Radiology

ORIGINAL RESEARCH • MUSCULOSKELETAL IMAGING

Deep Learning Detects Changes Indicative of Axial Spondyloarthritis at MRI of Sacroiliac Joints

Keno K. Bressem, MD • Lisa C. Adams, MD • Fabian Proft, MD • Kay Geert A. Hermann, MD • Torsten Diekhoff, MD • Laura Spiller, MD • Stefan M. Niehues, MD • Marcus R. Makowski, MD • Bernd Hamm, MD • Mikhail Protopopov, MD • Valeria Rios Rodriguez, MD • Hildurn Haibel, MD • Judith Rademacher, MD • Murat Torgutalp, MD • Robert G. Lambert, MD • Xenofon Baraliakos, MD • Walter P. Maksymowych, MD • Janis L. Vahldiek, MD • Denis Poddubnyy, MD**

From the Institute for Radiology (K.K.B., L.C.A., K.G.A.H., T.D., S.M.N., B.H., J.L.V.) and Department of Gastroenterology, Infectious Diseases and Rheumatology (including Nutrition Medicine) (F.P., L.S., M.P., V.R.R., H.H., J.R., M.T., D.P.), Charité–Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Hindenburgdamm 30, 12203 Berlin, Germany; Berlin Institute of Health at Charité–Universitätsmedizin Berlin, Berlin, Germany (K.K.B., L.C.A., J.R.); Department of Diagnostic and Interventional Radiology, Faculty of Medicine, Technical University of Munich, Munich, Germany (M.R.M.); Department of Medicine, University of Alberta, Edmonton, Alberta, Canada (R.G.L., W.P.M.); Rheumazentrum Ruhrgebiet Herne, Ruhr University Bochum, Germany (X.B.); and Epidemiology Unit, German Rheumatism Research Centre, Berlin, Germany (D.P.). Received October 15, 2021; revision requested December 7; revision received April 25, 2022; accepted June 2. **Address correspondence to** K.K.B. (email: *keno-kyrill.bressem@charite.de*).

The German Spondyloarthritis Inception Cohort (GESPIC)–Ankylosing Spondylitis has been financially supported by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung-BMBF) through the ArthroMark project (FKZ 01EC1401A). GESPIC-Crohn has been supported by the Clinical Research Unit grant from the Berlin Institute of Health. GESPIC-Uveitis has been supported by a research grant from AbbVie. Optimal Referral Strategy for Early Diagnosis of Axial Spondyloarthritis (OptiRef) has been supported by a research grant from Novartis. The Assessment of SpondyloArthritis international Society (ASAS) has supported the project with a research grant and provided access to the MRI scans from the ASAS classification cohort.

* J.L.V. and D.P. are co–senior authors.

Conflicts of interest are listed at the end of this article.

Radiology 2022; 000:1–11 • https://doi.org/10.1148/radiol.212526 • Content codes: MK MR

Background: MRI is frequently used for early diagnosis of axial spondyloarthritis (axSpA). However, evaluation is time-consuming and requires profound expertise because noninflammatory degenerative changes can mimic axSpA, and early signs may therefore be missed. Deep neural networks could function as assistance for axSpA detection.

Purpose: To create a deep neural network to detect MRI changes in sacroiliac joints indicative of axSpA.

Materials and Methods: This retrospective multicenter study included MRI examinations of five cohorts of patients with clinical suspicion of axSpA collected at university and community hospitals between January 2006 and September 2020. Data from four cohorts were used as the training set, and data from one cohort as the external test set. Each MRI examination in the training and test sets was scored by six and seven raters, respectively, for inflammatory changes (bone marrow edema, enthesitis) and structural changes (erosions, sclerosis). A deep learning tool to detect changes indicative of axSpA was developed. First, a neural network to homogenize the images, then a classification network were trained. Performance was evaluated with use of area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. $P < .05$ was considered indicative of statistically significant difference.

Results: Overall, 593 patients (mean age, 37 years ± 11 [SD]; 302 women) were studied. Inflammatory and structural changes were found in 197 of 477 patients (41%) and 244 of 477 (51%), respectively, in the training set and 25 of 116 patients (22%) and 26 of 116 (22%) in the test set. The AUCs were 0.94 (95% CI: 0.84, 0.97) for all inflammatory changes, 0.88 (95% CI: 0.80, 0.95) for inflammatory changes fulfilling the Assessment of SpondyloArthritis international Society definition, and 0.89 (95% CI: 0.81, 0.96) for structural changes indicative of axSpA. Sensitivity and specificity on the external test set were 22 of 25 patients (88%) and 65 of 91 patients (71%), respectively, for inflammatory changes and 22 of 26 patients (85%) and 70 of 90 patients (78%) for structural changes.

Conclusion: Deep neural networks can detect inflammatory or structural changes to the sacroiliac joint indicative of axial spondyloarthritis at MRI.

Online supplemental material is available for this article.

MRI is an important tool for the early diagnosis of axial spondyloarthritis (axSpA) and can reduce the time to diagnosis by depicting early inflammatory changes (1–3). Subchondral bone marrow edema in the sacroiliac joints is a characteristic imaging feature of spondyloarthritis and a required finding to fulfill the Assessment of SpondyloArthritis international Society (ASAS) criterion of positive MRI findings as a part of the classification criteria for axSpA. Other active inflammatory changes (eg, enthesitis, capsulitis, joint space enhancement, inflammation at the site of erosion, and joint space fluid) may also be present. Structural damage manifests as joint erosions, subchondral sclerosis, bone buds, ankylosis, fat metaplasia in an erosion cavity, or fatty lesions in bone (1,4). However, many of these signs individually are not specific for axSpA and may also occur in degenerative diseases (5,6).

Contextual interpretation of findings is therefore required for correct evaluation of MRI scans, which can be demanding, especially for radiologists or clinicians not specialized in axSpA (7,8). For example, in the French

Abbreviations

ASAS = Assessment of SpondyloArthritis international Society, AUC = area under the receiver operating characteristic curve, axSpA = axial spondyloarthritis, GESPIC = German Spondyloarthritis Inception Cohort

Summary

A deep learning tool was able to detect active inflammatory and structural changes indicative of axial spondyloarthritis at MRI of sacroiliac joints.

Key Results

- This retrospective study comprised 593 patients with suspected axial spondyloarthritis and centrally evaluated MRI examinations of the sacroiliac joints.
- The sensitivity and specificity for the detection of active inflammatory and structural changes were 96% and 76% and 95% and 75%, respectively, in the validation set and 88% and 71% and 85% and 78% in the test set.
- The areas under the receiver operating characteristic curve for the detection of active inflammatory and structural changes were comparable in the validation (0.92 and 0.90, respectively) and test (0.94 and 0.89, respectively) data sets (P = .38 and .13).

Cohort of Undifferentiated Spondyloarthritis (Devenir des Spondylarthropathies Indifférenciées Récentes), or DESIR, disagreement between two investigators occurred in 28% of MRI scans (9). Overall, these diagnostic challenges may lead to over- and underdiagnosis of axSpA. In addition, comprehensive reading of MRI data sets with use of a quantitative scoring system is a time-consuming process. Therefore, supportive tools that can help interpret MRI scans of the sacroiliac joints in patients with suspected axSpA are needed. In this context, deep learning has the potential to provide supportive tools to radiologists and clinicians (10).

Deep convolutional neural networks have previously been applied to MRI data, such as for classification of multisequence MRI knee scans (11), prostate cancer detection (12), automated contouring of tumor volumes in nasopharyngeal carcinoma (13), and brain lesion segmentation (14). The use of convolutional neural networks to detect changes associated with axSpA therefore appears promising. However, training convolutional neural networks, especially with MRI data from different institutions, can be challenging because of differences in sequence specifications, imaging protocols, reconstruction algorithms, and spatial resolution among the different types of scanners. In addition, MRI examinations can be corrupted by artifacts caused by patient motion, signal instabilities, partial failure of fat suppression, or magnetic field inhomogeneities (15).

