# Deep Learning Diagnosis and Classification of Rotator Cuff Tears on Shoulder MRI

*Dana J. Lin, MD,* Michael Schwier, PhD,† Bernhard Geiger, PhD,† Esther Raithel, PhD,‡ Heinrich von Busch, PhD,‡ Jan Fritz, MD,* Mitchell Kline, MD,* Michael Brooks, MD,* Kevin Dunham, MD,* Mehool Shukla, MD,* Erin F. Alaia, MD,* Mohammad Samim, MD,* Vivek Joshi, MD,* William R. Walter, MD,* Jutta M. Ellermann, MD,§ Hakan Ilaslan, MD,|| David Rubin, MD,*¶ Carl S. Winalski, MD,|| and Michael P. Recht, MD**



**Background:** Detection of rotator cuff tears, a common cause of shoulder disability, can be time-consuming and subject to reader variability. Deep learning (DL) has the potential to increase radiologist accuracy and consistency.

**Purpose:** The aim of this study was to develop a prototype DL model for detection and classification of rotator cuff tears on shoulder magnetic resonance imaging into no tear, partial-thickness tear, or full-thickness tear.

**Materials and Methods:** This Health Insurance Portability and Accountability Act-compliant, institutional review board–approved study included a total of 11,925 noncontrast shoulder magnetic resonance imaging scans from 2 institutions, with 11,405 for development and 520 dedicated for final testing. A DL ensemble algorithm was developed that used 4 series as input from each examination: fluid-sensitive sequences in 3 planes and a sagittal oblique T1-weighted sequence. Radiology reports served as ground truth for training with categories of no tear, partial tear, or full-thickness tear. A multireader study was conducted for the test set ground truth, which was determined by the majority vote of 3 readers per case. The ensemble comprised 4 parallel 3D ResNet50 convolutional neural network architectures trained via transfer learning and then adapted to the targeted domain. The final tear-type prediction was determined as the class with the highest probability, after averaging the class probabilities of the 4 individual models.

**Results:** The AUC overall for supraspinatus, infraspinatus, and subscapularis tendon tears was 0.93, 0.89, and 0.90, respectively. The model performed best for full-thickness supraspinatus, infraspinatus, and subscapularis tears with AUCs of 0.98, 0.99, and 0.95, respectively. Multisequence input demonstrated higher AUCs than single-sequence input for infraspinatus and subscapularis tendon tears, whereas coronal oblique fluid-sensitive and multisequence input showed similar AUCs for supraspinatus tendon tears. Model accuracy for tear types and overall accuracy were similar to that of the clinical readers.

**Conclusions:** Deep learning diagnosis of rotator cuff tears is feasible with excellent diagnostic performance, particularly for full-thickness tears, with model accuracy similar to subspecialty-trained musculoskeletal radiologists.

**Key Words:** deep learning, artificial intelligence, convolutional neural network, classification, rotator cuff, tendons, supraspinatus, infraspinatus, subscapularis, MRI

(*Invest Radiol* 2023;58: 405–412)

Rotator cuff tears are a common cause of shoulder disability, affecting nearly 40% of individuals older than 60 years and over half of individuals older than 80 years.[1] Magnetic resonance imaging (MRI) is widely used for rotator cuff tear detection, characterization, and to guide surgical versus nonsurgical management. Radiologist interpretation of rotator cuff tears, which can be time-consuming, consists of tear classification for each tendon, quantitative measurements, and qualitative descriptions of muscle and tendon status.[2] Rotator cuff interpretation is subject to reader variability,[3,4] particularly for partial-thickness tears and tears involving the subscapularis tendon.[5–8]

Computer-aided interpretation based on deep learning (DL) has the potential to increase the accuracy, consistency, and productivity of radiologists.[9,10] Applications of convolutional neural networks (CNNs) to the field of medical image analysis[11,12] and more recently to the field of musculoskeletal imaging[10,13–16] have shown that DL can assist the radiologist in the detection of fractures,[17,18] anterior cruciate ligament tears,[19,20] meniscal tears,[9] cartilage lesions,[21] and osteoarthritis.[22,23] To date, there are few reported investigations[24–26] that have applied DL to rotator cuff assessment on shoulder MRI, creating a significant opportunity to expand on this preliminary work.

The purpose of this study was to develop and test a DL model for rotator cuff tear diagnosis and classification from shoulder MRI.

## METHODS

### Data Set

This was a Health Insurance Portability and Accountability Act–compliant, institutional review board–approved study using data from 2 institutions. One institution searched the Radiology Information System for shoulder MRIs from April 2016 to 2018 and the electronic medical record for rotator cuff repair procedures performed before September 2018 with preoperative shoulder MRI obtained within 6 months prior.

The other institution searched the Radiology Information System from 2010 to 2018 for nonsurgical cases and the institution's customized orthopedic outcome database from 2015 to 2019 for patients with rotator cuff repair and preoperative shoulder MRI obtained within 6 months prior.

The searches yielded 13,118 shoulder MRIs (Fig. 1). Exclusion criteria were corrupt data, missing radiology reports, use of intra-articular contrast, and prior rotator cuff repair. This yielded 11,925 shoulder MRIs using Siemens, Philips, GE, Toshiba, and Hitachi systems (Table 1). Field strengths ranged from 0.3 T to 3 T (Table 1). Most examinations had 5 series, including coronal oblique, sagittal oblique, and axial fluid-sensitive (FS) (T2-weighted fat-suppressed, proton density [PD]–weighted fat-suppressed, or STIR) sequences as well as coronal oblique PD and sagittal oblique T1-weighted sequences. Example MR parameters as

Received for publication October 17, 2022; and accepted for publication, after revision, December 6, 2022.

From the *Department of Radiology, NYU Grossman School of Medicine, New York, NY; †Siemens Medical Solutions USA, Princeton, NJ; ‡Siemens Healthcare GmbH, Erlangen, Germany; §Department of Radiology, Center for Magnetic Resonance Research, University of Minnesota, Minneapolis, MN; ||Imaging Institute, Cleveland Clinic, Cleveland, OH; and ¶RadSource LLC, Brentwood, TN.

Co-first authorship: D.J.L. and M.S. contributed equally to this work.

Conflicts of interest and sources of funding: Industry collaboration with Siemens (authors' expertise as above), provision of funding for multireader study ground truth research reads.

Correspondence to: Dana J. Lin, MD, Center for Biomedical Imaging, Department of Radiology, NYU Grossman School of Medicine, 660 First Ave, 3rd Floor, New York, NY 10016. E-mail: dana.lin@nyulangone.org.

Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.investigativeradiology.com).



ISSN: 0020-9996/23/5806–0405
DOI: 10.1097/RLI.0000000000000951

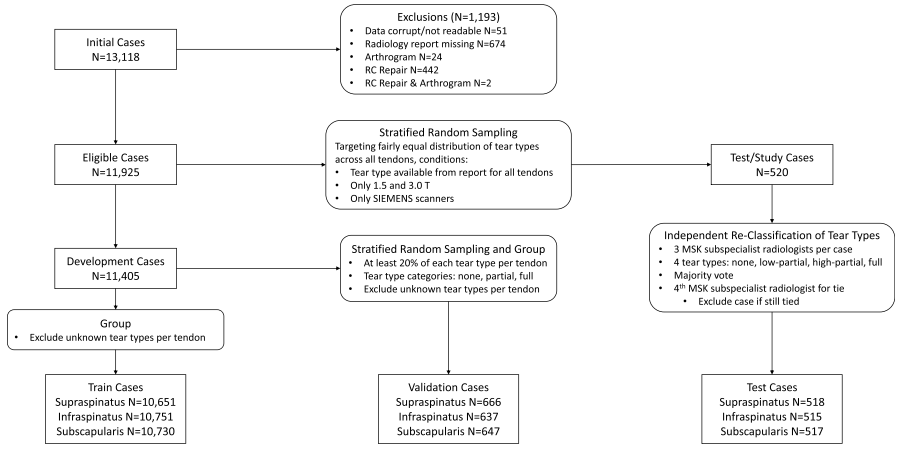


**FIGURE 1.** Study flowchart.

follows: axial PD-weighted fat-suppressed (TR = 2800+ milliseconds, TE = 30–40 milliseconds, FOV = 140 mm, matrix size = 320 × 272, slice thickness = 3 mm); coronal oblique T2-weighted fat-suppressed (TR = 3000+ milliseconds, FOV = 140 mm, matrix size = 320 × 256, slice thickness = 3 mm); sagittal oblique T1-weighted (TR = <800 milliseconds, TE = 10–15 milliseconds, FOV = 140 mm, matrix size = 320 × 240, slice thickness = 3 mm); coronal oblique PD-weighted (TR = 2300 + milliseconds, TE = 25–35 milliseconds, FOV = 140 mm, matrix size = 384 × 307); sagittal oblique T2-weighted fat-suppressed (TR = 3000+ milliseconds, TR = 50–60 milliseconds, FOV = 140 mm, matrix size = 320 × 240, slice thickness = 3 mm). All data were deidentified.

The data were split into training, validation, and test sets (Table 1, Fig. 1). The validation and test sets were selected via stratified random sampling to ensure that each tear type was represented in each set. The validation set was used for intermediate evaluation and parameter tuning during model development. The test set was reserved for the final model evaluation. Tear classifications for training and validation cases were derived from the radiology reports. At each institution, all initial interpretations were provided by subspecialty-trained musculoskeletal radiologists (1–34 years' experience), who had access to the clinical history and patient demographics for each case. Categories for tear classifications were as follows: no tear, partial tear, or full-thickness tear. If the tear type could not be determined, the case was excluded for that tendon; hence, data set sizes slightly differ between tendons. If the case was worded as high-grade partial tear with pinhole perforation or possible full-thickness, it was classified according to reader certainty, that is, high-grade partial tear.

## Establishing Test Set Ground Truth

Because of known reader variability in the interpretation of rotator cuff tears,[3,4,8] a multireader study was conducted to establish a ground truth classification for the test set. Before starting the study, 2 training sessions were held with prespecified "reading rules," training manual, and training cases provided to each reader.

Thirteen musculoskeletal radiologists from 4 academic institutions participated over a 16-week reading period. Each case in the test set was independently evaluated by 3 readers on a cloud-based informatics research platform (Flywheel, Minneapolis, MN) using PACS-type workstations in standard diagnostic reading conditions. The readers were blinded to each other's interpretations as well as to the clinical radiology reports.

The final ground truth for the test set was defined as the majority vote tear classification among the 3 readers. In cases where all 3 readers chose different tear classifications, a randomly assigned fourth reader adjudicated to break the tie.

## Training Data Augmentation

Data augmentation[27] is commonly used when training CNNs to artificially enrich the training data. The following data augmentation was applied to each image during training:

1. Random nonlinear histogram scaling to vary intensity and contrast.
2. Random noise was added to the image such that $I' = I + aN$, where $I$ represents the image intensities, $N$ is uniform random noise corresponding to the size of the image and generated in the range $[I_{min}, I_{max}]$, and $a$ an attenuation factor, which was chosen randomly in the range [0, 0.15].
3. The image was flipped along each of the $x$, $y$, $z$ axes with 50% probability.
4. Random rotation of the image between −45 and +45 degrees.

## Preprocessing

All image series were preprocessed by $z$-score normalization of the intensity values. Images were then resampled to a fixed image size of 256, 256, 32 in $(x, y, z)$ dimensions using trilinear interpolation. In the training phase, preprocessing was applied after augmentation.

## Model and Training

For each tendon, a dedicated ensemble was trained, where each ensemble comprised a set of parallel 3D ResNet50[28] CNN architectures. Each of these parallel models was trained independently on one sequence among the 3-plane FS, sagittal oblique T1-weighted, and coronal oblique PD sequences (Fig. 2). The ensemble setup allows handling of missing input sequences by ignoring the corresponding paths. Probabilities from the individual sequence models for each tear type were averaged into the final tear probabilities:

$$P_{avg}(t) = \sum_{s \in S} \frac{P_s(t)}{|S|}$$

for each tear-type $t \in T$, with $T := \{FullTear,\ PartialTear,\ NoTear\}$, $S \subseteq \{axialFS,\ coronalFS,\ sagittalFS,\ coronalPD,\ sagittalT1\}$ representing the available sequences, and hence $P_s(t)$ representing the predicted probability for $t$ from the model that was individually trained on sequence $s$. The final tear-type prediction was determined as the tear-type with the highest probability in the final probabilities:

$$t_{final} = \underset{t \in T}{\arg\max}\, P_{avg}(t).$$

Transfer learning[29–31] was used: models were initialized with weights pretrained on the Kinetics data set[32,33] and then further trained

