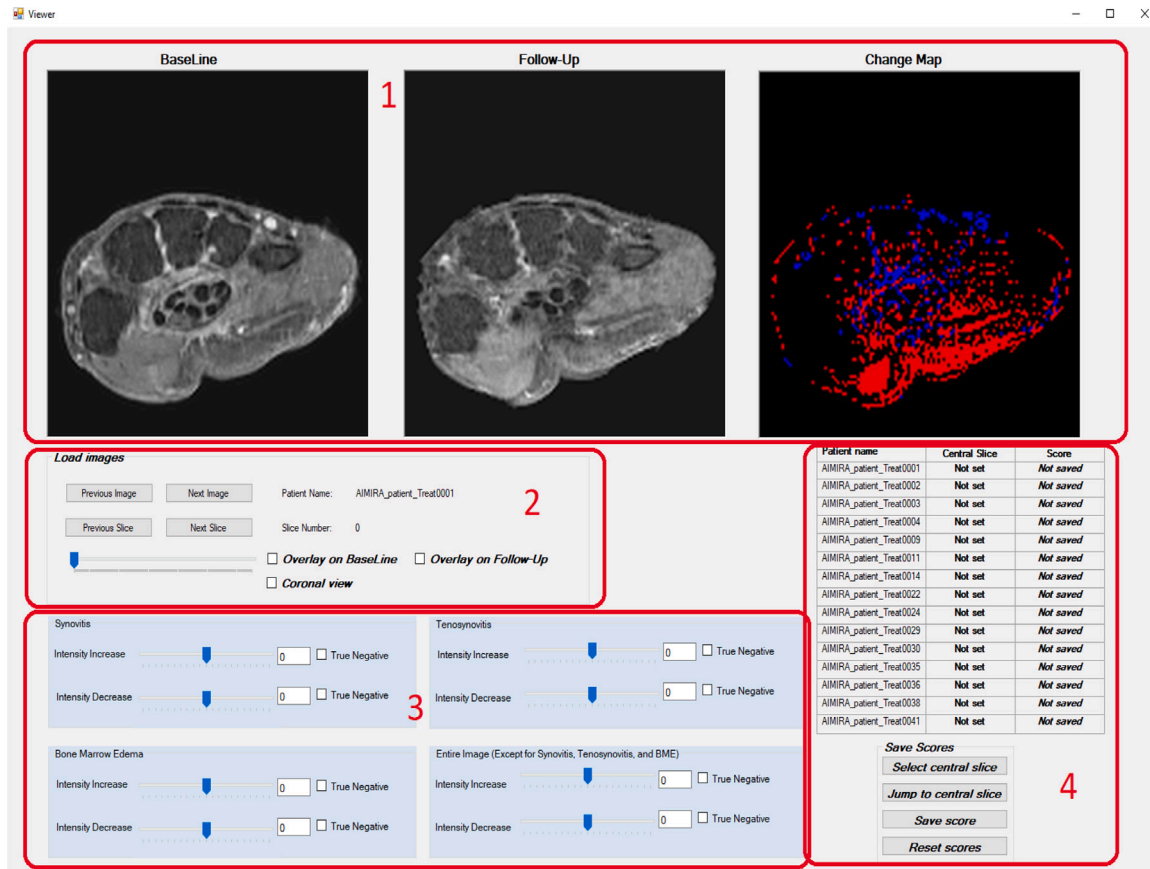


Fig. 6. The designed viewer for the observer study.

Table 4
Obtained scores from Reader 1.

Patient list	Synovium		Tenosynovium		Bone marrow		Other regions	
	Increase	Decrease	Increase	Decrease	Increase	Decrease	Increase	Decrease
01	TN	TN	TN	-3	TN	TN	TN	8
02	TN	1	-7	-2	-5	TN	-8	1
03	TN	TN	8	8	TN	TN	-2	7
04	TN	TN	9	TN	TN	TN	-1	TN
05	TN	TN	TN	-1	TN	TN	1	5
06	TN	TN	TN	9	TN	TN	-5	7
07	0	-1	-5	-3	TN	TN	6	-2
08	TN	9	TN	10	TN	TN	-1	9
09	9	TN	7	TN	TN	TN	5	-1
10	10	TN	-8	TN	0	0	3	-2
11	TN	-1	-5	TN	-1	TN	2	TN
12	-2	TN	-1	-1	0	TN	7	TN
13	TN	TN	7	10	TN	TN	7	TN
14	TN	TN	TN	TN	TN	TN	8	8
15	TN	9	0	8	TN	TN	8	9

model could find and represent them in the change maps with high accuracy. However in three cases (-1, -1, -2), the change maps showed minor changes, which were wrong, and in one case (0) the model could not recognise and present changes in the change map.