Our goal was to develop a deep neural network to detect inflammatory and structural changes in sacroiliac joints at MRI indicative of axSpA and to overcome the aforementioned challenges.

Materials and Methods

Study Sample

This retrospective multicenter study used MRI data from three multicenter studies: *(a)* the German Spondyloarthritis Inception Cohort (GESPIC), an ongoing German multicenter inception cohort study started in 2000 that includes university and community hospitals in Germany (16). Patients included in the current study were recruited between February 2015 and September 2020 at Charité–Universitätsmedizin Berlin. *(b)* Optimal Referral Strategy for Early Diagnosis of Axial Spondyloarthritis (or OptiRef), a cross-sectional study conducted at Charité–Universitätsmedizin Berlin between October 2016 and February 2018. *(c)* The ASAS classification cohort, a multinational and multicenter study conducted at university and community hospitals between November 2005 and October 2009. Patients included in the current study were recruited between January 2006 and December 2008.

All MRI scans were centrally evaluated between June 2019 and June 2021.

The inclusion criterion was the availability of semicoronal MRI scans of the sacroiliac joint with T1-weighted sequences and fluid-sensitive fat-suppressed sequences (Fig 1).

The GESPIC (Ankylosing Spondylitis, Uveitis, and Crohn groups) and OptiRef studies were approved by the ethics committee of Charité–Universitätsmedizin Berlin; the ASAS classification study was approved by the ethics committees of the individual study centers in accordance with the local laws and regulations. All studies were conducted in accordance with the Declaration of Helsinki and Good Clinical Practice. Appendixes E1 and E2 (online) provide further details.

MRI Sequence Parameters

Most examinations in the training set were performed with a 3.0-T MRI scanner (Skyra, Siemens Healthineers). The most common T1-weighted sequence was a turbo spin-echo sequence with repetition time of 652 msec, echo time 11 msec, and echo train length of four. The most used fluid-sensitive sequence with fat suppression was the fast spin-echo short tau inversion-recovery sequence with repetition time between 5000 and 7760 msec, echo time of 230 msec, inversion time of 200 msec, and echo train length of 12. Appendix E3 and Tables E1–E7 (online) provide more details.

Data Labeling

The training set included MRI scans of the sacroiliac joints (Fig 1) from the GESPIC–Ankylosing Spondylitis, GESPIC-Crohn, GESPIC-Uveitis, and OptiRef cohorts. Data from 73 of 477 patients in the training set (15%) were randomly selected as the validation set. Each MRI scan was evaluated by six trained raters (F.P., T.D., V.R.R., J.R., M.T., and D.P.) with 5–15 years of experience in axSpA. Raters were blinded to clinical data, as they evaluated pseudonymized versions of the MRI scans and did not have access to the other raters' assessments. This was ensured by the use of a specifically developed online tool. MRI scans from the ASAS classification cohort (4) were used as the external test set. MRI scans were evaluated by seven experienced raters (R.G.L., X.B., W.P.M., and four other raters) with more than 15 years of experience, as described previously (4).

All raters evaluated whether active inflammatory changes indicative of axSpA were present. If so, they evaluated specific changes, such as bone marrow edema fulfilling the ASAS definition, capsulitis, joint space enhancement, inflammation, and enthesitis. The same strategy was used for structural changes: the

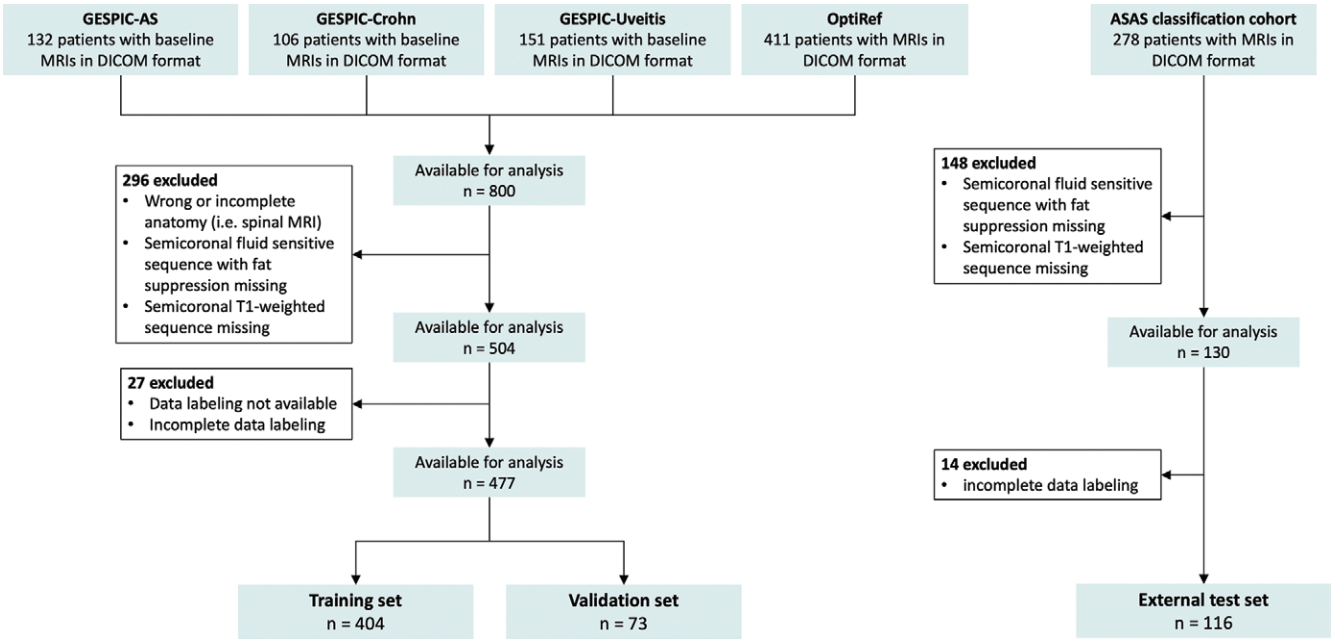

Figure 1: Data selection flowchart shows the process of data selection and the creation of the training set, validation set, and external test set. AS = ankylosing spondylitis, ASAS = Assessment of SpondyloArthritis international Society, DICOM = Digital Imaging and Communications in Medicine, GESPIC = German Spondyloarthritis Inception Cohort, OptiRef = Optimal Referral Strategy for Early Diagnosis of Axial Spondyloarthritis.

raters first assessed the presence or absence of structural changes compatible with axSpA globally, then specific changes (erosions, fat lesion, fat metaplasia in the erosion cavity, sclerosis, ankylosis, and bone buds).

In all sets, the reference standard for the global presence or absence of active inflammatory and structural lesions indicative for axSpA as well as global compatibility with the diagnosis of axSpA was defined as agreement among at least four raters. In the training set, if there was no majority decision, images were assessed in a consensus rater session.

To assess the performance of nonexpert raters, three board-certified radiologists (K.K.B., L.C.A., and J.L.V., with 5, 6, and 11 years of experience, respectively) with no special training in axSpA independently reviewed each MRI scan in the test set. Votes of these raters were not included in the ground truth. Appendix E4 (online) provides more details.

Model Training

The use of multiple scanners in this study introduced heterogeneity into the data, complicating the training of neural networks. Therefore, we first developed a three-dimensional U-Net architecture for MRI denoising, artifact reduction, and homogenization of intensity distribution between MRI scans (Fig 2). We then trained a three-dimensional dual-encoder residual neural network 101 to classify active inflammatory changes or structural changes indicative of axSpA (Fig 3). Gradient-weighted class activation mappings were used to provide visual explanations for the models' decision. The inference time (ie, the combined time of preprocessing and prediction) for the models was measured during testing. Appendixes E5–E7 (online) provide a detailed description of the methods. All code is available at *https://github.com/kbressem/spa*.