**TABLE 1.** Summary of Demographic Characteristics

| Characteristic | Supraspinatus | | | | | | Infraspinatus | | | | | | Subscapularis | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Training | | Validation | | | Test | Training | | Validation | | | Test | Training | | Validation | | | Test |
| No. cases | 10,651 | | 666 | | | 518 | 10,751 | | 637 | | | 515 | 10,730 | | 647 | | | 517 |
| | Mean | SD | Mean | SD | | Mean | SD | | Mean | SD | | Mean | SD | | Mean | SD | | Mean | SD |
| Age | 52.97 | 15.61 | 55.38 | 15.4 | | 59.42 | 14.43 | | 52.84 | 15.62 | | 57.72 | 14.64 | | 59.42 | 14.44 | | 52.85 | 15.61 | | 56.91 | 14.95 | | 59.36 | 14.43 |
| | n | | n | | | n | | | n | | | n | | | n | | | n | | | n | | | n | |
| Sex | | | | | | | | | | | | | | | | | | | | | | | | | |
| Female | 4693 | 44% | 317 | 48% | | 227 | 44% | | 4768 | 44% | | 281 | 44% | | 228 | 44% | | 4754 | 44% | | 276 | 43% | | 228 | 44% |
| Male | 5958 | 56% | 349 | 52% | | 291 | 56% | | 5983 | 56% | | 356 | 56% | | 287 | 56% | | 5976 | 56% | | 371 | 57% | | 289 | 56% |
| Field strength (T) | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0.3 | 1 | 0% | | | | | | | 1 | 0% | | | | | | | | 1 | 0% | | | | | | |
| 0.7 | 26 | 0% | 2 | 0% | | | | | 25 | 0% | | 2 | 0% | | | | | 24 | 0% | | 1 | 0% | | | |
| 1 | 261 | 2% | | | | | | | 260 | 2% | | | | | | | | 261 | 2% | | | | | | |
| 1.2 | 2 | 0% | | | | | | | 2 | 0% | | | | | | | | 2 | 0% | | | | | | |
| 1.5 | 5350 | 50% | 328 | 49% | | 186 | 36% | | 5368 | 50% | | 316 | 50% | | 184 | 36% | | 5363 | 50% | | 321 | 50% | | 187 | 36% |
| 3 | 5011 | 47% | 336 | 50% | | 332 | 64% | | 5095 | 47% | | 319 | 50% | | 331 | 64% | | 5079 | 47% | | 325 | 50% | | 330 | 64% |
| Manufacturer | | | | | | | | | | | | | | | | | | | | | | | | | |
| GE | 236 | 2% | 14 | 2% | | | | | 225 | 2% | | 17 | 3% | | | | | 228 | 2% | | 14 | 2% | | | |
| Hitachi | 77 | 1% | 4 | 1% | | | | | 72 | 1% | | 2 | 0% | | | | | 71 | 1% | | 1 | 0% | | | |
| Philips | 604 | 6% | 44 | 7% | | | | | 619 | 6% | | 45 | 7% | | | | | 615 | 6% | | 49 | 8% | | | |
| Siemens | 9729 | 91% | 601 | 90% | | 518 | 100% | | 9828 | 91% | | 571 | 90% | | 515 | 100% | | 9809 | 91% | | 581 | 90% | | 517 | 100% |
| Toshiba | 5 | 0% | 3 | 0% | | | | | 7 | 0% | | 2 | 0% | | | | | 7 | 0% | | 2 | 0% | | | |
| Tear type | | | | | | | | | | | | | | | | | | | | | | | | | |
| No tear | 4173 | 39% | 215 | 32% | | 133 | 26% | | 7398 | 69% | | 281 | 44% | | 187 | 36% | | 8120 | 76% | | 302 | 47% | | 191 | 37% |
| Partial | 3378 | 32% | 212 | 32% | | 231 | 45% | | 1881 | 17% | | 187 | 29% | | 219 | 43% | | 2064 | 19% | | 200 | 31% | | 254 | 49% |
| Full-thickness | 3100 | 29% | 239 | 36% | | 154 | 30% | | 1472 | 14% | | 169 | 27% | | 109 | 21% | | 546 | 5% | | 145 | 22% | | 72 | 14% |

SD, standard deviation.

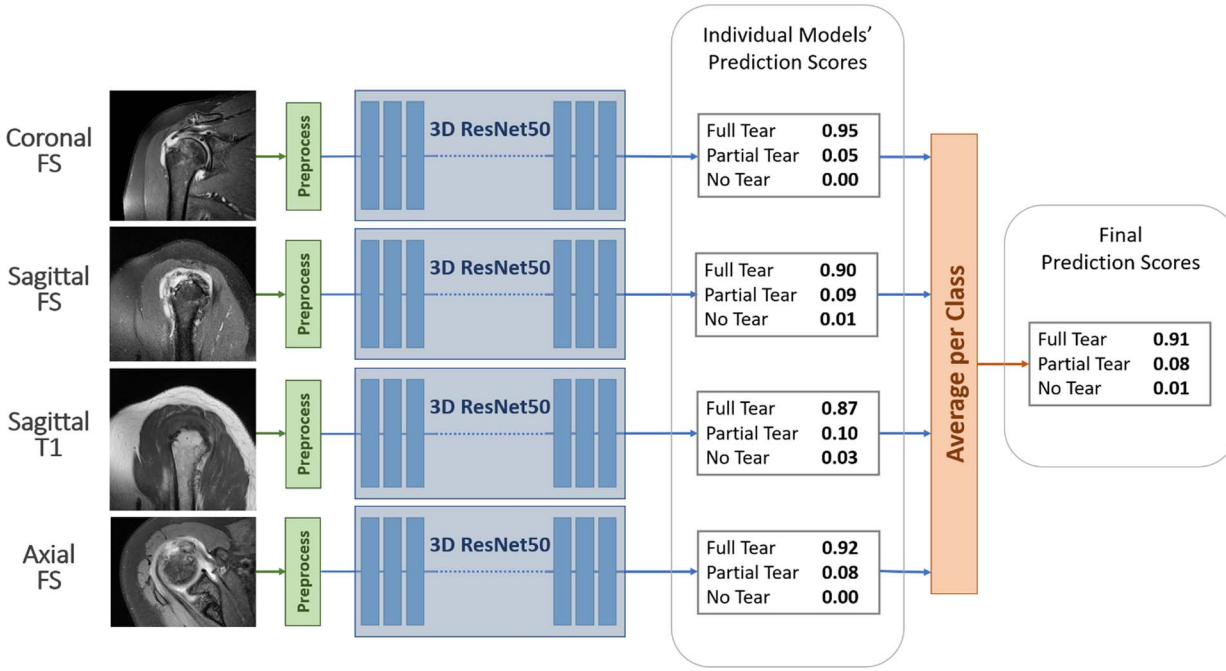


**FIGURE 2.** Four-view input architecture using an ensemble of 3D ResNet50 convolutional neural networks.

on the training cases to adapt the models to the targeted domain. Balanced data sampling ensured that each tear class was seen with equal frequency during each training epoch. Training was set up with Adam optimization, cross-entropy loss, and a learning rate of $10^{-5}$.

