Towards Accurate Detection of Axial Spondyloarthritis by Using Deep Learning to Capture Sacroiliac Joints on Plain Radiographs.

Nils Friedrich Grauhan (nils-friedrich.grauhan@charite.de)
Charité Universitätsmedizin Berlin: Charite Universitatsmedizin Berlin https://orcid.org/0000-0002-5682-5832

Keno Kyrill Bressem
Charite University Hospital Berlin: Charite Universitatsmedizin Berlin

Yves Nicolas Manzoni
Charite University Hospital Berlin: Charite Universitatsmedizin Berlin

Lisa Christine Adams
Charite University Hospital Berlin: Charite Universitatsmedizin Berlin

Valeria Rios-Rodgriguez
Charité Universitätsmedizin Berlin: Charite Universitatsmedizin Berlin

Fabian Nikolai Proft
Charité Universitätsmedizin Berlin: Charite Universitatsmedizin Berlin

Hildrun Haibel
Charite University Hospital Berlin: Charite Universitatsmedizin Berlin

Martin Rudwaleit
Charite University Hospital Berlin: Charite Universitatsmedizin Berlin

Stefan Markus Niehues
Charite University Hospital Berlin: Charite Universitatsmedizin Berlin

Denis Poddubnyy
Charite University Hospital Berlin: Charite Universitatsmedizin Berlin

Janis Lucas Vahldiek
Charite University Hospital Berlin: Charite Universitatsmedizin Berlin

---

Research article

Keywords: Axial spondyloarthritis, sacroiliac joint, radiograph, deep learning, convolutional neural network, object detection

Posted Date: April 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-379664/v1

Abstract

Background

Well-informed decisions about how to best treat patients with axial spondyloarthritis (SpA) regularly include an evaluation of the sacroiliac joints (SIJ) on plain radiographs. However, grading radiographic findings correctly has proven to be a considerable challenge to expert readers as well as to state-of-the-art convolutional neural networks (CNNs). A method to reduce image information to the clinically relevant core would undoubtedly lead to more accurate results. We, therefore, trained a CNN only to detect SIJs on radiographs and evaluated its potential as a preprocessing pipeline in the automated classification of SpA.

Materials and Methods

We employed a CNN of the RetinaNet architecture, which was trained on a total of 423 plain radiographs of the sacroiliac joints (SIJs). Images were taken from two completely independent datasets. Training and tuning were performed on image data from the Patients With Axial Spondyloarthritis (PROOF) study and testing was executed using images from the German Spondyloarthritis Inception Cohort (GESPIC). Performance was evaluated by manual review and standard object detection metrics from PASCAL and Microsoft COCO.

Results

The CNN produced excellent results in detecting SIJs on the tuning (n =106) and on the holdout dataset (n =140). Object detection metrics for the tuning data were AP = 0.996 and mAP = 0.538; values for the independent holdout data were AP = 0.981 and mAP = 0.515.

Conclusions

The developed CNN was highly accurate in detecting SIJs on radiographs. Such a model could increase the reliability of deep learning-based algorithms in detecting and grading SpA.

Introduction

Plain radiographs of the sacroiliac joints together with magnetic resonance imaging (MRI) play an important role in detecting axial spondyloarthritis (SpA) (1, 2). However, accuracy particularly in grading pathological findings in radiographs according to the modified New York criteria can differ significantly between observers (3, 4). Deep learning-based methods using convolutional neural networks (CNN) have recently demonstrated precision levels of human experts in reading medical images (5-7). They may therefore aid diagnosis of SpA as they have proven to excel when confronted with a specific task (8).

Such an approach was recently pursued by training a CNN to detect definite radiographic sacroiliitis on plain radiographs (9). The results demonstrated a high accuracy of the developed CNN. While these are

excellent results and comparable to the accuracy of domain experts, performance may still have been compromised: Firstly, computational capacity put a limit on the resolution of the radiographs included in the study, resulting in a significant downsizing. Although downsizing is a common practice in deep learning this entails the risk of losing crucial image information (10). Secondly, studies have shown that a CNN may not base its prediction solely on clinically relevant features (11). An excellent way to visually track a model's decision is to use Gradient-weighted Class Activation Mappings (Grad-CAMs) (12). Grad-CAM highlights regions that have a strong influence on the model's decision. In reviewing images that were incorrectly predicted in the holdout dataset of our previous study, several images were found where the model had based its decision on aspects unrelated to SpA features.

