




# Improving radiologist detection of meniscal abnormality on undersampled, deep learning reconstructed knee MRI

Natalia Konovalova , MS<sup>\*1</sup>, Aniket Tolpadi, PhD<sup>1,2</sup>, Felix Liu, MS<sup>1</sup>, Zehra Akkaya , MD<sup>1,3</sup>, Johanna Luitjens, MD<sup>1</sup>, Felix Gassert, MD<sup>1</sup>, Paula Giesler, MD<sup>1,4</sup>, Rupsa Bhattacharjee, PhD<sup>1</sup>, Misung Han, PhD<sup>1</sup>, Emma Bahroos, MS<sup>1</sup>, Sharmila Majumdar , PhD<sup>1</sup>, Valentina Pedoia, PhD<sup>1</sup>

<sup>1</sup>Radiology and Biomedical Imaging Department, University of California, San Francisco, San Francisco, CA, United States

<sup>2</sup>Bioengineering Department, University of California, Berkeley, Berkeley, CA, United States

<sup>3</sup>Faculty of Medicine, Radiology Department, Ankara University, Ankara, Turkey

<sup>4</sup>Faculty of Medicine, University of Freiburg Medical Center, Freiburg, Germany

\*Corresponding author: Natalia Konovalova, MS, Radiology and Biomedical Imaging Department, University of California, San Francisco, BH203, Byers Hall, 600 16th Street, San Francisco, CA 94158, United States (natalia.konovalova@ucsf.edu)

Present address: Natalia Konovalova, BA-34, San Francisco VA Medical Center, 4150 Clement Street, San Francisco, CA 94121, United States.

## Abstract

**Background:** Accurate interpretation of meniscal anomalies on knee MRI is critical for diagnosis and treatment planning, with artificial intelligence emerging as a promising tool to support and enhance this process through automated anomaly detection.

**Purpose:** To evaluate the impact of an artificial intelligence (AI) anomaly detection assistant on radiologists' interpretation of meniscal anomalies in undersampled, deep learning (DL)-reconstructed knee MRI and assess the relationship between reconstruction quality metrics and anomaly detection performance.

**Materials and Methods:** This retrospective study included 947 knee MRI examinations; 51 were excluded for poor image quality, leaving 896 participants (mean age, 44.7 ± 15.3 years; 472 women). Using 8-fold undersampled data, DL-based reconstructed images were generated. An object detection model was trained on original, fully sampled images and evaluated on 1 original and 14 DL-reconstructed test sets to identify meniscal lesions. Standard reconstruction metrics (normalized root mean square error, peak signal-to-noise ratio, and structural similarity index) and anomaly detection metrics (mean average precision, F1 score) were quantified and compared. Two radiologists independently reviewed a stratified sample of 50 examinations unassisted and assisted with AI-predicted anomaly boxes. McNemar's test evaluated differences in diagnostic performance; Cohen's kappa assessed interrater agreement.

**Results:** On the original images, the anomaly detection model achieved the following: 70.53% precision, 72.17% recall, 63.09% mAP, and a 71.34% F1 score. Comparing performance among the undersampled reconstruction datasets, box-based reconstruction metrics showed better correlation with detection performance than traditional image-based metrics (mAP to box-based SSIM,  $r=0.81$ ,  $P<.01$ ; mAP to image-based SSIM,  $r=0.64$ ,  $P=.01$ ). In 50 participants, AI assistance improved radiologists' accuracy on reconstructed images. Sensitivity increased from 77.27% (95% CI, 65.83-85.72; 51/66) to 80.30% (95% CI, 69.16-88.11; 53/66), and specificity improved from 88.46% (95% CI, 83.73-91.95; 207/234) to 90.60% (95% CI, 86.18-93.71; 212/234) ( $P<.05$ ).

**Conclusion:** AI-assisted meniscal anomaly detection enhanced radiologists' interpretation of undersampled, DL-reconstructed knee MRI. Anomaly detection may serve as a complementary tool alongside other reconstruction metrics to assess the preservation of clinically important features in reconstructed images, warranting further investigation.

**Keywords:** knee MRI, meniscal anomalies, deep learning reconstruction, anomaly detection, AI-assisted radiology, image quality

## Abbreviations

MRI = magnetic resonance imaging; AI = artificial intelligence; DL = deep-learning; CNN = convolutional neural network; SSIM = structural similarity index measure; PSNR = peak signal-to-noise ratio; nRMSE = normalized root mean squared error; mAP = mean average precision; MICCAI = Medical Image Computing and Computer Assisted Interventions; FSE = fast spin echo; DICOM = Digital Imaging and Communications in Medicine.

## Summary

AI-assisted meniscal anomaly detection aids radiologists in reading undersampled, deep learning-reconstructed knee MR images, enhancing diagnostic accuracy.

## Key Results

- Meniscal anomaly detection was performed similarly across 14 different undersampled, deep learning (DL)-reconstructed test sets of knee MRI.
- When evaluating DL-reconstructed images radiologists' performance improved with AI assistance (accuracy with and without AI, respectively, of 88.3% [84.21, 91.49] vs 86.0% [81.62, 89.47],  $P<.05$ ).
- Conventional reconstruction metrics showed only moderate correlation with anomaly detection performance compared to region-based metrics.

Received: November 8, 2024; Revised: February 25, 2025; Accepted: March 20, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of the Radiological Society of North America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

## Introduction

Meniscal lesions are a significant concern for individuals experiencing knee pain and represent a leading cause of orthopedic surgical interventions in the United States.<sup>1</sup> Accurate diagnosis of meniscal anomalies is essential for effective clinical management and improved patient outcomes.

MRI remains a clinically important non-invasive test to diagnose meniscal anomalies, providing high-resolution soft tissue contrast.<sup>2</sup> In recent years, deep learning (DL) models, including convolutional neural networks (CNNs), transformers, and diffusion models, exhibited remarkable capabilities in accelerating MR image reconstruction from undersampled data while preserving diagnostic quality.<sup>3–6</sup> Many efforts are now focused on integrating DL-based reconstruction into clinical workflows.<sup>7,8</sup> Simultaneously, object detection algorithms, such as Faster R-CNN and YOLO, have shown promise in medical imaging by automatically identifying and localizing pathologies, aiding clinical decision-making.<sup>9–12</sup> Recent studies have demonstrated that artificial intelligence (AI)-assisted detection can improve radiologists' sensitivity and reduce reading times across different levels of expertise.<sup>13</sup>

While DL-based reconstruction continues to evolve, its impact on AI-driven downstream tasks, like object detection and segmentation, remains understudied.<sup>14,15</sup> Traditional evaluation metrics for image reconstruction, such as normalized root mean square error (nRMSE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM), focus primarily on image fidelity for visual inspection by radiologists.<sup>16–18</sup> However, optimizing these metrics alone may not be sufficient to ensure that anomaly detection models designed to assist radiologists perform effectively on DL-reconstructed images. Additionally, prior studies reported that knee MRI reconstruction models optimized for these metrics may fail to preserve small meniscal lesions and other critical structures.<sup>19</sup>

Our feasibility study investigates the relationship between image reconstruction and object detection performance by assessing whether commonly used reconstruction metrics correlate with detection accuracy. We hypothesize that AI-assisted anomaly detection can enhance radiologists' performance when interpreting reconstructed images and serve as an additional evaluation tool to assess whether reconstruction techniques adequately preserve diagnostically important features in MR images.

## Materials and methods

### Dataset

This retrospective study was conducted in accordance with a process approved by the local institution review board with waived informed consent. In our study, we collected a dataset of knee MRI examinations from the clinical population at two University of California, San Francisco (UCSF) imaging sites between June 2021 and June 2022. These patients presented a variety of knee abnormalities, including bone, cartilage, and meniscal lesions, anterior and posterior cruciate ligament (ACL and PCL) tears, and ACL-reconstructed knees. No exclusion criteria were applied upon selection. A subset of 300 patients from this dataset was previously reported.<sup>15</sup> The prior article detailed results from the Medical Image Computing and Computer Assisted Interventions (MICCAI) challenge on developing simultaneous

reconstruction and segmentation pipelines. In this manuscript, we developed an automated anomaly detection pipeline for analyzing reconstructed images and assisting in radiological evaluations.