Training was stopped when no improvement in overall accuracy was observed on the validation data for 100 epochs. The weights from the epoch with the highest overall accuracy on the validation set were used in the final model.

## Results Analysis

Statistical analysis was performed using Python 3.7.7 (Python Software Foundation, https://www.python.org/) with pandas 1.0.5[34] and statsmodels 0.12.1.[35] For the purposes of this study, the "clinical reader" was the radiologist who rendered the original clinical radiology report and "study readers" were radiologists from the multireader study for the ground truth test set. Eight of the study readers were likely also clinical readers as faculty at the institution from which the data set originated. Accuracy of the model and clinical readers was determined by using the multireader study majority vote on the test set as the gold (reference) standard. The overall accuracy and tear-type accuracies of the 4-view model and of the clinical reader were compared for each tendon.

Overall accuracy is defined as

$$overall\ accuracy = \frac{TrueFull + TruePartial + TrueNone}{Full + Partial + None}$$

with *TrueFull*, *TruePartial*, and *TrueNone* representing the number of correct predictions for full, partial, and no tears, respectively, and *Full*, *Partial*, and *None* representing the actual number of full, partial, and no tears in the test data, respectively.

An informal survey of 10 sites in the United States and Europe showed that the coronal oblique PD sequence was less commonly obtained across sites, compared with the 3-plane FS plus sagittal T1 sequences. In the following, we will refer to the latter sequence combination as "4-view." A receiver operating characteristic (ROC) analysis was performed to assess the performance of the 4-view ensemble versus using each sequence alone.

The difference in overall accuracy between the 4-view DL ensemble model and the clinical reader was computed with McNemar test. A $P$ value of $<0.05$ was considered statistically significant. Linear-weighted Cohen $\kappa$ statistics were calculated for the multireader study. For testing statistical significance for classification accuracy differences between the 4-view DL

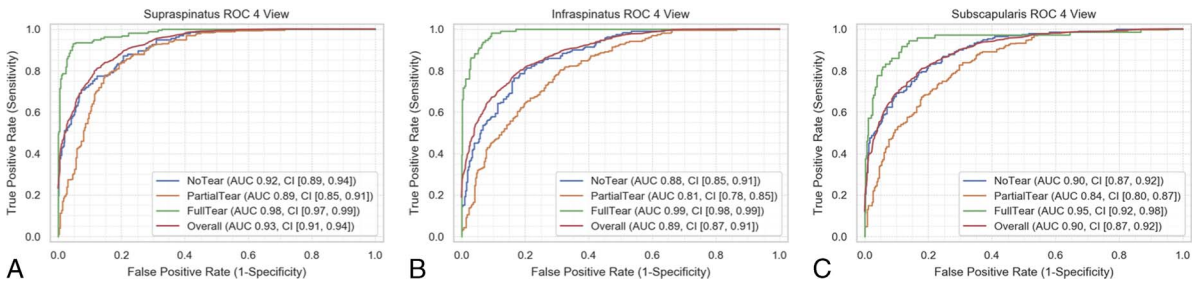


**FIGURE 3.** A–C, Receiver operating characteristic analysis of model performance on classifying supraspinatus, infraspinatus, and subscapularis tendon tear types using 4-view input. Overall ROC is computed as the macro average of the tear type ROCs.

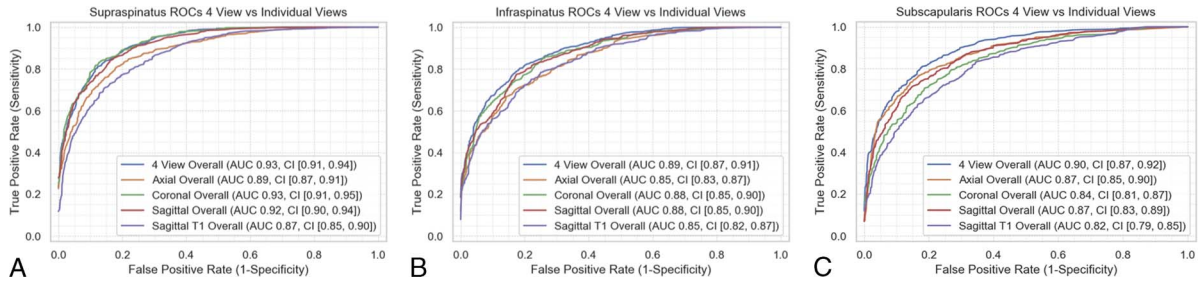


**FIGURE 4.** A–C, Receiver operating characteristic analysis of 4-view versus single sequence input by type of sequence for supraspinatus, infraspinatus, and subscapularis tendon tear classification.



ensemble model and single sequence models, a Holm-Bonferroni correction was applied to adjust the base significance level of $\alpha = 0.05$.

## RESULTS

### Overall Versus Tear Types

Using 4-view input, the AUC overall for supraspinatus, infraspinatus, and subscapularis tendon tears was 0.93 (95% confidence interval [CI], 0.91–0.94), 0.89 (95% CI, 0.87–0.91), and 0.90 (95% CI, 0.87–0.92), respectively. Among different tear types, the model demonstrated for all tendons higher AUCs for full-thickness tears, followed by no tears and partial-thickness tears (Fig. 3).

### Multisequence Versus Single Sequence

The AUC for 4-view input was higher than that for single sequence input for infraspinatus and subscapularis tears (Fig. 4). For supraspinatus, the AUCs for coronal oblique FS alone and 4-view input were similar: 0.93 (95% CI, 0.91–0.95) and 0.93 (95% CI, 0.91–0.94), respectively.

### Field Strength

Model performance was similar for 3 T and 1.5 T for all tendons (Fig. 5). For example, the AUCs for supraspinatus tears at 1.5 T versus 3 T were 0.93 (95% CI, 0.91–0.95) and 0.92 (95% CI, 0.89–0.95), respectively.

### Accuracy

The comparison between clinical readers and the model shows no statistically significant difference in overall accuracies for all tendons, with similar accuracies between the readers and the model across tear types (Table 2, Table 3). For example, the clinical reader overall accuracy for supraspinatus was 0.79 (95% CI, 0.75–0.82), with model overall accuracy of 0.81 (95% CI, 0.78–0.85).

