Incrementally semi-supervised classification of arthritis inflammation on a clinical dataset

Théodore Aouad, Clementina Lopez-Medina, Charlotte Martin-Peltier, Adrien Bordner, Sisi Yang, Anna Molto, Maxime Dougados, Antoine Feydy, Hugues Talbot

INCREMENTALLY SEMI-SUPERVISED CLASSIFICATION OF ARTHRITIS INFLAMMATION ON A CLINICAL DATASET

Theodore Aouad[1] *Clementina Lopez-Medina*[2] *Charlotte Martin-Peltier*[2] *Adrien Bordner*[2]
Sisi Yang[2] *Anna Molto*[2] *Maxime Dougados*[2] *Antoine Feydy*[2] *Hugues Talbot*[1]

[1]CentraleSupélec, Université Paris-Saclay, Inria, France [2]Cochin Hospital, France

ABSTRACT

For best medical imaging application results, learning-based approaches such as deep learning necessitate specific, extensive and precise annotations. Outside well-curated public benchmarks, these are rarely available in practice, and so it becomes necessary to use less-than-perfect annotations. One way of compensating for this is the embedding of anatomical knowledge. Complementing this, there is the incremental semi-supervised learning technique, whereby a small amount of annotations can be used to derive more and superior labels.

In this article, we illustrate this approach on a deep learning system to help radiologists and rheumatologists finely and interactively assess MRI scans of the sacro-iliac joint in order to correctly diagnose Axial Spondyloarthritis. Our model is trained initially on a relatively small set of images with promising results, on par with expert opinion and generalizable to new datasets.

***Index Terms*—** MRI, deep learning, medical imaging, diagnosis, follow-up.

1. INTRODUCTION

An objective of this article is to show how relatively rare diseases requiring a precise and difficult image-based diagnosis can benefit from modern machine learning techniques. These techniques can even be used in the presence of only weak and sparse annotations, which are typically used by human practitioners in a clinical context. As an example, we take the case of Axial Spondyloarthritis (axSpA).

This disease, which is the most common inflammatory rheumatism in young men, is painful and debilitating as it includes a long-term inflammation of the spine. The most affected area in the early stages of the disease is the joint where the spine meets the pelvis: the sacro-iliac joint (SIJ). It is still relatively rare, affecting approximately 0.6% of the population to a varying degree, and does not have a cure. However, treatments for axSpA can improve symptoms and prevent worsening if an early correct diagnosis is given. The Assessment of Spondyloarthritis International Society (ASAS) proposes a set of criteria for practitioners to detect and recognize early axSpA from MRI. However, expertise in the reading of MRI-SIJ is rare in both the rheumatology and radiology communities as it requires extensive experience.

One potential solution to all this is the use of image-based diagnosis through machine learning, and particularly deep learning which has recently made considerable progress [1, 2], [3, 4]. These systems can recognize the ASAS patterns in a quantitative fashion, may improve diagnosis workflow efficiency, and could be useful in the evaluation of MRI-SIJ images in patients with axSpA, similar to [5]. In fact, recent work by [6] has studied the use of machine learning to detect axSpA with different sequences in patients within the same hospital, with promising results.

We present a novel, ASAS-compliant medical imaging and AI-based system to assess MRI scans of the SIJ to help diagnose axSpA as early as possible. A proof-of-concept is available online: `https://github.com/TheodoreAouad/IAxSpA-demo`

2. DATA

The DESIR cohort [7] was built in order to study axSpA and its evolution over a period of at least 10 years. Around 700 people suspected of having axSpA were recruited across 25 different centers. Only the MRI scans of patients were used from the time of inclusion in the cohort. Each MRI scan includes between 12 and 18 semi-coronal slices of T1 and STIR sequences (see example figure 2). Scans were weakly annotated by 3 trained practitioners, called *readers*. Each reader chose and annotated 6 consecutive slices for inflammatory anomalies, and 6 slices for structural anomalies. In this study, only inflammatory anomalies were considered.

To avoid ambiguity, we only used the patients where all three readers agreed on the patient diagnosis and for which the STIR and T1 sequences were registered. This reduced the dataset to 288 patients.