### MRI acquisition

3D fast spin-echo (FSE) fat-suppressed images were acquired using the vendor-specific CUBE sequence (GE Healthcare) on a GE Discovery MR750 scanner with 18-channel knee transmit/receive coil and the following parameters: repetition time (TR)/echo time (TE), 1002/29 msec; field of view (FOV), 15 cm<sup>2</sup>; acquisition matrix, 256 × 256 × 200; slice thickness, 0.6 mm; echo train length, 36; readout bandwidth, ±62.5 kHz; acceleration, 4× ARC (Autocalibrating Reconstruction for Cartesian imaging), a parallel imaging technique that reduces acquisition time by undersampling k-space data while using coil sensitivities for reconstruction<sup>20</sup>; acquisition time, 4 min 58 s. Subsequently, an in-house pipeline was developed that leveraged GE Orchestra 1.10 and other post-processing tools to reconstruct images from ARC-undersampled k-space data and store them as Digital Imaging and Communications in Medicine (DICOM) files with uniform matrix dimensions of 512 × 512.

### Annotation

The data were anonymized by removing patient-sensitive information from the DICOM headers. Annotations were made using an online platform (MD.ai, New York, NY). Three radiologists (F.G. with 4 years of training, J.L. and P.G. both with 3 years of training) manually marked all knee anomalies by drawing bounding boxes on each sagittal image slice, as shown in [Figure S1](#). To calibrate the annotation procedure, the radiologists initially evaluated 15 cases together. The readers were then assigned nonoverlapping examinations and instructed to specify the anatomical location of the pathology when labeling the bounding boxes. This study focused on lesions in 6 meniscal compartments: the medial and lateral meniscal horns and bodies. The labels served as the ground truth for anomaly detection evaluation. The radiologists also identified cases with insufficient image quality for annotation. A full list of labels and counts is in [Table S1](#).

### Anomaly detection pipeline

A Faster R-CNN object detection model was used to detect anomalies on sagittal 2D image slices.<sup>11</sup> All meniscal anomaly labels were considered a single class for training purposes. The dataset was split into 80% training, 10% validation, and 10% testing partitions on the patient level, ensuring no data leaked between the splits. The training was conducted on two Tesla V100 32GB NVIDIA GPUs for a maximum of 30 epochs.

Detection performance was evaluated using precision, recall, mean average precision (mAP), and F1 score. True positive (TP) predictions were defined as those where the overlap between the predicted and actual anomaly regions, quantified by Intersection-over-Union (IoU), was at least 20%. Additionally, the confidence score for TPs was set above 70%, ensuring the model's certainty.

A detailed description of data fractionation, data augmentation,<sup>21</sup> normalization, bounding box upsampling, and the full list of adjustable training parameters are provided in [Appendix S1](#) and [Table S2](#). The source code is available at [https://github.com/konnatick/detection\\_project](https://github.com/konnatick/detection_project).

### Detection evaluation on reconstructed images

To assess the performance of anomaly detection on reconstructed images, we created 14 additional DL-reconstructed test sets using an in-house pipeline featuring 8× accelerated image reconstruction.<sup>22</sup> Each set included normalized reconstructed images for the same patients as in the original detection test set, representing the inference results of KIKI-inspired I-Net<sup>23</sup> and similar architecture UNet models trained with specific combinations of common loss functions. We also generated zero-filled and fully sampled sets for comparison. More details are available in [Appendix S1](#).

We used 3 standard metrics to evaluate reconstruction performance: nRMSE, PSNR, and SSIM. Reconstruction metrics were calculated using 2 approaches: a standard method based on the entire 3D image volume and an alternative method focused on pixels within predicted bounding boxes. The calculation algorithm is detailed in [Figure S2](#). Object detection metrics were also computed for all reconstructed sets to evaluate the possible link between reconstruction performance and object detection outcomes.

### AI-assisted reading

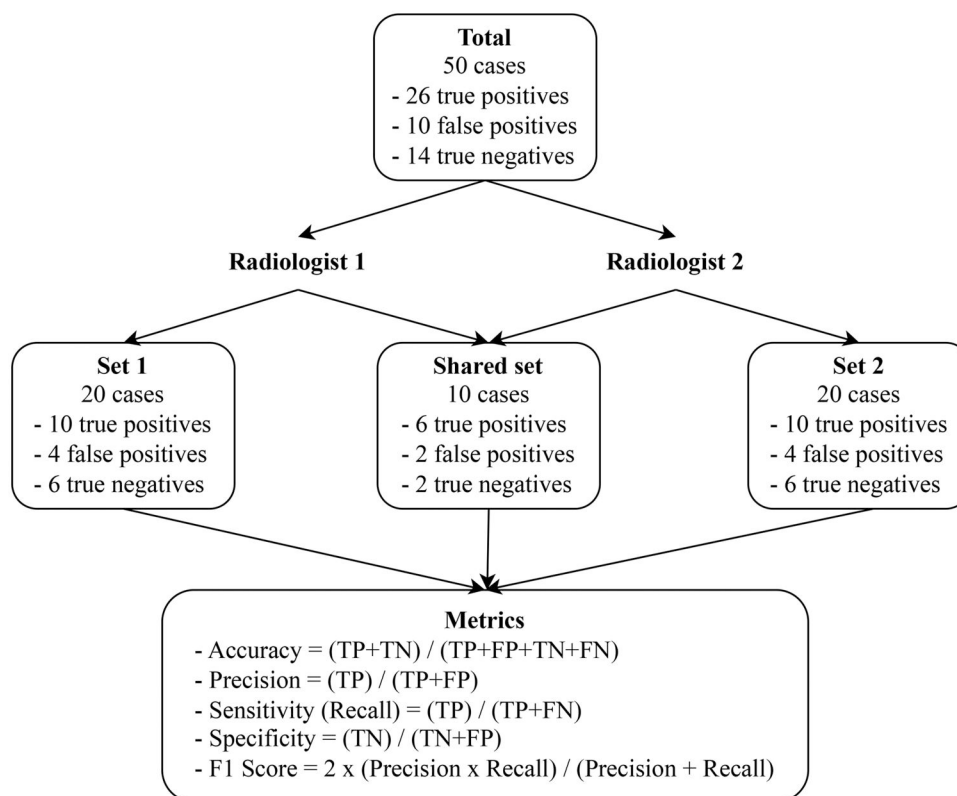
To assess the impact of an AI assistant on reading results, 2 radiologists (Z.A. with 10 years of clinical experience and J. L. with 3 years of experience) each independently evaluated 50 studies with 10 overlapping to assess interrater variability. The case selection scheme is shown in [Figure 1](#). Random stratification was applied to ensure a diverse distribution of predictions from the anomaly detection model in each set of

30 studies: 16 TPs with at least 1 lesion per patient, 6 false positives (FPs), and 8 anomaly-free cases.

Each radiologist was tasked with identifying pathologic or non-pathologic conditions in the 6 meniscal compartments under 4 conditions: original DICOM images (ground truth), reconstructed images without AI assistance, and original and reconstructed images with AI assistance using predicted anomaly boxes. The selection of the reconstructed image set for this analysis was based on the highest SSIM score. Readings were conducted using MD.ai, with a 2-week wash-out period between sessions. Additionally, 2 weeks after the readings, an expert radiologist (Z.A.) compared the results of readings performed on original images with and without boxes to identify any anomalies that may have been missed during the initial ground truth annotation. This experiment was conducted more than a year after the initial dataset annotation by J.L., which minimized the reader's bias.