For supraspinatus, there were 28 false-negatives, where in every case, the model predicted no tear and the ground truth was partial tear. An example case is shown in Figure 6. There were no cases where the model predicted no tear and the ground truth was full-thickness tear. The same was true for infraspinatus, where there were 30 false-negatives.

There were 39 false-positive supraspinatus cases. In all except 2, the model predicted partial tear when the ground truth was no tear. In 2 cases, the model predicted a full-thickness tear when the ground truth was no tear, and an example case of this is shown in Figure 6. In the other case, there was a full-thickness tear in the anterior infraspinatus that may have involved the posterior supraspinatus fibers at the supraspinatus-infraspinatus junction. There were 81 false-positive infraspinatus cases, all except 3 of which were instances where the model predicted a partial tear when the ground truth was no tear.

The negative predictive value of the 4-view model for rotator cuff tear detection is shown in Supplementary Material, Table S4 (http://links.lww.com/RLI/A791). As described previously in Methods, the study population was augmented with operative cases and therefore does not reflect the natural distribution or prevalence of tear types.

Table 2 illustrates model accuracy based on input sequences both overall and by tear type. To highlight just one such comparison, almost any combination of single, double, or more input sequences for the algorithm could detect a full-thickness supraspinatus tendon tear equally well, with accuracies ranging from 0.86 to 0.93. Table S1 in the Supplementary Material (http://links.lww.com/RLI/A791) shows the statistically significant differences in classification accuracy when comparing the 4-view model with different single sequence models by tendon.

There was no significant difference in model overall classification accuracy between 3T and 1.5 T (Supplementary Material, Table S2, http://links.lww.com/RLI/A791). Overall accuracy and accuracy by tear type and tendon for the clinical readers at 1.5 T versus 3T is shown in Table S3 (http://links.lww.com/RLI/A791).

### Processing Time

For classification of tendon status in all 3 tendons with an input of 4 imaging sequences, we measured a maximum memory consumption of 0.5 GB, and an average processing time of:

- 18 seconds per case in a CPU only test environment (Intel Xeon Gold 6128 CPU at 3.40 GHz).
- Under 1 second per case in a GPU test environment (NVIDIA Tesla V100 SXM2).

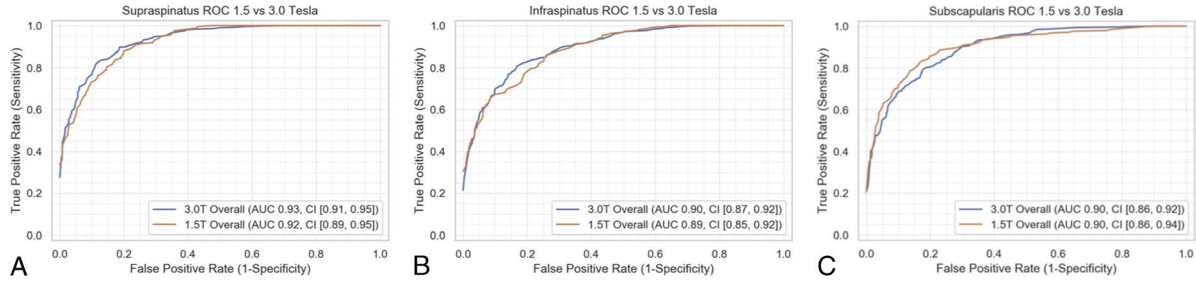


**FIGURE 5.** A-C, Receiver operating characteristic analysis by field strength of the 4-view ensemble.

**TABLE 2.** Tear Type and Overall Accuracy of Clinical Reader and Algorithms With Different Sequence Inputs*

| | Full Thickness | | Partial | | None | | Overall Accuracy | |
|---|---|---|---|---|---|---|---|---|
| Supraspinatus | | | | | | | | |
| Clinical reader | 0.88 | [0.82, 0.92] | 0.81 | [0.75, 0.85] | 0.65 | [0.57, 0.73] | 0.79 | [0.75, 0.82] |
| Algorithm [‘Ax,’ ‘Cor,’ ‘Sag,’ ‘SagT1’] | 0.93 | [0.88, 0.96] | 0.8 | [0.74, 0.85] | 0.71 | [0.62, 0.78] | 0.81 | [0.78, 0.85] |
| Algorithm [‘Ax,’ ‘Cor,’ ‘Sag,’ ‘SagT1,’ ‘CorPD’] | 0.92 | [0.86, 0.95] | 0.77 | [0.71, 0.82] | 0.69 | [0.61, 0.76] | 0.79 | [0.76, 0.83] |
| Algorithm [‘Ax’] | 0.92 | [0.86, 0.95] | 0.66 | [0.59, 0.72] | 0.65 | [0.56, 0.72] | 0.73 | [0.69, 0.77] |
| Algorithm [‘Cor’] | 0.9 | [0.84, 0.94] | 0.74 | [0.68, 0.79] | 0.76 | [0.68, 0.82] | 0.79 | [0.75, 0.82] |
| Algorithm [‘CorPD’] | 0.91 | [0.85, 0.95] | 0.71 | [0.65, 0.76] | 0.59 | [0.50, 0.67] | 0.74 | [0.70, 0.77] |
| Algorithm [‘Sag’] | 0.93 | [0.88, 0.96] | 0.75 | [0.69, 0.80] | 0.68 | [0.60, 0.76] | 0.79 | [0.75, 0.82] |
| Algorithm [‘SagT1’] | 0.86 | [0.80, 0.91] | 0.71 | [0.65, 0.77] | 0.6 | [0.52, 0.68] | 0.73 | [0.69, 0.77] |
| Infraspinatus | | | | | | | | |
| Clinical reader | 0.88 | [0.81, 0.93] | 0.79 | [0.73, 0.84] | 0.59 | [0.52, 0.66] | 0.74 | [0.70, 0.77] |
| Algorithm [‘Ax,’ ‘Cor,’ ‘Sag,’ ‘SagT1’] | 0.91 | [0.84, 0.95] | 0.77 | [0.71, 0.82] | 0.57 | [0.50, 0.64] | 0.73 | [0.69, 0.76] |
| Algorithm [‘Ax,’ ‘Cor,’ ‘Sag,’ ‘SagT1,’ ‘CorPD’] | 0.94 | [0.87, 0.97] | 0.79 | [0.73, 0.84] | 0.56 | [0.49, 0.63] | 0.74 | [0.70, 0.77] |
| Algorithm [‘Ax’] | 0.94 | [0.87, 0.97] | 0.57 | [0.50, 0.63] | 0.61 | [0.54, 0.68] | 0.66 | [0.62, 0.70] |
| Algorithm [‘Cor’] | 0.93 | [0.86, 0.96] | 0.65 | [0.59, 0.71] | 0.66 | [0.59, 0.72] | 0.71 | [0.67, 0.75] |
| Algorithm [‘CorPD’] | 0.95 | [0.90, 0.98] | 0.62 | [0.55, 0.68] | 0.51 | [0.44, 0.58] | 0.65 | [0.61, 0.69] |
| Algorithm [‘Sag’] | 0.8 | [0.71, 0.86] | 0.81 | [0.75, 0.85] | 0.5 | [0.43, 0.57] | 0.69 | [0.65, 0.73] |
| Algorithm [‘SagT1’] | 0.84 | [0.76, 0.90] | 0.73 | [0.67, 0.79] | 0.49 | [0.42, 0.56] | 0.67 | [0.62, 0.71] |
| Subscapularis | | | | | | | | |
| Clinical reader | 0.75 | [0.64, 0.84] | 0.79 | [0.73, 0.83] | 0.64 | [0.57, 0.70] | 0.73 | [0.69, 0.76] |
| Algorithm [‘Ax,’ ‘Cor,’ ‘Sag,’ ‘SagT1’] | 0.6 | [0.48, 0.70] | 0.81 | [0.75, 0.85] | 0.72 | [0.66, 0.78] | 0.75 | [0.71, 0.78] |
| Algorithm [‘Ax,’ ‘Cor,’ ‘Sag,’ ‘SagT1,’ ‘CorPD’] | 0.51 | [0.40, 0.63] | 0.8 | [0.75, 0.85] | 0.72 | [0.65, 0.78] | 0.73 | [0.69, 0.77] |
| Algorithm [‘Ax’] | 0.79 | [0.68, 0.87] | 0.79 | [0.73, 0.83] | 0.64 | [0.57, 0.71] | 0.74 | [0.70, 0.77] |
| Algorithm [‘Cor’] | 0.26 | [0.18, 0.38] | 0.86 | [0.81, 0.90] | 0.63 | [0.56, 0.69] | 0.69 | [0.65, 0.73] |
| Algorithm [‘CorPD’] | 0.03 | [0.01, 0.10] | 0.76 | [0.70, 0.80] | 0.6 | [0.53, 0.66] | 0.6 | [0.55, 0.64] |
| Algorithm [‘Sag’] | 0.51 | [0.40, 0.63] | 0.73 | [0.67, 0.78] | 0.79 | [0.72, 0.84] | 0.72 | [0.68, 0.76] |
| Algorithm [‘SagT1’] | 0.54 | [0.43, 0.65] | 0.65 | [0.59, 0.71] | 0.71 | [0.64, 0.77] | 0.66 | [0.62, 0.70] |

