Detecting radiographic sacroiliitis using deep learning with expert-level accuracy in axial spondyloarthritis

Keno K. Bressem,[+1,3] MD; Janis L. Vahldiek,*[+1] MD; Lisa C. Adams,[1,3] MD; Stefan M. Niehues,[1] MD; Hildrun Haibel,[2] MD; Valeria Rios Rodriguez,[2] MD; Murat Torgutalp,[2] MD; Mikhail Protopopov,[2] MD; Fabian Proft,[2] MD; Judith Rademacher,[2,3] MD; Joachim Sieper,[2] MD; Martin Rudwaleit,[4] MD; Bernd Hamm,[1] MD; Marcus R. Makowski,[1] MD; Kay Geert A. Hermann,[1] MD; Denis Poddubnyy,[2,5] MD

*Corresponding Author + contributed equally

1 Charité Universitätsmedizin Berlin, Department of Radiology, Hindenburgdamm 30, 12203 Berlin, Germany.

2 Charité Universitätsmedizin Berlin, Department of Gastroenterology, Infectious Diseases and Rheumatology, Hindenburgdamm 30, 12203 Berlin, Germany.

3 Berlin Institute of Health, BIH, Anna-Louisa-Karsch-Straße 2, 10178 Berlin, Germany.

4 Klinikum Bielefeld Rosenhöhe, Department of Internal Medicine and Rheumatology, An der Rosenhöhe 27, 33647 Bielefeld, Germany

5 German Rheumatism Research Centre, Department of Epidemiology, Charitéplatz 1, 10117 Berlin, Germany.

Materials and Correspondence

Dr. med. Janis L. Vahldiek

Charité – Universitätsmedizin Berlin, Klinik für Radiologie

Hindeburgdamm 30, 12203 Berlin, Germany

janis.vahldiek@charite.de

Tel.: +49 30 450 627 779

ABSTRACT

Conventional radiographs of sacroiliac joints from two independent cohorts of patients with axial spondyloarthritis (axSpA) were used to develop and validate an artificial neural network for the detection of definite radiographic sacroiliitis as a manifestation of the disease. The first cohort consisting of 1669 radiographs labelled by three experts was used for training. From the second cohort 100 radiographs labelled by two experts were randomly selected for the test dataset. The neural network achieved an excellent performance in detection of definite radiographic sacroiliitis with an area under the receiver operating characteristics curve of 0.97 and 0.96 for the validation and test datasets, respectively. Sensitivity and specificity for the cut-off weighting both measurements equally were 0.90 and 0.93 for the validation and 0.87 and 0.97 for the test set. The Cohen's kappa between the neural network and the reference judgements were 0.80 for both validation and test sets, and the absolute agreement on the classification yielded 91% and 90%, respectively.

KEYWORDS

axial spondyloarthritis, sacroiliitis, artificial intelligence, deep learning, machine learning

INTRODUCTION

Axial spondyloarthritis (axSpA) is a chronic inflammatory disorder primarily affecting axial skeleton – sacroiliac joints and spine. Detection of radiographic sacroiliitis has been for many years the only possibility to establish a definite diagnosis of the disease prior to development of structural damage in the spine. The presence of definite radiographic sacroiliitis (defined as sacroiliitis of at least grade 2 bilaterally or at least grade 3 unilaterally) is also an obligatory criterion in the modified New York criteria for ankylosing spondylitis (AS).[1] Although magnetic resonance imaging (MRI) of sacroiliac joints nowadays allows to diagnose axSpA earlier, definite radiographic sacroiliitis can be detected at the time-point of the diagnosis already in about 33% of the patients with symptom duration of up to one year and about 50% of the patients with symptom duration of 2 to 3 years.[2] Conventional radiography of the sacroiliac joints is still recommended as the first imaging method if axSpA is suspected.[3] Furthermore, radiographic sacroiliitis is included – together with sacroiliitis on MRI – in the Assessment of Spondyloarthritis International Society (ASAS) classification criteria for axSpA.[4] Depending on the presence or absence of definite radiographic sacroiliitis, axSpA can be classified either as radiographic axSpA (r-axSpA, synonymous to AS) or non-radiographic axSpA (nr-axSpA).[5] Such a classification might be relevant for both clinical practice (currently, labels for biological disease modifying antirheumatic drugs – bDMARDs are different for AS and nr-axSpA) and for research (i.e., stratification or selection of patients in a clinical trial).

Although conventional radiography of sacroiliac joints still plays an important role in clinical practice and for clinical trials, the reliability of radiographic sacroiliitis assessment has been reported in a number of studies as mostly poor, even if performed by expert readers.[6-10] Furthermore, it was shown that untrained local readers perform worse than expert readers specialised in SpA.[10] One possible solution to achieve a comparable high accuracy as an expert on detection of radiographic sacroiliitis, even in non-specialised clinics, could be development of an artificial intelligence-based model for analysis of radiographs.

Artificial intelligence-based or deep learning models have already accomplished remarkable results in the classification of medical and non-medical data. For example, neural networks have been trained to detect breast cancer in mammographs, to classify skin cancer or to label chest radiographs.[11-13] Common to all these studies was that they did not develop a de novo model, but rather applied a transfer learning approach using a pre-trained network. This allows knowledge of pre-trained models from non-medical fields to be used for a new visual task, effectively reducing the amount of data required for training while increasing the accuracy of the models.

In the present study, we therefore aimed to develop and to validate a deep learning artificial neural network for the detection of definite radiographic sacroiliitis using centrally scored images from two observational cohort studies.

METHODS

Cohort description

In this project, we used imaging data from two independent sources: 1) Patients With Axial Spondyloarthritis: Multicountry Registry of Clinical Characteristics (PROOF), and 2) German Spondyloarthritis Inception Cohort (GESPIC).

PROOF is an ongoing study conducted in clinical practices in 29 countries that includes 2170 adult patients diagnosed with axSpA (non-radiographic or radiographic) ≤12 months before study enrolment and fulfilling the ASAS classification criteria for axSpA. In 1669 patients, radiographs of sacroiliac joints were available for central reading.

GESPIC is a multicentre inception cohort study conducted in Germany that includes 646 patients with SpA.[14] In 525 patients, radiographs of sacroiliac joints were available for central reading.

