MRI-based Diagnosis of Rotator Cuff Tears using Deep Learning and Weighted Linear Combinations

Mijung Kim[*] MIJUNG.KIM@UGENT.BE
IDLab, ELIS
Ghent University, Ghent, Belgium

Ho-min Park HOMIN.PARK@UGENT.BE
IDLab, ELIS
Ghent University, Ghent, Belgium

Jae Yoon Kim, M.D., Ph.D.[*] KJYCJE@GMAIL.COM
Department of Orthopedic Surgery
Chung-Ang University Hospital, Seoul, Korea

Seong Hwan Kim, M.D., Ph.D. KSH170177@NATE.COM
Department of Orthopedic Surgery
Chung-Ang University Hospital, Seoul, Korea

Sofie Van Hoeke, Ph.D. SOFIE.VANHOEKE@UGENT.BE
IDLab, ELIS
Ghent University, Ghent, Belgium

Wesley De Neve, Ph.D. WESLEY.DENEVE@UGENT.BE
IDLab, ELIS
Ghent University, Ghent, Belgium

Editor: Editor's name

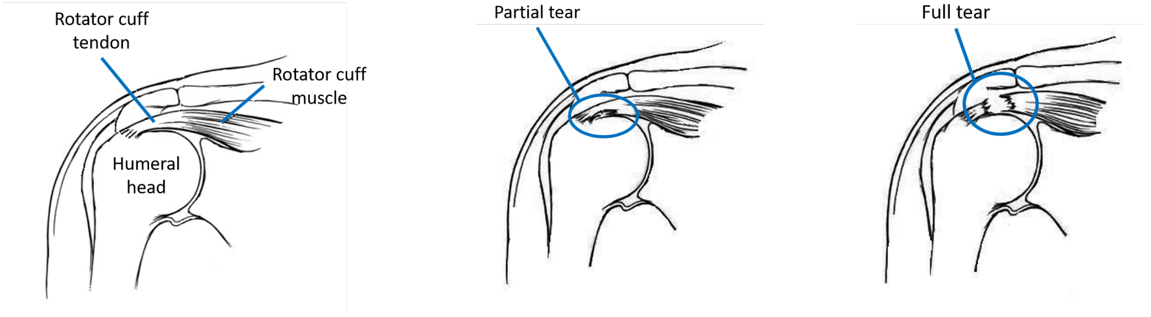
Abstract

Rotator Cuff Tears (RCTs) are a common injury among people who are middle-aged or older. For effective diagnosis of RCTs, orthopedic surgeons typically need to have access to both shoulder Magnetic Resonance Imaging (MRI) and proton density-weighted imaging. However, the generation and interpretation of such comprehensive image information is labor intensive, and thus time consuming and costly. Although computer-aided diagnosis can help in mitigating the aforementioned issues, no computational tools are currently available for diagnosing RCTs. Therefore, we introduce a computational approach towards RCT diagnosis in this paper, leveraging end-to-end learning by applying a deep convolutional neural network to shoulder MRI scans. Given that these shoulder MRI scans are 3-D by nature and highly biased towards normal shoulders, with only 6.6% of the available shoulder MRI scans containing partial-thickness tears, we made use of two tools to enhance our deep convolutional neural network. First, to enable the utilization of sequential information available in the 3-D MRI scans, we integrated a weighted linear combination layer. Second, to mitigate the presence of class imbalance, we adopted weighted cross-entropy loss. That way, we were able to obtain a diagnostic accuracy of 87% and an M-AUC score of 97%, outperforming a baseline of human annotators (diagnostic accuracy of 76% and an M-AUC score of 81%). In addition, we were able to outperform several approaches using

[*] Corresponding authors: Mijung Kim and Jae Yoon Kim

conventional machine learning techniques. Finally, to facilitate further research efforts and ease of benchmarking, we make our dataset of 2,447 shoulder MRI scans publicly available.

1. Introduction



(*a*) Normal (*b*) Partial-thickness tear (*c*) Full-thickness tear

Figure 1: Anatomical comparison between a normal and a torn tendon, showing where tears typically happen (in the blue circles) and the shape of these tears. If present, tears can largely be divided into two classes, depending on severity: partial-thickness tears and full-thickness tears.

