European Radiology (2023) 33:8310–8323
https://doi.org/10.1007/s00330-023-09704-y

European Radiology
EUROPEAN SOCIETY OF RADIOLOGY

MUSCULOSKELETAL

Neural network algorithm for detection of erosions and ankylosis on CT of the sacroiliac joints: multicentre development and validation of diagnostic accuracy

Thomas Van Den Berghe[1] · Danilo Babin[2] · Min Chen[3] · Martijn Callens[1] · Denim Brack[1] · Helena Maes[1] · Jan Lievens[1] · Marie Lammens[1] · Maxime Van Sumere[1] · Lieve Morbée[1] · Simon Hautekeete[1] · Stijn Schatteman[1] · Tom Jacobs[1] · Willem-Jan Thooft[1] · Nele Herregods[1] · Wouter Huysse[1] · Jacob L. Jaremko[4] · Robert Lambert[4] · Walter Maksymowych[4] · Frederiek Laloo[1] · Xenofon Baraliakos[5] · Ann-Sophie De Craemer[6,7] · Philippe Carron[6,7] · Filip Van den Bosch[6,7] · Dirk Elewaut[6,7] · Lennart Jans[1]

Received: 30 November 2022 / Revised: 3 March 2023 / Accepted: 25 March 2023 / Published online: 23 May 2023

Abstract

Objectives To evaluate the feasibility and diagnostic accuracy of a deep learning network for detection of structural lesions of sacroiliitis on multicentre pelvic CT scans.

Methods Pelvic CT scans of 145 patients (81 female, 121 Ghent University/24 Alberta University, 18–87 years old, mean 40 ± 13 years, 2005–2021) with a clinical suspicion of sacroiliitis were retrospectively included. After manual sacroiliac joint (SIJ) segmentation and structural lesion annotation, a U-Net for SIJ segmentation and two separate convolutional neural networks (CNN) for erosion and ankylosis detection were trained. In-training validation and tenfold validation testing (U-Net—$n = 10 \times 58$; CNN—$n = 10 \times 29$) on a test dataset were performed to assess performance on a slice-by-slice and patient level (dice coefficient/accuracy/sensitivity/specificity/positive and negative predictive value/ROC AUC). Patient-level skeletonisation was applied to increase the performance regarding predefined statistical metrics. Gradient-weighted class activation mapping (Grad-CAM++) heatmap explainability analysis highlighted image parts with statistically important regions for algorithmic decisions.

Results Regarding SIJ segmentation, a dice coefficient of 0.75 was obtained in the test dataset. For slice-by-slice structural lesion detection, a sensitivity/specificity/ROC AUC of 95%/89%/0.92 and 93%/91%/0.91 were obtained in the test dataset for erosion and ankylosis detection, respectively. For patient-level lesion detection after pipeline optimisation for predefined statistical metrics, a sensitivity/specificity of 95%/85% and 82%/97% were obtained for erosion and ankylosis detection, respectively. Grad-CAM++ explainability analysis highlighted cortical edges as focus for pipeline decisions.

Conclusions An optimised deep learning pipeline, including an explainability analysis, detects structural lesions of sacroiliitis on pelvic CT scans with excellent statistical performance on a slice-by-slice and patient level.

Clinical relevance statement An optimised deep learning pipeline, including a robust explainability analysis, detects structural lesions of sacroiliitis on pelvic CT scans with excellent statistical metrics on a slice-by-slice and patient level.

Study originated from: Department of Radiology and Medical Imaging, Ghent University Hospital, Corneel Heymanslaan 10, 9000 Ghent, Belgium

✉ Thomas Van Den Berghe
thovdnbe.vandenberghe@ugent.be;
thomasdanae@hotmail.com

[1] Department of Radiology and Medical Imaging, Ghent University Hospital, Corneel Heymanslaan 10, 9000 Ghent, Belgium

[2] Department of Telecommunication and Information Processing – Image Processing and Interpretation (TELIN-IPI), Faculty of Engineering and Architecture, Ghent University - IMEC, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium

[3] Department of Radiology, Peking University Shenzhen Hospital, Shenzhen 518036, China

[4] Department of Radiology and Diagnostic Imaging and Rheumatology, University of Alberta, 8440 122 Street NW, Edmonton, Alberta T6G 2B7, Canada

[5] Rheumazentrum Ruhrgebiet Herne, Ruhr-University Bochum, Claudiusstraße 45, 44649 Herne, Germany

[6] Department of Rheumatology, Ghent University Hospital, Corneel Heymanslaan 10, 9000 Ghent, Belgium

[7] Vlaams Instituut voor Biotechnologie (VIB) Centre for Inflammation Research (IRC), Ghent University, Technologiepark 927, 9052 Ghent, Belgium

Key Points

- *Structural lesions of sacroiliitis can be detected automatically in pelvic CT scans.*
- *Both automatic segmentation and disease detection yield excellent statistical outcome metrics.*
- *The algorithm takes decisions based on cortical edges, rendering an explainable solution.*

Keywords Artificial intelligence · Deep learning · Sacroiliitis · Spondyloarthropathies · Sacroiliac joint

Abbreviations

ADAM	Adaptive moment estimation
AUC	Area under curve
CNN	Convolutional neural network
Grad-CAM++	Gradient-weighted class activation mapping
NPV	Negative predictive value
PPV	Positive predictive value
ROC AUC	Receiver operating characteristic area under curve
SGD	Stochastic gradient descent
SIJ	Sacroiliac joint
TCL	Threshold for confidence level
TNW	Threshold for number of windows

Introduction

Axial spondyloarthritis is an inflammatory disease affecting the spine and sacroiliac joints (SIJ), with erosions, ankylosis and sclerosis as hallmark structural lesions detected by CT [1, 2]. Mostly beginning under the age of 40 years, spondyloarthritis is characterised by inflammatory low back pain [3]. As sacroiliitis mimics mechanical back pain, the average diagnostic delay is 7 years with about 50% of patients even remaining undiagnosed [3–8]. Early detection of disease opens opportunities for diagnostic workup and treatment, improving the patient's quality-of-life and avoiding a progression to irreversible structural damage, limited joint mobility and chronic low back pain [3, 9–15]. Structural joint lesions are present as incidental findings on many routine CT studies and gain importance in axial spondyloarthritis [16]. However, these subtle lesions are often overlooked in clinical practice.

Although SIJ bone marrow oedema on MRI remains the key finding for early detection, structural lesions have gained importance in spondyloarthritis diagnosis because of their high specificity [16–19]. As radiography has a low sensitivity for detecting early disease, CT remains the standard reference for structural bone lesion detection [18, 20]. Moreover, CT has a role in the diagnostic workflow of spondyloarthritis when MRI cannot be obtained and is not available or inconclusive [1, 2, 16, 21, 22].

Erosions of the SIJs are very specific for spondyloarthritis (90%) and are easy to depict on CT as cortical breaches [17, 18, 23]. Ankylosis is very sensitive and specific (97%) as a hallmark for end-stage spondyloarthritis with bony bridges, limiting joint mobility [17]. Ankylosis, in varying degrees, also occurs in diffuse idiopathic skeletal hyperostosis and following severe joint destruction, as seen in septic sacroiliitis or trauma [24, 25]. Sclerosis is commonly detected in spondyloarthritis but it lacks specificity (76%), as it is considered a hallmark of degenerative disease [17]. For this reason, only erosions and ankylosis were analysed in this study [26]. Structural bone lesions of the SIJs could be picked up in pelvic, abdominal or whole-body CT studies performed for other diagnostic purposes, thereby detecting undiagnosed spondyloarthritis patients in an opportunistic screening setting [25, 26]. As the time of disease onset is 20–40 years old and age-related SIJ degenerative erosions and ankylosis occur in older patients, the target population for opportunistic screening should be strictly defined to avoid overdiagnosis and excessive treatment initiation. Moreover, the opportunistic screening approach is restricted to patients within the predefined age range, encompassing approximately 25% of all patients undergoing a CT examination.

The aim of this study is to evaluate the feasibility and diagnostic performance of deep learning networks in automated detection of SIJ erosions and ankylosis on pelvic CT scans as signs of spondyloarthritis. This can be a first step towards a spondyloarthritis society screening in all CT scans depicting the SIJs [26].

Materials and methods

This study was approved by the local ethics committee and informed consent need was waived.