No anatomical information, including the precise location of inflammation was provided. Instead, each slice included the right side (R) and left side (L) of the SIJ. Each joint was divided into 4 quadrants. In each quadrant, readers indicated

Thanks to Société Française de Rhumatologie and to ASAS Research Grant for funding the study.

if an anomaly was present. Each reader gave an opinion concerning the ASAS positiveness of the patient. As a test set, we used an entirely separate ASAS dataset that follows the DESIR protocol, with 47 patients evaluated by 6 readers, as fully described in [8].

3. METHODOLOGY

Given the weak nature of the annotation, we propose to enrich our architecture with anatomical information. The diagnostic inflammation, if any, should occur near the sacro-iliac joint, which is our region of interest (ROI). In our two-step method, we first gather the anatomical data by training a pair of U-Nets [9] to segment iliac and sacrum bones, from which we derive the ROI. We subsequently train a Mask-RCNN [10] to learn both the ROI detection as well as whether or not the ROI is inflamed (binary classification). A summary of our method is in figure 1.

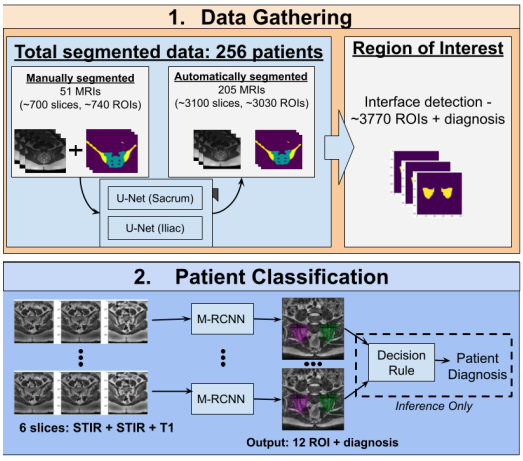

Fig. 1. Schema of our method. In step 1, we augment binary labels with ROI labels. In step 2, we train a Mask-RCNN using ROIs + binary slice label. At inference, we aggregate the slices of each patient to do a diagnosis.

3.1. Region-of-interest generation for training

3.1.1. Pre-processing

The raw data are MRI sequences of varying quality, from multiple sources and manufacturers, representing the clinical practice of 10 years ago. Older MRIs typically feature a global bias, which are low frequency variations that affect the gray-levels of the image [11]. To correct for this bias, we divide each slice by a Gaussian smoothing with a standard variation of $\sigma = 0.05 \times$ image size. We also apply histogram equalization using CLAHE [12] to improve contrast. We reduce noise with a median filter on 4-connected neighborhoods. The result of this preprocessing is shown in Fig. 2.

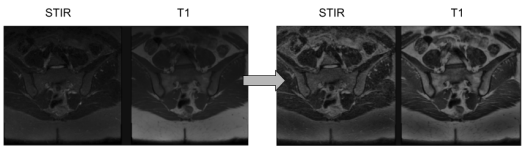

Fig. 2. Preprocessing applied to a slice.

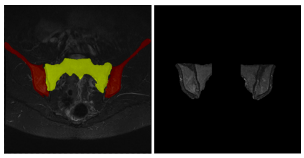

Fig. 3. Example of a segmentation and an obtained ROI after morphological post-processing.

3.1.2. Incremental segmentation of iliac and sacrum bones

To help locate the sacro-iliac joint, first we detect the sacrum and iliac bones. They are easy to annotate for any radiologist; by contrast, recognizing the ROIs necessitates expert knowledge about the disease. 51 patients are manually segmented by trained radiologists, yielding 706 segmented slices. These are then used to train a pair of U-Nets, one for each bone. For this step, we only use the T1 sequence, and apply the preprocessing described above in section 3.1.1.

During training, we augment the data with horizontal flip as well as random rotations within $[-15°, +15°]$. We use the RMSprop [13] optimizer with a learning rate of 10^{-4} and a weight decay of 10^{-8}. Additionally, we use a weighted cross entropy loss with weight $w_{bg} = 0.1$ for background and $w_{fg} = 1 - w_{bg} = 0.9$ for foreground (either the iliac or sacrum bone).

3.1.3. Finding regions of interest

In this paper, the ROI is the area around the joint between the sacrum and the iliac. We use classical mathematical morphology operators on the segmentation to find these regions in each slice. The procedure is given in algorithm 1: we first check all the potential joint parts, so all the iliac pixels that are close to a sacrum pixel, then we select the largest candidate. The iliac bone has so-called "wings", thin regions above the joint that are ignored. A morphological opening of the iliac is performed in order to separate it into two connected components, then the largest is retained. See figure 3 for the resulting ROI.