Among patients suffering from shoulder pain, rotator cuff disorders are most frequently implicated, being present in upto 86% of patients (Sharma et al., 2017). Specifically, Rotator Cuff Tears (RCTs), which are particularly prevalent in patients who are middle-aged or older (Kim et al., 2017), are the leading cause for shoulder surgery. Factors influencing the success of rotator cuff repair include age, chronicity, tear size, and tear shape. An illustration of different types of RCTs can be found in Figure 1.

To diagnose RCTs, orthopedic surgeons perform an examination by making use of Magnetic Resonance Imaging (MRI), leveraging T1- and T2-weighted sagittal, coronal, and axial images, and proton density-weighted (PD-weighted) imaging. However, generating and interpreting such comprehensive imaging information in support of medical decision making requires a considerable amount of human labor. Furthermore, orthopedic surgeons often still find it challenging to assess whether tears require surgical intervention.

The use of state-of-the-art deep learning approaches for Computer-Aided Diagnosis (CAD) of brain, lung, and cardiac diseases is an active area of research and development (Litjens et al., 2017). The resulting tools often come with a high diagnostic accuracy and high Area Under the Curve (AUC) scores, also making it possible to reduce the amount of time needed to examine medical images, thus facilitating more effective and faster decision taking by medical doctors. However, despite their significant impact on the quality of life of patients, to the best of our knowledge, no computer-aided tools are currently available for diagnosing RCTs. Therefore, in this study, we propose the first computational approach towards RCT diagnosis, leveraging an end-to-end learning approach through a deep Convolutional Neural Network (CNN). Specifically, given that our shoulder MRI scans are (1) 3-D by nature and (2) highly biased towards normal shoulders (that is, shoulders without tears make up for 66% of the total number of MRI scans at our disposal), the proposed approach uses, in combination with a deep CNN, a Weighted Linear Combination (WLC) layer to take advantage of the available 3-D MRI information and weighted cross-entropy loss in order to mitigate class imbalance issues.

Technical Significance As the first end-to-end CAD tool for RCT detection, our approach is able to take advantage of 3-D MRI information and to mitigate issues in terms of class imbalance. In addition, we only make use of T2-weighted coronal images, minimizing the amount of information necessary for performing a diagnosis, thus making the model more efficient. Working with three classes (normal, partial-thickness, and full-thickness), we obtained a diagnostic accuracy of 87% and an M-AUC score of 97%. To take into account class imbalance, we also made use of Precision-Recall (PR) curves and confusion matrices.

Clinical Relevance In this study, we introduce a fully automated CAD tool to detect RCTs in MRI scans, comparing the effectiveness of our predictive model to that of a human baseline. In the assumption that no shoulder-specialized surgeon or musculoskeletal-specialized radiologist is available, the use of our tool is expected to improve the accuracy of diagnosis when a general orthopedic surgeon is responsible for patient treatment, by for instance preventing under-diagnosis of RCTs.

In summary, our major contributions are as follows:

- Using a deep CNN and a WLC layer, we propose the first end-to-end learning approach for computer-aided diagnosis of RCTs, obtaining a diagnostic accuracy of 87% and an M-AUC score of 97%.
- We extensively compare our approach against several baselines, including an approach using human annotators and approaches based on traditional machine learning techniques.
- There is currently no shoulder MRI dataset publicly available for research purposes. By releasing our shoulder MRI dataset of 2,447 T2-weighted coronal scans, we hope to facilitate further research efforts and ease of benchmarks.

Generalizable Insights about Machine Learning in the Context of Healthcare

Given the research effort presented in this paper, we can put forward three generalizable insights into the use of machine learning in a healthcare context.

- A high class imbalance is an issue commonly associated with healthcare datasets. Paying more attention to this issue when developing and using (deep) machine learning approaches is expected to result in non-trivial diagnostic accuracy improvements.
- The increasing availability of computational approaches for determining disease severity is expected to contribute substantially to improved clinical decision taking.
- The need for developing effective and efficient (deep) machine learning approaches for dealing with 3-D datasets of MRI or Computed Tomography (CT) images is expected to become more prevalent, given that medical imaging equipment for generating 3-D datasets is more and more commonly used in a hospital setting.

The remainder of this paper is organized as follows. We review related work on MRI datasets in Section 2. In Section 3, we describe the proposed approach in more detail. We present our experimental setup and our experimental results in Section 4, also paying attention to the characteristics of our dataset. Finally, in Section 5, we provide concluding remarks, including a discussion of the limitations of our study. We also provide a number of suggestions for future research.