Statistical Analysis

Statistical analysis was performed using R (version 4.0.4, the R Foundation) and the "tidyverse," "irr," "boot," and "pROC" libraries (17–19). The classification models were evaluated on the validation and test sets. Receiver operating characteristic curves and the respective areas under the receiver operating characteristic curve (AUCs) were calculated. Sensitivity, specificity, and accuracy were calculated on the validation and test sets with use of an identical cutoff (previously calculated for the validation set). The 95% CIs for the AUCs were estimated by bootstrapping, and the CIs for metrics were calculated with use of the Clopper-Pearson test. Interrater agreement and agreement between models and raters were assessed using the Fleiss κ statistic. Bootstrapping was applied to compare two AUCs. The Fisher test was used to compare sensitivity and specificity and prevalence. $P < .05$ was considered indicative of statistically significant difference.

Results

Data Set Characteristics

The data set consisted of 389 patients from the three GESPIC groups, 411 patients from the OptiRef cohort, and 278 patients from the ASAS classification cohort. After exclusion of patients without the required MRI sequences (296 in the training and validation sets and 148 in the test set) and with incomplete labeling (27 in the training and validation sets and 14 in the test set), 369 patients (mean age, 38 years ± 12 [SD]; 194 men) from GESPIC, 108 patients (36 years ± 11; 57 women) from the OptiRef cohort, and 116 patients (36 years ± 9; 70 women) from the ASAS cohort were included in our analysis.

Active inflammatory changes consistent with axSpA were present in 197 of 477 patients (41%) in the training and validation

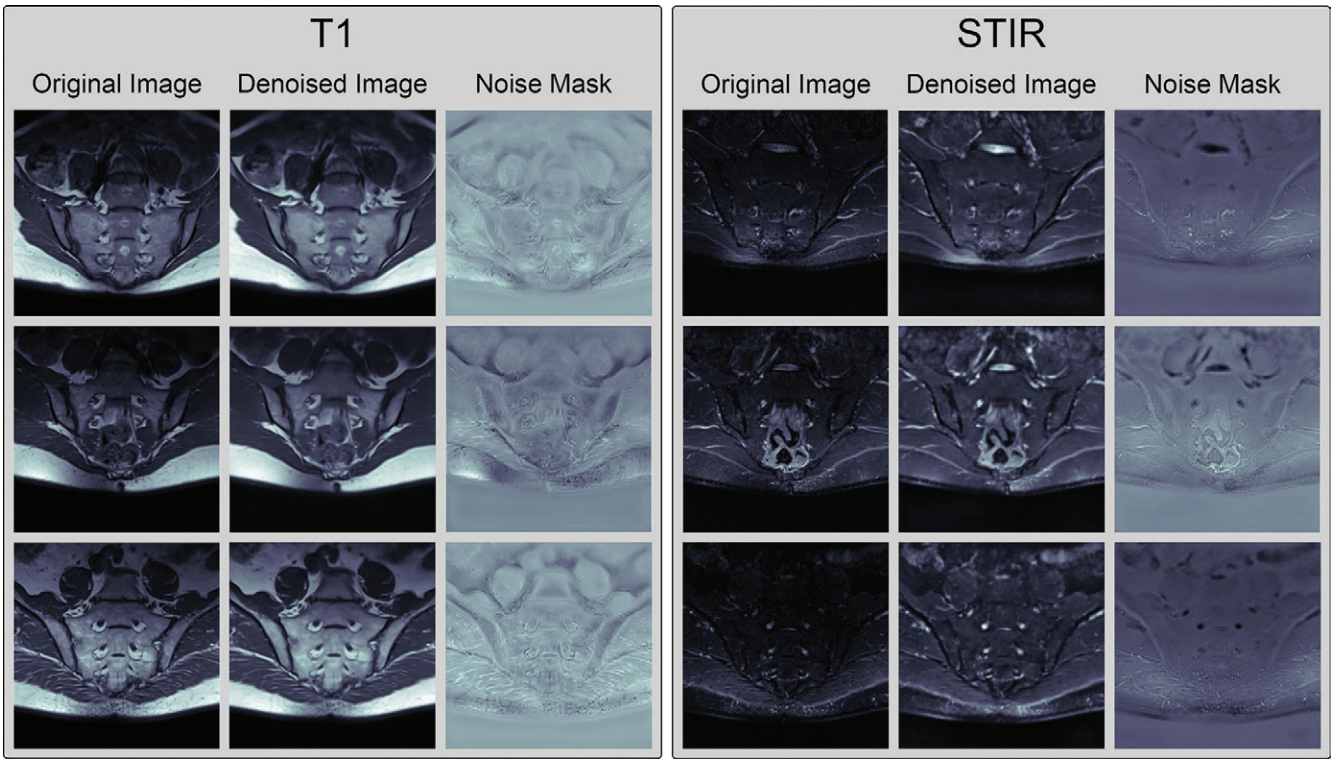

Figure 2: Results of the denoising U-Net. Examples of fluid-sensitive sequences with fat suppression and T1-weighted paracoronal images illustrate the effect of denoising by the U-Net. Original image refers to the raw image, which was resampled only to make the size and spacing uniform. Denoised image refers to the images created by the U-Net. The noise mask is the visualization of the image noise subtracted from the original image to create the denoised image. STIR = short tau inversion recovery.

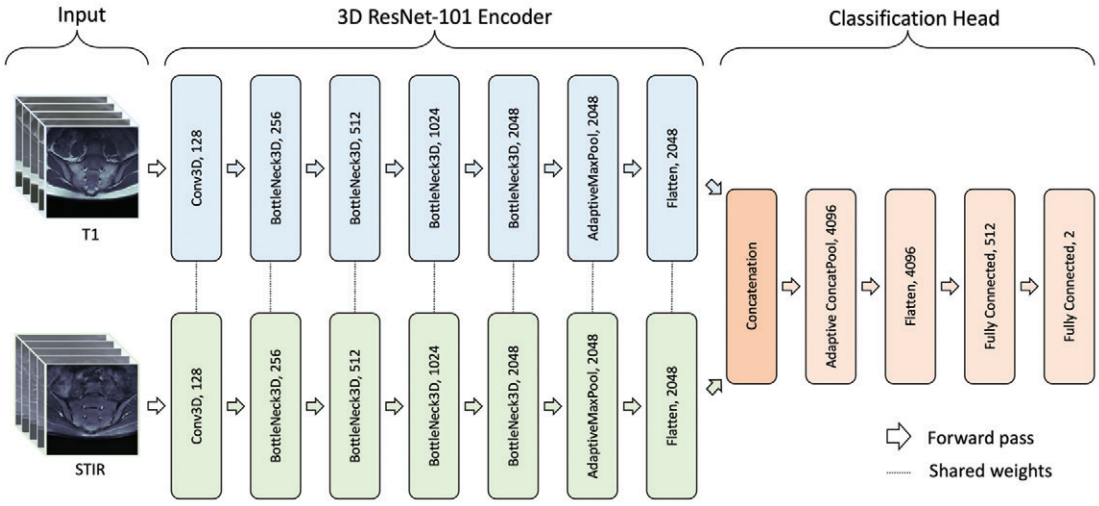

Figure 3: Diagram of the architecture of the classification models used. The input sequences are processed by two three-dimensional (3D) residual neural network 101 (ResNet-101) encoders that share their weights. The output of each encoder is then concatenated, pooled, and passed to the common classification head. The classification head consists of an adaptive concatenation pooling layer and two subsequent fully connected layers, each with batch normalization and dropout. Conv = convolutional layer, STIR = short tau inversion recovery.

sets, and 172 of 477 (36%) fulfilled the ASAS definition. In the test set, the prevalence was significantly lower, with 25 of 116 patients (22%) having inflammatory changes; 21 (18%) fulfilled the ASAS definition ($P = .03$). The prevalence of structural changes was also significantly different, with positive findings in 244 of 477 patients (51%) in the training and validation sets and in 26 of 116 patients (22%) in the test set ($P < .001$). The Table provides a detailed overview of the prevalence of active inflammatory and structural changes as well as characteristics of the patients included.

