Deep Learning-Enabled Intensity-based Segmentation (DEIS) of Inflammatory Lesions in Spondyloarthritis

Carolyna Hepburn¹, Hui Zhang¹, Juan Eugenio Iglesias²,³,⁴, Alexis Jones⁵, Alan Bainbridge⁶, Margaret A. Hall-Craggs⁷, and Timothy J.P. Bray⁷

¹Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK
²Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, London, UK
³Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, USA
⁴Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Boston, USA
⁵Department of Rheumatology, University College London Hospitals NHS Foundation Trust, London, UK
⁶Department of Medical Physics and Biomedical Engineering, University College London Hospitals NHS Foundation Trust, London, UK
⁷Centre for Medical Imaging, University College London, London, UK

Abstract

Short tau inversion recovery (STIR) MRI is routinely used in clinical practice to detect inflammation in patients with axial spondyloarthritis (axSpA). Several semi-quantitative scoring systems are used in clinical research. However, identification and quantification of inflammatory load is difficult and suffers from inter-reader variability. To assist radiological assessment, we propose a semi-automated workflow for segmentation of inflammatory lesions on STIR MRI of sacroiliac joints in adult patients with axSpA. This workflow mimics the human approach: it identifies regions, where inflammation typically occurs and designates voxels as inflammatory based on signal hyper-intensity. Specifically, the procedure consists of (1) segmentation of potentially inflamed subchondral bone and non-inflamed interforaminal sacral bone, (2) intensity threshold-based automatic segmentation of inflammatory lesions, i.e., voxels in the subchondral bone with outlying intensity with respect to the interforaminal sacral bone, and (3) human intervention to remove erroneously segmented areas based on knowledge of lesion morphology and location. We evaluated inter-reader agreement for the semi-automated approach using the Dice coefficient and compared it against baseline human performance (purely manual segmentation of inflammatory lesions). Inter-reader median Dice rose from 0.53 - 0.56 using the manual approach to 0.84 using the semi-automated segmentation, representing a 28-31% improvement in inter-reader agreement.

To reduce the need for human input, we investigated whether fully-automated segmentation of subchondral bone using deep learning was feasible. The model averaging ensemble showed an average Dice of 0.94 (the data, code and models are available at https://github.com/c-hepburn/Bone_MRI, indicating that automation of the potentially-inflamed region is achievable and providing a route towards wider implementation. The advantage of the proposed workflow is that it removes the need for intensity-based judgements to be made by the observer, eliminating a key source of subjectivity and variability in scan assessments and also increasing speed. The approach provides a path towards an interactive deep-learning enabled tool to assist radiological assessment of inflammatory load in axSpA.

1 Introduction

Axial spondyloarthritis (axSpA) is a rheumatic disease that primarily affects axial joints of the spine and causes pain, stiffness, and disability [1]. A hallmark of early disease is inflammation in the subchondral bone marrow of sacroiliac joints (SIJs), referred to as bone marrow edema (BME) [2]. In current clinical practice, inflammation is typically detected with short tau inversion recovery (STIR) magnetic resonance imaging (MRI) [3]. Accurate and precise identification and quantification of inflammation is necessary for therapeutic decision making to facilitate effective treatment, control pain and prevent disability.

At present, there are several methods for assessing inflammation on STIR MRI and these are based on visual assessment. For example, for research purposes several semi-quantitative scoring methods of BME have been proposed [4], [5], [6]. However, these methods only evaluate a limited number of slices, require the observer to make binary decisions about the presence or absence of inflammation (thereby losing information about lesion location and extent) and are laborious. Furthermore, visual assessment is a predominantly intensity-based approach: "the stronger the hyperintense signal the more likely it reflects active inflam-
However, this approach dictates that there is uncertainty around the presence of disease in areas of intermediate signal, making it difficult to define subtle lesions and lesion boundaries. In turn, this impacts on whether lesions meet the minimal areas to be determined as abnormal and causes variability in STIR MRI interpretation. In clinical practice, the extent of inflammation is described verbally with no numerical metric to enable quantification, but the same problems with interpreting signal abnormalities apply and can impact directly on the patient’s diagnosis and management.

One approach to address these problems in inflammation assessment is to automatically segment inflammatory lesions. In particular, the use of thresholding as a segmentation method can assist the observer with intensity-based judgments, reducing subjectivity in image interpretation. A benefit of segmentation is that lesion location is specified and inflammatory load can be estimated. Recently proposed semi-automated methods for detecting and quantifying only BME also addressed these issues. However, these approaches require the user either to identify lesion location - with the potential that some lesions might be missed - or to manually segment bone, which is time consuming and thus impractical in clinical practice. The drawback of manual segmentation of specific areas of bone was addressed in a recent study, which also aimed to demonstrate computational feasibility of a fully automated segmentation of BME.