2. Related Work

First, we briefly review a number of related research efforts that are mainly focusing on MRI data analysis, with the aim of identifying candidate baseline techniques. Next, we review a deep learning approach for analysis of musculoskeletal MRI data.

Machine learning for CAD in MRI began with brain disease diagnosis. Wang and Pham (2011) introduced a machine learning model for brain age prediction, whereas Usman and Rajpoot (2017) proposed the use of machine learning for brain tumor classification. Gurusamy and Subramaniam (2017) presented a model for brain tumor segmentation and classification, extracting features using the Discrete Wavelet Transform (DWT) and Principal Component Analysis (PCA), with PCA being explained in Abdi and Williams (2010). Using a biological image classification package developed by Shamir et al. (2008), Ashinsky et al. (2017) proposed a technique for osteoarthritis sign prediction using knee MRI. Note that all of the aforementioned machine learning techniques make use of hand-crafted feature extraction, given that raw images themselves typically contain too much information to be processed by traditional machine learning approaches.

In the field of machine learning, deep learning has attracted substantial attention during the past years. The use of deep learning has two distinct advantages over conventional machine learning approaches. A first advantage is learnable feature extraction. Unlike traditional machine learning using hand-crafted features, deep learning approaches define and extract features from the data themselves. In other words, when large amounts of labeled data are available, a deep learning model is able to find the optimal feature extraction method by itself. A second advantage of deep learning is its high accuracy. When plenty of data are available, deep learning approaches often surpass traditional machine learning approaches in terms of effectiveness, especially in the area of visual content understanding.

In recent years, different types of medical image sets have been released and numerous attempts have been made to gain insight into these medical image sets (see for instance the survey presented in Litjens et al. (2017)). In the musculoskeletal field in particular, a large number of knee MRI datasets have recently been made available for research purposes. Liu et al. (2018) proposed a model for segmentation of knee cartilage lesions, using U-Net (which is based on VGG-16) and 175 MRI slices. Roblot et al. (2019) conducted a study to find and classify meniscus tears using Fast R-CNN and Faster R-CNN, leveraging 1,823 MRI scans. Bien et al. (2018) presented MRNet, using a modified version of AlexNet in order to classify anterior cruciate ligament tears, meniscal tears, and abnormalities, leveraging 1,370 3-D MRI scans.

On the other hand, in the shoulder musculoskeletal field, a relatively low number of research efforts have thus far been pursued, mainly due to a lack of data. The authors of Liu et al. (2019) introduced Mask R-CNN to distinguish a glenoid head and a humeral

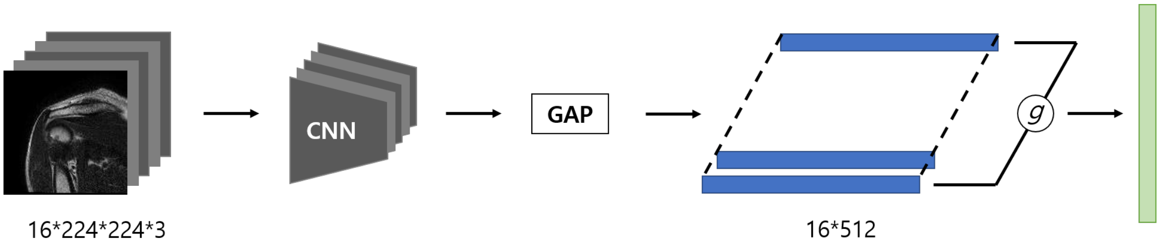


Figure 2: Augmented MRI scans of 16 slices (images) are used as an input. The final convolutional layer of VGG-16 is followed by A Global Average Pooling (GAP) layer. The output of this GAP layer is colored in blue. The weighted linear combination of the sixteen output vectors generates a 512×1 vector that has been colored in green.

head from other muscles in shoulder joint MRI. To that end, 50 sets of slices were used. The authors of Conze et al. (2019) presented a U-Net-based model for shoulder muscle segmentation. As an input, 24 pediatrics MRI series were used. Although the amount of data is small, each muscle part is finely divided. Kim et al. (2019) presented a study that is most similar to ours, analyzing a technique for detecting a fossa and muscle region using an encoder-decoder model based on VGG-19. As an input, 240 sagittal MRI series were used.