Interrater Agreement

No majority consensus was achieved in 51 of 477 MRI scans (11%) in the training and validation sets (15 of 125 scans [12%]

Baseline Characteristics of Patients in Different Study Cohorts after Data Selection

Characteristic	GESPIC-AS (*n* = 125)	GESPIC-Crohn (*n* = 102)	GESPIC-Uveitis (*n* = 142)	OptiRef (*n* = 108)	ASAS Cohort (*n* = 116)
Age (y)*	36 ± 10 (33 [28–45]) {19–67}	37 ± 13 (35 [27–47]) {15–73}	41 ± 13 (39 [31–50]) {19–72}	36 ± 11 (34 [28–43]) {16–57}	36 ± 9 (36 [21–41]) {21–71}
Sex					
M	79 (63)	47 (46)	68 (48)	51 (47)	46 (40)
F	46 (37)	55 (54)	74 (52)	57 (53)	70 (60)
Back pain, current	121 (97)	47 (46)	97 (68)	108 (100)	99 (85)
Inflammatory back pain, current	119 (95)	21 (21)	67 (47)	76 (70)	64 (55)
Duration of back pain (y)†‡	11 ± 9.7	8 ± 8	13 ± 10	7 ± 7	8 ± 7
Peripheral arthritis, current	21 (17)	3 (3)	9 (6)	4 (4)	43 (37)
Enthesitis, current	40 (32)	8 (8)	5 (4)	10 (9)	33 (28)
Dactylitis, current	0 (0)	1 (1)	0 (0)	0 (0)	8 (7)
Inflammatory bowel disease, ever	9 (7)	102 (100)	3 (2)	2 (2)	8 (7)
Acute anterior uveitis, ever	27 (22)	12 (12)	142 (100)	12 (11)	20 (17)
Psoriasis, ever	18 (14)	6 (6)	15 (11)	11 (10)	17 (15)
Family history of SpA	49 (39)	30 (29)	16 (11)	15 (14)	32 (28)
Human leukocyte antigen B27 positivity	109 (87)	12 (12)	111 (78)	64 (59)	47 (41)
C-reactive protein (mg/L)‡	13.1 ± 17.7	10.2 ± 25.4	4.7 ± 7.6	4.0 ± 6.5	3.3 ± 13.8
Diagnosed with axial spondyloarthritis	125 (100)	10 (10)	74 (52)	57 (53)	89 (77)
Active inflammatory changes indicative of SpA at MRI of sacroiliac joints	86 (67)	11 (11)	57 (40)	43 (40)	25 (22)
Active sacroiliitis at MRI according to the ASAS definition	75 (60)	11 (11)	47 (34)	39 (37)	21 (18)
Structural changes indicative of SpA at MRI of sacroiliac joints	124 (99)	13 (13)	64 (45)	43 (40)	26 (22)

Note.—Unless otherwise specified, data are numbers of patients, with percentages in parentheses. The training and validation sets consisted of data from the German Spondyloarthritis Inception Cohort (GESPIC) and Optimal Referral Strategy for Early Diagnosis of Axial Spondyloarthritis (OptiRef) cohort. The test set consisted only of data from the Assessment of SpondyloArthritis international Society (ASAS) cohort. AS = ankylosing spondylitis, SpA = spondyloarthritis.

* Data are means ± SDs, with the medians in parentheses, IQRs in brackets, and ranges in curly braces.

† In patients with back pain.

‡ Data are means ± SDs.

in GESPIC–Ankylosing Spondylitis, five of 102 scans [5%] in GESPIC-Crohn, 18 of 142 scans [13%] in GESPIC-Uveitis, and 13 of 108 scans [12%] in OptiRef), and these scans were adjudicated in a consensus reading session.

Overall, the raters achieved substantial agreement in both data sets evaluated. In the validation set, Fleiss κ values of 0.62 (95% CI: 0.50, 0.74) were achieved for active inflammatory changes, 0.61 (95% CI: 0.50, 0.71) for changes fulfilling the ASAS definition, and 0.71 (95% CI: 0.61, 0.80) for structural changes. In the test set, the agreement of the seven raters as measured by the Fleiss κ statistic was 0.63 (95% CI: 0.53, 0.74) for active inflammatory changes, 0.65 (95% CI: 0.54, 0.77) for the detection of changes compatible with the ASAS definition, and 0.73 (95% CI: 0.64, 0.83) for structural changes.