The reference standard for the clinical reader and the algorithms was the tear classification determined by the multireader study majority vote.
*95% confidence intervals shown in brackets.



## Reader Variability

In the multireader study, there was substantial agreement among the readers overall for the classification of supraspinatus and infraspinatus tendon tears with linear-weighted κ values of 0.68 (95% CI, 0.65–0.71) and 0.62 (95% CI, 0.59–0.66). There was moderate agreement among readers overall for subscapularis tears, with κ of 0.58 (95% CI, 0.54–0.61). However, κ values for pair-wise comparison between individual readers from the multireader study varied considerably for each tendon; for example, from a minimum of 0.18 (95% CI, 0.0–0.62) to a maximum of 0.93 (95% CI, 0.80–1.0) for infraspinatus tears (Table 4). On a per-tendon basis, approximately 7–8% of the test set cases required adjudication by a fourth reader.

## DISCUSSION

We have demonstrated that deep-learning diagnosis of rotator cuff tears is feasible with excellent diagnostic performance, particularly for full-thickness tears, followed by no tears and partial-thickness tears. The lower performance for partial-thickness tears for both the model as well as human readers corroborates the known challenges and reader variability for partial tears. The superior performance of the model for full-thickness tears compared with no tears may be due to the discrete fluid signal that can be more easily identified with full-thickness tears, whereas patients without tears may have varying degrees of signal abnormality and heterogeneity due to varying degrees of tendinosis.

Model AUCs were slightly higher for 4-view input versus single sequence for infraspinatus and subscapularis, whereas for supraspinatus, model performance was equal for coronal oblique FS and multisequence input. However, there was very little separation between the ROC curves for the supraspinatus and infraspinatus tendons in Figure 4, particularly between the 4-view, coronal oblique FS, and sagittal oblique FS inputs for supraspinatus. For full-thickness supraspinatus and infraspinatus tendon tears, the model demonstrated similar accuracies for virtually any combination of input sequences, whether 1, 2, or up to 5. This potentially means that the model could aid in the clinical interpretation of tendon tears with relative flexibility regarding protocols, available sequences, and in situations where examinations are incomplete or acquisitions are suboptimal due to artifacts. Algorithm performance was similar at 1.5 T and 3T, indicating the model does not require a higher field strength for adequate diagnosis, which is also advantageous to clinical application.

For all tendons, the 4-view model's overall accuracy and accuracy for each tear type were similar to that of the clinical reader. One proposed clinical implementation is to use the model as an aid in interpretation for the radiologist, similar to having a proficient resident or

**TABLE 3.** Statistical Significance Tests for Difference in Overall Accuracy Between the 4-View DL Ensemble Model and Clinical Reader Computed With McNemar Test

| Tendon | *P* | 95% CI |
|---|---|---|
| Supraspinatus | 0.246 | [−0.016, 0.066] |
| Infraspinatus | 0.751 | [−0.06, 0.04] |
| Subscapularis | 0.444 | [−0.027, 0.066] |