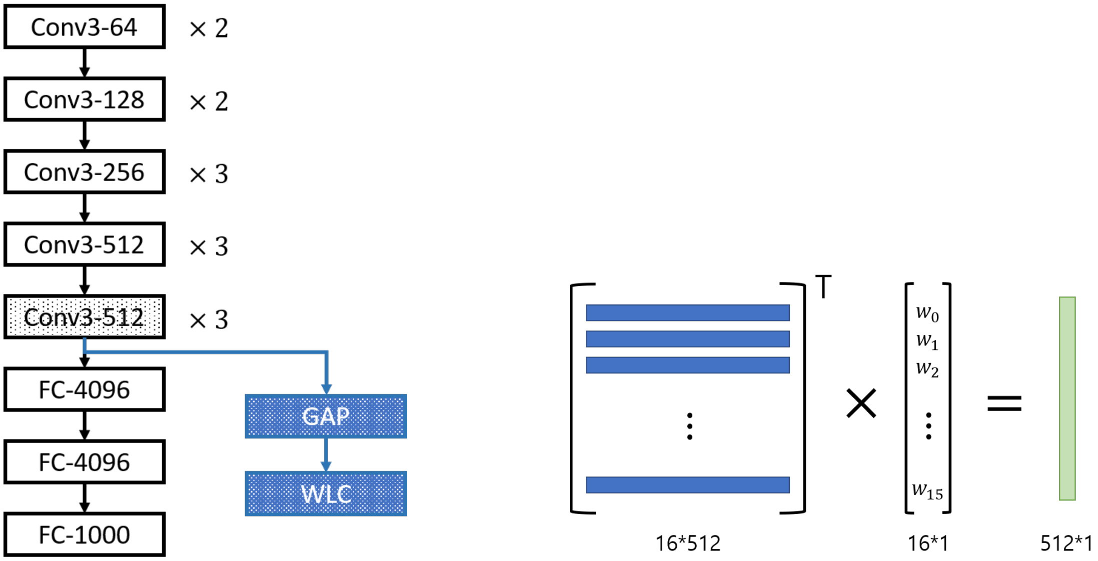
3. Method

In this section, we provide more details about the proposed approach, as shown in Figure 2. In particular, Section 3.1 introduces the model we built on top of VGG-16. Next, Section 3.2 describes the Weighted Linear Combination layer used. Lastly, Section 3.3 discusses the loss function and the optimization strategy used.

3.1. Embedding Function

One of the most effective approaches towards CAD is the use of convolutional neural networks. However, the latest high-performing CNNs are only able to take advantage of the spatial information available in the input images. As a result, a straightforward application of the aforementioned CNN-based models will not consider the sequential information present in 3-D MRI scans. In this study, to obtain more reliable diagnostic outcomes, we aim at leveraging both spatial and sequential tear information, given the observation that tears typically do not occur in a single slice but throughout a range of slices in a particular MRI scan. However, given this observation, we can also point that not all slices in an MRI scan are equally important. Therefore, as a first step towards effective RCT diagnosis, we employ a learnable weights layer after the last convolutional layer of a VGG-16 network (Simonyan and Zisserman, 2014).

Starting from a VGG-16 backbone, we built an embedding function for each slice in an MRI scan. To construct this embedding function, we removed all of the FC layers after the last convolutional layer of the VGG-16 backbone. By adding a Global Average Pooling (GAP) layer to flatten the last features maps, we obtained 16 feature vectors of size



(*a*) Network architecture

(*b*) Weighted linear combination

Figure 3: Proposed network architecture, incorporating a Weighted Linear Combination (WLC) layer. (a) Proposed simplified network architecture based on VGG-16. All the convolutional (Conv) layers have 3×3 kernels, with the number of channels in [64, 128, 256, and 512]. The notation $\times\{no.\}$ denotes the number of repetitions of each Conv layer. Except for the last Conv layer (dotted), each Conv layer is followed by a max pooling layer. Instead of the FC-[4096, 1000] layers in the original version of VGG-16, we used a GAP layer and a WLC layer. (b) The weighted matrix multiplication performed by the last layer.

512×1 per MRI scan, with each MRI scan containing 16 slices each. Dropout layers with a dropout rate of 0.5 are added before and after the GAP layer in order to reduce model overfitting. A visualization of our model can be found in Figure 3a. Upon availability of more computational resources, the base model can have deeper layers than the ones we used, so to enable an increase in diagnostic accuracy.