Study group

All patients (81 females, 18–87 years, 40 ± 13 years) who underwent a dual-energy or routine pelvic CT, with a clinical suspicion of sacroiliitis by a university rheumatologist, were retrospectively and consecutively identified and anonymised in two university hospitals (121 Ghent University and 24 Alberta University, 2005–2021). All of them experienced symptoms suspicious for axial spondyloarthritis (low back and/or buttock pain, morning stiffness) and/or had a positive family history and/or were human leukocyte antigen B27

Table 1 Patient and imaging characteristics

Patient characteristic	Ghent University	Alberta University	Total
Patients identified	138	24	162
Patients excluded	17	0	17
- Incomplete SIJ	9	0	9
- Blurred SIJ	2	0	2
- Metal artefacts	4	0	4
- Sacral tumours	2	0	2
Patients included	121	24	145
Sex			
- Female	73	8	81
- Male	48	16	64
Age range	18–87	19–78	18–87
Mean age (SD)			
- Female	40.0 ± 13.7	39.5 ± 9.8	40.0 ± 13.3
- Male	40.0 ± 11.7	35.4 ± 14.9	38.9 ± 12.8
- Total	40.3 ± 12.9	36.9 ± 13.6	39.7 ± 13.1
Final clinical diagnosis			
- Axial SpA	60	24	84
- No final diagnosis[a]	46	0	46
- Mechanical back pain	15	0	15
Scanner			
- DECT pelvis	79	0	79
- CT pelvis	42	24	66
Imaging characteristic	Ghent University	Alberta University	Total
Scanner type			
- Siemens[b]	119	12	131
- GE Medical Systems[c]	0	7	7
- Philips[d]	0	2	2
- Canon[e]	2	3	5
Scanner subtype Siemens			
- Somatom Definition Flash	84	0	84
- Somatom Definition AS	10	0	10
- Somatom Plus 4	25	0	25
- Sensation 16	0	8	8
- Sensation 64	0	2	2
- Sensation 4	0	1	1
- Definition	0	1	1
Scanner subtype GE Medical Systems			
- LightSpeed Plus	0	2	2
- HiSpeed CT/i	0	4	4
- LightSpeed QX/i	0	1	1
Scanner subtype Philips (Brilliance 16)	0	2	2
Scanner subtype Canon			
- Aquilion	2	1	3
- Asteion	0	1	1
- Aquilion ONE	0	1	1
Image dimensions	512 × 512		
Pixel spacing (mm)	0.250–0.977/0.250–0.977 (mean: 0.680)		
Field-of-view (mm)	*x*-dimension mean: 350; *y*-dimension mean: 350		
Number of slices	24–262		
Slice thickness (mm)	1–5		
Contrast	No		

Table 1 (continued)

kV standard CT imaging	120		
kV DECT imaging	80–140	N/A	80–140
Tube—detector angle	95°	N/A	95°

DECT, dual-energy CT; *kV*, kilovolt; *mm*, millimetre; *N/A*, not available; *SD*, standard deviation; *SIJ*, sacroiliac joint; *SpA*, spondyloarthritis

[a] Symptoms suspicious for axial spondyloarthritis (low back and/or buttock pain, morning stiffness) and/or positive family history and/or HLAB27 positivity and/or recurrent anterior uveitis and/or Crohn's disease but without final clinical diagnosis

[b] Siemens Healthineers; [c] GE Healthcare Systems; [d] Philips Medical Systems; [e] Canon Medical Systems

positive and/or experienced recurrent anterior uveitis and/or had Crohn's disease but without final diagnosis. Forty patients were previously included in a dual-energy CT versus MRI study [27]. Patients were excluded when images were incomplete and blurred and if metal artefacts or SIJ tumours were present. In this way, a heterogeneous dataset with a wide variety of patient and scanner characteristics was identified to increase generalizability of the algorithm.

Imaging protocol

Seventy-nine pelvic dual-energy CT scans and 66 conventional CT scans were included with a broad range of scanning parameters (pixel spacing, field-of-view, number of slices and slice thickness). Moreover, four different CT vendor types were included (131 patients: Siemens Healthineers — seven scanner subtypes; seven patients: GE Medical Systems — three subtypes; two patients: Philips Medical Systems — one subtype; five patients: Canon Medical Systems — three subtypes).

For dual-energy CT scans (Table 1), the weighted average 120 kV-like axial images from the 100–140 kV images were used. For routine CT scans, the unenhanced 120 kV axial images were used (Table 1). Automatic tube current modulation was performed.

Image reading and ground truth

On axial pelvic CT images in bone window, ground truth slice-by-slice manual segmentation and annotation steps were performed (ITK-SNAP 3.6.0, open-source) by three independent pretrained radiologists separately, with a consensus read in case of disagreement ((T.V.D.B., intraining/ M.C./L.B.O.J. with three/seven/17 years of experience), blinded for rheumatological and clinical information (Fig. 1) [28, 29].

The iliac and sacral SIJ surface was segmented as a single-pixel continuous line, creating SIJ segmentation masks. Annotation of erosions (single-pixel continuous line) and ankylosis (multi-pixel continuous region) was performed, creating structural lesion annotation masks. Masks were exported as NIfTI files [10, 30]. Erosions were defined as a cortical bone full thickness loss $\geq$1.0 mm. Ankylosis was defined as SIJ bridging $\geq$2.0 mm. A patient was defined positive for erosion/ankylosis if $\geq$1 structural lesion(s) was(were) present.

The reference standard was reader-assessed presence/absence of structural SIJ lesions, which were automatically detected with the algorithm pipeline.

Algorithm development pipeline

All steps were performed using Python 3.8.8 (Python Software Foundation), TensorFlow 2.10.1 /Keras 2.11.0/NumPy 1.11.3 (open-source, NumFOCUS).

Preprocessing

Preprocessing steps were performed to homogenise the heterogeneous original images from different CT scanners with different technical parameters (slice thickness, pixel spacing, etc.) to conform to identical and stable image quality to improve the statistical outcome performance of the segmentation and disease detection steps (Fig. 2).

Data preparation for detection prediction

The preprocessed 512×512 images were downsampled to 64×64 pixels for efficient training and values were normalised to $[-1,1]$ (Fig. 3).

Next, a U-Net was designed for SIJ segmentation, which is needed to restrict structural lesion detection in the next steps to the SIJs and to discard other anatomical areas. In this way, statistical performance of the disease detection is improved and false positives in areas outside the SIJs are avoided. Due to the small diseased areas, accurate segmentation methods to limit the search area and data balancing to facilitate learning were required. The training input dataset was images with SIJ masks to enhance the learning process. In-training validation (80/20 training/validation split; $n=70$: training/$n=17$: in-training validation) and testing (60/40 training/test split; $n=87$: training/$n=58$: testing) were performed. Regarding