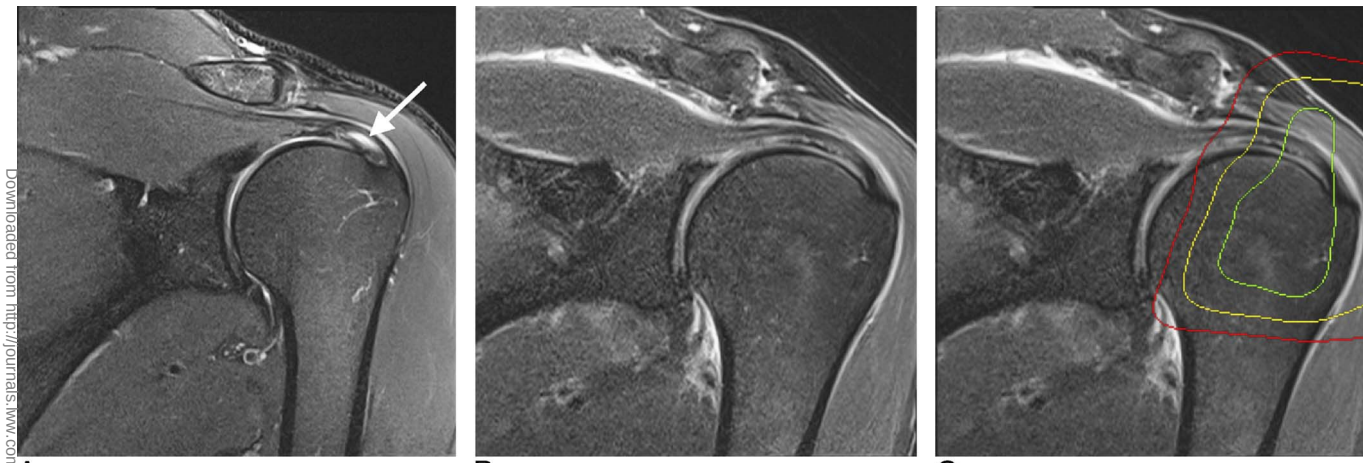

**FIGURE 6.** Example false-negative and false-positive supraspinatus cases. A, Coronal oblique PD-weighted fat-suppressed image of the supraspinatus tendon demonstrates an example false-negative case where the ground truth label was partial tear (arrow) and the model predicted no tear. B, Coronal oblique PD-weighted fat-suppressed image of the supraspinatus tendon demonstrates an example false-positive case where the ground truth label was no tear and the model predicted full-thickness tear. C, Activation map showing area outlined by green, the pixels within which most influence the model prediction, and a red outline, the pixels outside of which least influence the model prediction.

fellow who can detect, classify, measure rotator cuff tears, and provide a prepopulated report that would hopefully require minimal editing. Like clinical readers, the model performs better with certain tear types. Another potential application of the model is to leverage the variability of its performance characteristic for rapid triage of cases. Depending on the operating point chosen for the model, the relative accuracy between tear type classifications can be shifted. For example, based on the ROC and accuracy results, the model is proficient at detecting full-thickness tears (those more likely to require surgery) with high accuracy. The model could be used to highlight full-thickness tears and separate them from more challenging cases such as partial-thickness tears; this could increase radiologists' workstation efficiency. On the other hand, it may be preferable to diagnose or triage no tears or "normals" with higher accuracy, and the specificity of the model could be adjusted to do so.

Another potential advantage of DL models is their inherent reproducibility. Although there was substantial agreement among the study readers in general for supraspinatus and infraspinatus tear classification, the range of agreement between individual readers was large, from slight to almost perfect, even among a group of subspecialty-trained radiologists. The pairwise best and worst κ values reflect the real-life clinical variability in rotator cuff interpretation that can occur for individual patients, with concomitant management implications. As algorithms are less prone to variability, this is an attractive strength of using DL to improve consistency of rotator cuff interpretation. The DL model also provides a consistent level of expertise that is on par with a subspecialty-trained musculoskeletal radiologist. In practices that lack subspecialty-trained musculoskeletal radiology expertise, a DL model could potentially help by providing consistent, high-quality rotator cuff evaluations.

To our knowledge, 3 studies[24–26] have investigated automated rotator cuff tear classification using DL, 2 of which reported model performance superior to humans.[24,25] Both used significantly smaller data sets of around 2000 cases, compared with almost 12,000 in this study. Although Shim et al[24] had the advantage of arthroscopic ground truth for tear classification labeling, major limitations included lack of explicit tear localization by the algorithm, that is, specifying which rotator cuff tendon was torn and image interpretation performed by orthopedic surgeons.

Kim et al[25] reported that their model outperformed human readers; however, image interpretation was performed by medical students and orthopedic surgery residents. Similar to Shim et al,[24] their algorithm detects rotator cuff tears without specifying which tendon is involved. Their ground truth was based on the MRI interpretation of a single orthopedic surgeon with confirmation by a single musculoskeletal radiologist, which is less robust than arthroscopy or a multireader study. Algorithm input was limited to a single coronal T2-weighted acquisition, compared with the multisequence inputs in our study and Shim et al.[24] Finally, the authors excluded cases with fracture, dislocation, avascular necrosis, severe degenerative arthritis, and large calcific deposits. This limits the utility of a DL model, if there are specific input criteria and the model cannot interpret all types of shoulder MRI with common conditions but only a subset of all potential patients.

Recently, Yao et al[26] looked at the supraspinatus tendon alone and similarly found superior sensitivity for full-thickness tears than partial-thickness tears, as well as no significant difference in AUC between 1.5 T and 3T studies. The authors used a multistage approach composed of a slice selection network, segmentation network, and binary (tear present vs tear absent) classification network on a much smaller data set of 200 cases, 40 of which constituted the test set. A single coronal oblique T2-weighted sequence served as input and radiology reports as ground truth. The data were obtained on Siemens and GE systems from a single institution. Of note, their misclassified normal cases showed a significantly greater incidence of tendinosis than correctly classified cases.

**TABLE 4.** Minimum and Maximum κ Values Among Individual Study Reader Pairs and From All Study Readers by Tendon

| Tendon | Lowest κ Reader Pair | 95% CI | Highest κ Reader Pair | 95% CI | All Readers | 95% CI |
|---|---|---|---|---|---|---|
| Supraspinatus | 0.38 | [0.10, 0.65] | 0.92 | [0.76, 1.0] | 0.68 | [0.65, 0.71] |
| Infraspinatus | 0.18 | [0.0, 0.62] | 0.93 | [0.80, 1.0] | 0.62 | [0.59, 0.66] |
| Subscapularis | 0.05 | [0.0, 0.36] | 0.86 | [0.61, 1.0] | 0.58 | [0.54, 0.61] |

Our study has limitations. First, ground truth for training and validation was the clinical radiology report, and ground truth for the test set remained an imaging reference standard generated by a group of expert human readers. As a result, the study design precludes a conclusion of model performance superior to human readers. In addition, it is a reference standard with considerable variability, even with a large panel of expert readers. Using retrospective surgical ground truth, in our experience however, has been limited by incomplete information within operative reports. In addition, the surgical cases may be biased toward more severe tears with fewer patients having a partial-thickness tear or an intact cuff.

Second, our data set had a disproportionate number of examinations performed on MR systems from a single manufacturer, which may limit the generalizability of our model to shoulder MRs performed on non-Siemens systems. Although this study used training data from multiple institutions, the test set was composed of examinations from a single institution. Further testing of the model on data sets from other sites and vendors would evaluate its potential for broad, real-world application.