### Statistical analysis

Spearman's correlation and its corresponding *P*-value were employed to assess the relationships between reconstruction and detection metrics. Additionally, one-way ANOVA was performed to compare the means of metrics' distributions, followed by a post hoc pairwise *t*-test to determine any deviating groups. The statistical significance of AI-assisted reading results was evaluated using McNemar's test, which assesses changes in paired categorical data, making it well-suited for comparing radiologists' decisions with and without AI assistance. Cohen's kappa was used to assess interrater



**Figure 1.** Study design for the AI-assisted radiologist reading experiment. A total of 50 knee MRI cases were evaluated, including 26 true positives (TP), 10 false positives (FP), and 14 true negatives (TN). Each radiologist reviewed 30 cases: 20 unique to their set and 10 shared cases to assess interrater agreement. True positives (TP) represent lesions correctly detected by the model, while false positives (FP) correspond to incorrectly detected lesions that were not present in the ground truth. TN indicate cases correctly identified as anomaly-free. Radiologists' performance metrics, including accuracy, precision, sensitivity (recall), specificity, and F1 score, were calculated based on the confusion matrix.

reliability, with  $z$ -test applied to compare changes.<sup>24,25</sup> A post hoc power analysis using a two-proportion  $z$ -test was conducted to assess the sample size for the reader study.

## Results

### Dataset characteristics

A total of 947 knee MRI examinations were obtained; 51 were excluded due to poor image quality caused by motion artifacts or an incorrect imaging sequence, leaving a dataset of 896 examinations (175 492 slices). Of these, 406 patients had at least 1 meniscal abnormality, yielding a total of 18 059 bounding boxes drawn, and others had healthy menisci. The mean age was  $44.7 \pm 15.3$  years, the mean weight was  $74.8 \pm 15.8$  kg, and 52.7% (472 of 896) were females. The database flowchart is shown in Figure 2. Additional demographic characteristics of data partitions are summarized in Table 1.

### Anomaly detection and reconstruction performance

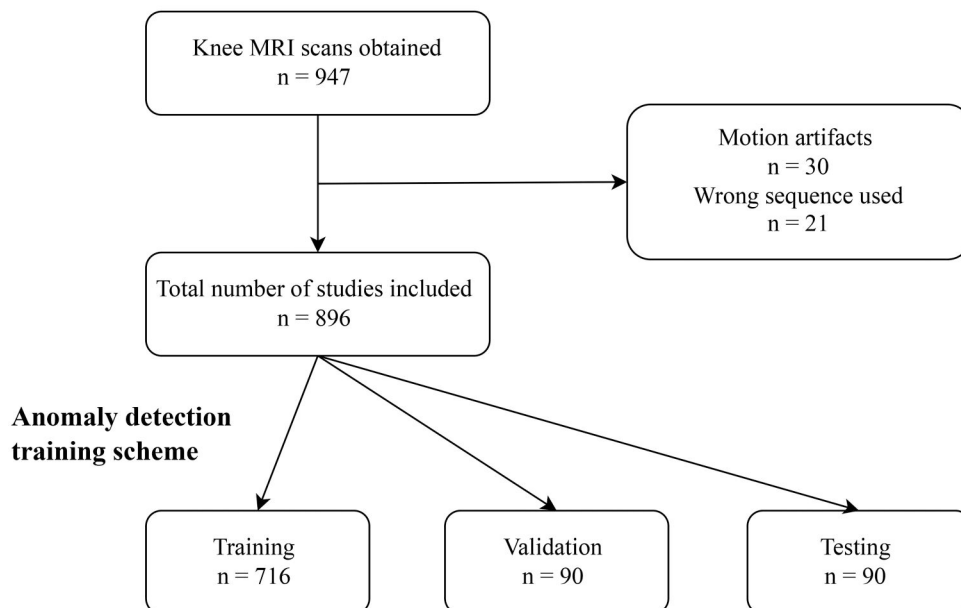
The anomaly detection model achieved the following results on original images: 70.53% precision, 72.17% recall, 63.09% mAP, and a 71.34% F1 score. For a summary and examples of its performance on reconstruction test sets, refer to Table 2 and Figure 3. Faster R-CNN performed well on reconstructed test sets, revealing strong correlations between box-based nRMSE, PSNR, and standard reconstruction metrics ( $r=1.00$ ,  $P<.05$  for both metrics). Box-based SSIM exhibited a robust correlation with image-wide SSIM ( $r=0.76$ ,  $P<.05$ ). Moderate to strong correlations were observed between anomaly detection performance metrics (mAP, F1 score) and reconstruction quality indicators such as nRMSE and PSNR. Notably, image-based SSIM demonstrated only a moderate or insignificant correlation with detection metrics (mAP:  $r=0.64$ ,  $P<.05$ ; F1:  $r=0.38$ ,  $P>.05$ ), whereas box-based SSIM exhibited a stronger correlation (mAP:  $r=0.81$ ,  $P<.05$ ; F1:  $r=0.65$ ,  $P<.05$ ). Further

details and visual representations of these relationships are provided in Figure S3. Additionally, the comparison of mean values of SSIM, PSNR, and nRMSE calculated separately for slices with TP, FP, and false negative (FN) predictions did not reveal any significant patterns that would suggest a strong predictor of detection performance. An example of this analysis is presented in Figure 4, with plots for other reconstruction models depicted in Figure S4. No significant correlation was observed between image-based reconstruction metrics and prediction confidence scores, as demonstrated in Table S3.

### AI-assisted reading results

The reconstructed images from the UNet with L1 loss were selected for assessment because the model showed top results in both reconstruction and detection evaluation. Reading results are reported in Table 3. Reading without AI assistance resulted in fair interrater agreement, as measured by Cohen's kappa, for both the original DICOM image set ( $k=0.41$ ) and the reconstructed image set ( $k=0.39$ ). The addition of anomaly boxes predicted by an AI assistant increased agreement for both sets of readings, resulting in  $k=0.60$  for the original image set and  $k=0.57$  for the reconstructed image set, respectively. However, the  $P$ -values from the  $z$ -test comparing the increase in Cohen's kappa were above 0.05 in both cases, indicating that the observed improvements were not statistically significant, likely due to the small sample size selected to assess interrater agreement. Radiologists' overall performance, as measured by accuracy, precision, recall, specificity, and F1 score, improved (McNemar's test  $P$ -values  $<.05$ ) when reading the reconstructed images assisted with boxes, as shown in Table 4 and Figure 5. A post hoc power analysis (80% power,  $\alpha = 0.05$ ) indicated that detecting a 3% improvement in radiologists' performance would require up to 3861 cases, while a 5% improvement would require up to 1367 cases, with the full analysis provided in Table S4.

### Study inclusion/exclusion



**Figure 2.** Data flow with study inclusion and exclusion criteria along with the anomaly detection model training scheme.



The comparison of results from reading original DICOM images with and without AI-predicted boxes led to the reclassification of FP cases, which were not reported during the initial annotation, as TPs and the subsequent addition of 17 new lesions to the dataset, with examples shown in Figure 6. Of note, due to random stratification, the reading set had a higher proportion of FP cases—instances where the model detected anomalies not annotated in the initial dataset. This was a deliberate decision to evaluate whether AI assistance could enhance the dataset annotation process.

## Discussion

In this study, we examine the relationship between image reconstruction and anomaly detection, specifically within the context of meniscal anomalies on undersampled knee MRI, aiming to enhance the assessment of reconstruction quality

**Table 1.** Patient demographics and data partitioning.

Parameter <sup>a</sup>	Training	Validation	Testing
Demographic characteristics			
Number of patients	716	90	90
Number of females (males)	376 (340)	48 (42)	48 (42)
Mean age (years)	44.2 ± 15.5	48.1 ± 14.8	44.7 ± 14.1
Mean weight (kg)	75.3 ± 15.9	72.7 ± 14.6	73.4 ± 16.0
Anomaly distribution			
Number of slices	140 202	17 640	17 650
Anomaly boxes	14 550	1 714	1 795
Slices with anomaly boxes	12 857	1 533	1 590

<sup>a</sup>Age and weight were calculated as mean ± standard deviation. Data partitions were 80% training, 10% validation, and 10% testing.

and to evaluate the utility of the detector for routine clinical interpretation of DL-based reconstructed images.