The aim of this study was therefore to train and validate a CNN in the detection and segmentation of both sacroiliac joints on plain radiographs by drawing bounding boxes around the region of interest. If proven to be accurate, this CNN could be integrated in a future preprocessing pipeline augmenting the actual detection or classification of SpA no longer restrained by low image resolution or irrelevant information. To test the generalizability of our results, image data for training and testing were taken from two completely independent datasets.

Methods

Data

In this study, we used plain radiographs retrieved from two completely separate image collections: a) Patients With Axial Spondyloarthritis: Multicountry Registry of Clinical Characteristics (PROOF) is a continuous study involving medical facilities in 29 countries. Overall 2170 adult patients diagnosed with SpA (non-radiographic or radiographic) ≤12 months before study enrolment and fulfilling the ASAS classification criteria for SpA are included (13). B) German Spondyloarthritis Inception Cohort (GESPIC) is a multicenter inception cohort study conducted in Germany that includes 646 patients with SpA (14).

Pre-processing and Segmentation

Both PROOF and GESPIC datasets contain radiographs of sacroiliac joints in DICOM format (Digital Imaging and Communications in Medicine). All grey-scale values were adjusted to a unified representation state using the Horos Project DICOM Viewer (version 4.0.0, www.horosproject.org). Afterward, all images were converted to the Tagged Image File Format (TIFF). Annotations for training, tuning and holdout datasets were done using the COCO Annotator (https://github.com/jsbroks/coco-annotator). Bounding boxes were placed by one expert radiologist (JLV) separately around each of the sacroiliac joints (SIJ) on every radiograph. A total of 529 annotated images from the PROOF dataset was available and then randomly split into a training dataset (80%, n=423) and a tuning dataset (20%, n=106). A total of 140 annotated images from GESPIC was available and taken to form the holdout dataset (see flowchart in Figure 1).

Model Training

We used a RetinaNet with a ResNet-50 backbone, pre-trained in ImageNet, specifically suited for object detection (15). Model training was performed using IceVision (https://airctic.com, version 0.5.2), which is an object-detection framework built on top of Python (https://www.python.org, version 3.8), as well as the Fastai (https://fast.ai, version 2.2.5) and PyTorch (https://pytorch.org, version 1.7.0) libraries. Training was executed on a dedicated Ubuntu 18.04 Workstation with two Nvidia GeForce RTX 2080ti Graphic cards. All input images were resized and cropped to 512 x 512 px. Random transformations were used to augment training data. During model training, an early stopping approach was utilized to avoid overfitting. As we used a pre-trained ResNet50 as a backbone, we disabled weight updates for the backbone during the first 53 training epochs. Then we trained for 27 more epochs, updating all weights of the model. The model with the best performance on the tuning data was exported and used for inference on the holdout data.

Statistical Analysis

Model performance was evaluated using two separate approaches: first, in a swift manual review, two expert readers (NFG, YNM) independently checked the results of the model on the holdout data and counted all images in which the model either partially or completely failed to detect the SIJs (see Table 1). Second, average precision (AP) and mean average precision (mAP) were calculated based on Intersection over Union (IoU). This metric is commonly used in object detection competitions such as the PASCAL Visual Object Classes 2012 (VOC 2012) challenge (16) or in the Microsoft Common objects in context (COCO) challenge (17). IoU is an expression for the overlap of the predicted area and the actual image (ground truth) divided by the area enclosed by both (see Figure 2). A standard requirement to identify a prediction as a true positive is a ratio of 0.5 (17).

Results

In a manual review of the inferred segmentations in the tuning data, both observers found no instances and only one instance in the holdout data in which the model had only partially captured the SIJ. In the tuning dataset, one instance was found in which the model had missed the SIJ, and seven instances in the holdout dataset. All remaining 105 (tuning data, 99.1%) and 132 (holdout data, 94.3%) instances were counted as fully captured by both observers. There was no disagreement between the two observers in assigning each image to the above categories.