## CONCLUSIONS

We have demonstrated that DL diagnosis of rotator cuff tears is feasible with excellent diagnostic performance, particularly for full-thickness tears, and with model accuracy similar to subspecialty-trained clinical readers. Potential future directions include external validation of the model on images from different vendors and sites, consideration of a surgical ground truth, and comparison of unassisted and DL-assisted radiologist interpretation to assess the model's effects on radiologist accuracy, variability, and efficiency.

## ACKNOWLEDGMENTS

*The authors acknowledge James Babb, PhD, for his statistical support. They would also like to acknowledge Mei Li, MD; Xiaojuan Li, PhD; Po-Hao (Howard) Chen, MD; Ceylan Colak, MD; Suri Surender and Greg Strnad from the Cleveland Clinic; and Kecheng Liu, PhD, MBA from Siemens Medical Solutions USA for their assistance with data acquisition.*

## REFERENCES

1. Tashjian RZ. Epidemiology, natural history, and indications for treatment of rotator cuff tears. *Clin Sports Med*. 2012;31:589–604.
2. Morag Y, Jacobson JA, Miller B, et al. MR imaging of rotator cuff injury: what the clinician needs to know. *Radiographics*. 2006;26:1045–1065.
3. Spencer EE Jr., Dunn WR, Wright RW, et al. Interobserver agreement in the classification of rotator cuff tears using magnetic resonance imaging. *Am J Sports Med*. 2008;36:99–103.
4. Robertson PL, Schweitzer ME, Mitchell DG, et al. Rotator cuff disorders: interobserver and intraobserver variation in diagnosis with MR imaging. *Radiology*. 1995;194:831–835.
5. Malavolta EA, Assunção JH, Gracitelli MEC, et al. Accuracy of magnetic resonance imaging (MRI) for subscapularis tear: a systematic review and meta-analysis of diagnostic studies. *Arch Orthop Trauma Surg*. 2019;139:659–667.
6. Malavolta EA, Assunção JH, Guglielmetti CLB, et al. Accuracy of preoperative MRI in the diagnosis of subscapularis tears. *Arch Orthop Trauma Surg*. 2016;136:1425–1430.
7. Kim HJ, Park JS, Kim JY, et al. Interstitial tears of the rotator cuff: difficulty in preoperative diagnosis. *J Shoulder Elbow Surg*. 2018;27:487–492.
8. Brockmeyer M, Schmitt C, Haupert A, et al. Limited diagnostic accuracy of magnetic resonance imaging and clinical tests for detecting partial-thickness tears of the rotator cuff. *Arch Orthop Trauma Surg*. 2017;137:1719–1724.
9. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med*. 2018;15:e1002699.
10. Fritz J, Kijowski R, Recht MP. Artificial intelligence in musculoskeletal imaging: a perspective on value propositions, clinical use, and obstacles. *Skeletal Radiol*. 2022;51:239–243.
11. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
12. Lin DJ, Walter SS, Fritz J. Artificial intelligence-driven ultra-fast superresolution MRI: 10-fold accelerated musculoskeletal turbo spin echo MRI within reach. *Invest Radiol*. 2022;58:28–42.
13. Germann C, Marbach G, Civardi F, et al. Deep convolutional neural network-based diagnosis of anterior cruciate ligament tears: performance comparison of homogenous versus heterogeneous knee MRI cohorts with different pulse sequence protocols and 1.5-T and 3-T magnetic field strengths. *Invest Radiol*. 2020;55:499–506.
14. Niemeyer F, Galbusera F, Tao Y, et al. A deep learning model for the accurate and reliable classification of disc degeneration based on MRI data. *Invest Radiol*. 2021;56:78–85.
15. Fritz B, Fritz J. Artificial intelligence for MRI diagnosis of joints: a scoping review of the current state-of-the-art of deep learning-based approaches. *Skeletal Radiol*. 2022;51:315–329.
16. Fritz B, Yi PH, Kijowski R, et al. Radiomics and deep learning for disease detection in musculoskeletal radiology: an overview of novel MRI- and CT-based approaches. *Invest Radiol*. 2022;58:3–13.
17. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A*. 2018;115:11591–11596.
18. Urakawa T, Tanaka Y, Goto S, et al. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol*. 2019;48:239–244.
19. Liu F, Guan B, Zhou Z, et al. Fully automated diagnosis of anterior cruciate ligament tears on knee MR images by using deep learning. *Radiol Artif Intell*. 2019;1:180091.
20. Chang PD, Wong TT, Rasiej MJ. Deep learning for detection of complete anterior cruciate ligament tear. *J Digit Imaging*. 2019;32:980–986.
21. Liu F, Zhou Z, Samsonov A, et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. *Radiology*. 2018;289:160–169.
22. Gyftopoulos S, Lin D, Knoll F, et al. Artificial intelligence in musculoskeletal imaging: current status and future directions. *Am J Roentgenol*. 2019;213:506–513.
23. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. *Skeletal Radiol*. 2020;49:183–197.
24. Shim E, Kim JY, Yoon JP, et al. Automated rotator cuff tear classification using 3D convolutional neural network. *Sci Rep*. 2020;10:15632.
25. Kim M, Park HM, Kim JY, et al. MRI-based diagnosis of rotator cuff tears using deep learning and weighted linear combinations. *Machine Learning for Healthcare Conference PMLR*. 2020. Available at: https://proceedings.mlr.press/v126/kim20a.html. Accessed January 21, 2021.
26. Yao J, Chepelev L, Nisha Y, et al. Evaluation of a deep learning method for the automated detection of supraspinatus tears on MRI. *Skeletal Radiol*. 2022;51:1765–1775.
27. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6:1–48.
28. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2016:770–778.
29. Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks? In: *Advances in Neural Information Processing Systems*. 2014. Available at: https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html. Accessed January 21, 2021.
30. Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. *J Big Data*. 2016;3:1–40.
31. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22:1345–1359.
32. Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2018:6546–6555.
33. Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset. *arXiv preprint arXiv*. 2017;1705.06950. Available at: https://arxiv.org/abs/1705.06950. Accessed January 21, 2021.
34. McKinney W. Data structures for statistical computing in Python. In: *Proceedings of the 9th Python in Science Conference*. 2010;445:51–56.
35. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. In: *Proceedings of the 9th Python in Science Conference*. 2010. Available at: https://pdfs.semanticscholar.org/3a27/6417e5350e29cb6bf04ea5a4785601d5a215.pdf. Accessed January 21, 2021.