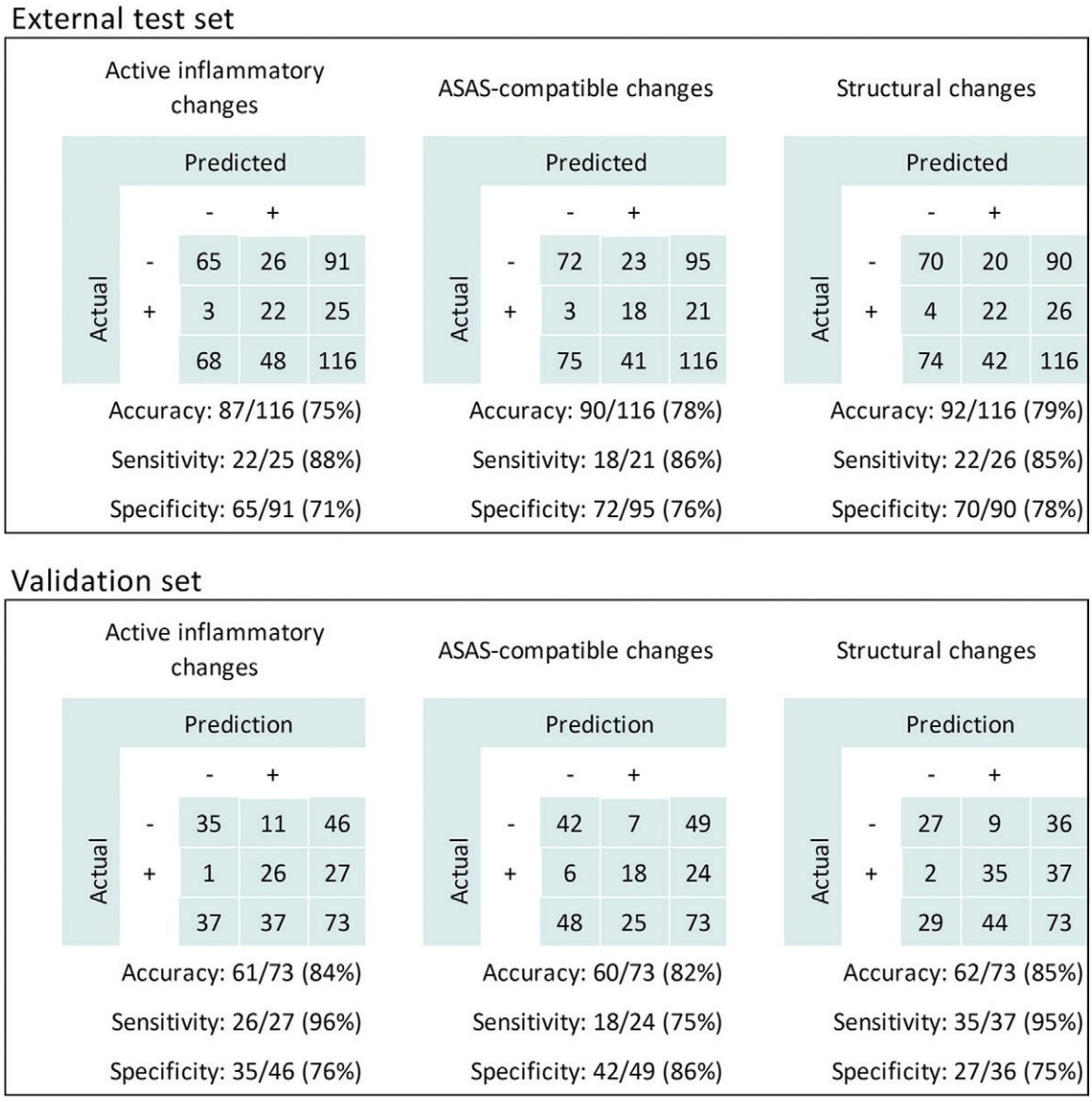

Figure 4: Confusion matrices for the performance of the classification models on the test and validation sets for each class compared with the reference standard (consensus of at least four experienced raters). Accuracy, sensitivity, and specificity of individual raters are reported in Tables E8–E10 (online). Analysis of discordant MRI scans for active inflammatory changes and ASAS-compatible changes showed that imaging artifacts, such as field bias and partial failure of fat suppression, might have led to some false-positive predictions ($n = 8$). The loss of detail after denoising could have led to the fact that small erosions were no longer delimitable, leading to false-negative predictions for structural changes. In some MRI scans ($n = 3$), false-positive predictions for structural changes also seemed to be influenced by degenerative changes.

Detection of Active Inflammatory Changes Indicative of axSpA

Compared with the reference standard (majority decision of all experienced raters), the model for the detection of active inflammatory changes showed an AUC of 0.92 (95% CI: 0.83, 0.97) and an accuracy of 61 of 73 patients (84% [95% CI: 73, 91]) in the validation set. In the test set, there was no significant difference in performance, with an AUC of 0.94 (95% CI: 0.84, 0.97) and an accuracy of 87 of 116 patients (75% [95% CI: 66, 83]; $P = .38$), despite a significantly lower prevalence ($P = .03$).

In the validation set, the model had a sensitivity of 26 of 27 patients (96% [95% CI: 81, 100]) and specificity of 35 of 46 patients (76% [95% CI: 61, 87]) for detecting active inflammatory changes. In the test set, the model sensitivity was 22 of 25 patients (88% [95% CI: 69, 97]) and specificity was 65 of 91 patients (71% [95% CI: 61, 80]).

For the detection of active inflammatory changes fulfilling the ASAS definition (bone marrow edema), the AUC and accuracy on the validation set were 0.86 (95% CI: 0.73, 0.90) and 60 of 73 patients (82% [95% CI: 71, 90]), respectively. In the test set, the model showed an AUC of 0.88 (95% CI: 0.80, 0.95) and an accuracy of 90 of 116 patients (78% [95% CI: 69, 85]; $P = .75$). The model's sensitivity in the test set was similar at 18 of 21 patients (86% [95% CI: 64, 97]; $P = .25$), with a corresponding specificity of 72 of 95 patients (76% [95% CI: 68, 86]; $P = .19$).

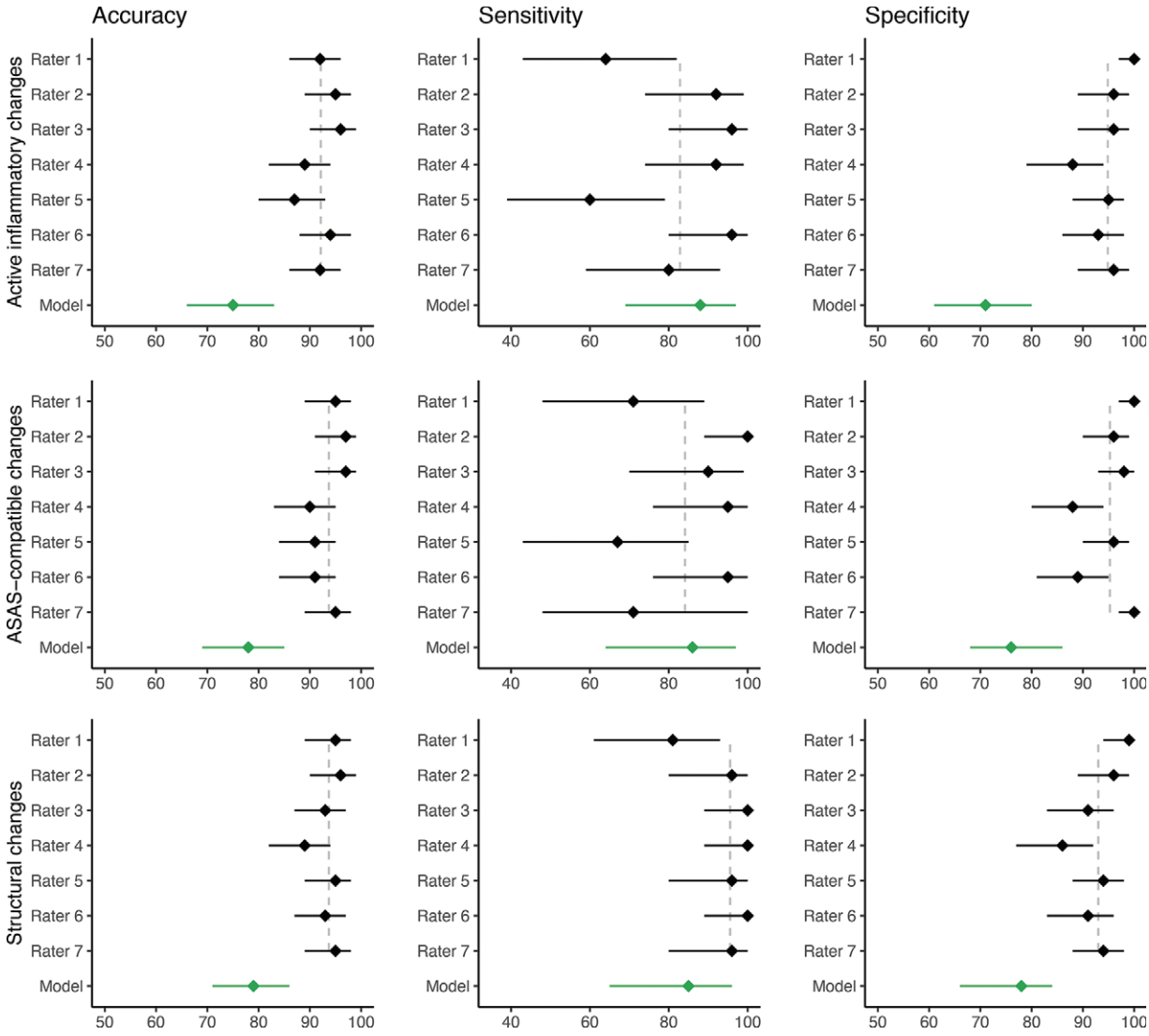

Figure 5: Diagnostic performance of individual raters and models. Dot plots show accuracy, sensitivity, and specificity (as percentages) for all seven raters on the test set and the performance of the classification models alongside 95% CIs (horizontal lines). The dashed lines indicate the mean of all raters. Individual values alongside *P* values for significance of differences between the human raters and the model are available in Table E8 (online). ASAS = Assessment of SpondyloArthritis international Society.

The confusion matrices of the diagnostic performance of the models for the test and validation groups are shown in Figure 4. Figure 5 compares the performance of the individual raters with that of the models. Figure 6 provides receiver operating characteristic curves of the model performance on the test and validation sets along with sensitivity and specificity estimates for the experienced human raters. Additional information on individual accuracy for each rater is given in Tables E8 and E9 (online).

Figures 7 and E1 (online) indicate example gradient-weighted class activation mappings to visualize the models' decisions and highlight image regions relevant to model predictions.

Detection of Structural Changes Indicative of axSpA

In the validation set, the model had an AUC of 0.90 (95% CI: 0.82, 0.96) for the detection of structural changes and an overall accuracy of 62 of 73 patients (85% [95% CI: 75, 92]). The associated sensitivity and specificity were 35 of 37 patients (95% [95% CI: 82, 99]) and 27 of 36 patients (75% [95% CI: 58, 94]), respectively.