The ‘Tenosynovium’ column from Table 4 shows that there were more inflammatory changes in the tenosynovium than in the synovium, a part of which has been detected accurately and represented in the final change maps. However, we can see more negative scores in these areas, which indicate that there were more FPs in this anatomical area. The ‘Bone Marrow’ column (see Table 4) indicates that BME did not occur frequently, which is recognised correctly. However, we had three cases, in which the proposed model could not find any changes in bone marrow (false negatives). Finally, the ‘Other Regions’ column represents scores related to the remaining areas (entire wrist excluding

synovium, tenosynovium and bone marrow), such as connective tissue, vessels and skin.

A similar pattern of scoring can be seen in Table 5, which was obtained from the second reader. In most of the cases, the scores from Reader 1 and Reader 2 were comparable. However there was a limited number of cases, where a completely opposite score was given.

In addition, four example images along with their scores are provided in Table 6, showing that the obtained change maps for the first two examples achieved high scores from both Readers. However, the two readers disagreed in evaluating the third example change map. The fourth example in Table 6, showed that both readers agreed that the obtained change maps could not precisely reflect the actual changes.

To summarise and easily interpret the results, a modified version of the confusion matrix was used (see Tables 7 and 8). As explained

Table 5
Obtained scores from Reader 2.

Patient list	Synovium		Tenosynovium		Bone marrow		Other regions	
	Increase	Decrease	Increase	Decrease	Increase	Decrease	Increase	Decrease
01	TN	TN	TN	-5	TN	TN	10	10
02	TN	TN	-10	-10	-10	TN	9	5
03	TN	TN	10	10	TN	TN	10	10
04	TN	TN	10	TN	TN	TN	-8	4
05	TN	TN	TN	-5	TN	TN	5	-5
06	TN	TN	-10	10	0	TN	10	10
07	5	-1	-10	-5	TN	TN	5	-10
08	TN	TN	TN	10	TN	TN	0	TN
09	-1	-1	10	TN	TN	TN	10	TN
10	10	TN	2	TN	TN	TN	-5	-2
11	TN	TN	-5	TN	TN	TN	-5	TN
12	-5	TN	TN	TN	TN	TN	5	TN
13	2	TN	-1	10	TN	TN	TN	TN
14	TN	TN	TN	TN	TN	TN	TN	TN
15	TN	10	TN	TN	TN	TN	10	5

Table 6
Four different examples along with the obtained scores from both readers.

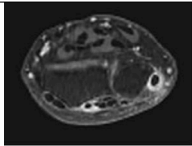
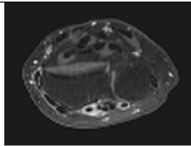
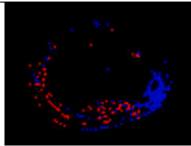
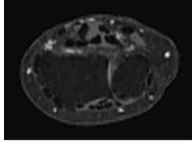
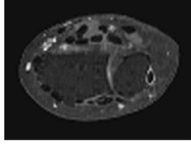
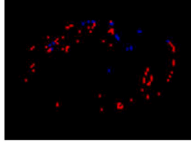
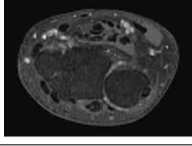
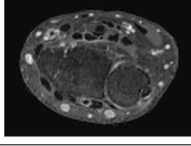
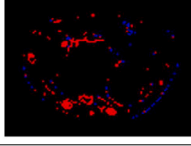
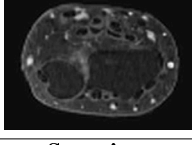
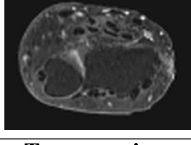
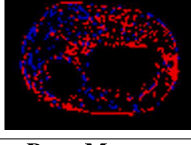
		Baseline	Follow-Up	Change map			
patient 03							
Scores	Region	Synovium		Tenosynovium		Bone Marrow	
	Intensity	Increase	Decrease	Increase	Decrease	Increase	Decrease
	Reader 1	TN	TN	8	8	TN	TN
	Reader 2	TN	TN	10	10	TN	TN
patient 04							
Scores	Region	Synovium		Tenosynovium		Bone Marrow	
	Intensity	Increase	Decrease	Increase	Decrease	Increase	Decrease
	Reader 1	TN	TN	9	TN	TN	TN
	Reader 2	TN	TN	10	TN	TN	TN
patient 09							
Scores	Region	Synovium		Tenosynovium		Bone Marrow	
	Intensity	Increase	Decrease	Increase	Decrease	Increase	Decrease
	Reader 1	9	TN	7	TN	TN	TN
	Reader 2	-1	-1	10	TN	TN	TN
patient 07							
Scores	Region	Synovium		Tenosynovium		Bone Marrow	
	Intensity	Increase	Decrease	Increase	Decrease	Increase	Decrease
	Reader 1	0	-1	-5	-3	TN	TN
	Reader 2	5	-1	-10	-5	TN	TN

Table 7

Confusion matrix of obtained scores associated with Reader 1. N: Negative, P: Positive, F: False, T: True.