Correlation analysis between structural similarity (SSIM) and object detection metrics highlights anomaly detection as a valuable tool for evaluating image reconstruction models, particularly for further image analysis. Our results showed that box-based SSIM exhibited a stronger correlation with detection metrics than image-based SSIM, suggesting that evaluating reconstruction quality in regions containing clinically relevant features may provide additional insights. Importantly, while conventional reconstruction metrics assess global image fidelity, anomaly detection directly evaluates the preservation of diagnostically significant structures.<sup>26</sup> To further explore this, we analyzed mean image-based SSIM values separately for TP, FP, and FN detection predictions. No consistent pattern was observed linking higher SSIM values to an increase in TP detections, though some models showed significant differences in the mean SSIM of FN predictions. These findings suggest that anomaly detection offers a unique perspective in reconstruction evaluation by assessing the retention of key diagnostic features.

Another discovery from our research is the beneficial impact of AI-assisted anomaly detection on the performance of radiologists interpreting reconstructed images. The use of AI-predicted anomaly boxes improved their performance in terms of accuracy, precision, recall, specificity, and F1 score. Implementing AI tools like anomaly detection algorithms is crucial in effectively integrating DL-based image reconstruction into clinical workflows. This may enhance the diagnostic accuracy and consensus among radiologists, which can be variable even among seasoned practitioners.<sup>27</sup>

**Table 2.** Detection and reconstruction performance on undersampled knee MRI.

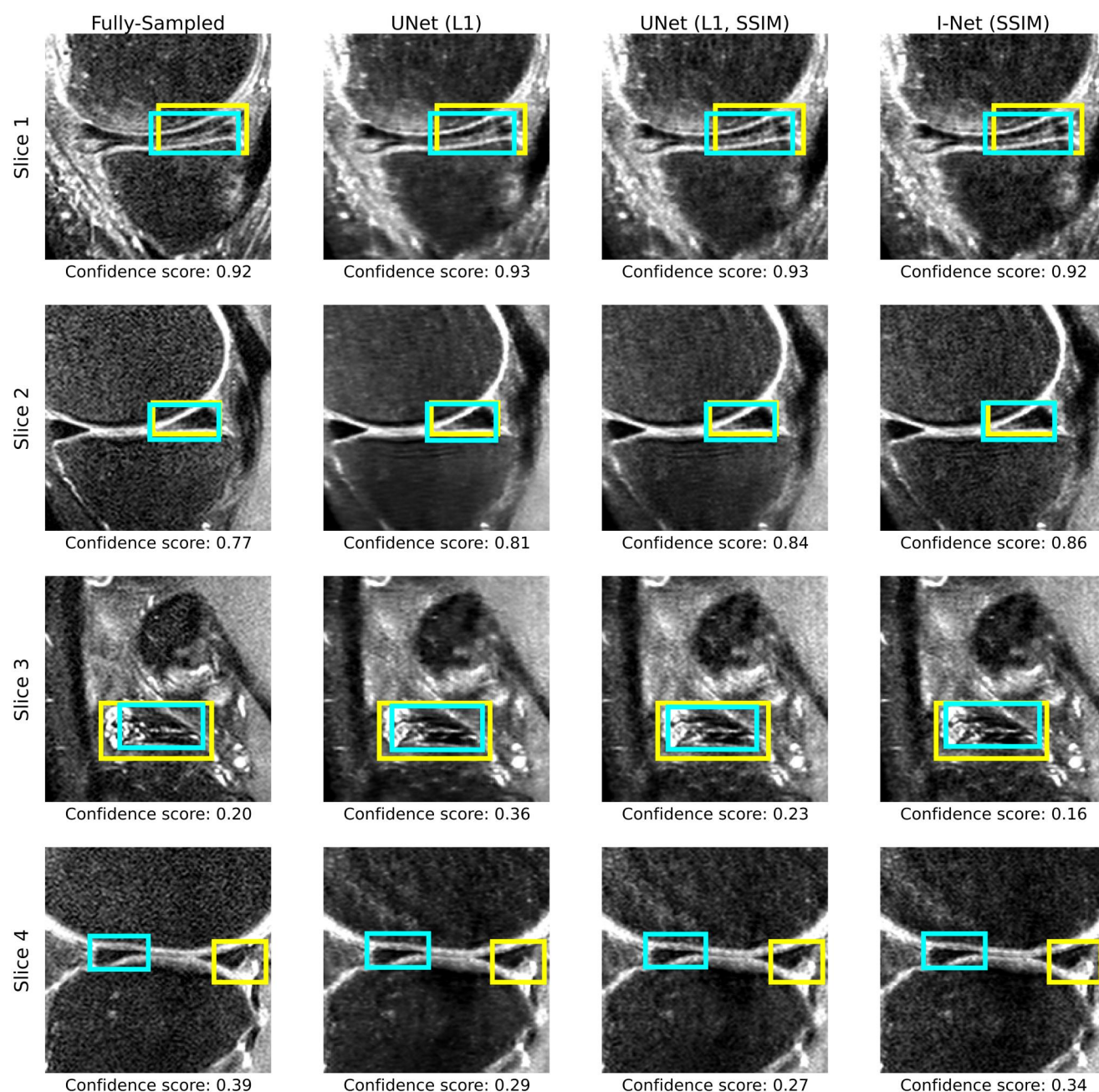
#	Deep learning reconstruction model	Recon image-based metrics <sup>a</sup>			Detection metrics <sup>b</sup>				Recon boxes-based metrics		
		nRMSE	PSNR	SSIM	Precision	Recall	mAP	F1	nRMSE	PSNR	SSIM
1	Zero-filled	0.45 ± 0.12	24.12 ± 1.97	29.36 ± 5.38	83.80	42.04	39.65	55.99	0.45 ± 0.12	12.71 ± 2.34	29.99 ± 10.67
2	UNet (k-space)	0.08 ± 0.05	31.84 ± 2.76	72.61 ± 4.69	61.83	73.06	61.12	66.98	0.08 ± 0.04	20.86 ± 2.47	67.13 ± 9.62
3	UNet (L1)	<b>0.02 ± 0.01</b>	<b>36.75 ± 1.97</b>	<b>81.63 ± 2.59<sup>c</sup></b>	<b>69.79</b>	<b>71.06</b>	<b>61.48</b>	<b>70.42</b>	<b>0.03 ± 0.02</b>	<b>24.48 ± 2.18</b>	<b>79.65 ± 5.79</b>
4	UNet (SSIM)	0.05 ± 0.02	34.16 ± 1.90	78.82 ± 2.78	67.67	66.29	56.92	66.97	0.05 ± 0.02	22.10 ± 2.19	76.50 ± 5.44
5	UNet (L1, k-space)	0.20 ± 0.09	27.93 ± 3.97	62.91 ± 8.54	71.96	66.43	58.84	69.08	0.20 ± 0.09	16.61 ± 2.93	57.53 ± 11.42
6	UNet (k-space, SSIM)	0.07 ± 0.04	33.09 ± 3.23	75.02 ± 4.79	69.45	70.16	60.60	69.80	0.07 ± 0.04	21.30 ± 2.88	75.54 ± 7.18
7	UNet (L1, SSIM)	<b>0.04 ± 0.02</b>	<b>35.35 ± 2.28</b>	<b>81.27 ± 2.62</b>	<b>67.81</b>	<b>71.47</b>	<b>61.43</b>	<b>69.59</b>	<b>0.04 ± 0.02</b>	<b>22.84 ± 2.25</b>	<b>79.64 ± 5.22</b>
8	UNet (L1, k-space, SSIM)	0.55 ± 0.10	23.09 ± 2.17	38.25 ± 7.48	83.63	49.94	46.02	62.54	0.52 ± 0.08	11.85 ± 2.00	25.34 ± 6.77
9	I-Net (k-space)	0.26 ± 0.13	27.04 ± 4.21	58.02 ± 10.48	68.92	65.26	56.12	67.04	0.26 ± 0.10	15.66 ± 2.90	47.37 ± 12.02
10	I-Net (L1)	0.03 ± 0.01	36.08 ± 2.07	74.43 ± 3.37	69.81	70.03	60.66	69.92	0.03 ± 0.01	24.21 ± 2.12	78.81 ± 5.94
11	I-Net (SSIM)	0.04 ± 0.02	34.54 ± 2.57	77.15 ± 3.10	<b>70.20</b>	<b>70.61</b>	<b>61.54</b>	<b>70.41</b>	0.05 ± 0.02	22.40 ± 2.12	77.89 ± 4.88
12	I-Net (L1, k-space)	0.03 ± 0.01	35.81 ± 2.44	76.54 ± 3.47	70.85	68.30	59.52	69.55	0.03 ± 0.01	24.03 ± 1.98	77.85 ± 5.94
13	I-Net (k-space, SSIM)	0.17 ± 0.11	29.32 ± 4.89	65.05 ± 9.27	73.88	66.68	58.51	70.10	0.16 ± 0.08	17.65 ± 3.63	64.94 ± 10.28
14	I-Net (L1, SSIM)	0.03 ± 0.01	35.44 ± 2.34	72.25 ± 3.81	68.42	71.08	61.02	69.72	0.04 ± 0.01	23.40 ± 2.05	79.15 ± 5.11
15	I-Net (L1, k-space, SSIM)	0.03 ± 0.01	35.66 ± 2.52	73.82 ± 3.63	<b>69.80</b>	<b>71.21</b>	<b>61.18</b>	<b>70.50</b>	0.04 ± 0.01	23.58 ± 2.08	79.44 ± 5.00
16	Fully sampled	—	—	—	74.70	66.88	59.29	70.57	—	—	—