Assessment of radiographic sacroiliitis

Baseline radiographs of sacroiliac joints were collected, digitized if necessary, anonymized and subsequently centrally scored by trained and calibrated readers according to the grading system of the modified New York criteria:[1]

Grade 0 normal;

Grade 1 suspicious changes;

Grade 2 minimal abnormality: small localized areas with erosion or sclerosis, without alteration in the joint width;

Grade 3 unequivocal abnormality: moderate or advanced sacroiliitis with erosions, evidence of sclerosis, widening, narrowing, or partial ankylosis;

Grade 4 severe abnormality: total ankylosis.

In the PROOF study, images were first assessed by the local readers, then by the central reader 1 (DP) who was blinded for the local assessment. In case of a disagreement on the presence of definite radiographic

sacroiliitis (grade ≥2 bilaterally or grade ≥3 unilaterally) between the local and central reader 1, the radiograph was evaluated by central reader 2 (HH) who was blinded to the previous assessments. The final decision on the presence of definite radiographic sacroiliitis and, therefore, on classification as nr-axSpA or r-axSpA was made based on the decision of two of the three readers.

In GESPIC, no local reading of radiographs was demanded; all collected images were scored independently by two trained and calibrated central readers (VRR and MT).

Image selection and pre-processing

The PROOF dataset consisted of 1669 radiographs of sacroiliac joints in DICOM (Digital Imaging and Communications in Medicine) format, varying in size, resolution and quality. 18 images were excluded due to very poor image quality (Figure 1). The Horos Project DICOM Viewer (version 4.0.0, www.horosproject.org) was used to manually adjust the greyscale-levels of all images and to convert them afterwards to the Tagged Image File Format (TIFF). If other body parts such as the thoracic spine were included in the image, it was manually cropped to the pelvis. The final data set for building the model consisted out of 1651 individual patient images and was split randomly into a training (85% - 1404 radiographs) and validation data set (15% - 247 radiographs).

For testing the generalisability across data sets, 100 radiographs of sacroiliac joints with agreement of both readers on the presence or absence of definite radiographic sacroiliitis were randomly selected from the 525 baseline radiographs of GESPIC and then pre-processed exactly as the training and validation data sets.

Model Training

Model training was performed on a dedicated Ubuntu 18.04 Workstation with two Nvidia GeForce RTX 2080ti Graphic cards as well as on a GPU node of the Berlin institute of health (BIH) high performance computing cluster using four Nvidia Tesla V100 Graphic Cards. All model training was mainly performed using Python (version 3.7), with the fastAI application programming interface which is built on top of PyTorch.[15,16]

As base model, a convolutional neural network (ResNet-50 architecture), pre-trained on the image net data set, consisting of over 14 million images, was used.[17] The images were augmented prior to training through various transformations consisting of flipping, rotation of up to 10°, magnification of up to 1.1, lighting variations and warping. We further utilised the mix-up method during training, originally introduced by Zhang et al,[18] in which images of different classes (nr-axSpA and r-axSpA) are combined during training to reduce memorisation of noisy labels and increase overall model robustness. As a loss function, we utilised cross entropy label smoothing, which reduced high-confidence predictions of the models, thus supporting regularisation and avoiding overfitting with subsequent better generalisation of the models on new data (e.g. test-data set). Model training was performed with discriminative learning rates and a progressive resizing approach, starting with images-sizes of 224x224 pixels (which is the default input-size for the ImageNet pretrained ResNet-50) and then increasing the resolution to 512x512 pixels and 768x768 pixels.[19] During training, first only the last two classification layers of the model were trained, with the weights of the other network-layers remaining frozen. A total of 100 epochs were trained, monitoring the area under the receiver-operating characteristics curve (AUROC) on the validation data set, saving the models weights on every improvement. After 100 epochs, the weights of the model with the highest AUROC value were re-loaded, the model was unfrozen and again trained for 100 epochs, while monitoring the AUROC and saving the weights at every improvement. This approach was repeated for all image resolutions. The size of the mini batches was 64 for 224x224 pixels, 32 for 512x512 pixels and 84 for 768x768 pixels. The training for lower resolutions could be performed at our local Workstation, while for 768x768 pixels the BIH high performance cluster was used. Overall, model training took approximately 24 hours on our local machine and about an additional 6 hours on the cluster. After training, activation maps were created to verify that the model used the sacroiliac joints to determine whether definite radiographic sacroiliitis was present.

Statistical analysis

Statistical analysis was performed using the R statistical environment (version 3.6), the "tidyverse", "ROCR" and "irr" libraries.[20-23] Raw predictions of the model on the validation data set as well as on the test data set using an image resolution of 768x768 pixels were exported from the python environment as comma separated values and imported into "R". ROC-curves and precision recall curves were plotted and the area under the curve was calculated. A ROC analysis was then performed to identify different prediction cut-offs. Sensitivity and specificity of the cut-offs favouring sensitivity, favouring specificity or balancing both were calculated. Confusion matrices were constructed using the prior defined cut-offs. Cohen's kappa and the percentage absolute agreement was used to assess the agreement between human and machine. 95% confidence intervals for calculated kappa values were estimated using bootstrapping with 1000 repetitions. A p-value of < 0.05 was considered statistically significant.

Patient and Public Involvement

This study as well as both parental cohorts were done without any formal patient/patient organization involvement in study design, development of patient relevant outcomes, interpretation of results, or the writing or editing of the manuscript.

Ethics approval

Both PROOF and GESPIC cohorts were approved by the local ethics committees of each study centre in accordance with the local laws and regulations and is being conducted in accordance with the Declaration of Helsinki and Good Clinical Practice. GESPIC was additionally approved by a central ethics committee of the coordinating centre. Written informed consent was obtained from all patients.

RESULTS

Definite radiographic sacroiliitis was present in 911 (64.9%) patients of the training set (PROOF, n=1404), in 160 (64.8%) patients of the validation set (PROOF, n=247) and in 54 (54%) patients of the independent test set (GESPIC, n=100). In a total of 392 (27.9%) and 67 (27.1%) patients in the training and validation sets, there was a discrepancy between the local and the central reader 1 that automatically resulted in the involvement of the central reader 2. A total of 160 (11.4%) and 39 (15.8%) patients in the training and validation sets were re-classified after the central reading, meaning the rating from both central readers differed from the rating of the local reader.