Regions	Synovium				Tenosynovium				Bone marrow			
	F	N	P	Level	F	N	P	Level	F	N	P	Level
Increase	1	1	0.2		1	5	2.6		2	2	0.6	
	11	2	1.9		5	4	3.1		11	0	X	
Decrease	0	2	0.2		0	5	1		1	0	X	
	10	3	1.9		5	5	4.5		14	0	X	
Total	1	3	0.4		1	10	3.6		3	2	0.6	
	21	5	3.8		10	9	7.6		25	0	X	

Table 8

Confusion matrix of obtained scores associated with Reader 2. N: Negative, P: Positive, F: False, T: True.

Regions	Synovitis				Tenosynovium				Bone marrow			
	F	N	P	Level	F	N	P	Level	F	N	P	Level
Increase	0	2	0.6		0	5	3.6		1	1	1	
	10	3	1.7		6	4	3.2		13	0	X	
Decrease	0	2	0.2		0	4	2.5		0	0	X	
	12	1	1		7	4	4		15	0	X	
Total	0	4	0.8		0	9	6.1		1	1	1	
	22	4	2.7		13	8	7.2		28	0	X	

above, False Positives and True Positives have not only been detected but also the level of (dis-)agreement has been rated with a range of values. Therefore, we added an additional column named Level, to show the level of agreement (for True Positives) or the error level (for False Positives). To analyse the performance of our proposed model in detecting changes in forward and backward directions, we calculated the confusion matrix for Increase (where changes showed an increase in intensity, as indicated by red pixels), Decrease (where changes showed a decrease in intensity, indicated by blue pixels), and a combination of both types of changes.

To calculate the level of (dis-)agreement, for example, in the Increase category for the Synovium from Reader 1, we counted the number of FNs, FPs, TNs and TPs as follows, $FN = 1$, $FP = 1$ (score = -2), $TN = 11$, $TP = 2$ (scores 9 and 10). Since the maximum range for the absolute scores is 10, we weighted FP and TP as follows, $FP = 2/10 = 0.2$, $TP = 9/10 + 10/10 = 1.9$. Then we put the weighed scores in the Level column. As can be seen from Tables 7 and 8, there were no considerable differences between the Increase and Decrease categories, which indicates that our proposed model had almost similar performance to find changes in the forward and backward direction. The results show also that we had a small number of False Negatives (from Reader 1’s perspective: five cases; and from the second Reader’s perspective just one case). This means that in only a very limited number of cases the proposed model was unable to detect changes. There are cases where there was no change between baseline and follow-up images, but by mistake our proposed model did show changes (False Positives). However, the error level was low, which indicate that detected false changes were small. On the other hand, the level of agreement of True Positives was high, which means that not only our proposed model could detect relevant changes correctly, but also in most cases could accurately represent them in the change maps.

5. Discussion and conclusion

In this paper, a deep learning-based model is proposed to detect inflammatory changes in Rheumatoid Arthritis from MRI scans. The proposed model is a combination of a convolutional neural network for image reconstruction and classical image processing techniques.

To do this, a U-Net-based model was developed for image reconstruction, where the model was able to reconstruct the follow-up MRI from the baseline MRI. In the proposed model, we found for the first time changes in both forward and backward direction. In this case, we could find both inflammation remission, and new or progressing inflammation over time.

In the end, the Otsu thresholding technique was applied to the obtained difference map between the baseline and follow-up image to show non-trivial changes over time.

Our algorithm uses the strengths of a joint U-Net to map the high intensity changes, contrast defects and textural changes from one image to another. As shown in a subjective comparison versus the Baseline model, less artefacts and noise appeared in the final change maps, as compared to simple subtraction.

In our proposed model, the accuracy of the final change map depends on the quality of image reconstruction. In some cases, where the reconstructed image was very similar to the original image without unique changes, the output change map was more accurate with less noise and fewer artefacts. To develop a better reconstruction model, we tried various settings for the U-Net, and also different augmentation methods. Since the follow-up image reconstruction is more difficult as compared to usual DL image reconstruction, where the same image as the input needs to be reconstructed, we found that resizing the image to 128×128 could help to improve the quality of the image reconstruction.