In the external test data, the AUC and accuracy were 0.89 (95% CI: 0.81, 0.96) and 92 of 116 patients (79% [95% CI: 71, 86]), respectively. Compared with the validation set, sensitivity and specificity were similar at 22 of 26 patients (85% [95% CI: 65, 96]; $P = .22$) and 70 of 90 patients (78% [95% CI: 66, 84]; $P = .82$), respectively.

Performance of Nonexpert Raters

The three board-certified radiologists who were not trained in axSpA had a mean sensitivity and specificity of 21 of 25 patients (83% [95% CI: 63, 91]) and 76 of 91 patients (84% [95% CI: 75, 91]), respectively, for detecting active inflammatory changes, 20 of 25 (79% [95% CI: 56, 93]) and 80 of 91 (87% [95% CI: 79, 93]) for active inflammatory changes

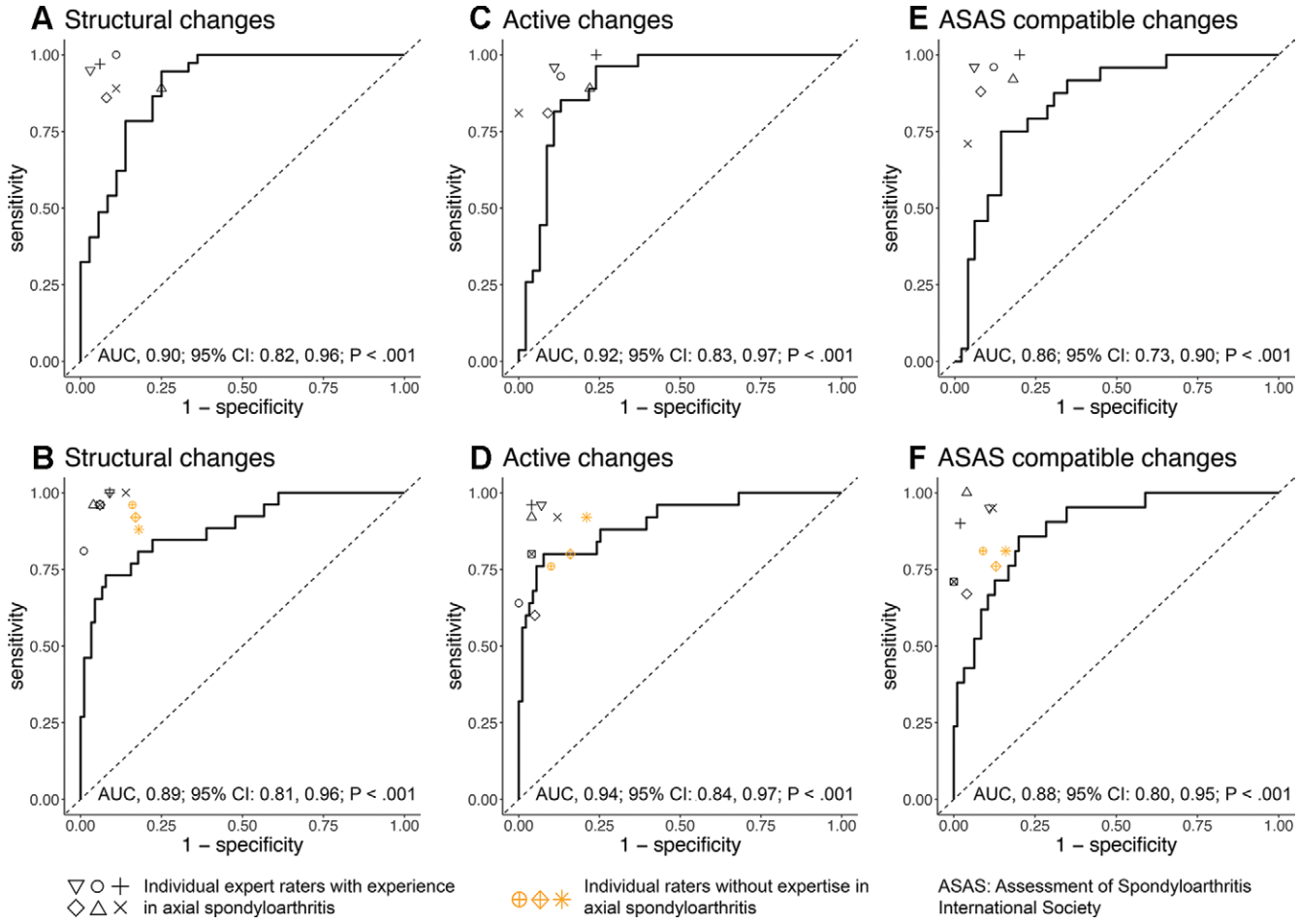

Figure 6: **(A–F)** Receiver operating characteristics curves and associated areas under the receiver operating characteristic curves (AUCs) for model performance together with estimates of the diagnostic accuracies compared with the individual human experts. **A**, **C**, and **E** represent the receiver operating characteristic curves for diagnostic performance in the validation set, while the curves for the test set are given in **B**, **D**, and **F**. The model performance did not exceed that of the trained experts but came close, especially for the detection of active inflammatory changes. Three board-certified radiologists who did not undergo specific training for axial spondyloarthritis imaging also scored each examination in the test set. Their performance is indicated by the orange symbols in **B**, **D**, and **F**.

fulfilling the ASAS definition, and 23 of 25 (92% [95% CI: 75, 90]) and 76 of 91 (83% [95% CI: 74, 90]) for structural changes indicative of axSpA. For active inflammatory changes, the model showed similar sensitivity (22 of 25 patients, 88% [95% CI: 69, 97]; P = .34) to that of the nonexpert radiologists. Details on the individual performance of the radiologists can be obtained from Table E10 (online).

Approximation of Inference Time

Image reading and preprocessing (resampling, bias correction) took an average of 16.2 seconds ± 0.0567 (SD) per image. Denoising with the U-Net took 1.25 seconds ± 0.0151 per MRI scan but had to be applied to T1-weighted MRI scans and fluid-sensitive, fat-suppressed MRI scans individually. The average prediction time for each MRI scan in the entire data set was 7.19 seconds ± 0.0673 on the validation set and 10.5 seconds ± 0.108 on the test set (approximately 100 msec per item). Overall, the estimated inference time for loading, preprocessing, and prediction for a single examination was 18.9 seconds.

Discussion

Interpretation of MRI scans associated with axial spondyloarthritis (axSpA) requires expertise in characteristic findings of the disease and is time-consuming. We developed a deep learning tool for detecting changes indicative of axSpA at MRI, which was evaluated against the performance of seven international experienced raters. The deep learning tool showed a sensitivity of 88% for detecting inflammatory changes indicative of axSpA.

MRI allows the detection of inflammatory changes at an earlier stage than radiography and is thus increasingly used (20,21). However, correctly identifying and interpreting changes to the sacroiliac joint as characteristic of axSpA can be challenging, especially for nonexperts, who have a lower reliability than experienced readers (8). One reason for this may be the overlap between inflammatory and degenerative findings. For example, bone marrow edema may occur in axSpA due to mechanical stress or in osteitis condensans ilii, making it difficult for nonexperts to make an accurate diagnosis (22,23). In our study, we also observed that radiologists without expertise in axSpA showed a lower diagnostic performance than domain experts did.