Model performance in the validation data set

On the validation data set, the model achieved an excellent performance. The ROC analysis showed an area under the curve (AUROC) of 0.971. For the precision-recall curve a value of 0.989 could be calculated. Both the local and central expert readers remained below the ROC and PR curves and were therefore outperformed by the accuracy of the model. We propose three cut-offs to convert the floating-point predictions into integer values, where a 1 represents the presence of definite radiographic sacroiliitis and 0 the absence. Cut-offs weighted sensitivity over specificity and specificity over sensitivity, aiming to find the optimal balance between both measurements (defined as maximal sum between sensitivity and specificity). The first cut-off value, which favours sensitivity over specificity, was calculated to be 0.606, resulting in a sensitivity of 0.994 and a specificity of 0.667 for the detection of radiographic SpA. The second cut-off, which favoured specificity over sensitivity, was 0.814, resulting in a sensitivity of 0.656 and a specificity of 1. The third cut-off was 0.706, resulting in a sensitivity of 0.9 and a specificity of 0.931. The human local reader had a sensitivity of 0.844 and a specificity of 0.839. The expert central reader had a higher sensitivity than the local reader with 0.975, but at the cost of a lower specificity of 0.724. The performance of the model is also shown in Figure 2 in the form of ROC curves and precision recall curves and Table 1 as confusion matrices with kappa-values and values of absolute agreement.

Table 1

Cut-off 1, favouring sensitivity over specificity			
	Reference negative	Reference positive	
Model negative	58	2	60
Model positive	29	158	187
	87	160	247
Cohen's Kappa:	0.704 (95% CI 0.611 to 0.794)	Absolute agreement:	n=216 (87.4%)

Cut-off 2, favouring specificity over sensitivity			
	Reference negative	Reference positive	
Model negative	87	56	143
Model positive	0	104	104
	87	160	247
Cohen's Kappa:	0.567 (95% CI 0.475 to 0.657)	Absolute agreement	n=191 (77.3%)

Cut-off 3, optimal relationship between sensitivity and specificity			
	Reference negative	Reference positive	
Model negative	81	17	98
Model positive	6	143	149
	87	160	247
Cohen's Kappa:	0.802 (95% CI 0.731 to 0.875)	Absolute agreement:	n=224 (90.7%)

Table 1 provides confusion matrices for the three proposed cut-offs for the model-predictions for the presence of definite radiographic sacroiliitis on the validation data set alongside the agreement between ratings measured by Cohen's Kappa and the absolute percentage agreement. Cut-off 1 weights sensitivity more than specificity, while cut-off 2 weights specificity more than sensitivity and cut-off 3 aims to be the optimal trade-off between sensitivity and specificity.

The reference was the presence of definite radiographic sacroiliitis according to the majority decision of three readers in the PROOF cohort.

With the cut-off with balanced sensitivity and specificity, the neural network achieved a kappa of 0.80 (95% CI 0.73 to 0.88) and an absolute agreement of 91% on the final classification. This was superior to the agreement between the local reader judgement and the final classification (Cohen's kappa = 0.66, 95%CI 0.57 to 0.76; absolute agreement 84%) and even better than the agreement between the central reader judgement and the final classification (Cohen's kappa = 0.74, 95%CI 0.65 to 0.83 absolute agreement 89%).

Model performance in the independent data set

In the test data set, the model performance was slightly worse than on the validation data set with an AUROC-value of 0.961 and an area under the precision recall curve (AUPRC) value of 0.979. Again, we propose three cut-offs: The first cut-off, which weighs sensitivity over specificity, was 0.58, yielding a sensitivity of 1 and a specificity of 0.478. The second cut-off, which weighs specificity over sensitivity, was 0.834, based on which a specificity of 1 and a sensitivity of 0.574 could be achieved. The optimal performance for both performance measurements could be achieved with a cut-off of 0.791, resulting in a sensitivity of 0.87 and specificity of 0.966. Figure 3 shows the ROC- and precision-recall curves for the model performance on the test-data set. Figure 4 demonstrates the different values for sensitivity and specificity defined for different cut-offs on the test and validation data set. Table 2 provides confusion matrices for the three proposed cut-offs on the test data set alongside evaluation of agreement in the form of Cohen's kappa and percentage agreement. Importantly, in this independent test set, almost the same performance of the model with balanced sensitivity and specificity could be achieved as in the validation data set: Cohen's kappa = 0.80, absolute agreement of 90%.

Figure 5 shows activation maps of the neural network for correct and incorrect predictions on the test- and validation data sets.

Table 2

Cut-off 1, favouring sensitivity over specificity			
	Reference negative	Reference positive	
Model negative	24	1	25
Model positive	22	53	75
	46	54	100
Cohens Kappa:	0.521 (95% CI 0.372-0.672)	Absolute agreement	n= 77 (77%)

Cut-off 2, favouring specificity over sensitivity			
	Reference negative	Reference positive	
Model negative	46	24	70
Model positive	0	30	30
	46	54	100
Cohens Kappa:	0.535 (95% CI 0.389-0.680)	Absolute agreement:	n=76 (76%)

Cut-off 3, optimal relationship between sensitivity and specificity			
	Reference negative	Reference positive	
Model negative	44	8	52
Model positive	2	46	48
	46	54	100
Cohens Kappa:	0.801 (95% CI 0.688-0.919)	Absolute agreement:	n=90 (90%)

Table 2 provides confusion matrices for the three proposed cut-offs for the raw model-predictions for the presence of definite radiographic sacroiliitis on the test data set alongside the agreement between ratings measured by Cohens Kappa and the absolute percentage agreement. Cut-off 1 weights sensitivity more than specificity, while cut-off 2 weights specificity more than sensitivity and cut-off 3 aims to be the optimal trade-off between sensitivity and specificity. The reference was the agreement of two central readers on the presence of definite radiographic sacroiliitis in the GESPIC cohort.

DISCUSSION

In this study, we successfully developed and tested an artificial intelligence model for the detection of radiographic sacroiliitis on conventional radiographs. With the developed model, we could even achieve a super-human accuracy on the validation data. Furthermore, we were able to demonstrate the generalisability of our model on a test data set of novel data, achieving a performance at least comparable to that of two human experts.