Object detection metrics calculated from IoU were at 0.538 for mean average precision and 0.996 for average precision in the tuning dataset. In the holdout dataset, values were 0.515 for mean average precision and 0.981 for average precision. The results are summarized in Table 1.

Discussion

Our model showed excellent results both in the manual review and displayed by high AP and mAP values. In addition, we would like to highlight that the manual review revealed only one instance of a partial capture of the SIJ. This underlines the fact that the model was very accurate and that a further increase in AP and mAP values would not necessarily reflect a more precise capture of the actual SIJ. Also, it should be noted that in the majority of cases where the model failed on holdout data radiographs included X-ray protective devices commonly absent in all the training and tuning data (see examples in Figure 3). All of this suggests that such a model could very well serve as a reliable pre-stage placed ahead of a second CNN charged with detecting specific features relevant to SpA.

Only a few studies have tried to use deep learning approaches to either detect SpA in a simple binary task or to grade them according to the modified New York Criteria. The study by Türk et al. is the only one known to us that relied on conventional radiographs to grade SpA (18). In their work authors report that while their model performed well compared to human observers accuracy still dropped when asked to differentiate between early stages of SpA (grade 0 and grade 1). Notably, no segmentation of the included images was performed. Therefore, we are confident that isolating SIJs on radiographs before predicting grades of SpA will boost model performance, especially in detecting these early stages.

We base this judgment on the possibility to reduce the need for downsizing images that contain rather subtle image information relevant to a correct prediction. By automatically cropping bounding boxes containing the SIJ and resizing the snippet to its original high resolution, we could achieve an efficient pipeline layout. Such an efficient pipeline is very important in grading SpA as the anatomy of SIJs is complex and features are often difficult to detect even for expert readers (19, 20). Also, lowering image resolution to facilitate model training can lead to unnecessary bias (10).

Deep learning models are known to be susceptible to various confounders in medical image data, such as the gender or age of the participants. (21, 22). Such confounders pose a serious problem as they can lead to overestimating the generalizability of the results (10). Furthermore, the frequently described “black box nature” of deep learning-based algorithms carries the risk that further systematic errors may be overlooked (23, 24). We are convinced that reducing image information to the clinically relevant core will help to decrease these errors.

An important prerequisite to verify the generalizability of the results is to check the performance of a CNN on holdout data from external datasets (25). As with our previous study, we have chosen two completely independent datasets for training and testing. Detection metrics used to evaluate the tuning and the holdout data (AP and mAP) were very close and therefore imply a robust performance of our model.

Beyond the application of detecting SpA, our study promises to be valuable in the broader field of digital applications in Rheumatology. Radiographic diagnostics in Rheumatology often rely on very subtle changes in radiographs such as small bony erosions, soft tissue swelling, or joint space narrowing. A similar approach could, therefore, potentially enhance any automated detection of more delicate findings on plain radiographs.

Conclusions

Convolutional neural networks are highly accurate in detecting sacroiliac joints on plain radiographs and may therefore aid in a more focused image analysis of these regions.

Abbreviations

AP: Average precision; ASAS: Assessment of Spondylarthritis international Society; CNN: Convolutional neural network; COCO: Microsoft Common objects in context; DICOM: Digital Imaging and Communications in Medicine; GESPIC: German Spondyloarthritis Inception Cohort; IoU: Intersection over union; mAP: mean Average precision; MRI: Magnetic Resonance Imaging; PASCAL VOC 2012: Pascal Visual Object Classes Challenge 2012; PROOF: Patients With Axial Spondyloarthritis (study); SIJ: Sacroiliac joint; SpA: axial Spondyloarthritis

Declarations

Ethical Approval and Consent to participate

Both PROOF and GESPIC cohorts were approved by the local ethics committees of each study center 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 center. Written informed consent to participate was obtained from all patients.

Consent for publication

Not applicable.

Availability of supporting data

The data that support the findings of this study are available from the corresponding author, N.F.G., upon reasonable request.

Competing Interests:

N.F.G, K.K.B., L.C.A., and Y.N.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 the German Research Foundation (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.

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.

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

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.

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

Funding:

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

Author’s contributions:

NFG and JLV: Data preparation, execution of model training and testing, statistical analysis, drafting and finalizing the manuscript.