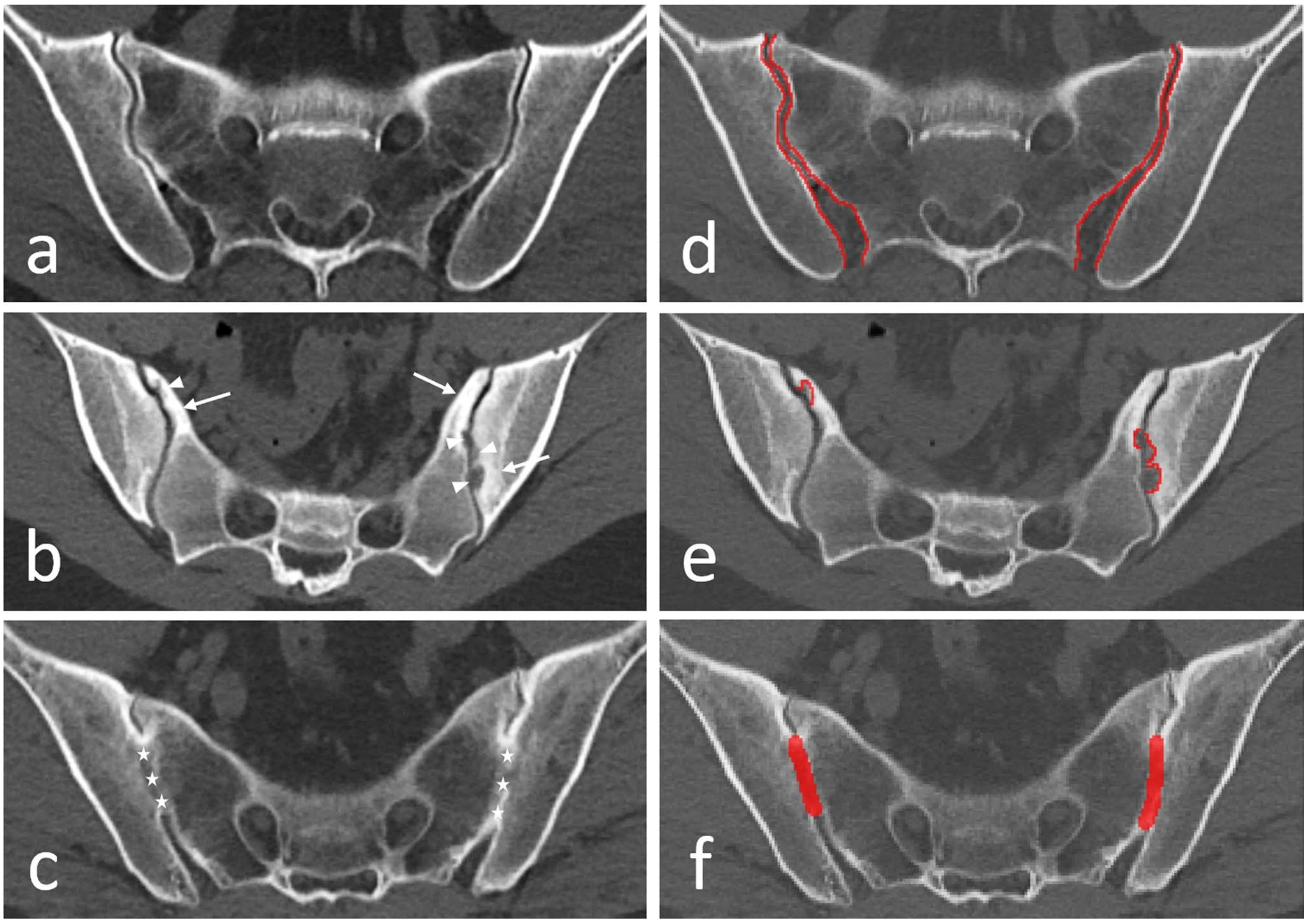

Fig. 1 Axial CT images of the sacroiliac joints. **a**–**c** Scoring principle on original DICOM data CT images. **a** Normal sacroiliac joint in a 44-year-old male with chronic mechanical low back pain due to disc degeneration. No erosion larger than 1.0 mm or ankylosis larger than 2.0 mm can be observed. Both the sacral and iliac cortex and adjacent trabecular bone are normal. **b** Erosions (arrow heads) larger than 1.0 mm are seen in an 18-year-old female with axial spondyloarthritis affecting the left and right sacral and left iliac bone with subchondral bone plate and adjacent trabecular bone destruction and surrounding sclerosis (arrows). **c** Ankylosis (asterisks) larger than 2.0 mm is observed in a 46-year-old male with a clinical diagnosis of axial spondyloarthritis. **d**–**f** Manual segmentation and annotation principle on raw data CT images. **d** Segmentation of the sacral and iliac surfaces of the sacroiliac joints (single-pixel continuous line). **e** Annotation of erosions larger than 1.0 mm (single-pixel continuous line). **f** Annotation of ankylosis larger than 2.0 mm (multi-pixel continuous zone)

technical parameters, batches of sizes 10, 15 and 20 tiles, different numbers of epochs and stochastic gradient descent (SGD) and adaptive moment estimation (ADAM) optimisers were evaluated (selection: batch size 256, 20 epochs, SGD, learning rate 0.7, binary cross-entropy loss function with dice similarity coefficient) (Fig. 3) (Eq. 1).

Equation 1: Dice similarity coefficient comparing the pixelwise similarity agreement of the manual and automatic predicted model sacroiliac joint segmentation. It ranges [0,1], with 1 signifying the greatest similarity between the predicted and the manual ground truth segmentation [30].

$$\text{Dice similarity coefficient} = \frac{2 \times |A \cap B|}{|A| \cup |B|}, \tag{1}$$

A and B are the manual and model segmentation

Subsequently, data preparation steps were performed to upsample the images to the original 512×512 images, to detect the centreline of the SIJs, sample the centreline and finally extract 16×16 and 32×32 images as training and test datasets to provide the lesion detection algorithms with stable input data (Fig. 3). In this way, the entire CT study is downscaled to smaller images which all contain parts of the SIJs, which can be used to train and test the disease detection algorithms.

Structural lesion detection and testing

The prepared 16×16 and 32×32 images were presented to two separate disease detection convolutional neural networks (CNN) for erosion and ankylosis detection (Fig. 3). The training data were balanced to a 50–50 ratio

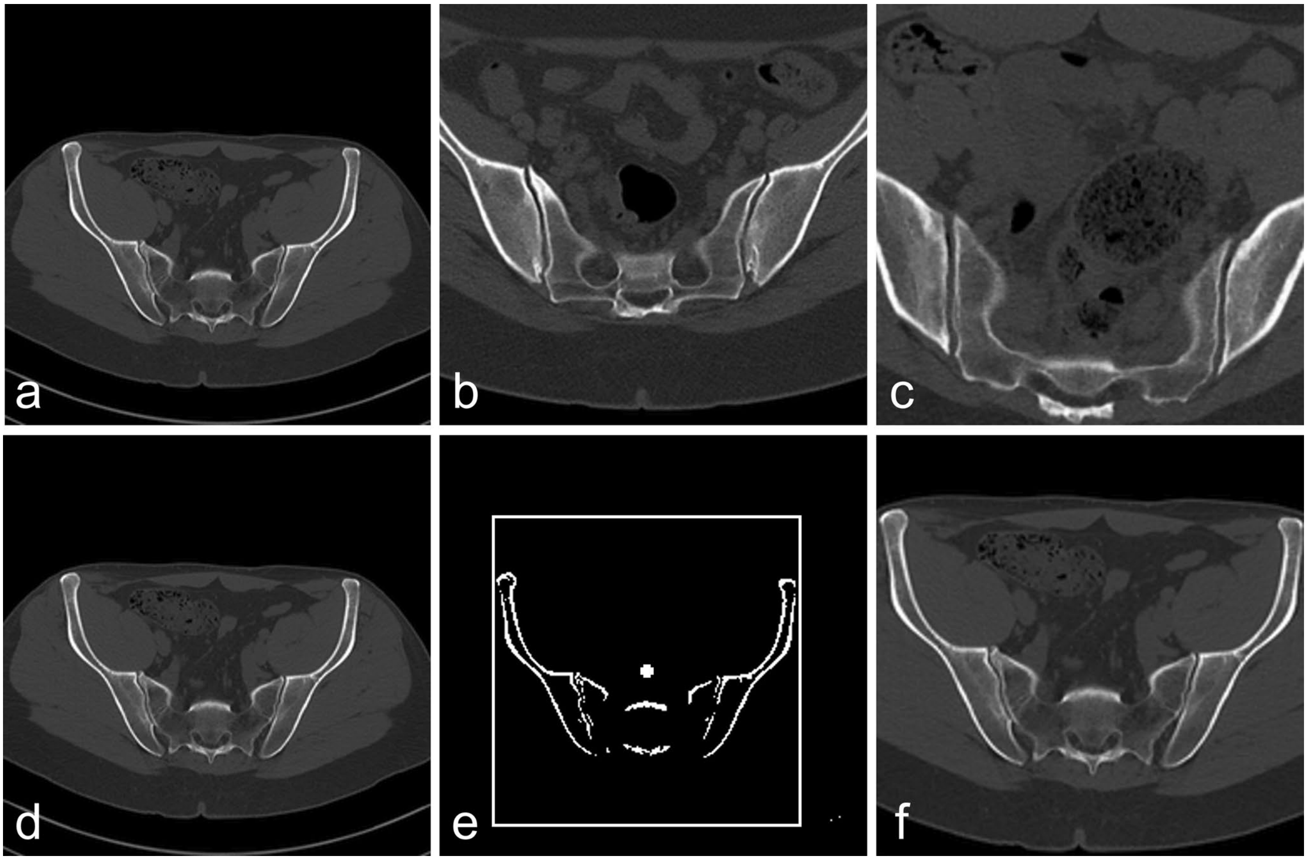

Fig. 2 Preprocessing pipeline for axial unenhanced pelvic CT images. **a**–**c** Axial CT images with different field-of-view, present in the dataset. In order to achieve optimal disease prediction results, the input images need to have equal spacing (pixel size) and field-of-view (dimensions) and need to be centred equally, so that the position of the sacroiliac joints is approximately equal over various slices. **d**–**f** Preprocessing pipeline. **d** Original axial CT image. **e** Axial CT threshold image used for centre of mass calculation to position the sacroiliac joints in the image centre for improved segmentation and annotation performance. The geometric median position of pixels per slice (white cross as centre of the crop window) whose Hounsfield unit values are within the range [0.4–0.6] of the maximum pixel value, is calculated to segment the bony structures and distribute them symmetrically in the image window. The range is empirically chosen to discard background (dark pixels), but also the highest intensity pixels, which proves to be valuable in case of metallic implants. The geometric median position is calculated per slice and the final coordinate of the centre of mass point is calculated by averaging the geometric median positions over all slices. Afterwards, each slice is centred around this reference point. **f** Rescaled (spacing 0.25 mm, lowest value in the dataset) and resized or cropped (dimensions 512 × 512, most frequent image dimension in the input dataset) image with the centre of mass as central point in the image window

(diseased-healthy) to enforce detection. In-training validation (75/25 training/validation split; $n = 87$: training/$n = 29$: in-training validation) and tenfold validation on a test dataset (80/20 training/test split; $n = 116$: training/$n = 29$: testing) were performed. Regarding statistical parameters, the average value E (Eq. 2) for the statistical parameter of interest of the ten-group test results is calculated as an estimate of the model accuracy and is used as a performance indicator of the tenfold validation model.