Although in industrialized countries magnetic resonance imaging is increasingly used for the detection of sacroiliitis, radiographs are still of importance. Especially in developing countries radiographs are the main imaging procedure due to a low availability and high associated costs of magnetic resonance imaging. Detection of definite radiographic sacroiliitis is important for both diagnosis and classification of axSpA. At the same time, a poor reliability of sacroiliitis detection by conventional radiographs is well known.[6-10] In the present study we used a large and unique dataset to train, to validate, and to test the model. The resulting performance was at least as good as (but most likely better than) the performance of an experienced reader with expertise in radiographic sacroiliitis assessment. The neural network could achieve almost the same level of performance in both validation and training sets indicating high level of reliability and robustness of the model. The developed model can be applied, therefore, as an additional diagnostic aid in clinical practice and as a classification tool in research projects involving patients with axSpA.

Neural networks have already been applied to a variety of medical imaging data, including radiographs but, to our knowledge, not for the detection of spondyloarthritis.[11-13,24] However, a low generalisability, i.e. poor performance of the models on new data, is an important challenge in training neural networks. A new meta-analysis on 'deep learning performance against healthcare professionals' by *Kim et al.* revealed methodological shortcomings that are present in many published studies on deep learning in medicine.[25] They criticized that many studies either did not compare the performance of their model to that of a

human domain expert, or evaluated the performance of their model on a dataset different from the dataset used to evaluate human performance, so that they achieve incorrectly high accuracies, mainly due to over-adaptation, and consequently have a low generalisability.[25] Similar observations were made by *Yao et al*., who showed that although 155 studies on deep learning in medicine have been published, they often lack external validation data.[26] However, the use of external validation data is an important measure to proof generalisability. It has been shown that medical computer vision models adapt poorly to the use of different scanners or imaging protocols, and the lack of external validation is likely to result in poor generalisability of the model to new data.[27] In a recent study, *McKinney et al* evaluated the performance of a neural network for the detection of breast cancer in mammographs, surpassing human performance.[11] They used different data sets from different studies to train and test their developed models and were thus able to demonstrate sufficient generalisability of their developed models.

Similar to their approach, we also used a heterogeneous training data set with radiographs from different imaging sites and achieved a good generalisability of the developed model, with the performance being only slightly inferior on the test data, which was independent in terms of patients and readers, as compared to the validation data.[18,28,29] While the heterogeneity of our training data set already reduced the risk of overfitting on systematic image noise, e.g. to device specific image features, we further increased generalisability by applying progressive resizing and the integration of mix-up as well label-smoothing into model-training.

Our study has some limitations. First, the reference for the training of the model was the judgement of a limited number (2 or, in the case of discrepancy, 2 out of 3) of human readers. Although both central readers in the PROOF study had many years of experience in reading of radiographs of sacroiliac joints, the complex anatomy of sacroiliac joints and heterogeneity of radiographic techniques and quality have brought an inevitable portion of uncertainty in the final classification used as a reference. In the independent dataset, we selected only cases where both readers agreed to be the reference standard for the evaluation of the model. This approach was chosen because we believe that these cases are most likely

true positive or true negative ones. Other approaches, such as evaluating all cases, in which the readers disagreed, to be negative, were discarded as we saw a risk of biasing the test data set and thereby distorting the evaluation of the model performance. Remarkably, despite all the uncertainty related to the assessment of radiographic sacroiliitis, a high level of agreement between the neural network's judgement and the consensus judgement by human could be achieved in both validation and test sets.

To further evaluate the model performance, we believe a prospective study is needed since the data we used is from two studies not primary designed for training of a deep neural network. Thus, a selection bias might be present in our data.

Conclusion

Convolutional neural networks can reliably detect definite radiographic sacroiliitis with a performance similar or even superior to that of a human expert reader.

FUNDING

GESPIC was initially supported by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung—BMBF). As consequence of the funding reduction by BMBF according to schedule in 2005 and stopped in 2007, complementary financial support has been obtained also from Abbott, Amgen, Centocor, Schering–Plough, and Wyeth. Starting from 2010, the core GESPIC cohort was supported by AbbVie. The PROOF study is funded by AbbVie. We would like to thank AbbVie for allowing the use of the PROOF dataset for the aim of the current study.

ACKNOWLEDGMENTS

LCA is grateful for her participation in the BIH Charité–Junior Clinician and Clinician Scientist Program and KKB is grateful for his participation in the BIH Charité Digital Clinician Scientist Program all funded by the Charité–Universitätsmedizin Berlin and the Berlin Institute of Health. We would like to thank the Berlin Institute of Health (BIH) for providing access to their high-performance computing cluster for our analysis.

COMPETING INTERESTS

K.K.B., L.C.A., J.R., M.T. and M.M. have nothing to disclose.

J.L.V reports non-financial support from Bayer, non-financial support from Guerbet, non-financial support from Medtronic, personal fees and non-financial support from Merit Medical, outside the submitted work

S.M.N. reports grants from Deutsche Forschungsgemeinschaft, personal fees from Bracco Imaging, personal fees from Canon Medical Systems, personal fees from Guerbet, personal fees from Teleflex Medical GmbH, personal fees from Bayer Vital GmbH, outside the submitted work.

H.H. reports personal fees from Pfizer, personal fees from Janssen, personal fees from Novartis, personal fees from Roche, personal fees from MSD, outside the submitted work.

V.R.R. reports personal fees from Abbvie, personal fees from Novartis, outside the submitted work.

M.P. reports personal fees from Novartis, personal fees from AbbVie, outside the submitted work.

F.P. reports personal fees from AbbVie, personal fees from AMGEN, personal fees from BMS, personal fees from Celgene, from MSD, grants and personal fees from Novartis, personal fees from Pfizer, from Roche, personal fees from UCB, outside the submitted work.

J.S. reports grants from AbbVie, during the conduct of the study; personal fees from AbbVie, personal fees from Novartis, personal fees from Pfizer, personal fees from Roche, personal fees from UCB, personal fees from Boehringer Ingelheim, personal fees from Janssen, personal fees from Merk, outside the submitted work.