To evaluate our proposed model, we used a subjective comparison versus the Baseline model, which uses the original SRR images to calculate change maps instead of using the reconstructed images. In this case, we noticed that our proposed model could successfully remove artefacts and noise from the final change maps. Also, we set up a reader study, where a radiologist and rheumatologist in training evaluated the final change maps. One slice was scored at the level of the Listers tubercle for each patients in the test set. We found that, synovitis was not frequently present in the Distal Radial Ulnar (DRU) region (on the selected slice), but in most of the cases if it appeared, it was detected and changes were represented correctly in the final change maps. BME was not frequently presented in the selected patients either, however the proposed model could successfully recognised them as True Negatives. Most importantly we could see minor intensity changes in the bone between baseline and follow-up images, however, our proposed model could learn these parts and they did not appeared in the final change map, which can be considered an advantage of our model. In addition, tenosynovitis was often well-recognised (both decreasing and increasing) by the model.

The other soft tissues (entire image except for synovitis, tenosynovitis, and BME) are also evaluated by the two readers. Therefore, there is a possibility to take a closer look at the changes in other regions (entire image excluding synovium, tenosynovium and bone marrow) to find possibly RA-related changes. For example, changes in vessels and/or skin might be related to RA. However, the readers believed that, when there are too many changes in this area, it makes scoring difficult and it can be distracting. However, changes related to vessels are recognised properly, but it remains to be seen if these changes are relevant to the development of RA. Sometimes many pixels with intensity changes undetectable by eye are presented in the final change maps. Also for these regions, further research is needed to evaluate whether these unexpected changes are relevant to RA development.

Because of the complexity of the scoring process, just one slice per patient was used for evaluation, Which may not properly reflect the overall quality of the change maps in some cases. Therefore, one of the concern is that the results are not shown as well as they should be. Therefore, the true results may be better than what the numbers have shown.

One of the drawbacks of our work is the lack of a proper ground truth. Ground truth change maps are, however, very time-consuming to produce. Therefore, it was not available for our large dataset, and we needed to use a subjective evaluation instead. Also, pixel-based

analysis can be prone to errors due to image noise. One of the solutions to address this problem can be region-based analysis. However, the wrist has a complicated anatomy with multiple anatomical sections, including bones, tendons, vessels, skin, and other tissue. Our preliminary experiment showed that we would then need to segment the wrist into 33 regions to conduct hypothesis-free experiments on the whole wrist. An accurate segmentation method for this purpose is not yet available

Furthermore, our proposed model finds all the changes between two MRI scans and a part of detected changes may be irrelevant to RA. For example, the changes could be because of fat suppression issues or wrist movement during MRI scanning. One of our plans is to label the changes into ‘relevant’ or ‘irrelevant’. In this case, we can train our model to learn just RA-related changes.

Overall, the obtained results were promising as a first step to detect inflammatory changes in rheumatoid arthritis using deep learning. In this work, we found changes in both forward and backward directions; therefore, we could distinguish between changes over time, detecting both progression and remission in different inflammatory regions.

CRedit authorship contribution statement

Tahereh Hassanzadeh: Conceptualization, Methodology, Software, Writing – original draft, Visualization, Data curation, Writing – review & editing. **Denis P. Shamonin:** Data curation. **Yanli Li:** Writing – review & editing. **Doortje I. Krijbolder:** Validation. **Monique Reijnierse:** Writing – review & editing, Validation, Funding acquisition. **Annette H.M. van der Helm-van Mil:** Writing – review & editing, Validation, Funding acquisition. **Berend C. Stoel:** Supervision, Project administration, Conceptualization, Writing – review & editing, Data curation, Funding acquisition.

Declaration of competing 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.

Data availability

The authors do not have permission to share data.

Acknowledgments

This project has been funded by the Dutch Research Council (NWO) Applied and Engineering Sciences (project number 17970). The TREAT-EARLIER trial has been funded by an NWO-ZonMW grant (project number 95104004). The Dutch Arthritis Society contributed financially to both grants.