3.2. Weighted Linear Combination

Each output $GAP(f_\theta(\mathbf{x}))$ is given as an input to a weighted linear combination g_ϕ, with θ and ϕ referring to weights. Using $\mathbf{x}$ to denote an MRI scan, then $\mathbf{x} \in \mathcal{R}^{16\times224\times224\times3}$. Furthermore, f_θ denotes the embedding function, consisting of the first 13 layers of VGG-16. The Weighted Linear Combination (WLC) layer learns which slices are important among all 16 slices available in an MRI scan by learning weights during training. As a result, g_ϕ can be defined as follows:

$$g_\phi(\mathbf{x}) = GAP(f_\theta(\mathbf{x}))^T \cdot \mathbf{w}_\phi. \tag{1}$$

The output of Equation 1 has the shape of a 512×1 feature vector. This feature vector is then given as an input to the softmax function in order to generate a final RCT diagnosis. This is also illustrated by Figure 3b.

Table 1: Summarizing statistics for the shoulder MRI datasets used for training, validation, and testing. The proportion of each class is approximately the same in each dataset.

Statistics	Training	Validation	Testing
Total number of examinations (%)	1,963 (100)	242 (100)	242 (100)
- Normal examinations (%)	1,308 (66.6)	160 (66.1)	160 (66.1)
- Partial-thickness tear examinations (%)	125 (6.4)	16 (6.6)	16 (6.6)
- Full-thickness tear examinations (%)	530 (27.0)	66 (27.3)	66 (27.3)
Total number of patients	1,847	231	228
- Female patients (%)	942 (51)	115 (49)	134 (58)
- Mean age of patients (±*std.*)	56 (±14.8)	57 (±14.9)	56 (±14.6)

3.3. Weighted Cross-Entropy Loss and Optimization

During the learning of the weight parameters $\mathbf{w}_\phi$, we calculate the loss by making use of the weighted cross-entropy loss function, so to be able to alleviate class imbalance issues:

$$\mathcal{L}_{weighted} = -\frac{1}{N}\sum_i \alpha_i \cdot y_i \cdot \log(\hat{y}_i), \tag{2}$$

where α_i is the weight assigned to class i, N is the number of classes used, y_i is the true class, and $\hat{y}_i$ is the predicted class. The weights $\alpha_\mathbf{i}$ are inversely proportional to the number of training examples available per class. Optimization was conducted by making use of Adaptive Moment Estimation (Adam) (Kingma and Ba, 2014).

4. Experiments

4.1. Patient Cohort

For a total of 2,492 subjects, shoulder MRI scans were obtained at Chung-Ang University Hospital in Korea, from March 2010 to October 2018. Specifically, the MRI scans were first captured by making use of a 3.0 Tesla Achieva system (Phillips, the Netherlands) from March 2010 to June 2017. This MRI system was then replaced with a 3.0 Tesla Skyra system (Siemens AG Healthcare, Germany) from July 2017 to October 2018. The patients were placed in the supine position and the humerus in a neutral position. All scans in the entire dataset only consist of T2-weighted coronal slices, coming with a slice thickness of 2 mm.

Among all subjects with a shoulder MRI scan available, only patients with rotator cuff pathology were included for this study. The exclusion criteria were (1) a history of prior shoulder surgery, (2) bone or joint destruction due to infections or tumorous conditions, (3) severe degenerative arthritis or avascular necrosis, (4) fractures with/without dislocations, and (5) large calcific deposits in the supraspinatus tendon (over 1 cm). As a result, 2,447 patients who met our inclusion/exclusion criteria were included. The included patients were categorized as follows: (1) normal or partial thickness tears with a thickness less than 50% of the tendon thickness, (2) partial thickness tears with a thickness more than 50% of the tendon thickness, and (3) full-thickness tears (Osti et al., 2017; Katthagen et al., 2018).

In the presence of a full-thickness tear or in the presence of a partial thickness tear that is exceeding more than 50% of the tendon thickness, an image is annotated as belonging to the tear class. Otherwise, it is annotated as belonging to the normal class. The total number of patients is 2,447: 1,628 normal patients and 819 tear patients. For each patient, the dataset contains an MRI scan consisting of 16 shoulder slices. We split the dataset into a training set and a test set, using a ratio of 9:1. We also used one tenth of the training images for validation purposes. The patient cohort statistics are summarized in Table 1. Our annotated dataset is publicly available.