3.2. Automatic diagnosis of axSpA

3.2.1. Detection of inflammation

We perform joint detection and classification for each ROI by training an end-to-end Mask-RCNN [10] with a ResNet-50 backbone. We use the ROI generated in 3.1.3 as ground-truth

Algorithm 1: Joint Extraction

Result: A mask of the image containing the joint
for *all slices* **do**
  When slice contains both iliac *Il* and sacrum *Sac*;
  for *all iliac connected components* **do**
    Check if there is a sacrum next to it;
    Delete wings of the iliac bone;
    Dilate iliac with a disk $\rightarrow D(\text{Il})$;
    Dilate iliac horizontally $\rightarrow D_H(\text{Il})$;
    Take joint: $J = D(\text{Il}) \cup (D_H(\text{Il}) \cap \text{Sac})$;
    Close the joint: $J = Closing(J)$;
    Fill Holes;
  end
  for *each side* **do**
    Take the largest of the potential joints
  end
end

mask, and the readers' opinions for the label of the ROI, then apply the same pre-processing as in section 3.1.1 to the input image. Additionally, we resample the images so that pixels represent $(0.5 \text{ mm}, 0.5 \text{ mm})$ on all images.

We use a network pre-trained on ImageNet, which imposes a three-channel input. The first channel and second channels are the T1 and STIR sequence, the latter being duplicated. The STIR channel highlights the inflammation information, while the T1 channel is used for more anatomical knowledge. The target is described in figure 1. During training, for each image, we provide the two ROIs as binary masks, and the inflammation information. An example of an ROI is given in figure 3. The standard Mask-RCNN loss [10] and the Detectron2 implementation [14] are used and we apply random rotation with angles between $[-15°, +15°]$, and random horizontal flip as augmentation. We use the standard solver, with a constant learning rate of $2.5 * 10^{-4}$.

3.2.2. Aggregation of all half-slices

At inference, having detected the presence of inflammation on the half-slices of a patient, we aggregate these results to provide a global response for each patient. We choose a criterion based on clinical practice: if the number of positive ROIs is higher or equal to 2, the patient is considered positive. Note that it is possible for the model to predict multiple bounding boxes per half-slice: we sum the results of all bounding boxes.

4. RESULTS

4.1. Incremental segmentation

We manually segmented 51 patients, uniformly chosen amongst the different centers and keeping the same proportion of positive patients of the DESIR dataset at baseline. We split them into 31 for training, 6 for validation and 14 for evaluation. Each U-net took 23 minutes of training on a Nvidia Geforce GTX 1080ti and used 9.5 GB of GPU memory. On the evaluation set, we report a DICE [15] of 0.67 for the sacrum, and 0.84 for the iliac bone.

Note that this number of patients is too low for training a deep classification network with good performance. Patients with good segmentations on most slices (according to clinicians on our team) were selected so we could incrementally augment the number of segmented patients. Out of the 237 remaining patients, 205 patients were deemed successfully segmented (86%). We declare a failure when our pipeline (described in section 3.1) does not detect the SIJ on the slices provided by the readers on two or more half-slices. An example of success is in figure 3. Because of this, the segmented patients base is augmented to 256 $(= 205 + 51)$.

4.2. Classification results

We train 10 models with identical hyperparameters that estimate uncertainty on the performance of our network. Additionally, we are able to propose an ensembling model that can perform a vote. In terms of the annotation, we consider *consensus* to be the majority vote between all readers.

The train - validation - evaluation patient split is as follows, with half-slices in parenthesis: 178 patients (2035), 25 patients (281) and 53 patients (597). The number of half-slices is not exactly equal to $(2 * 6 * |\text{patients}|)$, since some half-slices may not contain a joint. We train the model over 8000 iterations (48 min on a RTX 6000), keeping only the weights with the lowest validation loss, and used 1.3 GB of GPU memory.