Ultimately, a tool that can automatically detect and quantify inflammation in a reproducible, fast manner is desirable. With advances in artificial intelligence (AI), it is natural to consider supervised deep learning to automate segmentation. However, training an algorithm directly on manual segmentations poses a problem because manual segmentation is highly variable, meaning that a robust reference standard is difficult to obtain. Therefore, the goal of this study was to develop a semi-automated workflow, abbreviated as DEIS (deep learning-enabled intensity-based segmentation), inspired by the human approach, to improve the precision and objectivity of radiological assessment.

## 2 Materials and Methods

### 2.1 Available Data

Data were taken from a completed prospective longitudinal study (16 females, 14 males; mean age 42.7 years). The study was conducted between April 2018 and July 2019 with the aim of evaluating responsiveness and response prediction using quantitative imaging biomarkers and performed with institutional review board approval (REC reference 15/LO/1475). Patients diagnosed with axSpA according to 2009 ASAS criteria underwent pre- and post-biologic therapy MRI scans of SIJs (n=30). Patients were excluded if they had a contraindication to MRI scanning. The protocol consisted of STIR and T1-weighted (T1W) turbo spin echo sequences, acquired in an oblique coronal plane with the same FOV on a 3T Philips Ingenia scanner (Philips Healthcare, Best, Netherlands). MRI parameters are detailed below.

- STIR: repetition time (TR) 5316ms, echo time (TE) 60ms, inversion time 210ms, echo train length (ETL) 21, flip angle 90°, slice thickness 3mm, spacing between slices 3.3mm, pixel spacing 0.59 × 0.59 (mm), image matrix 336 × 336, number of slices 23-25;

- T1-weighted TSE: TR 1012ms, TE 8ms, ETL 3, flip angle 90°, slice thickness 3mm, spacing between slices 3.3mm, pixel spacing 0.59 × 0.59 (mm), image matrix 336 × 336, number of slices 23-25;

### 2.2 Data Selection

13 subjects with pre- and post-biological treatment scans were selected from the available pool of MRI. Images were assessed visually and scans were selected by an experienced radiologist to represent a spectrum of inflammatory lesion loads and locations in

- 1. subchondral bone marrow abutting the sacroiliac joints;
- 2. the joint space;
- 3. the adjacent lumbar vertebra (L5, mainly seen in the endplates and facet joints);

Bone above the L4/5 disc was not considered as it was not consistently included in the field of view; 5 scans were used during development of the workflow procedures and manual segmentation guidelines, 8 scans were used for the actual study (fig. 1 fig. 2).

### 2.3 Overview of AI-enabled workflow

To segment inflammatory lesions a human observer first identifies potentially inflamed regions and then performs judgment based on signal hyper-intensity as compared to a reference signal. Analogously, the proposed approach identifies sites of potential inflammation and region of normal bone, from which intensity thresholds are determined, and designates voxels within the sites as inflamed whenever their intensity is above the determined thresholds. The procedure comprises the following steps (fig. 1 a detailed description is given in the section 2.3.2 section 2.3.5):

(i) **Reference region segmentation**
Bone in the interforaminal region of the sacrum is segmented. This is typically spared from inflammation, i.e., this is normal bone (section 2.3.2).

(ii) **Estimation of STIR intensity thresholds**
Upper and lower thresholds are determined from the upper end of the normal bone intensity distribution, enabling separation of normal from inflamed marrow (section 2.3.3).

(iii) **Disease region segmentation**
The disease region includes all potential sites of in-
flammation, i.e., all bone in the imaged pelvis and SIJs (section 2.3.4).

(iv) **Thresholding within the disease region**
Voxels in the disease region are assigned labels of 0 if voxel intensity is below the lower intensity limit, 1 if between lower and upper intensity limits and 2 if equal or above the upper limit.

(v) **Automatic removal of very small regions**
I.e., regions, containing <4 pixels (=1.39mm²), on each slice, which commonly arise due to noise and small vessels crossing the surface of, or within, the bone.

(vi) **Manual correction of the final segmentation of inflammatory lesions by a human observer**
Correction is based on morphology and anatomical location (section 2.3.5).

Step (iii) was consequently automated via supervised deep learning to ensure the practicality of the proposed pipeline (section 2.4).

---

Figure 1: Steps of the proposed workflow (left) and their schematic representation, i.e., image slice with a superimposed mask (right).
2.3.1 Manual Segmentation and Annotators

Manual segmentation of inflammatory lesions – to assess baseline human performance – was performed by a consultant radiologist and a musculoskeletal radiology fellow with over 25 and 5 years of musculoskeletal MRI experience, respectively. A trained non-radiologist reader assisted with segmentation of disease and reference regions: these preliminary segmentations were verified and corrected by the two radiologists. Information about all manual segmentations performed for this study can be found in the supplemental material (fig. S11). The fig. 2 describes in detail the data flow and how it was used in the study. We used ITK SNAP software (version 3.8.0) for all segmentations [13], [www.itksnap.org](http://www.itksnap.org).