M.R. received honoraria and/or consulting fees from AbbVie, BMS, Celgene, Janssen, Eli Lilly, MSD, Novartis, Pfizer, Roche, UCB Pharma.

B.H. reports grants from Abbot, grants from Actelion Pharmaceuticals, grants from Bayer Schering Pharma, grants from Bayer Vital, grants from BRACCO Group, grants from Bristol-Myers Squibb, grants from Charite Research Organisation GmbH, grants from Deutsche Krebshilfe, grants from Essex Pharma, grants from Guerbet, grants from INC Research, grants from lnSightec Ud, grants from IPSEN Pharma, grants from

Kendlel MorphoSys AG, grants from Lilly GmbH, grants from MeVis Medical Solutions AG, grants from Nexus Oncology, grants from Novartis, grants from Parexel Clinical Research Organisation Service, grants from Pfizer GmbH, grants from Philipps, grants from Sanofis-Aventis, grants from Siemens, grants from Teruma Medical Corporation, grants from Toshiba, grants from Zukunftsfond Berlin , grants from Amgen, grants from AO Foundation, grants from BARD, grants from BBraun, grants from Boehring Ingelheimer, grants from Brainsgate, grants from CELLACT Pharma, grants from CeloNova Bio-Sciences, grants from GlaxoSmithKline, grants from Jansen, grants from Roehe, grants from Sehumaeher GmbH , grants from Medronic, grants from Pluristem, grants from Quintiles, grants from Roehe, grants from Astellas, grants from Chiltern, grants from Respicardia, grants from TEVA, grants from Abbvie, grants from AstraZenaca , grants from Galmed Research and Development Ltd, , outside the submitted work.

K.G.H. reports personal fees from AbbVie, personal fees from Pfizer, personal fees from MSD, personal fees from Roche, outside the submitted work.

D.P. reports grants and personal fees from AbbVie, during the conduct of the study; grants and personal fees from AbbVie, personal fees from BMS, personal fees from Celgene, grants and personal fees from Lilly, grants and personal fees from MSD, grants and personal fees from Novartis, grants and personal fees from Pfizer, personal fees from Roche, personal fees from UCB, outside the submitted work.

AUTHOR CONTRIBUTIONS

Conceived and designed the study: KKB, JLV and DP; Data processing and extraction: KKB, JLV and DP, HH, JS, MR, KGAH; Analyzed the data: KKB, JLV, MR, JS, HH and DP; Training of deep neural networks: KKB, JLV; Drafting of the manuscript: KKB, JLV, LCA, DP; Commented on the manuscript, suggested/made revisions of important intellectual content and approved the final version: KKB, JLV, LCA, SMN, HH, VRR, MT, MP, FP, JR, JS, MR, BH, MRM, KGAH, DP; Administrative and technical support: DP, SMN, BH and MRM.

REFERENCES

1. van der Linden, S., Valkenburg, H.A. & Cats, A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. *Arthritis Rheum* **27**, 361-368 (1984).
2. Poddubnyy, D.*, et al.* The frequency of non-radiographic axial spondyloarthritis in relation to symptom duration in patients referred because of chronic back pain: results from the Berlin early spondyloarthritis clinic. *Annals of the rheumatic diseases* **71**, 1998-2001 (2012).
3. Mandl, P.*, et al.* EULAR recommendations for the use of imaging in the diagnosis and management of spondyloarthritis in clinical practice. *Annals of the rheumatic diseases* **74**, 1327-1339 (2015).
4. Rudwaleit, M.*, et al.* The development of Assessment of SpondyloArthritis international Society classification criteria for axial spondyloarthritis (part II): validation and final selection. *Annals of the rheumatic diseases* **68**, 777-783 (2009).
5. Boel, A.*, et al.* Do patients with axial spondyloarthritis with radiographic sacroiliitis fulfil both the modified New York criteria and the ASAS axial spondyloarthritis criteria? Results from eight cohorts. *Annals of the rheumatic diseases* **78**, 1545-1549 (2019).
6. Spoorenberg, A.*, et al.* Radiological scoring methods in ankylosing spondylitis. Reliability and change over 1 and 2 years. *J Rheumatol* **31**, 125-132 (2004).
7. Christiansen, A.A.*, et al.* Limited Reliability of Radiographic Assessment of Sacroiliac Joints in Patients with Suspected Early Spondyloarthritis. *J Rheumatol* **44**, 70-77 (2017).
8. Yazici, H.*, et al.* Observer variation in grading sacroiliac radiographs might be a cause of'sacroiliitis' reported in certain disease states. *Annals of the rheumatic diseases* **46**, 139-145 (1987).
9. Poddubnyy, D.*, et al.* Rates and predictors of radiographic sacroiliitis progression over 2 years in patients with axial spondyloarthritis. *Annals of the rheumatic diseases* **70**, 1369-1374 (2011).

10. Van den Berg, R.*, et al.* Agreement between clinical practice and trained central reading in reading of sacroiliac joints on plain pelvic radiographs: results from the DESIR cohort. *Arthritis Rheum* **66**, 2403-2411 (2014).

11. McKinney, S.M.*, et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89-94 (2020).

12. Esteva, A.*, et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118 (2017).

13. Irvin, J.*, et al.* CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Thirty-Third Aaai Conference on Artificial Intelligence / Thirty-First Innovative Applications of Artificial Intelligence Conference / Ninth Aaai Symposium on Educational Advances in Artificial Intelligence*, 590-597 (2019).

14. Rudwaleit, M.*, et al.* The early disease stage in axial spondylarthritis: Results from the german spondyloarthritis inception cohort. *Arthritis Rheum* **60**, 717-727 (2009).

15. Paszke, A.*, et al.* PyTorch: An imperative style, high-performance deep learning library. in *Advances in Neural Information Processing Systems 32* (eds. Wallach, H.*, et al.*) 8024-8035 (Curran Associates, Inc., 2019).

16. Howard, J. & Gugger, S. fastai: A Layered API for Deep Learning. *Information* **11**, 108 (2020).

17. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

18. Zhang, H., Cisse, M., Dauphin, Y.N. & Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. in *International Conference on Learning Representations* (2017).