Results are reported as the Matthews Correlation Coefficient (MCC) instead of the accuracy metric due to class imbalance. The results for ROIs detection are shown on table 1. For around 10% of half-slices, the Mask-RCNN fails to predict the correct number of regions. In the metrics computation, for each half-slice with the wrong number of bounding boxes, we assume failure. The aggregated results for patient diagnosis are shown in table 2.

To measure the impact of the addition of the incremental segmentation described in 4.1 and the added preprocessing, we performed an ablation study and report the results on the ASAS dataset in table 4.

5. DISCUSSION

5.1. Annotation augmentation

While this step is only used to augment annotations for our training set, we observe that using two independent U-Nets for each bone yields better results than a single U-Net. This is due to the fact that some slices do not contain any sacrum. For these, the single U-Net fails to differentiate the iliac from the sacrum. Although the segmentation results for the sacrum

Table 1. Matthews Correlation Coefficient (x100 for clarity). Cross results on half-slices on the 53 evaluation patients of multiple sources. How to read: annotations from reader 1 vs those from reader 2: MCC= 0.76. Consensus GT vs results from M-RCNN: MCC=0.59 ± 0.03. Underlined: worst results. **bold**: best results.

Readers	1	2	3	M-RCNN (k=10)
1	-	**76.34**	63.40	$\mathbf{61.28 \pm 1.87}$
2	-	-	66.96	60.04 ± 3.34
3	-	-	-	44.34 ± 2.5
Consensus	**85.14**	63.41	75.24	59.16 ± 2.59

Table 2. (MCC $\times$ 100). Results on DESIR evaluation set on patient diagnosis (N=53). We also give the MCC of readers against each other, on the 689 original patients. Decision rule: |pos half slices| $\geq$ 2. Consensus is not given as all readers agree on the evaluation set.

Readers	1 vs 2	1 vs 3	2 vs 3	M-RCNN	ensembling
mcc	**89.06**	74.28	74.48	84.78 ± 3.05	90.24

Table 3. (MCC$\times$100). Results on patient diagnosis on ASAS dataset (N=47). We also give the MCC of readers against each other. Decision rule: |pos half slices| $\geq$ 2.

Readers	1	2	4	5	6	7	M-RCNN	ensembling
1	-	74	65	74	65	81	$\mathbf{71.01 \pm 5.34}$	**80.89**
2	-	-	65	80	89	74	53.34 ± 3.82	58.01
4	-	-	-	58	55	65	45.73 ± 3.92	50.71
5	-	-	-	-	71	**91**	64.8 ± 4.95	73.54
6	-	-	-	-	-	65	47.19 ± 3.4	50.71
7	-	-	-	-	-	-	$\mathbf{71.01 \pm 5.34}$	**80.89**
Consensus	79	**94**	71	86	83	79	52.59 ± 7.67	62.40

Table 4. (MCC $\times$ 100). Ablation study on ASAS (N=47) and DESIR (N=53) datasets. We show the results on the ensembling model.

Ablation	ASAS (MCC*100)	DESIR (MCC*100)
All included	62	**90**
No Constant Pixel	46	**90**
No Remove Gaussian Bias	53	85
No Local Median	**63**	**90**
No CLAHE	54	**90**
No Data Augmentation	53	85
Only manually	21	69

are not as good as for the iliac, this is not critical since we only need an approximate prediction of the sacroiliac joint.

5.2. Classification discussion

Our model succeeds in providing good classifications results. Tables 1 and 2 show the correlation of the readers' opinions in the DESIR dataset for both the half-slices and the patients, respectively. It also shows the inter-reader concordance. By performing a vote on 10 trained Mask-RCNN, referred to here as *ensembling*, we achieve an MCC on the evaluation set of 0.90, which is on par with the comparison of readers opinions against each other. Note however, due to our inclusion criterion explained in section 2, that all three readers agreed on the 53 patients used for the evaluation of our model whereas the correlations for the readers were estimated using the whole 689 patients of our cohort. Nonetheless, this is a very promising result, especially when coupled with the accuracy of this end-to-end approach, which is very high at 96%, with only 2 misclassified patients out of 53, in our test set.

These results translate acceptably well when we test with a different dataset without retraining. As shown in table 3, we still achieve an MCC of 0.62 with the consensus between readers, even in the presence of higher inter-reader variability. The model yields a valid opinion even on a completely new dataset with no retraining.