KKB, YNM, DP, SMN and LCA: Essential contributions to the concept of the study, editing and correcting the manuscript.

VRR, FP, HH, and MR: Editing and correcting the manuscript.

Authors' information

1 Charité – Universitätsmedizin Berlin, Department of Radiology, Berlin, Germany

2 Berlin Institute of Health (BIH), Berlin, Germany

3 Charité – Universitätsmedizin Berlin, Department of Gastroenterology, Infectiology and Rheumatology, Berlin, Germany

4 Department of Internal Medicine and Rheumatology, Klinikum Bielefeld Rosenhöhe, Bielefeld, Germany

Acknowledgements:

LCA is participant in the BIH-Charité Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health. KKB is participant in the BIH-Charité Digital Clinician Scientist Program funded by the Charité –Universitätsmedizin Berlin and the Berlin Institute of Health.

References

1. Slobodin G, Hussein H, Rosner I, Eshed I. Sacroiliitis–early diagnosis is key. Journal of inflammation research. 2018;11:339.
2. Heuft-Dorenbosch L, Landewe R, Weijers R, Wanders A, Houben H, Van Der Linden S, et al. Combining information obtained from magnetic resonance imaging and conventional radiographs to detect sacroiliitis in patients with recent onset inflammatory back pain. Ann Rheum Dis. 2006;65(6):804–8.
3. van den Berg R, Lenczner G, Feydy A, van der Heijde D, Reijnierse M, Saraux A, 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 rheumatology. 2014;66(9):2403–11.
4. Weiss PF, Xiao R, Brandon TG, Biko DM, Maksymowych WP, Lambert RG, et al. Radiographs in screening for sacroiliitis in children: what is the value? Arthritis research therapy. 2018;20(1):141.
5. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
6. Bressem KK, Adams LC, Erxleben C, Hamm B, Niehues SM, Vahldiek JL. Comparing different deep learning architectures for classification of chest radiographs. Scientific reports. 2020;10(1):1–16.
7. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol. 2019;48(2):239–44.
8. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. Skeletal radiology. 2020:1–15.

9. Bressem KK, Vahldiek JL, Adams LC, Niehues SM, Haibel H, Rodriguez VR, et al. Detecting radiographic sacroiliitis using deep learning with expert-level accuracy in axial spondyloarthritis. medRxiv. 2020.
10. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. arXiv preprint arXiv:180700431. 2018.
11. Ferrari E, Retico A, Bacciu D. Measuring the effects of confounders in medical supervised classification problems: the Confounding Index (CI). Artif Intell Med. 2020;103:101804.
12. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, editors. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision; 2017.
13. Poddubnyy D, Inman RD, Sieper J, Ganz F, Hojnik M. Region-Specific Differences in Clinical Presentation of Patients with Axial Spondyloarthritis – Results from a Large Multinational Cohort Study [abstract]. Arthritis & Rheumatology. 2018;70.
14. Rudwaleit M, Haibel H, Baraliakos X, Listing J, Märker-Hermann E, Zeidler H, et al. The early disease stage in axial spondylarthritis: results from the German Spondyloarthritis Inception Cohort. Arthritis Rheumatism: Official Journal of the American College of Rheumatology. 2009;60(3):717–27.
15. Lin T-Y, Goyal P, Girshick R, He K, Dollár P, editors. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision; 2017.
16. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. Int J Comput Vision. 2010;88(2):303–38.
17. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D et al, editors. Microsoft coco: Common objects in context. European conference on computer vision; 2014: Springer.
18. Türk S, Cukur T, Hepdurgun C, Dar S, Karabulut A, Tamsel I et al, editors. Deep Learning Algorithm for Sacroiliac Joint Evaluation: Grading with Radiographs2020: European Congress of Radiology 2020.
19. Spoorenberg A, De Vlam K, van der Linden S, Dougados M, Mielants H, van de Tempel H, et al. Radiological scoring methods in ankylosing spondylitis. Reliability and change over 1 and 2 years. The Journal of Rheumatology. 2004;31(1):125–32.
20. Van Tubergen A, Heuft-Dorenbosch L, Schulpen G, Landewe R, Wijers R, Van Der Heijde D, et al. Radiographic assessment of sacroiliitis by radiologists and rheumatologists: does training improve quality? Annals of the rheumatic diseases. 2003;62(6):519–25.
21. Rao A, Monteiro JM, Mourao-Miranda J, Initiative AsD. Predictive modelling using neuroimaging data in the presence of confounds. NeuroImage. 2017;150:23–49.
22. Brown MR, Sidhu GS, Greiner R, Asgarian N, Bastani M, Silverstone PH, et al. ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. Front Syst Neurosci. 2012;6:69.
23. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. Jama. 2017;318(6):517–8.

24. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. Jama. 2017;318(22):2211–23.
25. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683.

Tables

Table 1 summarizes values for Average Precision (AP) and mean Average Precision (mAP) in both the tuning and holdout dataset as well as results from a manual review by two expert readers in categories of fully or partially captured SIJs as well as missed predictions. Judgment of both reviewers was identical and is therefore not further differentiated in this summary.

	SIJ captured (Manual review)			Calculated metrics	
	Fully	Partially	Missed	AP	mAP
Tuning data (n=106)	105	0	1	0.996	0.538
Holdout data (n=140)	132	1	7	0.981	0.515

Figures

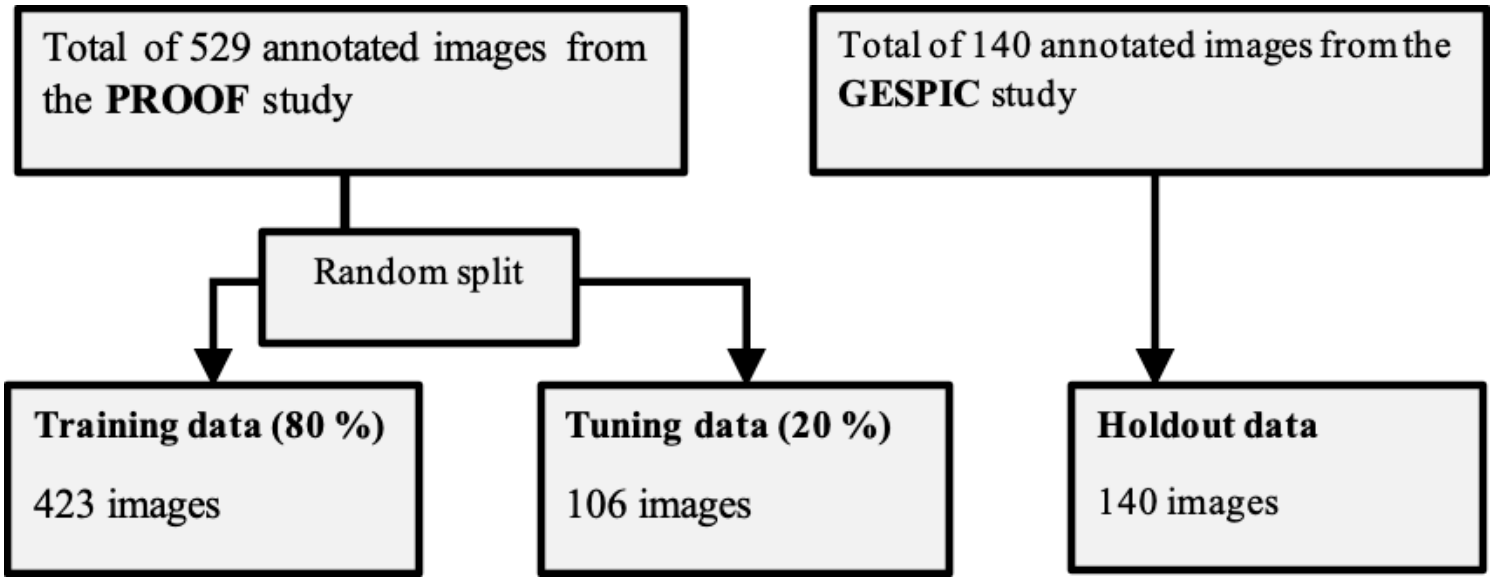

Figure 1

demonstrates the allocation of all images used in this study. Training and tuning data were retrieved from the PROOF dataset and split randomly as shown above. The holdout data was exclusively taken from the GESPIC dataset.