![Diagram of data flow and use in the study.](image)

Figure 2: Detailed data flow and use in the study.
2.3.2 Reference Region Segmentation

The reference region was segmented manually on STIR images to ensure that entirely normal regions of marrow were selected whilst artefacts and blood vessels were avoided. The interforaminal sacral bone region (fig. 3) was segmented on multiple slices and included the largest possible area of normal bone to minimise sampling error and more accurately represent its intensity distribution.

2.3.3 Estimation of Intensity Thresholds from Reference Region

Intensities corresponding to outliers and rare observations of the normal bone intensity distribution are attributed either to noise or increased water content due to natural variation in tissue composition. Therefore, voxels at or above these intensity levels are designated abnormal. Specifically, an upper limit, \( L_{\text{upper}} \) was defined as the maximum intensity, \( I_{\text{max}} \) of the distribution. A lower limit, \( L_{\text{lower}} \) was computed as the sum of the upper quartile, \( Q_{\text{U}} \) and a multiple, \( n \) of the inter-quartile range, \( IQR \) of the distribution: \( L_{\text{lower}} = Q_{\text{U}} + n \cdot IQR \). The multiple was determined automatically as the empirical lower limit value of \( L_{\text{lower}} = Q_{\text{U}} + 1.5 \cdot IQR \) resulted in obvious over-segmentation in certain cases (fig. S12, fig. S13). Starting with the value of 1.5, the multiple \( n \) was incremented by 0.05 until the difference between upper and lower limits was less than the half of the interquartile range, \( 0 < L_{\text{upper}} - L_{\text{lower}} < IQR/2 \). If the condition was initially satisfied, no incrementation was performed.

2.3.4 Disease Region Segmentation

The disease region was segmented manually on T1W images and included all potential sites of inflammation: all bone in the imaged pelvis, the sacroiliac and facet joint spaces. The neural foramina, spinal fluid and intervertebral discs were excluded (fig. 3, fig. S14). In cases when identification of the border between vertebra and discs was difficult, STIR images were consulted and the segmentation was adjusted accordingly. The T1W and STIR images were acquired in the same plane with the same FOV, enabling usage of disease region mask for consequent segmentation on STIR image.

Figure 3: Top: STIR MRI slice in oblique coronal plane (left) with super-imposed manual segmentation of reference (interforaminal sacral bone) region (middle) and disease region (right). Bottom: T1W MRI slice in oblique coronal plane (left), used to manually segment the disease region, with the corresponding segmentation (right).
2.3.5 Correction of the Final Segmentation of Inflammatory Lesions by a Human Observer

Regions which were deemed non-inflammatory – for example due to the presence of vessels or artefact – were removed by readers based on morphology and anatomical location. Lesions were either left in place or removed altogether, i.e., the boundaries of lesions were not modified, except when the posterior part of the joint or foramen were segmented along with a potential lesion. Lesions located above the L4/5 disc were removed in consistency with manual segmentations. T1W images were used to assist readers in identification of anatomical structures and regions of increased fat content. The two readers discussed and agreed upon the procedure using automatically segmented images from two subjects (fig. 2).

2.4 Feasibility of Automating Segmentation of Disease Region via Supervised Deep Learning

To reduce the need for human input within the workflow, we investigated the possible performance of a supervised deep learning algorithm for automatic segmentation of the disease region.

2.4.1 Reference Standard and Data Partitions

248 T1W image slices from 10 subjects were used for model training with four-fold cross validation and testing (fig. 2). The reference standard for algorithm training is described in the section 2.3.4. Data was partitioned at subject level at random into two sets: 200 image slices (8 subjects) for training with four-fold cross validation [14] to find the optimal set of hyper-parameters (section 2.4.3) and 48 image slices (2 subjects) for testing. Data partition into subsets for training (150 image slices, 6 subjects) and folds for validation (50 image slices, 2 subjects) was performed at subject level to avoid correlation between image slices, used to train and validate a model (fig. 2).