Data augmentation was implemented using randomization, using a probability of 0.5, selecting all parameter values in an empirical way. For rotation, the angle was randomly selected in the range of [-15, 15] degrees. The kernel size for Gaussian blur was also randomly selected in the range of [1, 3, 5, 7, and 9]. Gaussian noise was normalized with mean 0 and variance 0.1.

4.2. Implementation Details

Baseline 1: Human We selected ten people knowledgeable about the medical domain (medical school senior students and university hospital residents). They diagnosed the images in our test set.

Baseline 2: Wavelet feature extraction and classification To investigate the effectiveness of our model, we adopted the feature extraction method of (Nayak et al., 2016) as a baseline. In particular, the authors of (Nayak et al., 2016) applied a 3-level wavelet decomposition and Probabilistic PCA (PPCA) to brain MRI images, extracting a 13-D feature vector per image. Furthermore, they created an AdaBoost classification model. Given that this approach was able to classify brain lesions with a promising effectiveness, we extracted features by applying the same techniques to the shoulder MRI scans at our disposal. However, unlike the model of Nayak et al., 2016, we have to classify the presence or absence of tears in 16 slices instead of a single image. As such, by applying a 3-level wavelet decomposition and PPCA to the slices available, we created a total of 208 ($= 13 \times 16$) feature vectors, which were then classified using (1) a k-nearest neighbor model ($k = 9$) and (2) AdaBoost (Schapire, 2013) (10 estimators), with the latter obtaining the highest accuracy in Nayak et al., 2016.

Proposed approach We implemented our approach using Python 2.7 and PyTorch 1.0, leveraging a VGG-16 network pretrained on ImageNet (Deng et al., 2009). Execution was done on two Intel(R) Xeon(R) E5-2620 2.4GHz CPUs and an NVIDIA GeForce GTX TITAN X GPU. The batch size was 32, the learning rate was 1e-4, and the overall number of epochs needed to reach training error convergence was 35.

4.3. Results

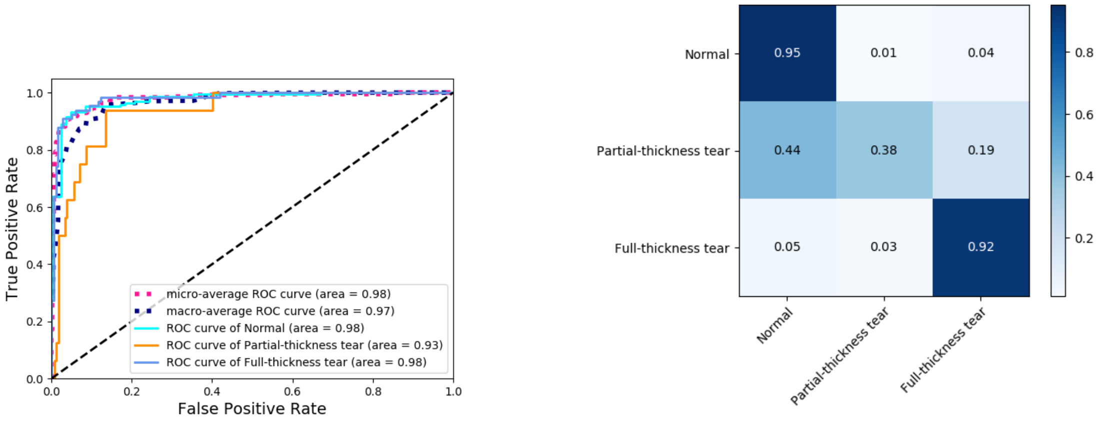
We evaluated the effectiveness of diagnosis using accuracy and the AUC score. Other than that, the model robustness was quantified using recall (*a.k.a. sensitivity*), precision, and the F1 score. Together with the sensitivity, specificity was also calculated in support of a clinical assessment.

Table 2: Overall comparison of the effectiveness of the different classifiers used. M-AUC denotes the macro-averaged AUC score and m-AUC denotes the micro-averaged AUC score.