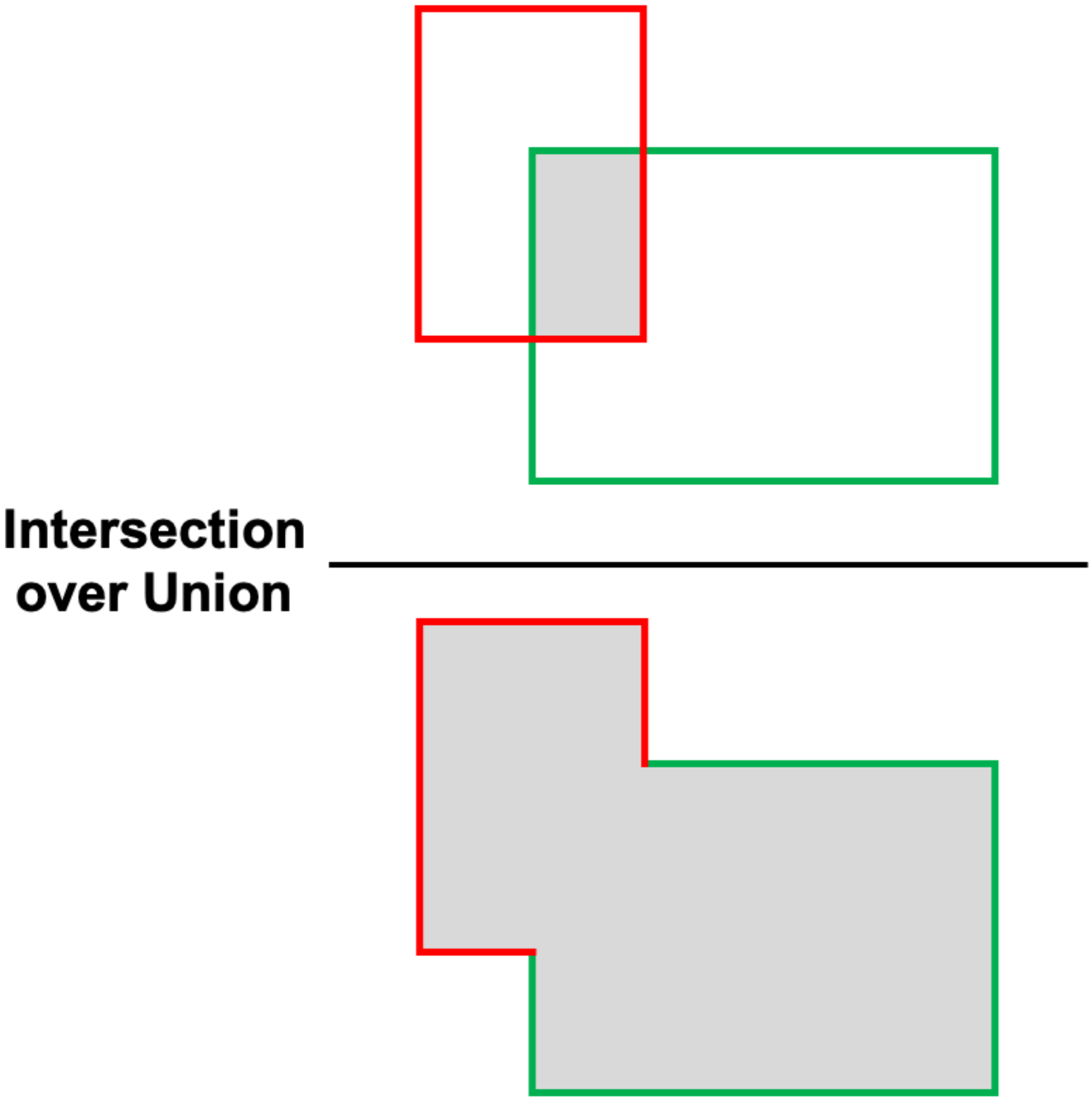

Figure 2

demonstrates the mathematical concept of the term Intersection over Union (IoU). The area of intersection (grey) between the predicted area (red) and the ground truth (green) is divided by the area enclosed (union) by both the predicted area (red) and the ground truth (green).