2.4.2 Data Pre-processing and Augmentation

Images were normalized by three standard deviation of image intensity distribution, allowing the same intensity scale between subjects and consistency in intensity levels of voxels representing the same tissue across the whole image volume for each subject (fig. S15). Each pre-processed image and corresponding mask slices undergone elastic deformation (https://github.com/gvtulder/elasticdeform/tree/v0.4.9), affine transformation (rotation, scaling, shearing) and random flip with 0.5 probability, i.e., the order of image array elements along the left-right axis was reversed (fig. S16). To make the network robust against between-subject variations in the intensity level of bone voxels, intensities were raised to a random power. All transformation parameters (rotation angle, scaling and shearing factors, power) were randomly sampled from uniform distributions of pre-defined ranges (table 1).

| Range       | Rotation  | Scaling factor | Shearing factor | Power     |
|-------------|-----------|----------------|-----------------|-----------|
| (-5,5)°     | (0.6, 1.2)| (-0.2, 0.2)    | (0.9, 1.2)      | (0.6, 1.2) |

Table 1: Ranges for uniform distributions of different transformation parameters.

2.4.3 Model Training

A convolutional neural network with two-dimensional U-Net architecture [15] (fig. S17) was trained on mini batches by optimizing binary cross entropy loss [16] using the Adam optimizer [17]. The architecture included batch normalization to keep the distribution of convolution layers outputs fixed, allowing faster convergence [18]. The network was trained with pre-processed data augmented on the fly, which allowed a substantial increase of the diversity in the training samples. At each training epoch 350 augmentation steps were performed [19]. At each step, a random batch was selected from the available pool, augmented, and fed into the model. Data shuffling ensured that the same batch contained different image slices every epoch. The following pseudo-code captures the training process:

Algorithm 1 Pseudo-code

1. for k in range (epochs) do
2. for i in range (augmentation steps) do
3. select a mini batch from training data at random
4. perform mini batch augmentation
5. prediction = model(augmented mini batch)
6. loss = loss function(prediction, augmented target batch)
7. compute loss function gradients
8. update model parameters
Optimal hyper-parameters were identified through training with four-fold cross validation, specifically, (i) the number of epochs (60, 100), (ii) the number of resolution levels (2, 4, 6), where a level represents all feature maps between two max-pooling or two up-sampling operations (fig. S17) and (iii) convolution kernel size (3 × 3, 5 × 5). The batch size (four) and learning rate (0.001) were kept constant. The model parameters were initialized using Pytorch default initialization scheme. A publicly available implementation of the U-Net in Pytorch was used (https://github.com/jvanvugt/pytorch-unet).

2.4.4 Model Evaluation

The section 2.5 describes evaluation metrics used in this study.

To reduce individual models’ errors, the optimal model (optimal set of hyper-parameters) was re-trained three times and performance of the model averaging ensemble was evaluated on the test data – average prediction from three models was computed, then rounded (fig. 2). Finally, to enable a further, larger-scale performance assessment (since the number of slices available for the former assessment was limited by the time taken for manual segmentation) the optimal model was re-trained three times using 248 image slices (10 subjects), for which manual segmentations were available, and the final model averaging ensemble was evaluated, based on visual assessment (satisfactory vs non-satisfactory), on the remaining 1000 image slices (20 subjects, 40 scans), for which manual segmentation was not available (fig. 2). Note that in this case each subject is seen by the model averaging ensemble twice due to availability of pre- and post-treatment scans for each subject. The visual assessment was performed by a postdoctoral researcher with two years of MRI experience who had received training in interpretation of the relevant anatomy by a radiologist, who also reviewed a subset of the cases to ensure that these anatomical assessments were accurate.

2.5 Evaluation Metrics

The similarity of a pair of binary segmentations was evaluated with the Dice coefficient, defined for the class of interest (abnormal or background) as the ratio of number of pixels (voxels) having identical location in both segmentations to the average of number of pixels (voxels) in each segmentation [20]:

\[
\text{Dice} = \frac{2|S_1 \cap S_2|}{|S_1| + |S_2|} \tag{1}
\]

\(S_{i, i \in \{1, 2\}}\) represents a point set, containing pixel (voxel) coordinates, subscript refers to the first or second segmentation and \(S_1 \cap S_2\) is the intersection of the sets. We refer to the Dice coefficient as the area or volume overlap, depending on whether two segmented areas or volumes were compared.

To evaluate model performance during training, a soft, differentiable generalization of Dice was used, implemented as [21]:

\[
\text{Dice} = \frac{2 \sum_i r_i p_i}{\sum_i r_i^2 + \sum_i p_i^2} \tag{2}
\]

where summation runs over pixels of reference standard, \(r_i \in R\) and network probability map, \(p_i \in P\).