Abbreviations: nRMSE = normalized root mean square error, PSNR = peak signal-to-noise ratio, SSIM = structural similarity index measure, reported in %. All detection metrics are reported in %. mAP = mean average precision, F1 = F1 score.

<sup>a</sup>nRMSE quantifies the average intensity error between reconstructed and reference images, normalized by the image intensity range. PSNR measures the ratio of maximum signal intensity to noise, expressed in decibels, with higher values indicating better reconstruction quality. SSIM evaluates perceptual similarity by comparing luminance, contrast, and structure between images, where a value closer to 1 indicates higher similarity.

<sup>b</sup>Precision measures the proportion of correctly identified anomalies (true positives) among all identified anomalies. Recall represents the proportion of correctly identified anomalies relative to all actual anomalies in the dataset. Mean average precision (mAP) summarizes detection accuracy across a range of Intersection-over-Union (IoU) thresholds, reflecting how well the predicted anomaly regions overlap with the actual regions. F1 score is the harmonic mean of precision and recall, balancing these 2 aspects of performance.

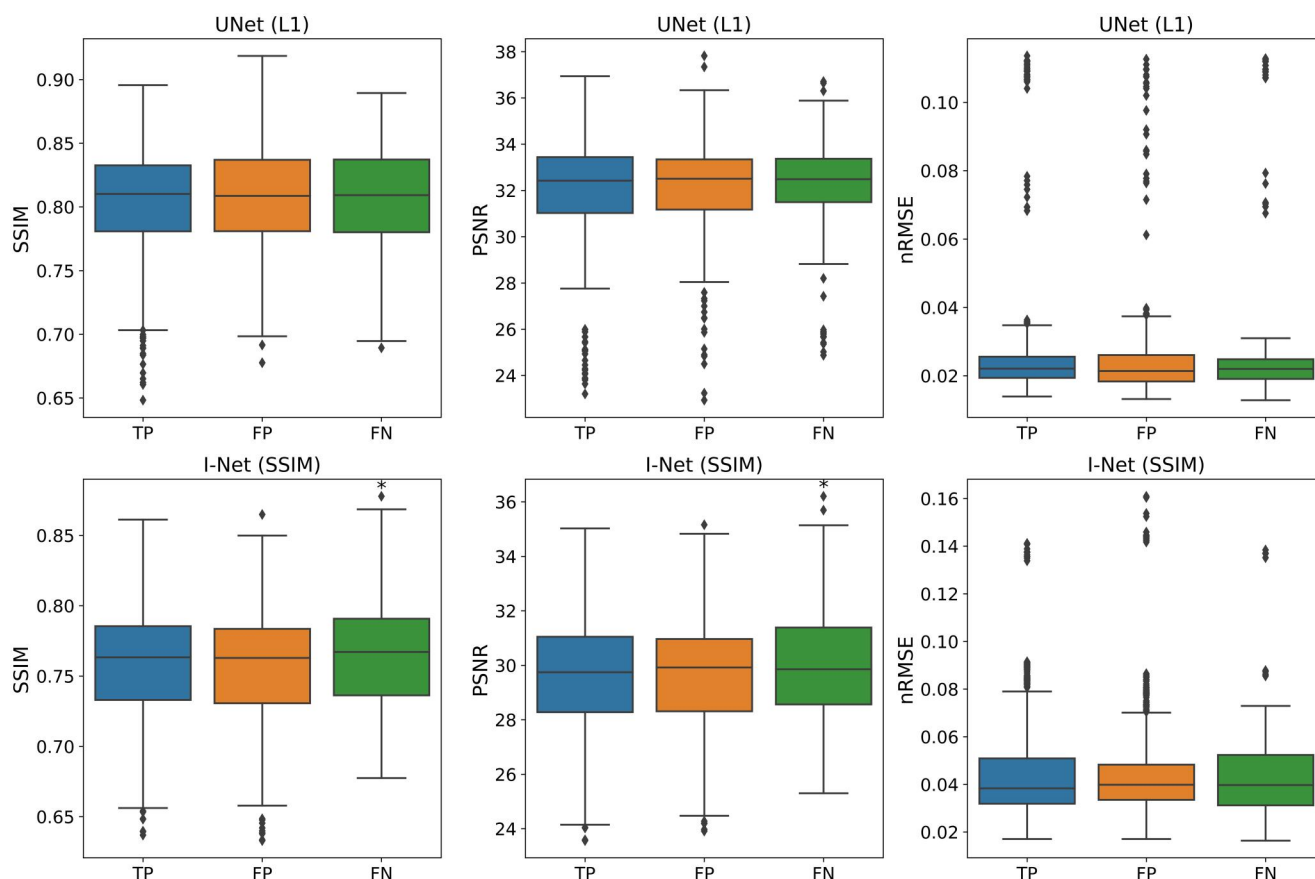
<sup>c</sup>Bold numbers indicate the 2 best-performing reconstruction models based on nRMSE, PSNR, and SSIM, and the 4 best-performing reconstruction models based on object detection metrics. Note that the UNet with the combined L1 and SSIM loss function was not the second-best performer in the detection experiment.



**Figure 3.** Anomaly detection on undersampled knee MRI datasets. Rows represent 4 different sagittal slices; columns show fully sampled and top 3 reconstruction models' outcomes. Ground truth boxes are in yellow, and predicted boxes are in blue. Slice 1: the lesion is visible, and the detection model yielded high-confidence true positive predictions. Slice 2: the lesion is less visible, and predictions are still within the confidence threshold of 0.75 and considered true positive. Slice 3: the lesion is visible; however, the model failed to predict with high confidence. Slice 4: the detection model fails to predict the lesion's correct location.

Many factors can affect annotation quality for model training, including the visualization software and hardware used, the availability of images from different MRI sequences, the experience level of the annotators, and potential desensitization from reviewing large datasets. AI-assisted anomaly detection can help mitigate these challenges by refining and validating annotations, ultimately improving dataset quality. In our study, AI-assisted review led to the addition of 17 previously unmarked lesions to the initial dataset. This demonstrates AI's potential as both a diagnostic aid and a tool for enhancing ground truth accuracy and assisting in radiology education.<sup>28–30</sup>

Our study has several limitations. We focused narrowly on knee MRI examinations featuring meniscal anomalies. Thus, our findings might not extend as effectively to different anatomical areas or other types of anomalies, such as cartilage abnormalities. There is a need for future research to test the wider applicability of our results in various clinical contexts. Our analysis was conducted on knee MRI data acquired using a single scanner, protocol, and field strength, which limits the generalizability of our results. Specifically, we employed a 3D FSE fat-suppressed CUBE MRI sequence that, despite its effectiveness in our research, might not be the most common in diverse clinical settings.<sup>31</sup>



**Figure 4.** The comparison of mean slice-based SSIM with detection performance. Each slice was classified as having true positive (TP), false positive (FP), or false negative (FN) predictions. A predicted box with a confidence score  $>0.70$  and  $\text{IoU} > 0.20$  was considered a TP. If a slice had multiple types of predictions, its SSIM was included in each of the 3 groups. One-way ANOVA tests, followed by paired  $t$ -tests, were used to determine significant differences in means, marked with asterisks. The plot displays results for the 2 best-performing reconstruction models based on classic SSIM, with additional plots shown in Figure S4. The results revealed no specific pattern indicating that classic SSIM significantly influenced TP, FP, or FN predictions. Among the 14 models analyzed, 8 showed significantly higher mean PSNR for the FP group compared to TP and FN, while 7 models exhibited lower nRMSE for the FP group. Abbreviations: nRMSE = normalized root mean square error, PSNR = peak signal-to-noise ratio, SSIM = structural similarity index measure.