Equation 2: Average of the ten-group validation test result.

$$E = \frac{1}{10}\sum_{i=1}^{10} \tag{2}$$

$Ei, Ei =$ cross-validation of the ith group in tenfold validation

Data sampling in nearby locations produced near-to-exact image copies causing overfitting. To avoid this, these were excluded in the test dataset to prevent overly optimistic statistical results [31, 32]. Regarding technical parameters, batches of sizes 16, 32 and 64 tiles, different numbers of epochs and SGD/ADAM optimisers were evaluated (selection: batch size 512, 30 epochs, ADAM, binary cross-entropy loss function).

Gradient-weighted class activation mapping (Grad-CAM++) heatmaps highlighted image parts with statistically important regions for algorithmic decisions as explainability analysis [33]. In this way, it is possible to assess whether the disease detection algorithms decide on the presence of structural lesions based on the correct pixels in the images or not.

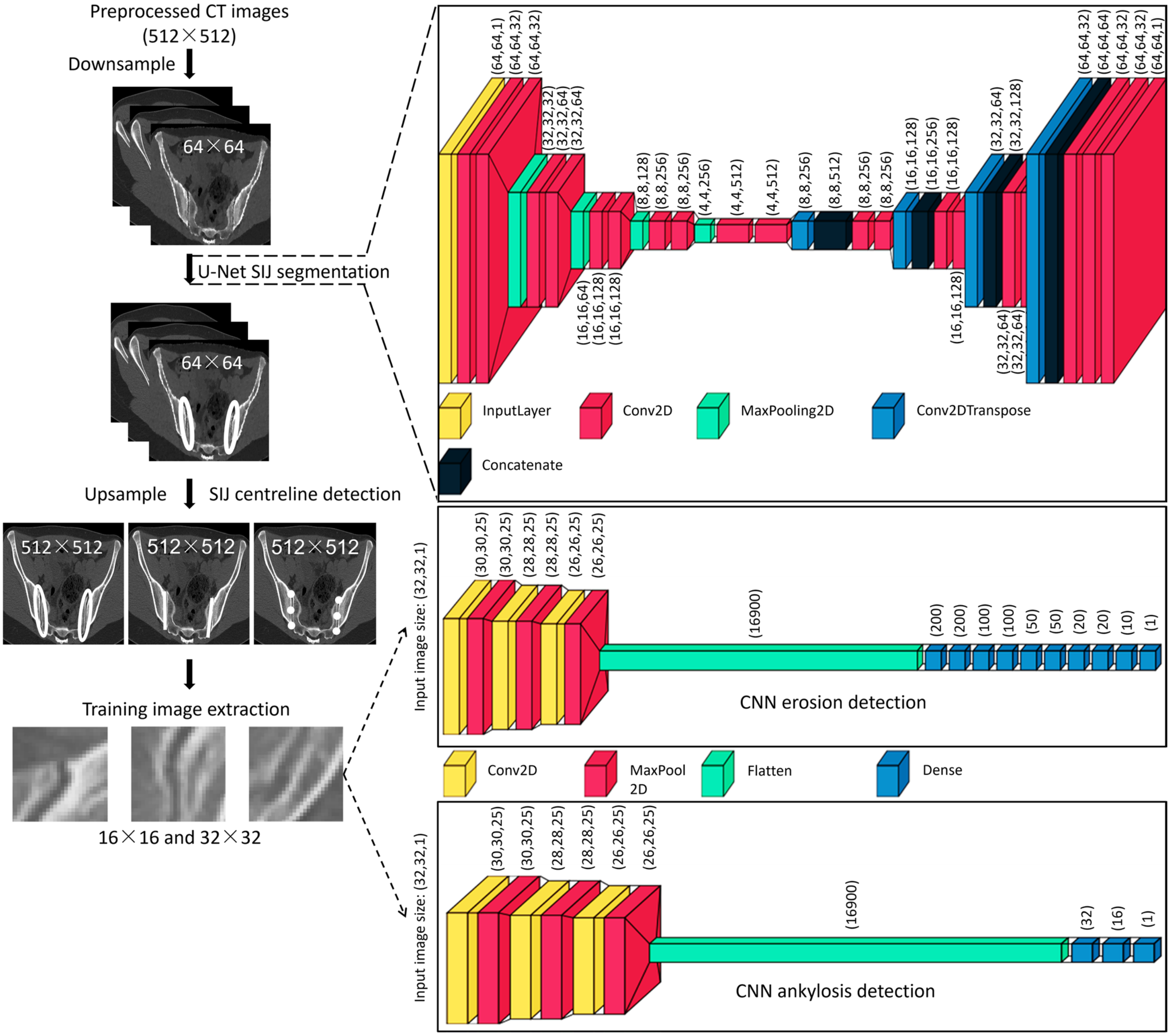

Fig. 3 Data preparation for disease detection prediction on axial CT images with sacroiliac joint segmentation and disease detection convolutional neural networks. Preprocessed images are downsampled to low resolution 64×64 pixel images. Next, a U-Net is used for slice-by-slice sacroiliac joint segmentation in low resolution images (64×64). After U-Net segmentation, morphological closing in three dimensions is applied to encompass as much of the joint region as possible and joints are extracted by using the two largest connected components. Mapping/upsampling of the segmentation to the high resolution images (512×512) is performed to enhance anatomical detail. Next, sacroiliac joint centreline detection (per slice by two-dimensional thinning-based skeletonisation) and centreline equidistant three-dimensional sampling point extraction is performed (sampling distance equals the size of the windows to be extracted). Finally, at sampling points, training sample images are extracted (16×16 and 32×32), which are further processed in the disease detection steps by use of erosion and ankylosis detection convolutional neural network architectures. Compared to the convolutional neural network model for erosion detection, the ankylosis detection network has fewer dense layers (while achieving better detection results) because ankylosis is less unbalanced due to the larger area of sacroiliac joints that is affected. *2D*, two-dimensional; *CNN*, convolutional neural network; *Conv*, convolution; *Max*, maximum; *SIJ*, sacroiliac joint

The algorithm's performance was evaluated in the training and test datasets based on accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), receiver operating characteristic (ROC) curve and 95% confidence interval on a lesion and patient level.