The similarity between volumes derived from two binary segmentations is defined in terms of absolute volume difference divided by the sum of the compared volumes [20]:

\[
\text{VS} = 1 - \frac{|S_1| - |S_2|}{|S_1| + |S_2|} \tag{3}
\]
3 Results

3.1 Baseline human performance

In order to establish a baseline for the evaluation of the semi-automated algorithm, the human performance in manual segmentation of inflammatory lesions from two readers was first evaluated. The fig. 4 shows the 3D rendering of example manual segmentation trials of inflammatory lesions from both readers and provides a qualitative demonstration of differences in location, fragmentation and volume of segmented lesions between trials. Readers’ self-consistency and agreement in terms of volume overlap are shown in fig. 6a, b. Overlap values ranged from 0.28 to 0.87. Intra- and inter-reader median overlap values were 0.63 and 0.69 for reader 1 and 2 and in the range 0.53-0.56, respectively. There were occasions when readers agreed on overall inflammatory load but disagreed on lesion location (fig. 6d, fig. S18).

3.2 Improvement in agreement for semi-automated method

The fig. 5 shows corrected automatic segmentations by the two readers and the sum of manual segmentation trials for the same subjects (examples of automatic segmentation via thresholding with no correction by readers are presented in the fig. S19). The between-reader agreement for the cleaned segmentations (section 2.3.5) was evaluated in terms volume overlap. Note that for the evaluation we considered segmentation at lower threshold only. Median value was 0.84, which represents an increase of 28-31% compared to pure manual segmentation (fig. 6c). There was one outlier where the agreement was reduced; review of the images indicated that the disagreement mostly related to the presence of inflammation in the joint space, where blood vessels can be misinterpreted (fig. 7). Inflammation assessment using the semi-automated pipeline was quicker than with purely manual segmentation by 13.34 and 3.12 minutes on average for readers 1 and 2 respectively (fig. S20).
Figure 5: STIR MRI slices in an oblique coronal plane for the same subjects as in Figure 4 (top to bottom). Corrected automatic segmentations by the two readers are shown in the column 2 and 3. Pixels with intensity between the lower and upper thresholds are marked in yellow and equal or above the upper threshold are marked in red. The sum of manual segmentation trials (column 4) indicates pixels that were labelled by readers as abnormal once (dark blue), twice (light blue), trice (yellow) and four times (red).
Figure 6: Volume overlap evaluated from comparison of different manual segmentation trials within reader (a), between readers (b) and for corrected automatic segmentations (c). \( R_{ij} \) stands for reader with the first subscript corresponding to the reader and second - to manual segmentation trial, when applicable. The scatter plot of overlap vs volume similarity from comparison of some trials from both readers is shown in (d). The vertical line corresponds to the similarity value when one volume, \( V \), is twice as large as the other, \( V^* \).
Figure 7: STIR image slices in oblique coronal plane for one subject with super-imposed automatic segmentation, corrected by the two readers. Figure illustrates disagreement between readers on the potential inflammation in / around the joint space of SIJs.
3.3 Feasibility of Automatic Segmentation of Disease Region

The optimal model was trained until the average area overlap (section 2.5), evaluated on training data, reached a plateau (fig. 8a). The model did not overfit; however, there were fluctuations in its performance on the fourth validation fold (fig. 8b). This may be attributed to the fact that one scan was acquired with different angulation with respect to the other scans.

To reduce individual model errors the optimal model was re-trained three times using all training data (200 T1W image slices). From each run a model yielding the highest average area overlap was chosen: these models averaging ensemble yielded 0.94 average area overlap on the test data (48 T1W image slices), ranging from 0.85 to 0.98 (fig. 8c). Examples of automatically segmented disease region and corresponding reference standard are shown in the fig. 9.

Next, for a larger-scale performance assessment the optimal model was re-trained three times using 248 image slices (10 subjects) and the final model averaging ensemble was evaluated based on visual assessment of segmented image slices (satisfactory vs non-satisfactory), using the remaining 20 subjects, 40 scans (section 2.4.4, fig. 2). Model performance was either perfect or subject to minor corrections for 17 subjects (fig. 10). Model failed or its performance was worse for three subjects with very abnormal bone, i.e., very high fat content or strong sclerosis; such features were absent in the training data (fig. S21). Consequently, the subjects with very abnormal bone were added to the training data, the optimal model was retrained three times for public access and re-evaluated based on visual assessment. Note that in this case we added a diversity to the training data. Optimal hyper-parameters and number of model parameters updates remained fixed: such model flexibility and training procedure did not show overfitting.