**Table 3.** Confusion matrix for AI-assisted reading of meniscal abnormality on undersampled deep learning reconstructed knee-MRI.

Counts <sup>a</sup>	Combined		Radiologist 1		Radiologist 2	
	Without boxes	With boxes	Without boxes	With boxes	Without boxes	With boxes
Total labels	300	300	180	180	180	180
Anomaly labels	66	66	30	30	50	50
No anomaly labels	234	234	150	150	130	130
True positives (TP)	51	53	18	23	46	43
False positives (FP)	27	22	11	12	24	17
True negatives (TN)	207	212	139	138	106	113
False negatives (FN)	15	13	12	7	4	7

<sup>a</sup>Reading was performed on 50 cases, with each radiologist evaluating 30 cases, including an overlap of 10 shared cases. Each radiologist assessed the presence (1) or absence (0) of a lesion in 6 meniscal compartments: medial anterior horn, medial body, medial posterior horn, lateral anterior horn, lateral body, and lateral posterior horn. The confusion matrix was generated based on these assessments. TPs are correctly identified lesions, FPs are incorrectly identified lesions, TNs are correctly identified cases without lesions, and FNs are missed lesions.

While our findings demonstrated successful detection on both original and reconstructed images, broader validation across different imaging vendors, magnet strengths, and acquisition settings is necessary. Additionally, since reconstruction methods inherently alter image characteristics, further research is needed to determine whether performance differences arise from reconstruction itself or the model's ability to generalize. Future studies should evaluate anomaly detection models across diverse reconstruction techniques, assess

training models directly on reconstructed images, and test performance on datasets with varying degrees of image fidelity.

Furthermore, our study relied on older reconstruction models to integrate image reconstruction with detection pipelines. Newer methodologies, such as variational networks and transformers, warrant investigation for potentially deeper insights.<sup>3,32</sup> Recent literature also indicates that the loss of fine but clinically significant features could stem from

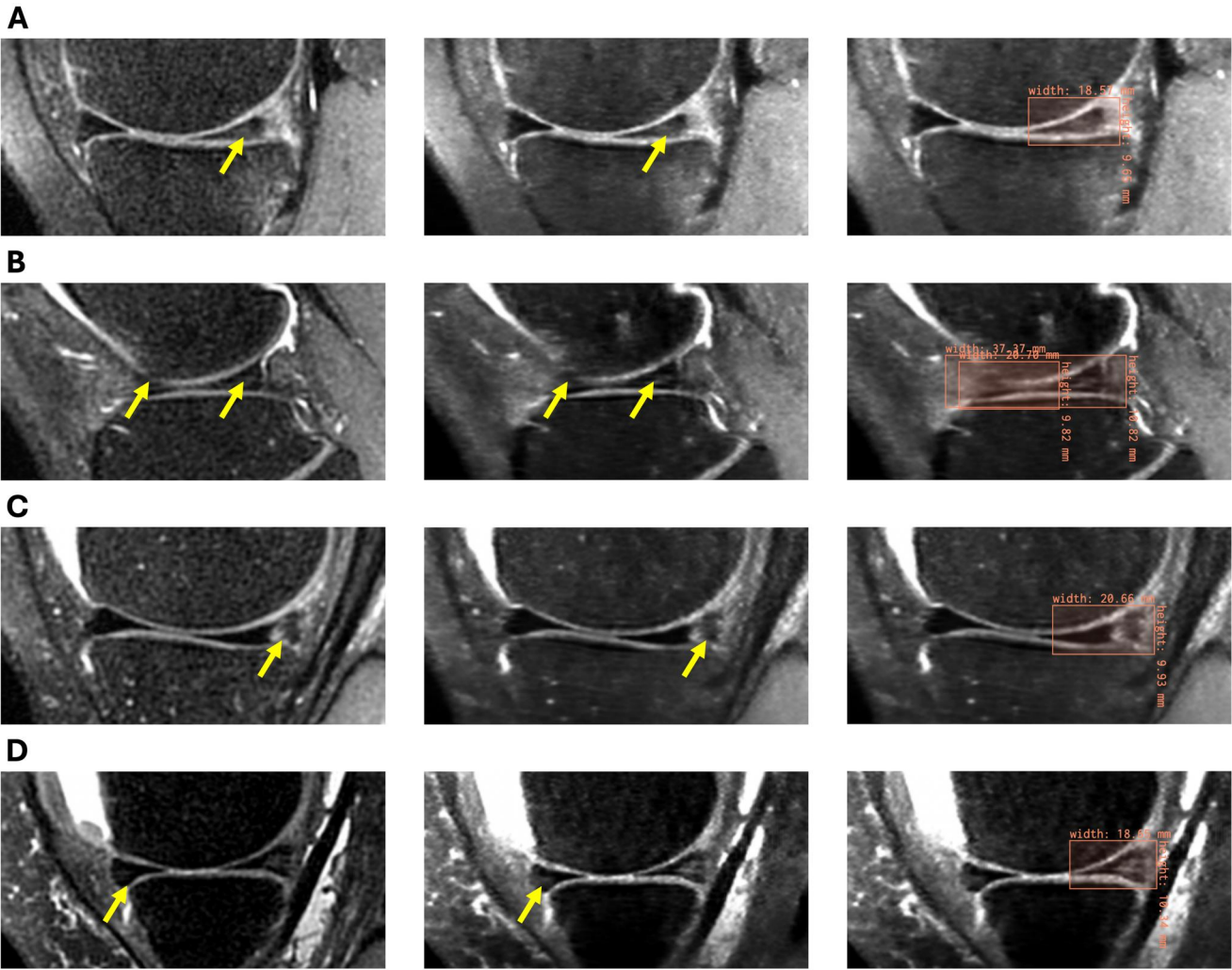


**Table 4.** Radiologist detection of meniscal abnormality on undersampled deep learning reconstructed knee-MRI with AI assistance.

Metric <sup>a</sup>	Combined [95% CI]		Radiologist 1 [95% CI]		Radiologist 2 [95% CI]	
	Without boxes	With boxes	Without boxes	With boxes	Without boxes	With boxes
Accuracy, %	86.00 [81.62, 89.47]	88.33 [84.21, 91.49]	87.22 [81.56, 91.33]	89.44 [84.10, 93.14]	84.44 [78.44, 89.01]	86.67 [80.93, 90.87]
Precision, %	65.39 [54.33, 75.00]	70.67 [59.56, 79.76]	62.07 [44.00, 77.31]	65.71 [49.15, 79.17]	65.71 [54.04, 75.75]	71.67 [59.23, 81.49]
Sensitivity (recall), %	77.27 [65.83, 85.72]	80.30 [69.16, 88.11]	60.00 [42.32, 75.41]	76.67 [59.07, 88.21]	92.00 [81.16, 96.85]	86.00 [73.81, 93.05]
Specificity, %	88.46 [83.73, 91.95]	90.60 [86.18, 93.71]	92.67 [87.35, 95.86]	92.00 [86.54, 95.37]	81.54 [74.00, 87.27]	86.92 [80.05, 91.67]
F1 score <sup>b</sup> , %	70.83	75.18	61.02	70.77	76.67	78.18
McNemar's <i>P</i> -value of the confusion matrix	<0.05	<0.05	>0.05	>0.05	<0.05	>0.05
Cohen's kappa	0.39 [0.12, 0.67]	0.57 [0.33, 0.81]	—	—	—	—