Model	Accuracy	Precision	Recall	F1 score	M-AUC	m-AUC
Logistic Regression	0.72	0.67	0.72	0.67	0.59	0.79
AdaBoost	0.66	0.44	0.66	0.53	0.50	0.75
K-Nearest Neighbors	0.67	0.61	0.67	0.63	0.56	0.76
Decision Tree	0.68	0.63	0.68	0.65	0.60	0.76
Random Forest	0.73	0.72	0.73	0.66	0.58	0.80
Multilayer Perceptrons	0.71	0.66	0.71	0.66	0.58	0.79
Gaussian NB	0.52	0.67	0.52	0.56	0.61	0.64
Quadratic Discriminant Analysis	0.57	0.55	0.57	0.56	0.54	0.68
Gaussian Process	0.61	0.60	0.61	0.60	0.57	0.71
XGBoost	0.74	0.69	0.74	0.69	0.61	0.80
Our approach	**0.87**	**0.81**	**0.87**	**0.84**	**0.97**	**0.98**

Table 3: Results obtained by human annotators, with Test 1 to Test 10 representing an anonymized individual annotator.

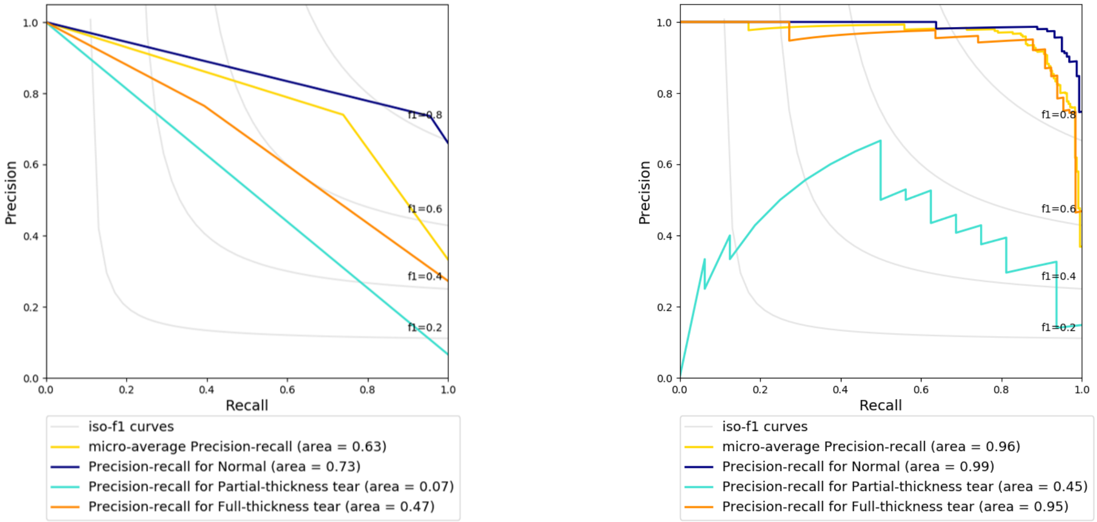
	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Average
Accuracy	78%	70%	74%	78%	83%	75%	80%	69%	67%	85%	76%
M-AUC	82%	76%	81%	81%	86%	81%	84%	72%	74%	84%	80%
Recall	92%	91%	99%	91%	92%	99%	96%	78%	92%	80%	91%
Specificity	71%	61%	63%	72%	80%	64%	72%	65%	56%	88%	69%
Precision	59%	51%	54%	59%	67%	55%	61%	50%	48%	75%	58%
F1 score	72%	65%	70%	72%	77%	71%	74%	61%	63%	77%	70%



(*a*) ROC curves

(*b*) Confusion matrix

Figure 4: ROC curves and confusion matrix for our approach.