Figure 8: Mean area overlap vs training epoch for different training data subsets (a) and validation folds (b). Each point represents area overlap averaged over (i) classes (foreground & background), (ii) samples in a mini batch and (iii) 350 augmentation steps. Area overlap from pair-wise comparison of reference standard and rounded prediction on the test data from models averaging ensemble (three runs using all training data, 200 T1W image slices) (c).
Figure 9: T1W MRI slices in oblique coronal plane for two 'test' subjects (top to bottom). Reference standard (middle) and models averaging ensemble (rounded) prediction of disease region (right).
Figure 10: T1W image slices in oblique coronal plane (top) for four subjects with super-imposed models averaging ensemble rounded prediction (bottom), for which the reference standard was not available.
4 Discussion

To improve the precision and objectivity of radiological assessment of inflammation in axSpA we propose a semi-automated procedure that mimics human approach: it identifies regions, where inflammation typically occurs and designates voxels as inflammatory based on signal hyperintensity. Therefore, the procedure consists of the following steps (1) segmentation of subchondral (potentially inflamed) bone and interforaminal sacral bone (non-inflamed), (2) automatic segmentation of inflammatory lesions within the subchondral bone via two thresholds, determined from the intensity distribution of ‘interforaminal bone’ voxels, and (3) human intervention to correct for erroneously segmented areas. We showed that this procedure produced a 28-31% improvement (difference in median dice scores) in segmentation consistency compared to manual segmentation and also produced time savings. Furthermore, automated segmentation of subchondral bone region via supervised deep learning (an average Dice of 0.94) showed that it is possible to substantially reduce the need for human input.

Several previous studies have also investigated the use of threshold-based methods for quantifying inflammation [10], [22]. However, these studies relied on manual segmentation to identify an optimal threshold, whereas our data show that using manual segmentation as a “gold standard” is problematic and may lead to inconsistent interpretation especially in cases when inflammation is subtle or precise lesion boundary cannot be identified. To highlight this point, a recent study [11] revised the threshold value developed in earlier work [10], finding an optimal threshold of 1 compared to 1.5 in the prior study. Clearly, a threshold which depends on reference standard provided by human observers is not desirable. In contrast, the approach proposed in this work removes the need for intensity-based judgements to be made by the observer. The use of an intensity-based threshold derived from normal marrow means that the choice of voxels is primarily influenced by the physical properties of the tissue, specifically, the extent to which the intensity in each voxel deviates from the intensity observed in normal marrow. The normal bone region effectively serves as a reference region and means that the judgment around which voxels are hyperintense is tailored to each individual and each scan of the individual.

In general, accurate detection and quantification of inflammatory load on STIR MRI is a difficult problem, partly because there is no defined intensity threshold between normal and abnormal water content resulting from increase in cellularity and/or expansion of the extracellular space [23], [24]. Furthermore, several factors influence the way we interpret the signal: small increases in water content indistinguishable from natural variations in bone marrow composition, confounding factors such as sclerosis and fat metaplasia masking abnormal signal, quality of fat suppression, sensitivity of the receiving coil, noise etc. These give rise to ambiguity in definition of any type of inflammatory lesion visible on STIR MRI: "The stronger the hyperintense signal the more likely it reflects active inflammation (intensity of the hyperintense signal is similar to that of blood vessels or spinal fluid)" [7]. "At present it is not possible to give a more precise definition of the minimum size (area) of the BMO which is necessary for it to be described as 'positive' " [2]. The reader’s image interpretation relies on a subjective perception of brightness, which is at the origin of variability either in manual segmentation (section 3.1) or scoring [25]. The use of quantitative imaging methods may help to improve the consistency of segmentation [26].

Limitations

One of the limitations of the proposed approach is the fixed value of the lower threshold. In practice, the observer could adjust the lower intensity limit. This, however, would introduce further variability, that needs to be quantified. On the other hand, retaining a degree of user control of the algorithm can be beneficial for achieving successful implementation in the clinic [27]. Secondly, although the approach mitigates the variability, it does not eliminate it, which is a consequence of a limitation of qualitative imaging modality itself (as discussed above). Finally, the primary focus of this study was development of the described workflow, and the dataset used here is relatively small. Future work could focus on larger-scale clinical validation and achieving greater automation of the method. In particular, training and validation in larger datasets may help to segment interforaminal and subchondral bone across a broad range of scanners. Development of an interactive prototype software tool could help to streamline the process of image interpretation for observers using this approach.

Conclusion

We have proposed a workflow that improves the objectivity and consistency of radiological assessment of inflammation on qualitative MRI of sacroiliac joints. This could facilitate more precise and consistent disease assessment and in turn support diagnosis and disease monitoring for patients with spondyloarthritis.
References

[1] Sieper J, Braun J, Dougados M, Baeten D. Axial spondyloarthritis. Nat Rev Dis Prim. 2015;1(7):1-17;

[2] Rudwaleit M, Jurik AG, Hermann KGA, Landewé R, Van Der Heijde D, Baraliakos X, et al. Defining active sacroilitis on magnetic resonance imaging (MRI) for classification of axial spondyloarthritis: A consensual approach by the ASAS/OMERACT MRI group. Ann Rheum Dis. 2009;68(10):1520–7;

[3] Maksymowych WP. The role of imaging in the diagnosis and management of axial spondyloarthritis. Nat Rev Rheumatol. 2019;15(11):657–72;