19. Howard, J. & Ruder, S. Universal Language Model Fine-tuning for Text Classification. in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 328-339 (Association for Computational Linguistics, 2018).

20. R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria, 2013).

21. Wickham H*, et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

22. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940-3941 (2005).

23. Gamer, M., Lemon, J., Fellows, I. & Singh, P. irr: Various Coefficients of Interrater Reliability and Agreement. in *R package version 0.84.1* (R package version 0.84.1, 2010).

24. Liu, X.*, et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* **1**, e271-e297 (2019).

25. Kim, D.W., Jang, H.Y., Kim, K.W., Shin, Y. & Park, S.H. 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* **20**, 405-410 (2019).

26. Yao, A.D., Cheng, D.L., Pan, I. & Kitamura, F. Deep Learning in Neuroradiology: A Systematic Review of Current Algorithms and Approaches for the New Wave of Imaging Technology. *Radiology: Artificial Intelligence* **2**, e190026 (2020).

27. Shin, H.-C.*, et al.* Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* **35**, 1285-1298 (2016).

28. He, T.*, et al.* Bag of tricks for image classification with convolutional neural networks. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 558-567 (Long Beach, CA, USA, 2019).

29. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2818-2826 (Las Vegas, NV, 2016).

FIGURE CAPTIONS

Figure 1

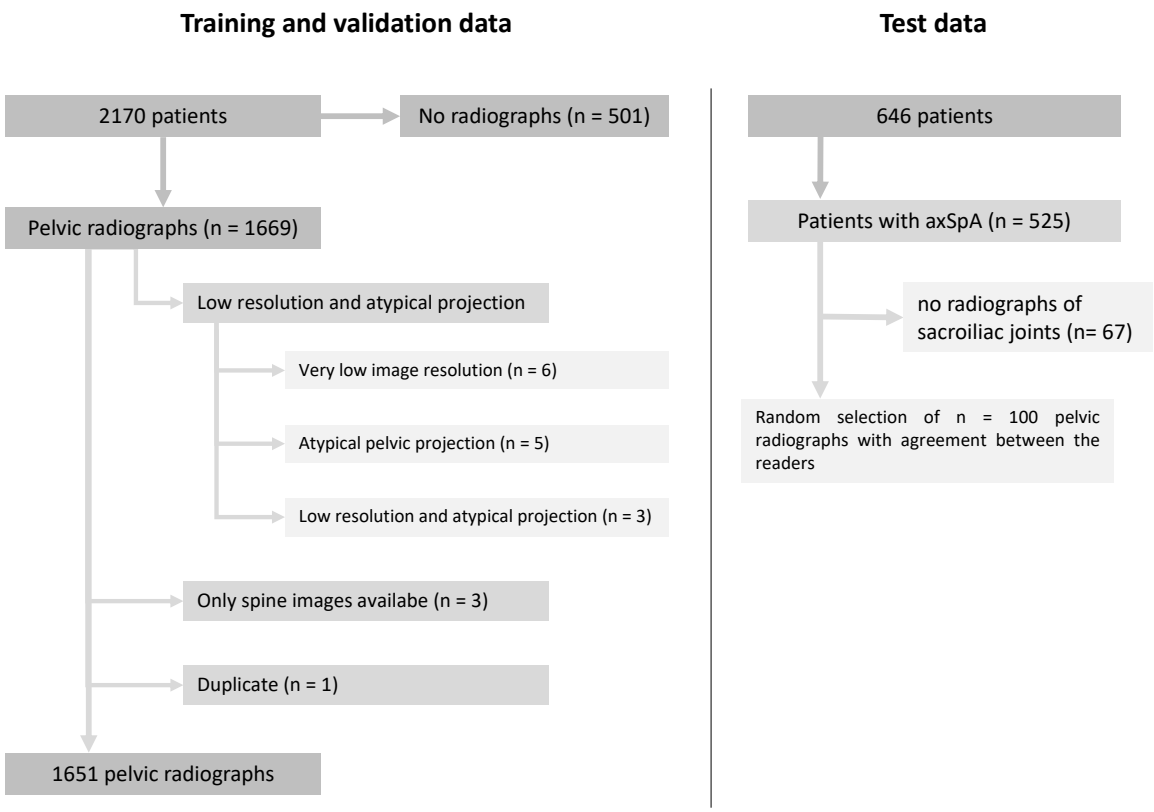

Figure 1 shows a flowchart for selection of cases from the PROOF (training and validation set) and GESPIC (test set) studies. Overall, 18 radiographs had to be excluded from the PROOF study, mostly because of low image resolution or atypical pelvic projections of the radiographs, which were not in accordance to standardised pelvic radiographs. For the test data set, 100 images with the agreement on the presence or absence of definite radiographic sacroiliitis were randomly selected and subsequently used for the final model evaluation.