On a patient level, two optimisation parameters were used to generate a disease detection to interpret the unbalanced data for one patient where most of the extracted images did not contain disease: threshold for confidence level (TCL) (window classification prediction, range [0,1], confidence of the algorithm to classify a single CT slice as positive/negative for the presence of structural lesions) and threshold for number of windows (TNW) for which the confidence is above the TCL (percentage of the total

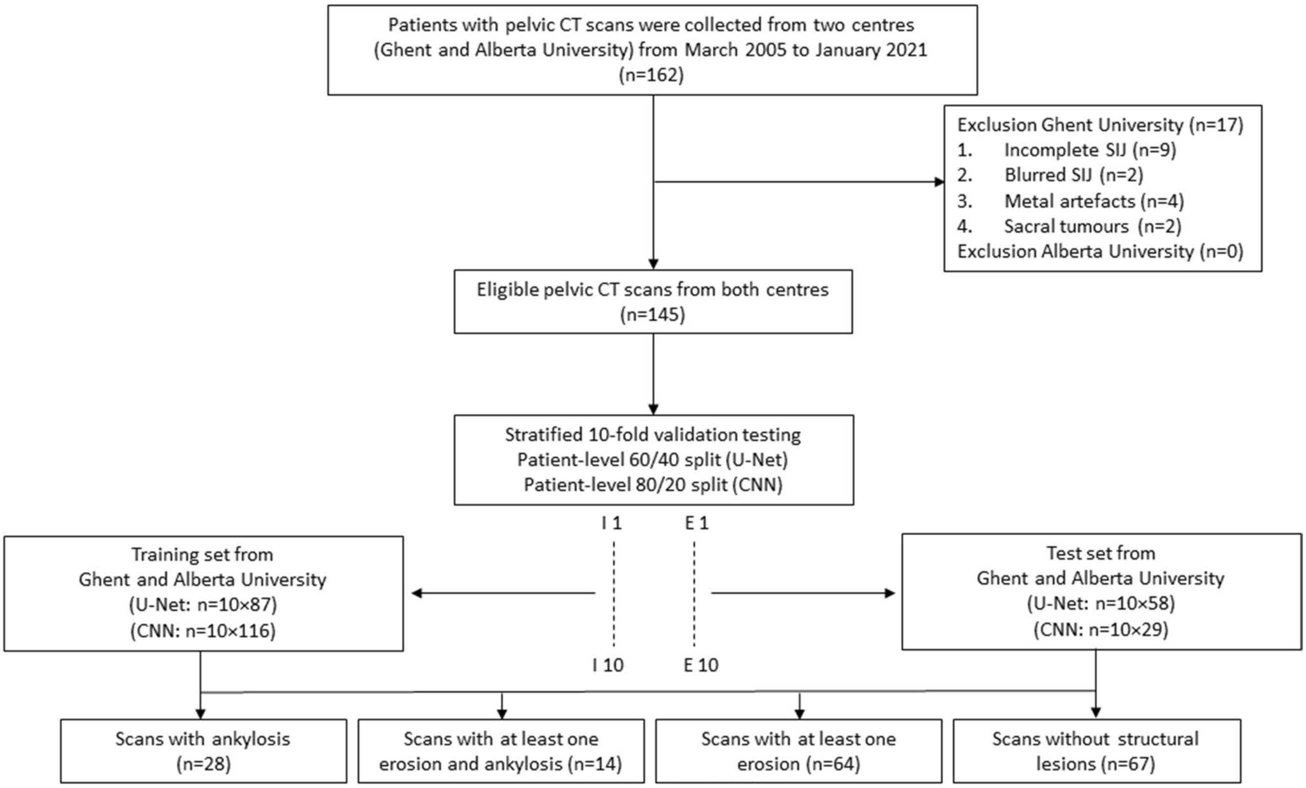
Patients with pelvic CT scans were collected from two centres (Ghent and Alberta University) from March 2005 to January 2021 (n=162)
Exclusion Ghent University (n=17)
1. Incomplete SIJ (n=9)
2. Blurred SIJ (n=2)
3. Metal artefacts (n=4)
4. Sacral tumours (n=2)
Exclusion Alberta University (n=0)
Eligible pelvic CT scans from both centres (n=145)
Stratified 10-fold validation testing
Patient-level 60/40 split (U-Net)
Patient-level 80/20 split (CNN)
I 1
E 1
I 10
E 10
Training set from Ghent and Alberta University (U-Net: n=10×87) (CNN: n=10×116)
Test set from Ghent and Alberta University (U-Net: n=10×58) (CNN: n=10×29)
Scans with ankylosis (n=28)
Scans with at least one erosion and ankylosis (n=14)
Scans with at least one erosion (n=64)
Scans without structural lesions (n=67)

Fig. 4 Patient inclusion, exclusion and split in training and test datasets. For sacroiliac joint segmentation (U-Net), a 60/40 patient-level split was performed between the training and test dataset. For disease detection (convolutional neural networks), an 80/20 patient-level split was performed between the training and test dataset. *CNN*, convolutional neural network; *Ei*, validation result i (error of the cross-validation iteration step Ii); *Ii*, cross-validation iteration step i; *n*, number; *SIJ*, sacroiliac joint

number of windows/slices, range [1,100]). Metrics to optimise for were accuracy, weighted accuracy towards less false positives or false negatives and weighted accuracy towards exclusion of false positives. Statistical analysis was performed using IBM SPSS 26.

Results

Study group

In total, 162 patients were retrospectively identified. Seventeen CT scans were excluded, resulting in 145 included patients. Patient demographics, inclusion, exclusion, training versus in-training validation versus test dataset split on a patient level and imaging characteristics are illustrated in Fig. 4 and Table 1. Eighty-four patients had a final diagnosis of spondyloarthritis, made by a university rheumatologist. In 15 patients, a mechanical origin of low back and/or buttock pain was identified. In 46 patients, no clear origin of the pain was identified.

Image reading and ground truth

Sixty-four (44%) of patients had SIJ erosions and 28 (19%) had ankylosis. Fourteen (10%) had both erosion(s) and ankylosis and 67 (46%) had no structural lesions. SIJ masks were mostly uniformly distributed over the three-dimensional images, with 24% of slices containing the SIJs. Considering only slices where the SIJ mask existed, 354 pixels were segmented for the SIJ per 512×512 slice window, on average (0.14%). Erosion mask volumes were, on average, nine times smaller than ankylosis mask volumes. The relative occurrence per SIJ mask (number of pixels of annotated structural lesions divided by number of SIJ segmentation pixels) was 4.2% for erosion and 34.3% for ankylosis (ratio eight/nine).

Algorithm pipeline performance

Data preparation for detection prediction

For the U-Net SIJ segmentation and assuming an average region-of-interest class/non-region-of-interest class ratio

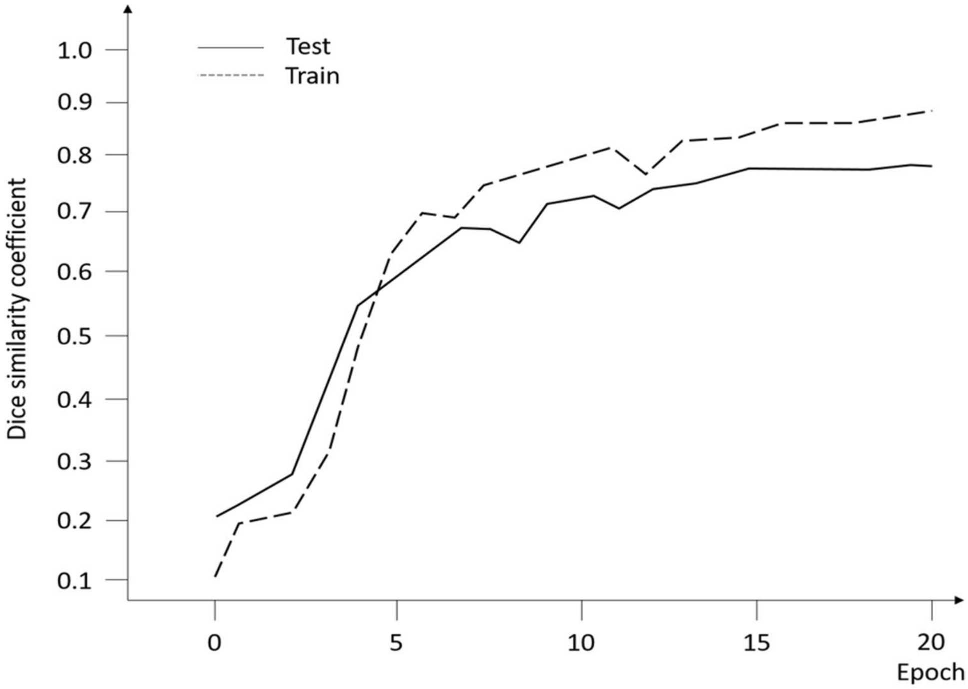

Fig. 5 U-Net sacroiliac joint segmentation dice similarity coefficient as a function of the number of epochs in the training dataset (dotted line) and in the test dataset (full line)

of 1/3, a training and test dataset dice similarity coefficient (Eq. 1) of 0.89 and 0.75 ± 0.03 were obtained after 20 epochs on tenfold validation testing (Fig. 5). In this way, CT scans were approximately reduced to one-third of the total number of slices that were present in the original CT scan before region-of-interest selection and the SIJs were segmented properly.

Slice-by-slice structural lesion detection and testing

Results of the two-dimensional slice-by-slice structural lesion detection on 16 × 16 and 32 × 32 images in the test dataset are presented in Table 2 and Fig. 6a, b.