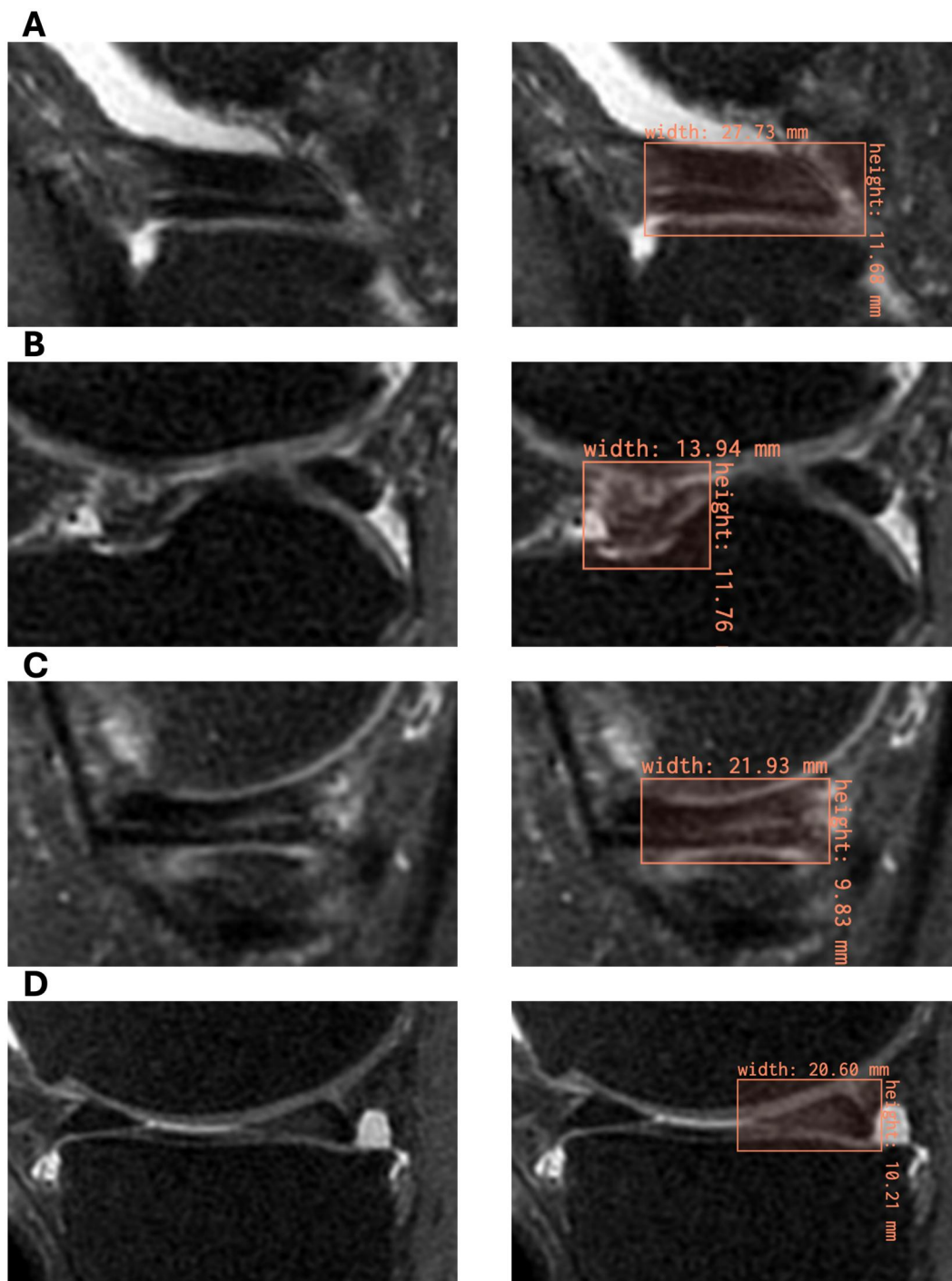
<sup>a</sup>Accuracy measures the proportion of correctly classified cases out of the total number of cases. Precision indicates the proportion of correctly identified lesions among all cases classified as lesions. Sensitivity (recall) measures the proportion of actual lesions that were correctly identified. Specificity reflects the proportion of correctly identified cases without lesions. F1 score is the harmonic mean of precision and recall, providing a balance between false positives and false negatives. The statistical significance of AI-assisted reading results was determined using McNemar's test applied to the confusion matrix, which assesses differences across accuracy, precision, sensitivity, and specificity. The *P*-values in the table reflect whether the observed variations in performance were likely due to chance. Cohen's kappa was calculated based on 10 shared cases between the readers' datasets to assess inter-reader agreement.

<sup>b</sup>Confidence intervals are not reported for F1 score as it is a derived metric from precision and recall, which already have their own confidence intervals.



**Figure 5.** Knee MRI cases where AI assistance improved reader sensitivity and specificity for meniscal lesions. The first column shows the original images, while the second and third columns display the same reconstructed image without and with model-predicted boxes, respectively. All lesions shown were already present in the ground truth dataset. An arrow indicates meniscal abnormality. (A) A lesion in the medial posterior meniscal horn was missed when reviewing the reconstructed image without annotation but was correctly identified with AI assistance. (B) Lesions spanning the lateral anterior horn and lateral meniscal body were initially overlooked in the reconstructed image but detected when AI predictions were available. (C) A lesion in the medial posterior horn was missed without AI assistance but recognized when reading the same reconstructed image with predicted boxes. (D) A lesion was initially marked in both the medial anterior and posterior horns when reading the reconstructed image without boxes. However, with AI assistance, the medial anterior horn was reclassified as lesion-free. Upon later review, the radiologist confirmed that no lesion was present in the anterior horn, attributing the shadow to an image artifact rather than a true abnormality.





**Figure 6.** Meniscal lesions on knee MRI that were initially missed from the ground truth original DICOM dataset and later added after reviewing AI predictions. The first column shows the original image, while the second column displays model-predicted anomalies. (A) A lesion in the lateral meniscal body was not included in the original annotations but was later added after AI-assisted review of the original images. (B) A lesion in the lateral anterior horn was missed in the initial dataset and later incorporated based on AI predictions. (C) A lesion in the medial meniscal body, originally absent from the ground truth annotations, was added after reviewing predicted anomalies. (D) A lesion in the medial posterior horn, initially overlooked, was included in the dataset following model-assisted review.

the chosen undersampling technique, not solely the reconstruction algorithm, as different random sampling seeds can influence the performance of reconstruction models.<sup>33</sup> These findings emphasize the need for further research to determine the most effective anomaly detection strategies in various reconstruction settings.

Finally, the potential role of AI-assisted anomaly detection in refining dataset annotation should be further explored on

larger datasets. In our study, AI-assisted review led to the reclassification of initially unmarked lesions, highlighting its potential for refining ground truth annotations. However, while McNemar's test indicated statistically significant improvements in radiologists' performance with AI assistance, the post hoc power analysis revealed that the study was underpowered to detect small effect sizes, emphasizing the need for larger validation cohorts. Additionally, due to

the small sample size, the increase in the interrater agreement in our study did not reach the statistical significance. Systematic evaluation of AI's role in annotation workflows across more diverse datasets is necessary to assess its impact on training data quality and model performance.

Our feasibility study underscores the critical role of anomaly detection in the assessment of DL-based image reconstruction models, particularly when aiming to maintain clinically significant features. Incorporating anomaly detection and box-based reconstruction metrics is likely important in evaluating reconstruction models for downstream applications. Looking ahead, we suggest investigating reconstruction and anomaly detection pipelines that operate concurrently, using the detector as a penalizing factor, and exploring the use of object detection to refine undersampling patterns to optimize reconstruction results. Moreover, our research points to the beneficial effects of AI-assisted detection on the interpretive accuracy of radiologists, thereby expanding the potential for clinical use.