References

- [1] Vikas Majithia, Stephen A. Geraci, Rheumatoid arthritis: diagnosis and management, *Amer. J. Med.* 120 (11) (2007) 936–939.
- [2] Mikkel Østergaard, Charles Peterfy, Philip Conaghan, Fiona McQueen, Paul Bird, Bo Ejbjerg, Ron Shnier, Philip O’Connor, Mette Klarlund, Paul Emery, et al., OMERACT rheumatoid arthritis magnetic resonance imaging studies. Core set of MRI acquisitions, joint pathology definitions, and the OMERACT RA-MRI scoring system, *J. Rheumatol.* 30 (6) (2003) 1385–1386.
- [3] Fan Xiao, James F Griffith, Andrea L Hilken, Jason CS Leung, Jiang Yue, Ryan KL Lee, David KW Yeung, Lai-Shan Tam, ERAMRS: a new MR scoring system for early rheumatoid arthritis of the wrist, *Eur. Radiol.* 29 (10) (2019) 5646–5654.
- [4] Evgeni Aizenberg, Denis P Shamonin, Monique Reijnierse, Annette HM van der Helm-van, Berend C Stoel, et al., Automatic quantification of tenosynovitis on MRI of the wrist in patients with early arthritis: a feasibility study, *Eur. Radiol.* 29 (8) (2019) 4477–4484.
- [5] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234–241.
- [6] Nobuyuki Otsu, A threshold selection method from gray-level histograms, *IEEE Trans. Syst. Man Cybern.* 9 (1) (1979) 62–66.
- [7] Doortje I Krijbolder, Marloes Verstappen, Bastiaan T van Dijk, Youssa J Dakkak, Leonie E Burgers, Aleid C Boer, Yune Jung Park, Marianne E de Witt-Luth, Karen Visser, Marc R Kok, et al., Intervention with methotrexate in patients with arthralgia at risk of rheumatoid arthritis to reduce the development of persistent arthritis and its disease burden (TREAT EARLIER): a randomised, double-blind, placebo-controlled, proof-of-concept trial, *Lancet* 400 (10348) (2022) 283–294.
- [8] Richard J Radke, Srinivas Andra, Omar Al-Kofahi, Badrinath Roysam, Image change detection algorithms: a systematic survey, *IEEE Trans. Image Process.* 14 (3) (2005) 294–307.
- [9] Julia Patriarche, Bradley Erickson, A review of the automated detection of change in serial imaging studies of the brain, *J. Digit. Imaging* 17 (3) (2004) 158–174.
- [10] Julia W. Patriarche, Bradley J. Erickson, Change detection & characterization: A new tool for imaging informatics and cancer research, *Cancer Inform.* 4 (2007) 117693510700400002.
- [11] Alexander Naitsat, Emil Saucan, Yehoshua Zeevi, A differential geometry approach for change detection in medical images, in: *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2017, pp. 85–88.
- [12] Julia Willamena Patriarche, Bradley James Erickson, Part 1. Automated change detection and characterization in serial MR studies of brain-tumor patients, *J. Digit. Imaging* 20 (3) (2007) 203–222.
- [13] Hae Jong Seo, Peyman Milanfar, A non-parametric approach to automatic change detection in MRI images of the brain, in: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, IEEE, 2009, pp. 245–248.
- [14] Varvara Nika, P. Babyn, Zhu, Change detection of medical images for three dimensional volumetric data, *J. Theor. Comput. Sci.* 2.
- [15] Guillaume Dupont, Ekaterina Kalinicheva, Jérémie Sublime, Florence Rossant, Michel Pâques, Analyzing age-related macular degeneration progression in patients with geographic atrophy using joint autoencoders for unsupervised change detection, *J. Imaging* 6 (7) (2020) 57.
- [16] Ekaterina Kalinicheva, Jérémie Sublime, Maria Trocan, Change detection in satellite images using reconstruction errors of joint autoencoders, in: *International Conference on Artificial Neural Networks*, Springer, 2019, pp. 637–648.
- [17] Dirk H.J. Poot, Vincent Van Meir, Jan Sijbers, General and efficient super-resolution method for multi-slice MRI, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2010, pp. 615–622.
- [18] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, Josien PW Pluim, Elastix: a toolbox for intensity-based medical image registration, *IEEE Trans. Med. Imaging* 29 (1) (2009) 196–205.
- [19] Denis P Shamonin, Esther E Bron, Boudewijn PF Lelieveldt, Marion Smits, Stefan Klein, Marius Staring, Fast parallel image registration on CPU and GPU for diagnostic classification of alzheimer’s disease, *Front. Neuroinform.* 7 (2014) 50.
- [20] Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, Klaus H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, *Nature Methods* 18 (1548–7105) (2021) 203–211.
- [21] Francois Chollet, et al., Keras, 2015.
- [22] Espen A Haavardsholm, Mikkel Østergaard, Bo J Ejbjerg, Nils P Kvan, Tore K Kvien, Introduction of a novel magnetic resonance imaging tenosynovitis score for rheumatoid arthritis: reliability in a multireader longitudinal study, *Ann. Rheum. Dis.* 66 (9) (2007) 1216–1220.