In summary, the maximal accuracy to detect erosions was equal in both the 16 × 16 and 32 × 32 test images,

Table 2 Two-dimensional slice-by-slice structural lesion detection in the test dataset for 16 × 16 and 32 × 32 images

Parameter	16 × 16 images		32 × 32 images	
	Erosion	Ankylosis	Erosion	Ankylosis
Lesion present (*n*)	1224	2756	1224	2756
Accuracy (%)	89 (2168/2448)[a]	87 (4822/5512)[b]	89 (2193/2458)[a]	92 (5061/5512)[b]
FN (%)	5 (60/1224)	15 (412/2756)	10 (127/1224)	7 (191/2756)
TP (%)	95 (1164/1224)	85 (2344/2756)	90 (1097/1224)	93 (2565/2756)
FP (%)	18 (220/1224)	10 (278/2756)	11 (138/1224)	9 (260/2756)
TN (%)	82 (1004/1224)	90 (2478/2756)	90 (1096/1224)	91 (2496/2756)
Sensitivity (%)	95 (1164/1224)	85 (2344/2756)	90 (1097/1224)	93 (2565/2756)
Specificity (%)	82 (1004/1224)	90 (2478/2756)	89 (1096/1234)	91 (2496/2756)
NPV (%)	94 (1004/1064)	86 (2478/2890)	90 (1096/1223)	93 (2496/2687)
PPV (%)	84 (1164/1384)	89 (2344/2622)	89 (1097/1235)	91 (2565/2825)

FN, false negative; *FP*, false positive; *n*, number; *NPV*, negative predictive value; *PPV*, positive predictive value; *TN*, true negative; *TP*, true positive

An equal number of 1224 and 2756 slices did and did not contain erosions and ankylosis, respectively. In between brackets: the numerators and denominators for each specified metric

[a]Baseline = 0.50. Accuracy interval for confidence 95%: 1.2%; [b]Baseline = 0.50. Accuracy interval for confidence 95%: 0.8%

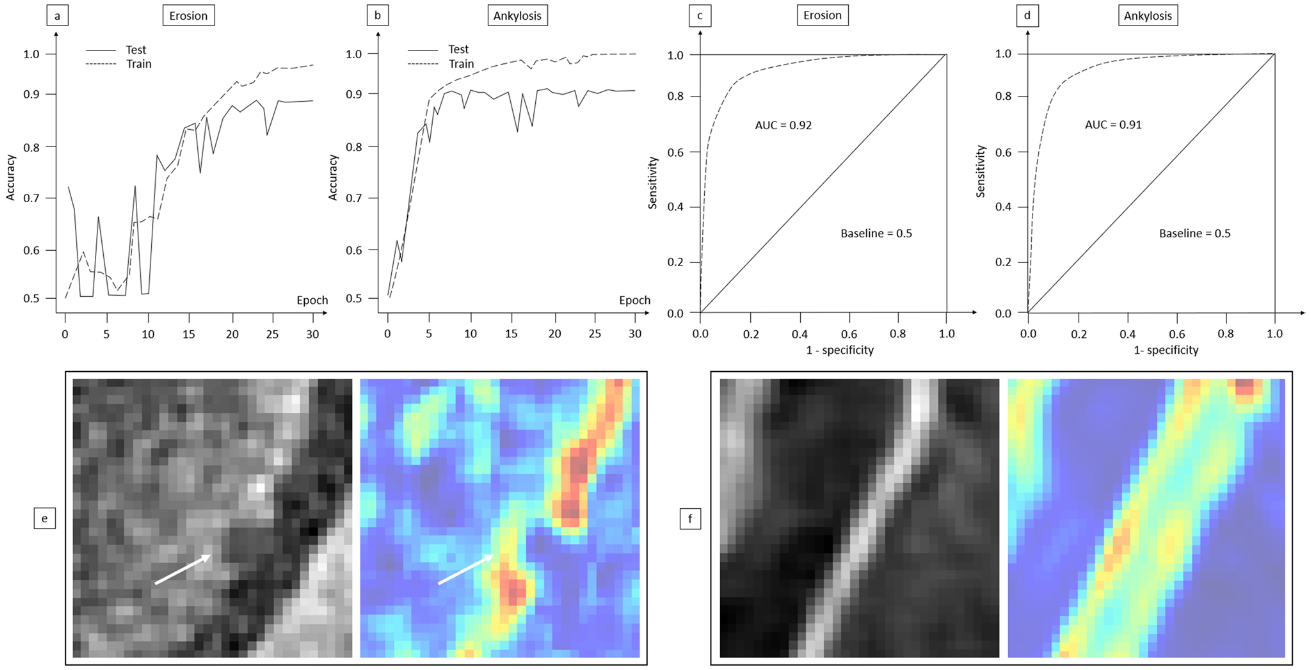

Fig. 6 Disease detection module statistical performance and explainability by gradient-weighted class activation mapping (Grad-CAM++). **a**, **b** Two-dimensional slice-by-slice algorithm accuracy in detecting (**a**) erosion and (**b**) ankylosis of the sacroiliac joints as a function of the number of epochs in the training dataset (dotted line) and in the test dataset (full line). **c**, **d** Two-dimensional slice-by-slice receiver operating characteristic curve (dotted line; full line as random guess no-discrimination reference) for the detection of (**c**) erosion and (**d**) ankylosis of the sacroiliac joints in the test dataset showing the balance between the true positive rate and the true negative rate of findings. **e**, **f** Gradient-weighted class activation mapping (Grad-CAM++) analysis with heat maps clearly shows that the focus of the convolutional neural network in the decision-making for both (**e**) erosion (white arrows) and (**f**) ankylosis detection is on the cortical edges of the sacroiliac joints, as intended. For ankylosis detection, compared to the heat maps of erosion detection, it is observed that in some cases wider areas of the window are considered, which is in accordance with the fact that ankylosis is present in wider areas of the sacroiliac joint cortex as compared to erosions. *AUC*, area under curve

while in the 32 × 32 test images, higher values for ankylosis were obtained. For erosions, the sensitivity and the NPV were higher in the 16 × 16 test images, while in the 32 × 32 test images, higher values for ankylosis were obtained. For erosions and ankylosis, the specificity and PPV were higher in the 32 × 32 test images.

The ROC curves, depicting the relationship between the true positive and true negative rate, rendered an area under curve (AUC) of 0.92 for erosion detection and 0.91 for ankylosis detection in the test dataset (Fig. 6c, d).

Grad-CAM++ analysis with heatmaps showed that the focus of the CNN in the decision-making to detect structural lesions is on the cortical edges of the SIJs, as intended (Fig. 6e and f) [33].

Patient-level structural lesion detection, optimisation and testing

For a three-dimensional patient-level lesion detection, the trained balanced CNN to predict per extracted 32 × 32 window was used as a basis and optimised with the TCL and TNW parameters. The results in the test dataset can be found in Table 3.

Maximal accuracy optimisation rendered an accuracy of 74% for erosion detection with a sensitivity of 87% and a NPV of 86%. For ankylosis, an accuracy of 88% with a specificity of 90% and a NPV of 95% were obtained.

Optimisation to reduce false negatives rendered a sensitivity of 95% and a NPV of 94% for erosion detection and a sensitivity of 82%, a specificity of 90% and a NPV of 95% for ankylosis detection.

Optimisation to reduce false positives rendered a specificity of 85% and a PPV of 73% for erosion detection and a specificity of 97%, a PPV of 75% and a NPV of 88% for ankylosis detection.