[4] Maksymowych WP, Inman RD, Salonen D, Dhillon SS, Williams M, Stone M, et al. Spondyloarthritis Research Consortium of Canada magnetic resonance imaging index for assessment of sacroiliac joint inflammation in ankylosing spondylitis. Arthritis Care Res. 2005;53(5):703–9;

[5] Maksymowych WP, Dhillon SS, Chiowchanwisawakit P, Pedersen SJ, Martinez B, Østergaard M et al. Development and Validation of Web-based Training Modules for Systematic Evaluation of Active Inflammatory Lesions in the Spine and Sacroiliac Joints in Spondyloarthritis. J Rheumatol. 2009;84:48-57;

[6] Navallas M, Ares J, Beltrán B, Lisbona MP, Maymó J, Solano A. Sacroilitis associated with axial spondyloarthropathy: New concepts and latest trends. Radiographics. 2013;33(4):933–56;

[7] Sieper J, Rudwaleit M, Baraliakos X, Brandt J, Braun J, Burgos-Vargas R, et al. The Assessment of SpondyloArthritis international Society (ASAS) handbook: A guide to assess spondyloarthritis. Ann Rheum Dis. 2009;68:i1–44;

[8] Mayerhoefer ME, Breitenseher M, Hofmann S, Aigner N, Meizer R, Siedentop H, et al. Computer-assisted quantitative analysis of bone marrow edema of the knee: Initial experience with a new method. Am J Roentgenol. 2004;182(6):1399–403;

[9] Zarco P, Almodóvar R, Bueno Á, Molinero LM, Moreno M, Juanola X, et al. Development and validation of SCAIIS, a tool for semi-automated quantification of sacroilitis by magnetic resonance in spondyloarthritis. Rheumatol Int. 2018;38(10):1919–26;

[10] Kucybała I, Tabor Z, Polak J, Urbanik A, Wojciechowski W. The semi-automated algorithm for the detection of bone marrow oedema lesions in patients with axial spondyloarthritis. Rheumatol Int. 2020;40(4):625–33;

[11] Rzecki K, Kucybała I, Gut D, Jarosz A, Nabaglo T, Tabor Z, Wojciechowski W. Fully automated algorithm for the detection of bone marrow oedema lesions in patients with axial spondyloarthritis – Feasibility study. Biocybernetics and Biomedical Engineering 2021;41(2):833-53;

[12] Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys. 2019;29(2):102–27;

[13] Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 2006;31(3):1116–28;

[14] Casella G, Fienberg S, Olkin I. An Introduction to Statistical Learning. New York: Springer; 2013. 176–184 p.;

[15] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. MICCAI 2015;9351:234–41;

[16] Bishop CM. Neural Network for Pattern Recognition. Oxford: Clarendon Press; 1995. 230–232 p.;

[17] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv:14126980v9. 2015;

[18] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 32nd Int Conf Mach Learn ICML 2015. 2015;1:448–56;

[19] Billot B, Bocchetta M, Todd E, Dalca AV, Rohrer JD, Iglesias JE. Automated segmentation of the hypothalamus and associated subunits in brain MRI. Neuroimage. 2020;223;

[20] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging. 2015;15:29;

[21] Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proc - 2016 4th Int Conf 3D Vision, 3DV 2016. 2016;565–71;

[22] Chronaiou I, Thomsen RS, Huuse EM, Euceda LR, Pedersen SJ, Hoff M, et al. Quantifying bone marrow inflammatory edema in the spine and sacroiliac joints with thresholding. BMC Musculoskelet Disord. 2017;18(1):1–8;

[23] Bollow M, Braun J, Hamm B, Eggens U, Schilling A, König H, et al. Early sacroilitis in patients with spondyloarthopathy: Evaluation with dynamic gadolinium-enhanced MR imaging. Radiology. 1995;194(2):529–36;

[24] Bollow M, Fischer T, Reifshauer H, Backhaus M, Sieper J, Hamm B, et al. Quantitative analyses of sacroiliac biopsies in spondyloarthropathies: T cells and
macrophages predominate in early and active sacroiliitis - Cellularity correlates with the degree of enhancement detected by magnetic resonance imaging. Ann Rheum Dis. 2000;59(2):135–40;

[25] Landewé RBM, Hermann KGA, Van Der Heijde DMFM, Baraliakos X, Jurik AG, Lambert RG, et al. Scoring sacroiliac joints by magnetic resonance imaging. A multiple-reader reliability experiment. J Rheumatol. 2005;32(10):2050–5.