Figure 2

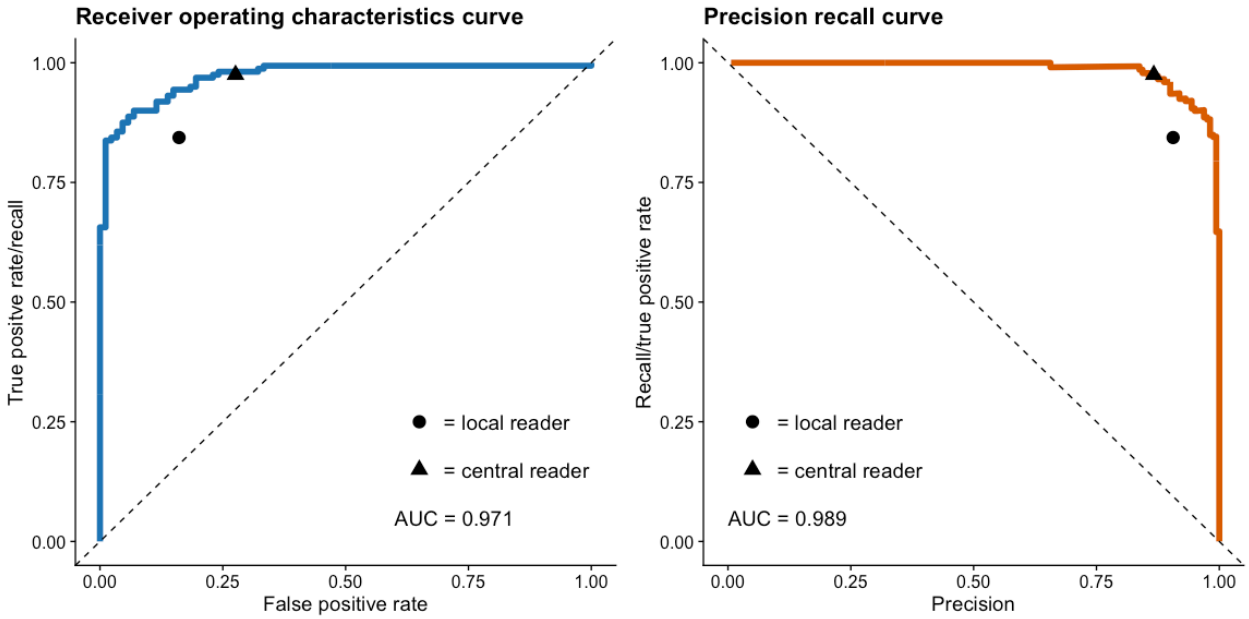

Figure 2 shows the receiver operation characteristics curve and the precision recall curve for the model performance in detection of definite radiographic sacroiliitis (classification as non-radiographic or radiographic axSpA) on the validation data set as well as the corresponding area under the curve (AUC). Individual values for both expert readers are displayed as a triangle or period.

Figure 3

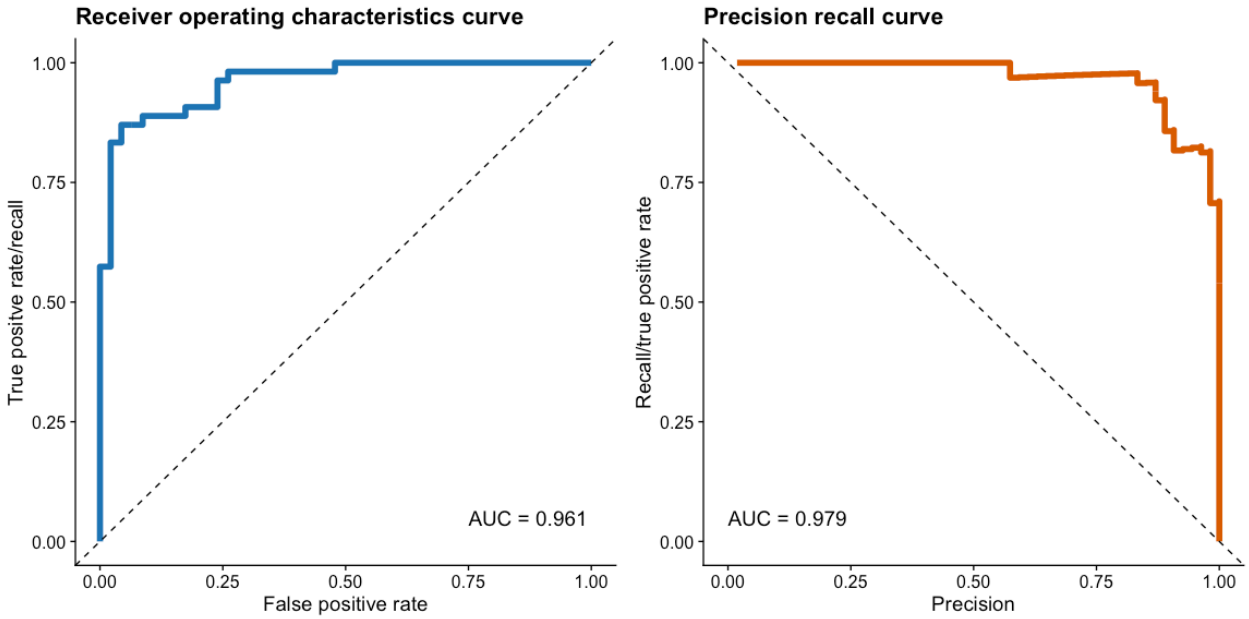

Figure 3 shows the receiver operation characteristics curve and the precision recall curve for the model performance in detection of definite radiographic sacroiliitis (classification as non-radiographic or radiographic axSpA) on the test data set alongside respective area under the curve (AUC). Since the reference standard here was the agreement of two independent readers, a presentation of the accuracy of these readers was not performed.