Finally, for the ablation study (table 4),the importance of performing the annotation augmentation with our segmentation approach is highlighted. For the DESIR dataset, this approach significantly improves the MCC from 0.69 to 0.9. On the ASAS dataset, the improvement is much higher, with a model three times as effective. Also note that while the multiple preprocessing had limited effects on the evaluation set of DESIR, they were really useful on the ASAS dataset. This was expected as we applied the pre-processing for normalizing purposes. When the data is from a similar distribution, it has little effect, but when the data comes from a different source, it improves results. Applying the local median had limited effect in all cases.

6. CONCLUSION

We present an approach to solve a challenging but useful medical imaging classification problem. The dataset is realistic since it is typical in age and size for what is currently available, with only weak annotations designed for clinicians. We proposed a semi-supervised deep learning approach with minimal anatomical annotation that not only demonstrates the benefits of augmenting annotation via a segmentation approach, but was also able to learn and detect clinically relevant inflammation. In the final analysis, our end-to-end approach was able to classify Axial Spondyloarthritis with a MCC similar to a trained radiologist, thus providing a useful opinion, even on a totally new dataset. In ongoing work, larger datasets are being collected to confirm these results and improve accuracy up to the level of human experts. We are also working on the explainability of these models by highlighting inflamed regions to clinicians and plan to perform a longitudinal study of the disease by using the DESIR state of patients after 5 and 10 years. This research is paving the way for further acceptance of Machine Learning methods in clinical practice, allowing useful, highly specific expertise to become more readily available in the near future.

7. REFERENCES

[1] Gabriel Chartrand, Phillip M Cheng, Eugene Vorontsov, Michal Drozdzal, Simon Turcotte, Christopher J Pal, Samuel Kadoury, and An Tang, "Deep learning: a primer for radiologists," *Radiographics*, vol. 37, no. 7, pp. 2113–2131, 2017.

[2] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al., "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, pp. 20170387, 2018.

[3] Geoffrey Hinton, "Deep learning—a technology with the potential to transform health care," *Jama*, vol. 320, no. 11, pp. 1101–1102, 2018.

[4] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[5] Keno-Kyrill Bressem, Janis Lucas Vahldiek, Lisa Adams, Stefan M Niehues, Hildrun Haibel, Valeria Rios Rodriguez, Murat Torgutalp, Mikhail Protopopov, Fabian Proft, Judith Rademacher, et al., "Development and validation of an artificial intelligence approach for the detection of radiographic sacroiliitis," *medRxiv*, 2020.

[6] Kang Hee Lee, Sang Tae Choi, Guen Young Lee, You Jung Ha, and Sang-Il Choi, "Method for diagnosing the bone marrow edema of sacroiliac joint in patients with axial spondyloarthritis using magnetic resonance image analysis based on deep learning," *Diagnostics*, vol. 11, no. 7, pp. 1156, 2021.

[7] Maxime Dougados, Maria-Antonietta D'Agostino, Joëlle Benessiano, Francis Berenbaum, Maxime Breban, et al., "The DESIR cohort: a 10-year follow-up of early inflammatory back pain in France: study design and baseline characteristics of the 708 recruited patients.," *Joint Bone Spine*, vol. 78, no. 6, pp. 598–603, Dec. 2011.

[8] Walter P Maksymowych, Robert GW Lambert, Mikkel Østergaard, Susanne Juhl Pedersen, Pedro M Machado, Ulrich Weber, Alexander N Bennett, Juergen Braun, Ruben Burgos-Vargas, Manouk De Hooge, et al., "Mri lesions in the sacroiliac joints of patients with spondyloarthritis: an update of definitions and validation by the asas mri working group," *Annals of the rheumatic diseases*, vol. 78, no. 11, pp. 1550–1558, 2019.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[11] Jaber Juntu, Jan Sijbers, Dirk Van Dyck, and Jan Gielen, "Bias field correction for mri images," in *Computer Recognition Systems*, Marek Kurzyński, Edward Puchała, Michał Woźniak, and Andrzej żołnierek, Eds., Berlin, Heidelberg, 2005, pp. 543–551, Springer Berlin Heidelberg.

[12] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[13] T. Tieleman and G. Hinton, "Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural Networks for Machine Learning, 2012.

[14] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick, "Detectron2," `https://github.com/facebookresearch/detectron2`, 2019.

[15] Lee R Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.