[26] Bray TJP, Bainbridge A, Punwani S, Ioannou Y, Hall-Craggs MA. Simultaneous Quantification of Bone Edema/Adiposity and Structure in Inflamed Bone Using Chemical Shift-Encoded MRI in Spondyloarthritis. Magn Reson Med. 2018;79(2):1031–42;

[27] Brouwer CL, Boukerrouri D, Oliveira J, Looney P, Steenbakkers RJHM, Langendijk JA, et al. Assessment of manual adjustment performed in clinical practice following deep learning contouring for head and neck organs at risk in radiotherapy. Phys Imaging Radiat Oncol. 2020;16(9):54–60;
## Supplemental Material

| Object               | Image | Number of scans | Annotator                              | Purpose                                                                                           | Number of trials |
|----------------------|-------|-----------------|----------------------------------------|---------------------------------------------------------------------------------------------------|------------------|
| Reference region     | STIR  | 10              | Trained non-radiologist (verified by radiologist) | To determine thresholds from the intensity distribution of normal bone voxels                     | 1                |
| Disease region       | T1W   | 10              | Trained non-radiologist (verified by radiologist) | To identify the area of potential inflammation; To create a dataset for deep learning algorithm training | 1                |
| Inflammatory lesions | STIR  | 8               | Consultant radiologist Radiology fellow | To establish baseline of human performance in manual segmentation of inflammation for consequent comparison with the semi-automated approach | 2 independent trials per reader |

Figure S11: Description of manual segmentation performed for the study.

Figure S12: Boxplot of normal bone STIR intensity for ten subjects, showing different degree of skewness (a). Histogram of normal bone intensity for one subject (b). Vertical lines represent different threshold values; resulting segmentations are presented in the fig. S13.
Figure S13: STIR MRI slice in oblique coronal plane for one subject (left) with super-imposed automatic segmentation via three different thresholds: $Q_U + 1.5 \cdot IQR$ (blue), $Q_U + 2.55 \cdot IQR$ (orange) and maximum intensity (red) of the normal bone intensity distribution (fig. S12), with $Q_U$ upper quartile and $IQR$ inter-quartile range.

Figure S14: T1W MRI slices in oblique coronal plane for one subject (top) with corresponding super-imposed manual segmentations of disease region (bottom). 3D rendering reveals the segmented structure (right).
Figure S15: T1W image slice in oblique coronal plane for two subjects: raw (a) and normalized data (b) with corresponding (slice) intensity histograms.
Figure S16: Original (pre-processed) T1W image slice in oblique coronal plane with corresponding mask (a). Transformed image and mask slices (b). Corresponding (slice) intensity histograms indicate change in intensity levels population after transformation (c).
Figure S17: Schematic representation of the optimal U-Net architecture. Number of channels and matrix size are, respectively, on the top and in the low left corner of a grey shaded box, representing an individual input image slice or a feature map. Network takes as input a mini batch, containing four different T1W image slices, and outputs a batch of corresponding probability maps. Figure details the operations and illustrates how the size of an *individual* image slice or a feature map changes, as the slice / map propagates through the network.
Figure S18: STIR image slices in oblique coronal plane with super-imposed manual segmentation trials from both readers for two subjects (left). 3D rendering reveals segmented structures (right). Comparison of these trials yielded volume overlap and similarity values: 0.59 and 0.97 (top), 0.53 and 0.89 (bottom).
Figure S19: STIR image slices in oblique coronal plane without/with super-imposed automatic segmentation via thresholding with no correction by readers (six subjects). Pixels with intensity between the lower and upper thresholds are marked in yellow and equal or above the upper threshold are marked in red.
| Scan  | **Reader 1** | | **Reader 2** | |
|-------|--------------|-----------------|------------------|---------------------|
|       | Manual delineation only, 1\textsuperscript{st} trial | Residues’ removal | Manual segmentation, 1\textsuperscript{st} trial | Residues’ removal |
| 1003  | 20           | 27              | 15               | 20                  |
| 1007  | -            | 22              | 10               | 15                  |
| 1012  | 18           | 11              | 25               | 15                  |
| 1013  | 56           | 51              | 25               | 20                  |
| 1016  | 60           | 36              | 20               | 30                  |
| 1234  | -            | 20              | 25               | 15                  |
| 3002  | 50           | 19              | 25               | 15                  |
| 4368  | 34           | 14              | 25               | 15                  |
| **Average** | **39.67** | **26.33** | **21.25** | **18.13** |

Figure S20: Time in minutes, taken to manually delineate or segment inflammatory lesions on STIR MRI and to remove residues on automatic segmentations.
Figure S21: T1W image slices in oblique coronal plane (top) for three subjects with super-imposed models averaging ensemble rounded prediction (bottom), for which the reference standard was not available. Subjects exhibit very abnormal bone: high fat content (hyperintense signal; left, middle) and sclerosis (hypointense signal; right) were partially missed by the models ensemble.