AI-Assisted MRI Interpretation Improves Surgeon Performance in Diagnosing Bankart Lesions

Sahil Sethi[1], Sai Kashyap Reddy, Mansi Sakarvadia, Jordan Serotte, Darlington Nwaudo, Sherwin S W Ho, Nicholas Henry Maassen, Lewis L Shi
[1]Pritzker School of Medicine

INTRODUCTION:

Bankart lesions can cause significant pain and shoulder dysfunction. These lesions can be difficult to detect on standard non-contrast MRI, often leading to the use of MRI arthrography (MRA) to improve diagnostic sensitivity. However, MRAs are invasive, more expensive, and can be painful for patients. Deep learning (DL) is a subset of machine learning that recognizes patterns and makes decisions by automatically extracting the relevant measurements or features in large datasets. DL has improved diagnostic accuracy in several medical domains and is well suited for addressing diagnostic challenges in imaging. This study evaluates a deep learning (DL) model trained to detect Bankart lesions on both MRI and MRA and assesses its interpretability and clinical utility through a multi-reader study involving orthopaedic surgeons.

METHODS: The dataset included 586 shoulder MRIs (335 standard, 251 MRA) from 546 patients at a single academic institution who underwent arthroscopy within one year after imaging. Exclusion criteria included prior ipsilateral surgery and poor image quality. Operative reports were reviewed to determine the presence of Bankart lesions, which served as the reference standard. We evaluated several DL architectures, including convolutional neural networks and transformer-based models, using axial, sagittal, and coronal sequences as input. The final model was trained on 410 MRIs (224 standard, 186 MRA), tuned on 59 MRIs (40 standard, 19 MRA), and tested on 117 MRIs (71 standard, 46 MRA). To assess model interpretability, we generated gradient-weighted class activation maps (Grad-CAM) on correctly classified test cases (Figure 1). To evaluate the clinical impact of model assistance, a follow-up reader study was conducted with four orthopaedic clinicians: two shoulder/elbow fellowship-trained surgeons and two orthopaedic surgery residents (one PGY-3 and one PGY-4). Each reader reviewed all 117 test MRIs/MRAs in two phases: once without model predictions (unaided), and again after a 60-day washout period with DL-generated predictions shown (aided). Readers recorded binary Bankart diagnoses and confidence scores (scale 1–10) for each case. Differences in confidence between phases were analyzed using paired t-tests.

RESULTS:

Out of 586 MRIs, 109 (18.5%) had arthroscopically confirmed Bankart lesions. On the held-out test set, the model achieved 90.14% accuracy, 83.33% sensitivity, 90.77% specificity, and an area under the receiver operating characteristic curve (AUROC) of 0.9051 for detecting Bankart lesions on standard MRI (see Table 1). In comparison, radiology reports on the same cases showed substantially lower sensitivity (16.7%), similar specificity (86.2%), and lower overall accuracy (80.3%). The model's sensitivity on standard MRI also exceeded literature-reported ranges (52–55%). On MRA, the model achieved 89.13% accuracy, 94.12% sensitivity, 86.21% specificity, and an AUROC of 0.9256. Radiology reports on these MRAs demonstrated lower sensitivity (82.4%) with the same specificity (86.2%), and lower accuracy overall (84.8%). The model's sensitivity on MRA also matched published radiologist ranges (74–96%) and maintained comparable specificity. Grad-CAM visualizations indicated that the model consistently focused on the anterior labrum across both modalities, aligning with expert annotations (see Figure 1). In the reader study, mean sensitivity improved from 39.1% to 76.1% when clinicians were provided with model predictions, with all four readers demonstrating improved sensitivity across both standard MRIs and MRAs in the aided phase. Specificity decreased slightly from 91.0% to 86.7%, and diagnostic accuracy increased from 80.8% to 84.6%. Average confidence increased by 1.78 points on a 10-point scale ($p < 0.001$).

DISCUSSION AND CONCLUSION: Across both modalities, the model exceeded radiologist sensitivity while maintaining similar specificity. Further, it achieved comparable performance on standard MRI to radiologist performance on MRA. This suggests that deep learning may help close the diagnostic gap between non-contrast and contrast-enhanced imaging, avoiding the added cost, discomfort, and invasiveness of MRA. Grad-CAM visualizations confirmed that predictions were based on appropriate anatomic regions. The reader study demonstrates that surgeons and residents benefited from structured diagnostic support—achieving greater sensitivity and confidence when aided by the model. These results highlight the potential for DL tools to supplement clinical interpretation. Future work will focus on integrating such models into real-time workflows.

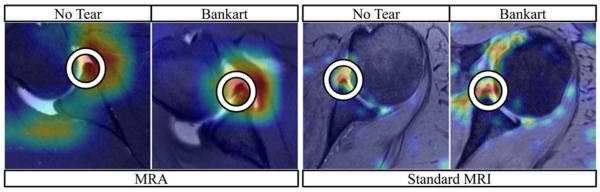

Figure 1. Gradient-weighted class activation mapping (Grad-CAM) visualizations for Bankart lesion detection on MRAs (left) and standard MRIs (right) for the axial view. Cases with and without Bankart lesions are presented. The model correctly classified all four cases. White circles highlight the anterior labrum (the region of interest), annotated by a shoulder/elbow fellowship-trained orthopaedic surgeon. Heatmaps indicate regions most influential to the model's prediction, with warmer colors signifying higher relevance (i.e., the model "looks" at the areas in red/yellow when making predictions).

Table 1. Model performance compared to original radiology reports and literature benchmarks.

	Accuracy	**Sensitivity**	**Specificity**	**AUROC**
Standard MRIs (n=71)				
Model	**90.14% (64/71)**	**83.33% (5/6)**	**90.77% (59/65)**	**0.9051**
Radiology Reports	80.28% (57/71)	16.67% (1/6)	86.15% (56/65)	-
Literature Radiologists	-	52-55%	89-100%	-
MRAs (n=46)				
Model	**89.13% (41/46)**	**94.12% (16/17)**	**86.21% (25/29)**	**0.9256**
Radiology Reports	84.78% (39/46)	82.35% (14/17)	86.21% (25/29)	-
Literature Radiologists	-	74-96%	91-98%	-