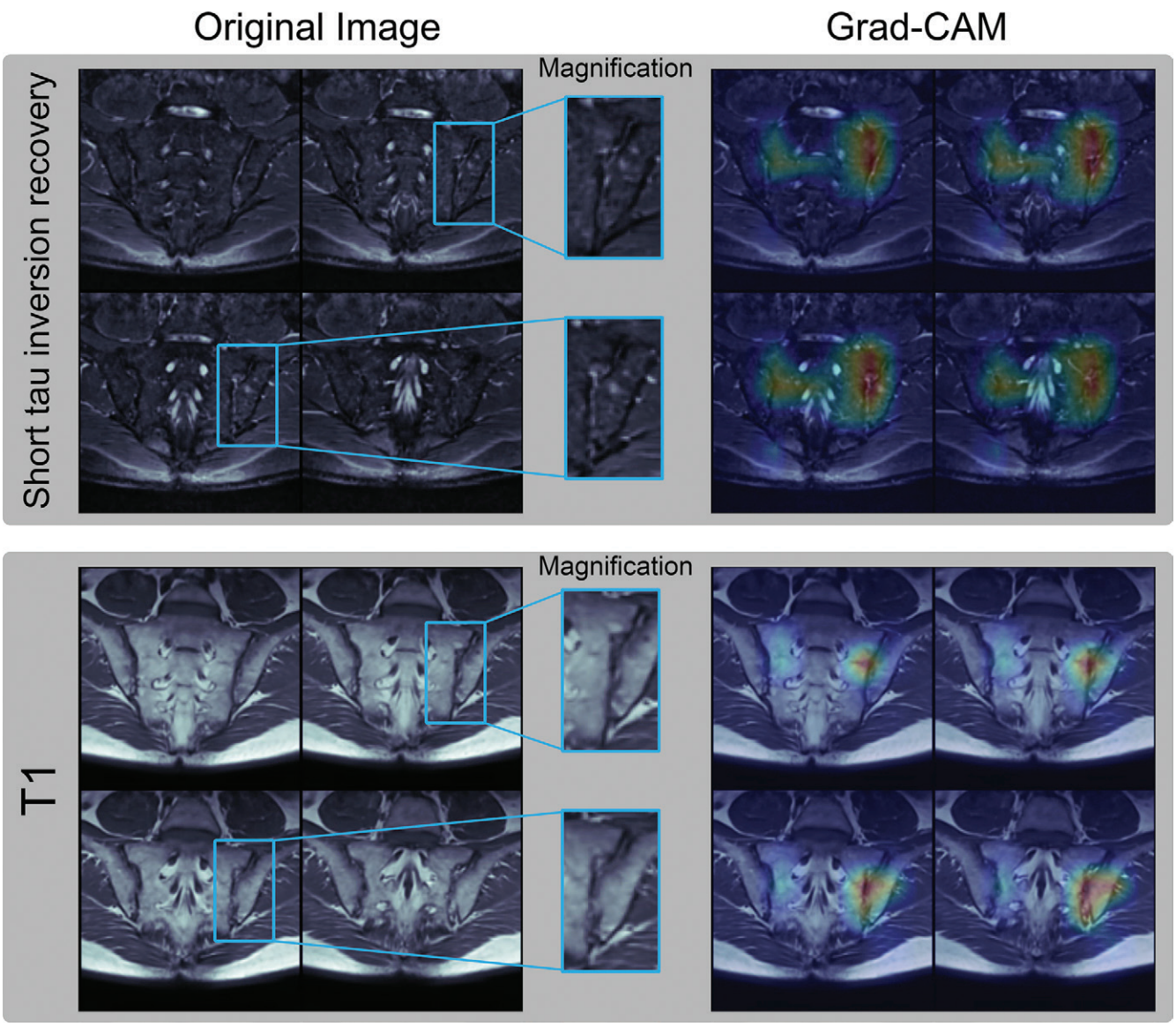

Figure 7: Example of gradient-weighted class activation mapping (Grad-CAM) for the classification model. In this MRI scan in a 41-year-old woman with confirmed axial spondyloarthritis, the model correctly predicted the presence of structural changes, active inflammatory changes, and Assessment of SpondyloArthritis international Society–compatible changes. The scan consisted of a semicoronal T1-weighted and a semicoronal short tau inversion-recovery sequence. Changes are best seen in the region of the left sacroiliac joint, where the model also showed the strongest activations. The magnification is approximately ×2.

A strength of our study is the inclusion of MRI scans acquired on different machines with different settings (thus reflecting real clinical practice), the central standardized evaluation of images by experts, and the use of an external test set. External test sets allow for more realistic evaluation of deep learning models, as recently emphasized in two meta-analyses by Liu et al (24) and Kim et al (25). Our test set had a significantly lower prevalence of axSpA, with the use of different scanners and the deployment of different raters. Nevertheless, our developed models generalized to the new data without a significant decrease in performance.

The generalizability of our models could be due to the large heterogeneity of the training data. Mårtensson et al (26) recently investigated the importance of heterogeneous training. A criticism they presented was that training with curated research data does not represent a realistic clinical setting and will lead to lower performance when the models are applied to data acquired with different scanners and protocols. However, heterogeneous data can also complicate training and require more extensive preprocessing to normalize the images, which is time-consuming. Ran et al (27) and Chauhan and Choi (28) have therefore proposed the use of neural networks to preprocess MRI scans, similar to the approach presented in our study. Our study demonstrates the effectiveness of this approach and the positive impact on generalizability, as shown by the ablation presented in Appendix E8 (online).

Our study has several limitations. First, the consensus of multiple experts was used as the reference standard because there is no true reference standard for the presence of axSpA. Nevertheless, the consensus of multiple readers can still be wrong, possibly introducing noise into the data set and affecting model performance. Second, the prevalence of axSpA in the test set was low, which may introduce uncertainty in the assessment of

performance. Third, in GESPIC-Uveitis and OptiRef, MRI was performed in only a subset of patients, which could introduce selection bias. Fourth, our models were trained with semicoronal images only, so different orientations could lead to model failure. Fifth, we chose global labels for model training and did not provide a quadrant-based analysis of the sacroiliac joints, which would have allowed a more spatially accurate assessment of different joint regions. Sixth, the diagnostic performance of the models is not evidence of their clinical utility. Further trials are necessary to evaluate if the use of the models translates into a benefit for patients. Finally, because of the variety of scanners and protocols used, we were unable to provide imaging parameters for all MRI scans, which limits the reproducibility of our data. Nevertheless, we believe that the approach used in our study will generalize to new scanner protocols, as it generalized to the test set.

In conclusion, a deep learning tool was developed for the detection of axial spondyloarthritis (axSpA)–associated abnormalities at MRI. The deep learning tool could help clinicians detect inflammation earlier and properly to initiate appropriate treatment in patients with axSpA. In addition, it could serve as a classification tool in clinical trials. Because accurate diagnosis of axSpA depends on experience, our tool could be particularly helpful for hospitals without specialization in axSpA. However, future research is needed to evaluate the clinical value of the gain in accuracy with our deep learning tool and the impact on therapy, ideally in the form of a prospective study.

Acknowledgments: We thank our colleagues who performed annotation of the images from the ASAS classification cohort: Pedro Machado, MD, PhD; Mikkel Østergaard, MD; Susanne Juhl Pedersen, MD, PhD; and Ulrich Weber, MD. Further, we thank Torsten Karge, Dipl Wi-Ing, for the development of the MRI reading interface for GESPIC and OptiRef images and Joel Paschke, BSc, for the development of the scoring interface for ASAS images. L.C.A. is grateful for her participation in the Berlin Institute of Health (BIH) Charité–Junior Clinician and Clinician Scientist Program, and K.K.B. is grateful for his participation in the BIH Charité–Digital Clinician Scientist Program, all funded by the Charité–Universitätsmedizin Berlin and the BIH. J.R. is grateful for her participation in the BIH Charité–Junior Clinician and Clinician Scientist Program.