Discussion

Spondyloarthritis, with a long diagnostic delay and a 40–50% underdiagnosis, has a significant burden if detected lately [4–9]. Therefore, this study aimed to develop an algorithm to automatically detect SIJ erosions and ankylosis and to evaluate its

Table 3 Three-dimensional patient-level structural lesion detection, optimisation and testing for 32 × 32 images in the test dataset

Optimisation parameter	TCL	TNW (%)	Acc. (%)	Sens. (%)	Spec. (%)	NPV (%)	PPV (%)
Max. accuracy							
Erosion	0.98	15	74* (105/142)	87 (54/62)	64 (51/80)	86 (51/59)	65 (54/83)
Ankylosis	0.70	27	88*[1] (125/142)	78 (21/27)	90 (104/115)	95 (104/110)	66 (21/32)
Reduce FN							
Erosion	0.97	11	72*[2] (102/142)	95 (59/62)	54 (43/80)	94 (43/46)	61 (59/96)
Ankylosis	0.97	12	88*[3] (125/142)	82 (22/27)	90 (103/115)	95 (103/108)	65 (22/34)
Without FP							
Erosion	0.43	97	58*[4] (83/142)	5 (3/62)	100 (80/80)	58 (80/139)	100 (3/3)
Ankylosis	0.98	38	82*[5] (117/142)	7 (2/27)	100 (115/115)	82 (115/140)	100 (2/2)
Reduce FP							
Erosion	0.99	40	70*[6] (100/142)	52 (32/62)	85 (68/80)	69 (68/98)	73 (32/44)
Ankylosis	0.39	44	87*[7] (123/142)	44 (12/27)	97 (111/115)	88 (111/126)	75 (12/16)

Acc., accuracy; *FN*, false negative; *FP*, false positive; *Max.*, maximal; *NPV*, negative predictive value; *PPV*, positive predictive value; *Sens.*, sensitivity; *Spec.*, specificity; *TCL*, threshold for confidence level; *TNW*, threshold for number of windows

In between brackets: the numerators and denominators for each specified metric

Accuracy interval for confidence 95%: *7.2%; *[1]5.3%; *[2]7.4%; *[3]5.3%; *[4]8.1%; *[5]6.3%; *[6]7.5%; *[7]5.7%

performance on CT pelvis as a first step towards expansion of the algorithm pipeline to all CT types depicting the SIJ. This feasibility study confirmed similar outcomes in region-of-interest extraction/SIJ segmentation and lesion detection on a slice-by-slice and patient level as compared to manual pixel-by-pixel expert-based ground truth segmentation and annotation.

Regarding segmentation performance, a dice coefficient of 0.75 was obtained. Substantial variability and uncertainty about which exact pixels to segment existed in manual segmentation on low-quality 64 × 64 images. The only possible alternative was rough bounding box segmentation. If an adjacent pixel was segmented automatically as compared to manually, the dice coefficient would drop significantly in pixelwise comparison. Moreover, only rough segmentation was required to feed the disease detection algorithms with anatomical data. As a comparison, rough SIJ bounding box localisation (instead of pixelwise segmentation in this study) reached a validation dice coefficient of 0.82 (Shenkman et al) but this was a different approach and easier to achieve because the region-of-interest that needed to be recognised was larger [3]. Detailed 512 × 512 image segmentation would render more details but this would require more computer memory usage and was not necessary for further analysis.

Regarding disease detection, variation in performance between statistical outcome metrics existed. Firstly, the performance for ankylosis detection exceeded that for erosion for every statistical outcome parameter except for sensitivity, regardless of the optimisation metric. The higher abundance of slices with ankylosis and the larger ankylosed region per slice rendered more training data, thus reducing data imbalance and thus artificially increasing the detection probability. Secondly, the two-dimensional slice-by-slice performance was more accurate as compared to patient-level analysis because a larger data imbalance existed in patient-level analysis as the majority of slices and pixels did not contain disease. optimisation steps increased statistical performance of the three-dimensional patient-level analysis.

To our knowledge, this study was the first to successfully develop a pipeline to detect structural lesions based on ground truth pixelwise segmentations/annotations with three-dimensional patient-level optimisation instead of qualitative scoring. Shenkman et al developed a qualitative algorithm to grade sacroiliitis on CT scans and a joint classification system with an accuracy of 92%/86%, sensitivity of 95%/82% and AUC of 0.97/0.57. Due to the different approach, comparison of statistical metrics is impossible as structural lesion detection was not an outcome measure. Moreover, Geijer et al confirmed that qualitative scores were obsolete in CT scans and that only erosions were a valid diagnostic feature of spondyloarthritis [34]. Nevertheless, our statistical outcome measures mirrored those in this study, but in a quantitative way on a slice and patient level. A deep learning algorithm by Bressem et al detected sacroiliitis on conventional radiography with a ROC AUC of 0.94 for the test dataset [35]. A study by Castro-Zunti et al achieved excellent statistical metrics but did not use entire CT studies but cropped images to reduce the disbalance between diseased versus normal pixels, thus artificially improving the performance [36].

Early structural lesion detection may lead to preventive/timely treatment. In this way, irreversible joint damage could be avoided, significantly improving patients' quality-of-life and reducing societal costs [10–15]. However, early structural lesions are often missed in clinical practice, as early disease changes may be subtle and the SIJs are often overlooked. An automated analysis of CT scans could be used as a screening tool [3, 25]. Chan et al and Hermann et al stated that presence of ≥ 3 erosions or ankylosis of the middle or dorsal part of the SIJ is sufficient for identifying patients at risk of sacroiliitis, thus supporting the hypothesis that opportunistic detection of structural SIJ lesions can be correlated with final clinical diagnosis [5, 26, 37]. A potential drawback of a screening strategy is the risk of overdiagnosis and excessive treatment initiation, especially in elderly patients where age-related degenerative erosions and ankylosis can be present without suffering from axial spondyloarthritis. Therefore, a strict age range for opportunistic screening should be defined before clinical application starts. As a guide, the age of disease onset of 20–40 years and mean diagnostic delay of 7.5 years in men and 9.5 years in women can be used [5]. Moreover, the opportunistic automatic screening algorithms should only flag the presence of structural lesions and cannot be used to make the final clinical diagnosis of axial spondyloarthritis, which is the responsibility of the radiologist and rheumatologist after extensive clinical, MRI, genetic and laboratory testing to assess inflammatory low back or buttock pain, the presence of bone marrow oedema and the inflammatory status. In future developments, automatic patient-level structural SIJ lesion detection should be adapted to clinical needs in order to avoid overdiagnosis, which can be performed by adapting the TCL and TNW parameters.

Our feasibility study had several limitations. Firstly, it included a limited number of pelvic CTs. Secondly, only pelvic CTs were included; thus, future work is required to expand to other CT types for deployment of the pipeline as a screening tool [25, 34]. Thirdly, both centres were tertiary university hospitals and no patients from private hospitals were included. Moreover, no external testing was performed with data from other institutions. Lastly, a consensus joint segmentation and lesion annotation was used without intra- or interobserver variability analysis. To address the limitations, the robustness of the algorithm's performance was evaluated by including heterogeneous international datasets to avoid selection bias. Moreover, preprocessing steps and tenfold validation testing on a separate dataset were performed to assess performance, reduce overfitting and improve generalizability [31, 32].

In conclusion, we developed a deep learning-based algorithm with promising outcome metrics for automatic detection of structural lesions in sacroiliitis on pelvic CT scans on a slice-by-slice and patient level.

Acknowledgements Patient data were provided by the Ghent University Hospital, Ghent, Belgium and the University of Alberta Hospital, Edmonton, Canada. The authors would like to thank Heleen Coreelman (MSc in Biochemistry and Biotechnology, MSc in Forensic Science) for proofreading the final version of the manuscript and Griet Alleman (BSc Radiographer, Study Coordinator Ghent University Hospital) for data collection and study management.

Funding This study was funded by multiple scientific international organizations:

ESSR (European Society of Musculoskeletal Radiology) Young Researcher Grant 2022 (Thomas Van Den Berghe, MD), European Validation, 2500 euros.

King Baudouin Foundation Scientific Research on Rheumatology Fund 2022 (Thomas Van Den Berghe, MD/Lennart Jans, MD, PhD), Belgian Federal Validation, 15000 euros.

Research Foundation Flanders Fundamental Clinical Mandate 2022 (Lennart B.O. Jans, MD, PhD), 100000 euros.

Declarations

Guarantor The scientific guarantors of this publication are Thomas Van Den Berghe (MD) and Lennart Jans (MD, PhD, professor of Radiology) from the Ghent University Hospital, Ghent, Belgium.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry Danilo Babin (Ir., PhD) from the Ghent University (Department of Telecommunication and Information Processing, Image Processing and Interpretation, Faculty of Engineering and Architecture, IMEC, Ghent, Belgium) kindly provided statistical advice for this manuscript.

One of the authors (Danilo Babin, Ir., PhD) has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap Forty of the 145 patients were previously included in a dual-energy CT versus MRI study on bone marrow oedema detection in another setting. In the present study, algorithms based on artificial intelligence (CNN/U-Net) are being built on CT scans of the pelvis to detect structural lesions (erosions and ankylosis), which is a totally different study design.

Reference study (attached): Chen M, Herregods N, Jaremko JL, Carron P, Elewaut D, Van den Bosch F, et al Bone marrow oedema in sacroiliitis: detection with dual-energy CT. Eur Radiol. 2020;30(6):3393–400.