### Author contributions

Natalia Konovalova (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Aniket Tolpadi (Data curation, Investigation, Methodology, Software, Validation, Writing—review & editing), Felix Liu (Methodology, Supervision, Writing—review & editing), Zehra Akkaya (Data curation, Investigation, Validation, Writing—review & editing), Johanna Luitjens (Data curation, Investigation, Validation, Writing—review & editing), Felix Gassert (Data curation, Writing—review & editing), Paula Giesler (Data curation, Writing—review & editing), Rupsa Bhattacharjee (Data curation, Writing—review & editing), Misung Han (Data curation, Writing—review & editing), Emma Bahroos (Data curation, Project administration, Writing—review & editing), and Sharmila Majumdar (Conceptualization, Project administration, Resources, Supervision, Writing—review & editing), and Valentina Podoia (Conceptualization, Project administration, Resources, Supervision, Writing—review & editing)

### Supplementary material

Supplementary material is available at *Radiology Advances* online.

### Funding

This work was funded the National Institutes of Health (NIH) under grant number R01AR078762 (April 1, 2021–March 31, 2026).

### Conflicts of interest

Please see ICMJE form(s) for author conflicts of interest. These have been provided as [supplementary materials](#). The authors have no conflict of interest to disclose.

### Data availability

The data underlying this article cannot be shared publicly due to patient privacy and confidentiality concerns, in

compliance with Health Insurance Portability and Accountability Act (HIPAA) regulations and institutional guidelines. As the dataset includes medical clinical images containing sensitive patient information, it cannot be made publicly available. Data access may be granted upon reasonable request to the corresponding author, subject to review and approval by the relevant ethical and privacy committees.

### References

1. Murphy CA, Garg AK, Silva-Correia J, Reis RL, Oliveira JM, Collins MN. The meniscus in normal and osteoarthritic tissues: facing the structure property challenges and current treatment trends. *Annu Rev Biomed Eng*. 2019;21:495–521.
2. Trunz LM, Morrison WB. MRI of the knee meniscus. *Magn Reson Imaging Clin N Am*. 2022;30(2):307–324.
3. Hammernik K, Klatzer T, Kobler E, et al. Learning a variational network for reconstruction of accelerated MRI data. *Magn Reson Med*. 2018;79(6):3055–3071.
4. Chandra SS, Bran Lorenzana M, Liu X, Liu S, Bollmann S, Crozier S. Deep learning in magnetic resonance image reconstruction. *J Med Imaging Radiat Oncol*. 2021;65(5):564–577.
5. Wu Z, Liao W, Yan C, et al. Deep learning based MRI reconstruction with transformer. *Comput Methods Programs Biomed*. 2023; 233:107452.
6. Cao C, Cui Z-X, Wang Y, et al. High-frequency space diffusion model for accelerated MRI. *IEEE Trans Med Imaging*. 2024;43 (5):1853–1865.
7. Yang A, Finkelstein M, Koo C, Doshi AH. Impact of deep learning image reconstruction methods on MRI throughput. *Radiol Artif Intell*. 2024;6(3):e230181.
8. Heckel R, Jacob M, Chaudhari A, Perlman O, Shimron E. Deep learning for accelerated and robust MRI reconstruction. *MAGMA*. 2024;37(3):335–368.
9. Astuto B, Flament I, Namiri NK, et al. Automatic deep learning-assisted detection and grading of abnormalities in knee MRI studies. *Radiol Artif Intell*. 2021;3(3):e219001.
10. Roblot V, Giret Y, Bou Antoun M, et al. Artificial intelligence to diagnose meniscus tears on MRI. *Diagn Interv Imaging*. 2019;100 (4):243–249.
11. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(6):1137–1149.
12. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; Jun 26–Jul 1, 2016, Las Vegas, NV, USA. IEEE; 2016:779–788.
13. Guerazi A, Omoumi P, Tordjman M, et al. How AI may transform musculoskeletal imaging. *Radiology*. 2024;310(1):e230764.
14. Caliva F, Ardila D, Rakshit S, et al. Breaking speed limits with simultaneous ultra-fast MRI reconstruction and tissue segmentation. In: *Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL)*, Jul 6–8, 2020, Montréal, QC, Canada. Proceedings of Machine Learning Research (PMLR); 2020:113–126.
15. Tolpadi AA, Bharadwaj U, Gao KT, et al. K2S challenge: from undersampled K-space to automatic segmentation. *Bioengineering*. 2023;10(2):267.
16. Fienup JR. Invariant error metrics for image reconstruction. *Appl Opt*. 1997;36(32):8352–8357.
17. Horé A, Ziou D. Is there a relationship between peak-signal-to-noise ratio and structural similarity index measure? *IET Image Process*. 2013;7(1):12–24.
18. Dosselmann R, Yang XD. A comprehensive assessment of the structural similarity index. *Signal Image Video Process*. 2011;5 (1):81–91.
19. Knoll F, Murrell T, Sriram A, et al. Advancing machine learning for MR image reconstruction with an open competition: overview

- of the 2019 fastMRI challenge. *Magn Reson Med*. 2020; 84(6):3054-3070.
20. Brau ACS, Beatty PJ, Skare S, Bammer R. Comparison of reconstruction accuracy and efficiency among autocalibrating data-driven parallel imaging methods. *Magn Reson Med*. 2008; 59(2):382-395.
  21. Zoph B, Cubuk ED, Ghiasi G, Lin TY, Shlens J, Le QV. Learning data augmentation strategies for object detection. In: A Vedaldi, H Bischof, T Brox, JM Frahm, eds. *Computer Vision – ECCV 2020. Part VII. Lecture Notes in Computer Science*, vol 12352. Cham: Springer; 2020:566-583.
  22. Tolpadi A, Calivà F, Han M, et al. A cartilage-specific loss function improves image reconstruction performance in multiple tissues of clinical interest. In: *Proceedings of the 31st Joint Annual Meeting ISMRM-ESMRMB ISMRT*, May 7–12, 2022, London, United Kingdom.
  23. Eo T, Jun Y, Kim T, Jang J, Lee HJ, Hwang D. KIKI-net: cross-domain convolutional neural networks for reconstructing under-sampled magnetic resonance images. *Magn Reson Med*. 2018;80(5):2188-2201.
  24. Crewson PE. Reader agreement studies. *Am J Roentgenol*. 2005; 184(5):1391-1397.
  25. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychol Bull*. 1969;72(5):323-327.
  26. Zhao R, Zhang Y, Yaman B, Lungren MP, Hansen MS. End-to-end AI-based MRI reconstruction and lesion detection pipeline for evaluation of deep learning image reconstruction. arXiv 11524. <https://doi.org/10.48550/arXiv.2109.11524>, September 23, 2021, preprint: not peer reviewed.
  27. Kim SH, Lee HJ, Jang YH, Chun KJ, Park YB. Diagnostic accuracy of magnetic resonance imaging in the detection of type and location of meniscus tears: comparison with arthroscopic findings. *J Clin Med*. 2021;10(4):606.
  28. Philbrick KA, Weston AD, Akkus Z, et al. RIL-Contour: a medical imaging dataset annotation tool for and with deep learning. *J Digit Imaging*. 2019;32(4):571-581.
  29. Peng S, Liu Y, Lv W, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *Lancet Digit Health*. 2021; 3(4):e250-e259.
  30. Pennig L, Hoyer UCI, Krauskopf A, et al. Deep learning assistance increases the detection sensitivity of radiologists for secondary intracranial aneurysms in subarachnoid hemorrhage. *Neuroradiology*. 2021;63(12):1985-1994.
  31. Garwood ER, Recht MP, White LM. Advanced imaging techniques in the knee: benefits and limitations of new rapid acquisition strategies for routine knee MRI. *Am J Roentgenol*. 2017; 209(3):552-560.
  32. Feng C-M, Yan Y, Chen G, et al. Multimodal transformer for accelerated MR imaging. *IEEE Trans Med Imaging*. 2023; 42(10):2804-2816.
  33. Johnson PM, Jeong G, Hammernik K, et al. Evaluation of the robustness of learned MR image reconstruction to systematic deviations between training and test data for the models from the fastMRI challenge. In: de Bruijne M, Cattin P, Cotin S, et al., eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Part IV. Lecture Notes in Computer Science*, vol 12264. Cham: Springer; 2021:280-290.