Author contributions: Guarantors of integrity of entire study, **K.K.B., D.P.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **K.K.B., L.C.A., K.G.A.H., S.M.N., J.L.V., D.P.**; clinical studies, **F.P., K.G.A.H., T.D., L.S., M.P., M.T., J.L.V., D.P.**; experimental studies, **K.K.B., S.M.N., J.L.V.**; statistical analysis, **K.K.B., J.L.V., D.P.**; and manuscript editing, **K.K.B., L.C.A., F.P., K.G.A.H., T.D., S.M.N., M.R.M., B.H., M.P., V.R.R., X.B., W.P.M., J.L.V., D.P.**

Disclosures of conflicts of interest: **K.K.B.** No relevant relationships. **L.C.A.** No relevant relationships. **F.P.** Grants to institution from UCB, Novartis, and Lilly; consulting fees from AbbVie; Celgene, Janssen, Novartis, and UCB; payment for lectures from Amgen, AbbVie, Bristol Myers Squibb, Celgene, Janssen, MSD, Novartis, Pfizer, Roche, and UCB; support for attending meetings or travel from Celgene; Janssen, Pfizer, and Novartis; participation on a data safety monitoring board or advisory board for AbbVie, Celgene, Janssen, Novartis, and UCB; leadership or fiduciary role with ASAS, Y-ASAS, GRAPPA, Y-GRAPPA, EULAR/EMEUNET, DGRh/AGJR, DEGUM, and Rheumazentrum Berlin; receipt of research material from Aidan. **K.G.A.H.** Consulting fees from AbbVie; payment for lectures from MSD, Pfizer, and Novartis; cofounder of BerlinFlame. **T.D.** Grant from the Berlin Institute of Health; payment for lectures from Novartis, MSD, Canon Medical Systems, and AbbVie; participation in an educational program for Lilly. **L.S.** No relevant relationships. **S.M.N.** Grants from the Berlin Institute of Health and Deutsche Forschungsgemeinschaft; payment for lectures from Bayer, Bracco Imaging, Canon Medical Systems, Guerbet, Teleflex/Vidacare, and Vital Images. **M.R.M.** No relevant relationships. **B.H.** No relevant relationships. **M.P.** Support for attending meetings or travel from UCB; participation on a data safety monitoring board or advisory board from Novartis. **V.R.R.** No relevant relationships. **H.H.** Payment for lectures from AbbVie, MSD, Janssen, Roche, Pfizer, and Sobi; support for attending meetings or travel from AbbVie, Novartis, and UCB; participation on a data safety monitoring board or advisory board for Janssen, Sobi, and Novartis. **J.R.** Support for attending meetings or travel from AbbVie, Novartis, and UCB. **M.T.** No relevant relationships. **R.G.L.** Consulting fees from CARE Arthritis Image Analysis Group and Parexel; honoraria for lectures from the Dr Sulaiman Al Habib Medical Group. **X.B.** No relevant relationships. **W.P.M.** No relevant relationships. **J.L.V.** Nonfinancial support from Bayer, Guerbet, Medtronic, and Merit Medical; personal fees from Merit Medical. **D.P.** Research grants to institution from AbbVie, Lilly, MSD, Novartis, and Pfizer; consulting fees from AbbVie, Biocad, Lilly, Gilead, GlaxoSmithKline, Janssen, MSD, Novartis, Pfizer, Samsung Bioepis, and UCB; participation on a data safety monitoring board or advisory board for AbbVie, Lilly, Gilead, GlaxoSmithKline, Janssen, MSD, Novartis, Pfizer, and UCB.

References

1. Sieper J, Poddubnyy D. Axial spondyloarthritis. Lancet 2017; 390(10089):73–84.
2. Sieper J, Rudwaleit M, Baraliakos X, et al. The Assessment of SpondyloArthritis international Society (ASAS) handbook: a guide to assess spondyloarthritis. Ann Rheum Dis 2009;68(suppl 2):ii1–ii44.
3. Poddubnyy D, Rudwaleit M, Haibel H, et al. Rates and predictors of radiographic sacroiliitis progression over 2 years in patients with axial spondyloarthritis. Ann Rheum Dis 2011;70(8):1369–1374.
4. Maksymowych WP, Lambert RG, Østergaard M, et al. MRI lesions in the sacroiliac joints of patients with spondyloarthritis: an update of definitions and validation by the ASAS MRI working group. Ann Rheum Dis 2019;78(11):1550–1558.
5. Baraliakos X, Richter A, Schmidt CO, Braun J. Response to: 'Correspondence on 'Which factors are associated with bone marrow oedema suspicious of axial spondyloarthritis as detected by MRI in the sacroiliac joints and the spine in the general population?' by Su et al. Ann Rheum Dis 2021;80(4):469–474.
6. Poddubnyy D, Weineck H, Diekhoff T, et al. Clinical and imaging characteristics of osteitis condensans ilii as compared with axial spondyloarthritis. Rheumatology (Oxford) 2020;59(12):3798–3806.
7. Cereser L, Zabotti A, Zancan G, et al. Magnetic resonance imaging assessment of ASAS-defined active sacroiliitis in patients with inflammatory back pain and suspected axial spondyloarthritis: a study of reliability. Clin Exp Rheumatol 2021;39(6):1331–1337.
8. van den Berg R, Lenczner G, Thévenin F, et al. Classification of axial SpA based on positive imaging (radiographs and/or MRI of the sacroiliac joints) by local rheumatologists or radiologists versus central trained readers in the DESIR cohort. Ann Rheum Dis 2015;74(11):2016–2021.
9. Bakker PA, van den Berg R, Lenczner G, et al. Can we use structural lesions seen on MRI of the sacroiliac joints reliably for the classification of patients according to the ASAS axial spondyloarthritis criteria? Data from the DESIR cohort. Ann Rheum Dis 2017;76(2):392–398.
10. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. Radiology 2019;290(3):590–606.
11. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med 2018;15(11):e1002699.
12. Song Y, Zhang YD, Yan X, et al. Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI. J Magn Reson Imaging 2018;48(6):1570–1577.
13. Lin L, Dou Q, Jin YM, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. Radiology 2019;291(3):677–686.
14. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 2017;36:61–78.
15. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310–1320.
16. Rudwaleit M, Haibel H, Baraliakos X, et al. The early disease stage in axial spondylarthritis: results from the German Spondyloarthritis Inception Cohort. Arthritis Rheum 2009;60(3):717–727.
17. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(1):77.

18. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. J Open Source Softw 2019;4(43):1686.
19. Gamer M. irr: Various coefficients of interrater reliability and agreement. https://cran.r-project.org/web/packages/irr/irr.pdf. Published 2010. Accessed September 2021.
20. Rudwaleit M, Jurik AG, Hermann KG, et al. Defining active sacroiliitis 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):1520–1527.
21. van der Heijde D, Rudwaleit M, Landewé RB, Sieper J. Justification for including MRI as a tool in the diagnosis of axial SpA. Nat Rev Rheumatol 2010;6(11):670–672.
22. de Winter J, de Hooge M, van de Sande M, et al. Magnetic resonance imaging of the sacroiliac joints indicating sacroiliitis according to the Assessment of SpondyloArthritis international Society definition in healthy individuals, runners, and women with postpartum back pain. Arthritis Rheumatol 2018;70(7):1042–1048.
23. Weber U, Jurik AG, Zejden A, et al. Frequency and anatomic distribution of magnetic resonance imaging features in the sacroiliac joints of young athletes: exploring "background noise" toward a data-driven definition of sacroiliitis in early spondyloarthritis. Arthritis Rheumatol 2018;70(5):736–745.
24. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019;1(6):e271–e297.
25. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019;20(3):405–410.
26. Mårtensson G, Ferreira D, Granberg T, et al. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. Med Image Anal 2020;66:101714.
27. Ran M, Hu J, Chen Y, et al. Denoising of 3D magnetic resonance images using a residual encoder-decoder Wasserstein generative adversarial network. Med Image Anal 2019;55:165–180.
28. Chauhan N, Choi BJ. Denoising approaches using fuzzy logic and convolutional autoencoders for human brain MRI image. Int J Fuzzy Logic Intell Syst 2019;19(3):135–139.