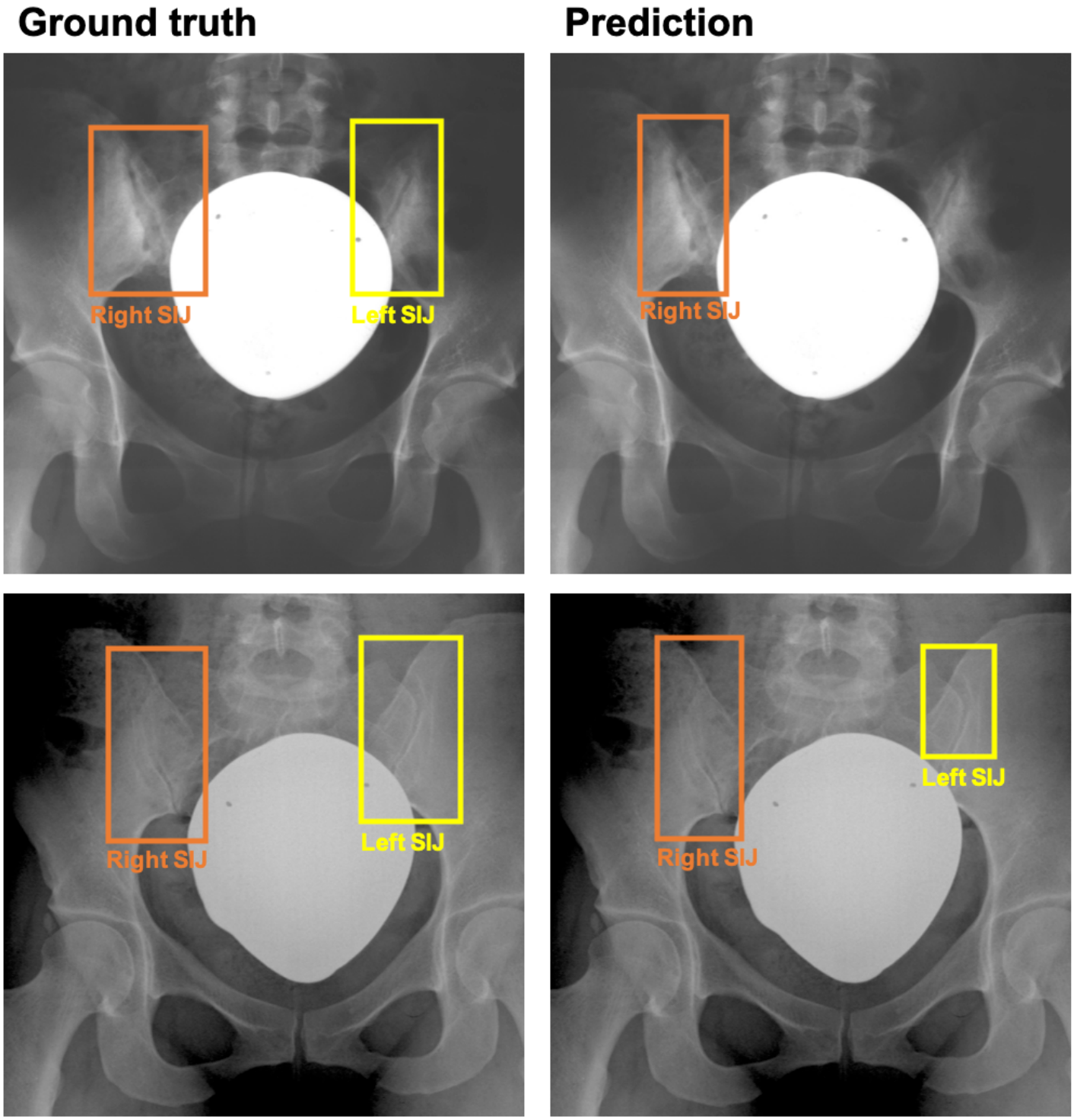

Figure 3

shows two examples from the holdout data. The top row images display ground truth (left) and prediction (right) of a case in which the left SIJ was missed. The bottom row images demonstrate an example where the left SIJ was only partially captured.