Methodology

- retrospective
- diagnostic or prognostic study
- multicentre study

References

1. Dougados M, Baeten D (2011) Spondyloarthritis. Lancet 377:2127–2137. https://doi.org/10.1016/S0140-6736(11)60071-8
2. Sieper J, Poddubnyy D (2017) Axial spondyloarthritis. Lancet 390:73–84. https://doi.org/10.1016/S0140-6736(16)31591-4
3. Shenkman Y, Qutteineh B, Joskowicz L et al (2019) Automatic detection and diagnosis of sacroiliitis in CT scans as incidental findings. Med Image Anal 57:165–175. https://doi.org/10.1016/j.media.2019.07.007
4. Song I-H, Sieper J, Rudwaleit M (2007) Diagnosing early ankylosing spondylitis. Curr Rheumatol Rep 9:367–374
5. Khan S, Shridharmurthy D, Lapane KL et al (2022) The disease burden of axial spondyloarthritis: through a gendered lens. Clin Rheumatol 41:1115–1124
6. Wendling D, Claudepierre P, Prati C (2013) Early diagnosis and management are crucial in spondyloarthritis. Joint Bone Spine 80:582–585. https://doi.org/10.1016/j.jbspin.2013.03.003
7. Danve A, Deodhar A (2019) Axial spondyloarthritis in the USA: diagnostic challenges and missed opportunities. Clin Rheumatol 38:625–634
8. Poddubnyy D, Rudwaleit M (2012) Early Spondyloarthritis. Rheum Dis Clin North Am 38:387–403. https://doi.org/10.1016/j.rdc.2012.04.007
9. Strand V, Singh JA (2017) Patient burden of axial spondyloarthritis. J Clin Rheumatol 23:383–391. https://doi.org/10.1097/RHU.0000000000000589
10. Kotsis K, Voulgari PV, Drosos AA et al (2014) Health-related quality of life in patients with ankylosing spondylitis: a comprehensive review. Expert Rev Pharmacoecon Outcomes Res 14:857–872. https://doi.org/10.1586/14737167.2014.957679
11. Boonen A, Sieper J, van der Heijde D et al (2015) The burden of non-radiographic axial spondyloarthritis. Semin Arthritis Rheum 44:556–562. https://doi.org/10.1016/j.semarthrit.2014.10.009
12. Kobelt G, Andlin-Sobocki P, Brophy S et al (2004) The burden of ankylosing spondylitis and the cost-effectiveness of treatment with infliximab (Remicade ®). Rheumatology 43:1158–1166. https://doi.org/10.1093/rheumatology/keh271
13. Lee JS, Oh B-L, Lee HY, Song YW, Lee EY (2018) Comorbidity, disability, and healthcare expenditure of ankylosing spondylitis in Korea: A population-based study. PLoS One 13(2):e0192524
14. Tu L, Rai JC, Cao S et al (2014) Costs and work limitation of patients with ankylosing spondylitis in China. Clin Exp Rheumatol 32:661–666
15. Sieper J, Holbrook T, Black CM et al (2016) Burden of illness associated with non-radiographic axial spondyloarthritis: a multiperspective European cross-sectional observational study. Clin Exp Rheumatol 34:975–983
16. Maksymowych WP, Lambert RG, Baraliakos X et al (2021) Data-driven definitions for active and structural MRI lesions in the sacroiliac joint in spondyloarthritis and their predictive utility. Rheumatology (Oxford) 60:4778–4789. https://doi.org/10.1093/rheumatology/keab099
17. Sieper J, Rudwaleit M, Baraliakos X et al (2009) The Assessment of SpondyloArthritis international Society (ASAS) handbook: A guide to assess spondyloarthritis. Ann Rheum Dis 68(Suppl II):ii1–ii44. https://doi.org/10.1136/ard.2008.104018
18. Jans L, Egund N, Eshed I et al (2018) Sacroiliitis in Axial Spondyloarthritis: Assessing Morphology and Activity. Semin Musculoskelet Radiol 22:180–188. https://doi.org/10.1055/s-0038-1639470
19. Pialat JB, Di Marco L, Feydy A et al (2016) Sacroiliac joints imaging in axial spondyloarthritis. Diagn Interv Imaging 97:697–708. https://doi.org/10.1016/j.diii.2016.02.013
20. Weber U, Lambert RGW, Østergaard M et al (2010) The diagnostic utility of magnetic resonance imaging in spondylarthritis: An international multicenter evaluation of one hundred eighty-seven subjects. Arthritis Rheum 62:3048–3058. https://doi.org/10.1002/art.27571
21. Melchior J, Azraq Y, Chary-Valckenaere I et al (2017) Radiography and abdominal CT compared with sacroiliac joint CT in the diagnosis of sacroiliitis. Acta Radiol 58:1252–1259. https://doi.org/10.1177/0284185116688377
22. Lambert RGW, Hermann KGA, Diekhoff T (2021) Low-dose computed tomography for axial spondyloarthritis: update on use and limitations. Curr Opin Rheumatol 33:326–332. https://doi.org/10.1097/BOR.0000000000000803
23. Diekhoff T, Hermann KGA, Greese J et al (2017) Comparison of MRI with radiography for detecting structural lesions of the sacroiliac joint using CT as standard of reference: results from the SIMACT study. Ann Rheum Dis 76:1502–1508. https://doi.org/10.1136/annrheumdis-2016-210640
24. Althoff CE, Sieper J, Song IH et al (2013) Active inflammation and structural change in early active axial spondyloarthritis as detected by whole-body MRI. Ann Rheum Dis 72:967–973. https://doi.org/10.1136/annrheumdis-2012-201545
25. Leibushor N, Slonimsky E, Aharoni D et al (2017) CT abnormalities in the sacroiliac joints of patients with diffuse idiopathic skeletal hyperostosis. AJR Am J Roentgenol 208:834–837. https://doi.org/10.2214/AJR.16.16994
26. Hermann KGA, Ziegeler K, Kreutzinger V et al (2022) What amount of structural damage defines sacroiliitis: a CT study. RMD Open 8:e001939. https://doi.org/10.1136/rmdopen-2021-001939
27. Chen M, Herregods N, Jaremko JL et al (2020) Bone marrow oedema in sacroiliitis: detection with dual-energy CT. Eur Radiol 30:3393–3400. https://doi.org/10.1007/s00330-020-06670-7
28. Slobodin G, Croitoru S, Starikov N et al (2012) Incidental computed tomography sacroiliitis: Clinical significance and inappropriateness of the New York radiological grading criteria for the diagnosis. Clin Rheumatol 31:425–428. https://doi.org/10.1007/s10067-011-1871-6
29. Yushkevich PA, Piven J, Hazlett HC et al (2006) User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31:1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015
30. Zou KH, Warfield SK, Bharatha A et al (2004) Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 11:178–189. https://doi.org/10.1016/S1076-6332(03)00671-8
31. Yu AC, Eng J (2020) One algorithm may not fit all: How selection bias affects machine learning performance. Radiographics 40:1932–1937. https://doi.org/10.1148/rg.2020200040
32. Bluemke DA, Moy L, Bredella MA et al (2020) Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers-from the Radiology Editorial Board. Radiology 294:487–489. https://doi.org/10.1148/radiol.2019192515
33. Selvaraju RR, Cogswell M, Das A et al (2020) Grad-CAM: visual explanations from deep networks via gradient-based localisation. Int J Comput Vis 128:336–359. https://doi.org/10.1007/s11263-019-01228-7
34. Geijer M, Gadeholt Gothlin G, Gothlin JH (2009) The validity of the new york radiological grading criteria in diagnosing

sacroiliitis by computed tomography. Acta Radiol 50:664–673. https://doi.org/10.1080/02841850902914099

35. Bressem KK, Vahldiek JL, Adams L et al (2021) Deep learning for detection of radiographic sacroiliitis: achieving expert-level performance. Arthritis Res Ther 23:1–10. https://doi.org/10.1186/s13075-021-02484-0
36. Castro-Zunti R, Park EH, Choi Y et al (2020) Early detection of ankylosing spondylitis using texture features and statistical machine learning, and deep learning, with some patient age analysis. Comput Med Imaging Graph 82:101718. https://doi.org/10.1016/j.compmedimag.2020.101718
37. Chan J, Sari I, Salonen D et al (2016) Development of a screening tool for the identification of sacroiliitis in computed tomography scans of the abdomen. J Rheumatol 43:1687–1694. https://doi.org/10.3899/jrheum.150939

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.