Figure 4

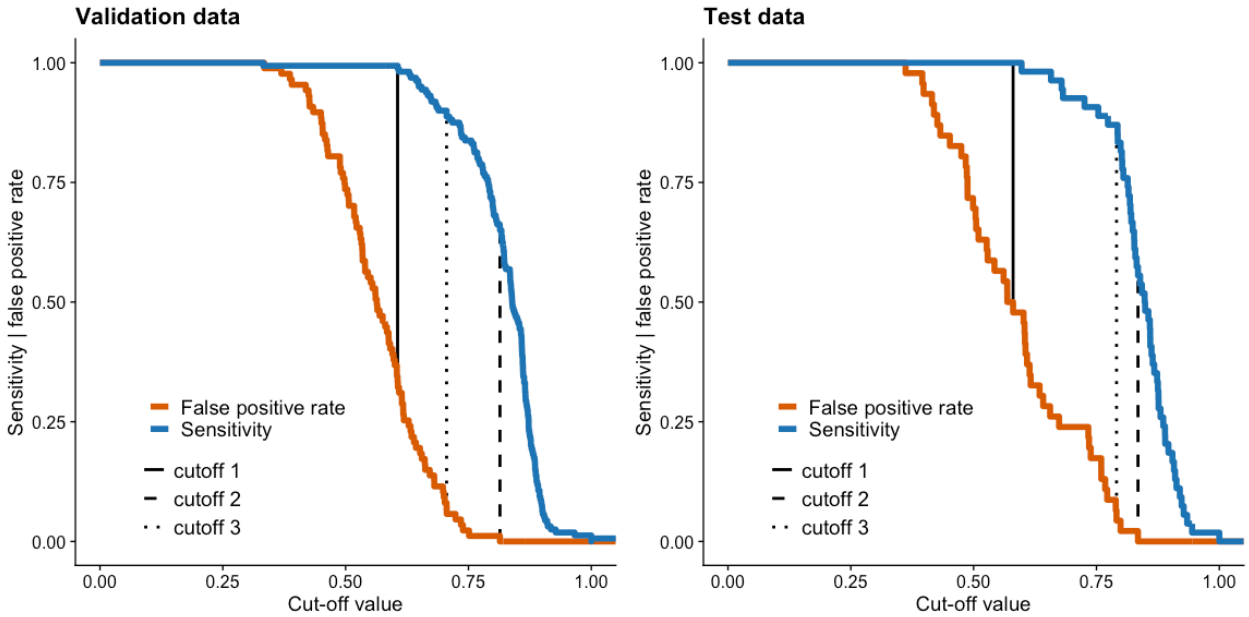

Figure 4 shows the sensitivity and 1-specificity (false positive rate) on the test and validation data set with varying cut-off values for the model predictions on the presence of radiographic sacroiliitis (classification as non-radiographic or radiographic axSpA). We analysed three cut-off values, indicated by vertical dashed lines. Cut-off 1 weights sensitivity over specificity, cut-off 2 weights specificity over sensitivity and cut-off 3 aims to be the optimal balance between the two performance measurements.

Figure 5

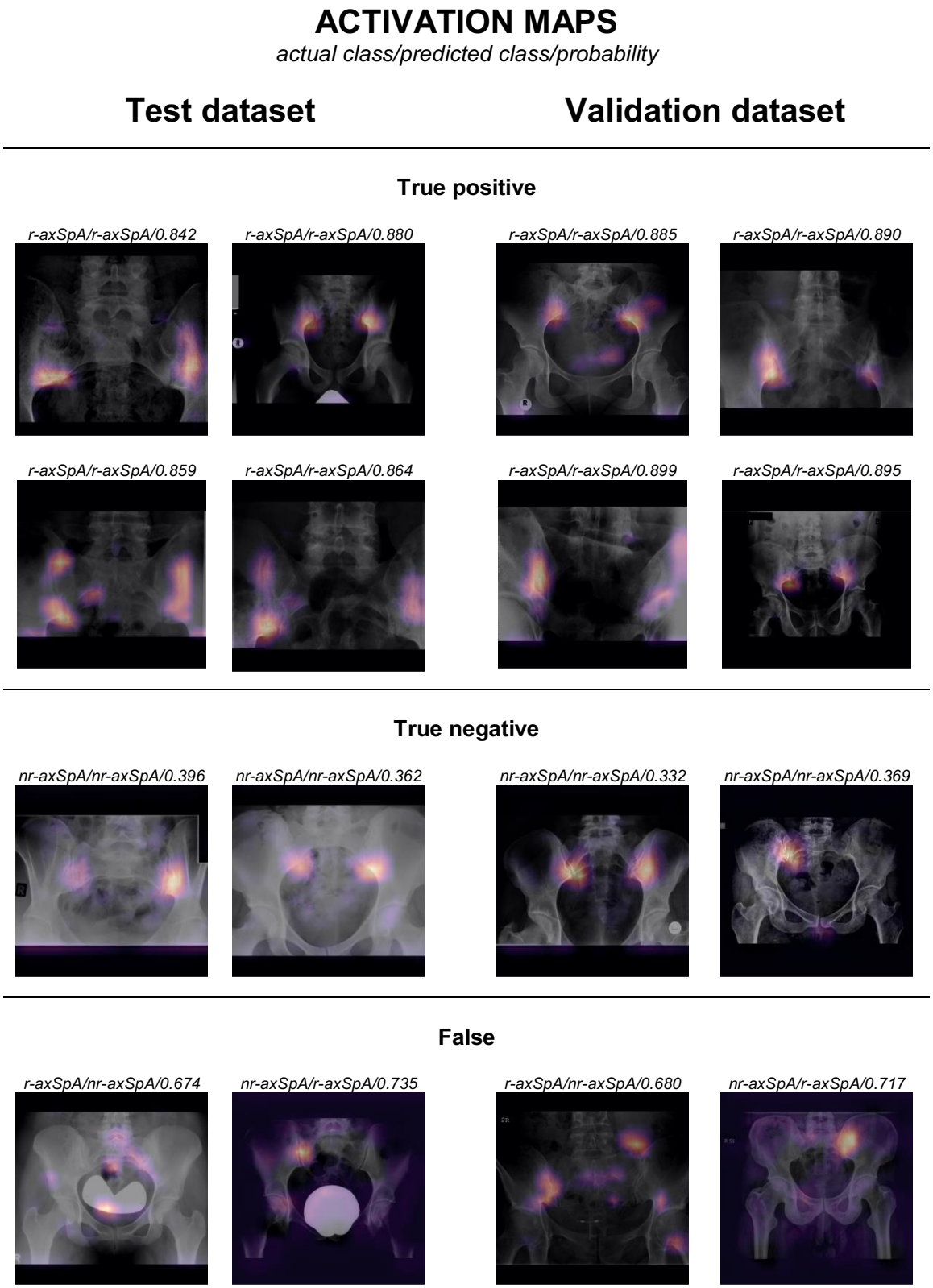
ACTIVATION MAPS
actual class/predicted class/probability
Test dataset
Validation dataset
True positive
r-axSpA/r-axSpA/0.842
r-axSpA/r-axSpA/0.880
r-axSpA/r-axSpA/0.885
r-axSpA/r-axSpA/0.890
r-axSpA/r-axSpA/0.859
r-axSpA/r-axSpA/0.864
r-axSpA/r-axSpA/0.899
r-axSpA/r-axSpA/0.895
True negative
nr-axSpA/nr-axSpA/0.396
nr-axSpA/nr-axSpA/0.362
nr-axSpA/nr-axSpA/0.332
nr-axSpA/nr-axSpA/0.369
False
r-axSpA/nr-axSpA/0.674
nr-axSpA/r-axSpA/0.735
r-axSpA/nr-axSpA/0.680
nr-axSpA/r-axSpA/0.717

Figure 5 shows activation maps of the model on example images taken from the validation and test data sets. It can be seen, that the sacroiliac joints are the most decisive for the model's predictions, but the lumbo-sacral joint and the hip joints also appear to be partially relevant for some predictions.