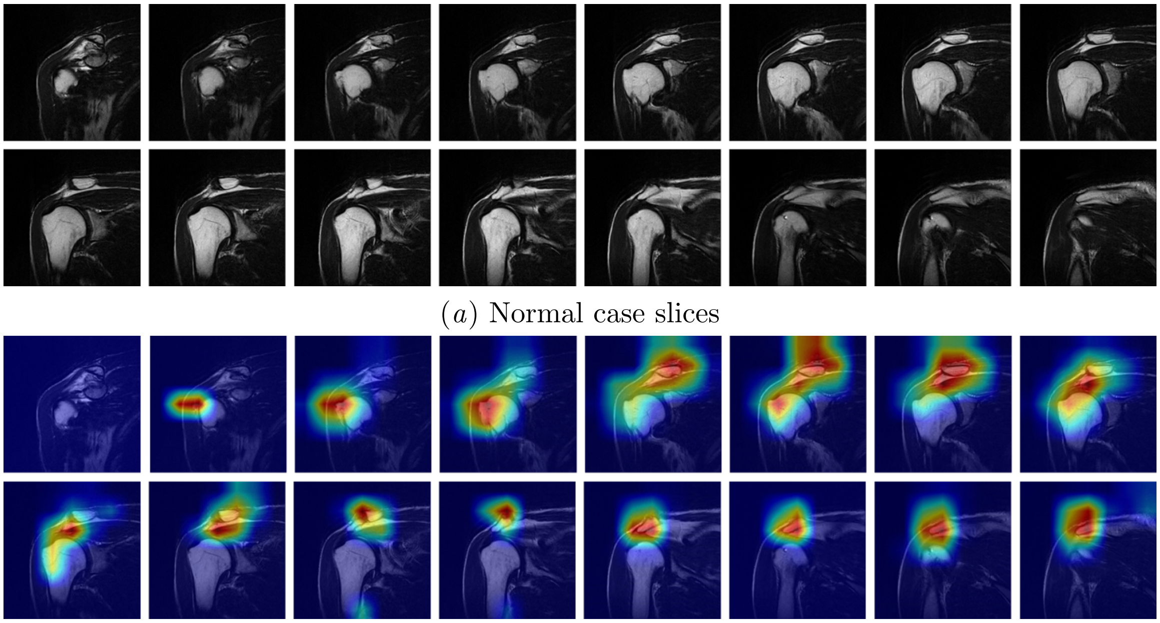
(*a*) PR curves for XGBoost

(*b*) PR curves for our approach

Figure 5: PR curves for XGBoost and our approach.

Table 3 shows that the diagnosis accuracy of the human annotators is 76% on average, with the individual values varying between 67% to 85%.

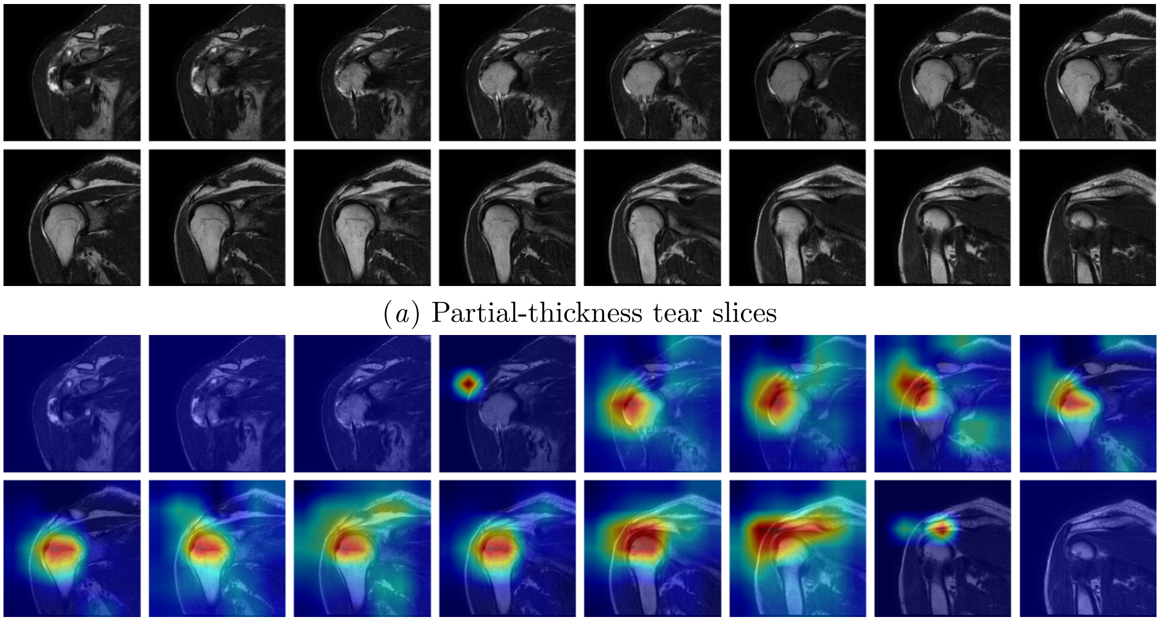
Table 2 shows that our approach towards RCT diagnosis achieved an accuracy of 87%, outperforming the accuracy of the different baselines by at least 13%. The accuracy values achieved by the conventional machine learning approaches are even lower than the accuracy values achieved by the human annotators. Even though the wavelet transform feature extractor can extract spatial information from the given data, hereby achieving good accuracy



(a) Normal case slices

(b) Grad-CAM output

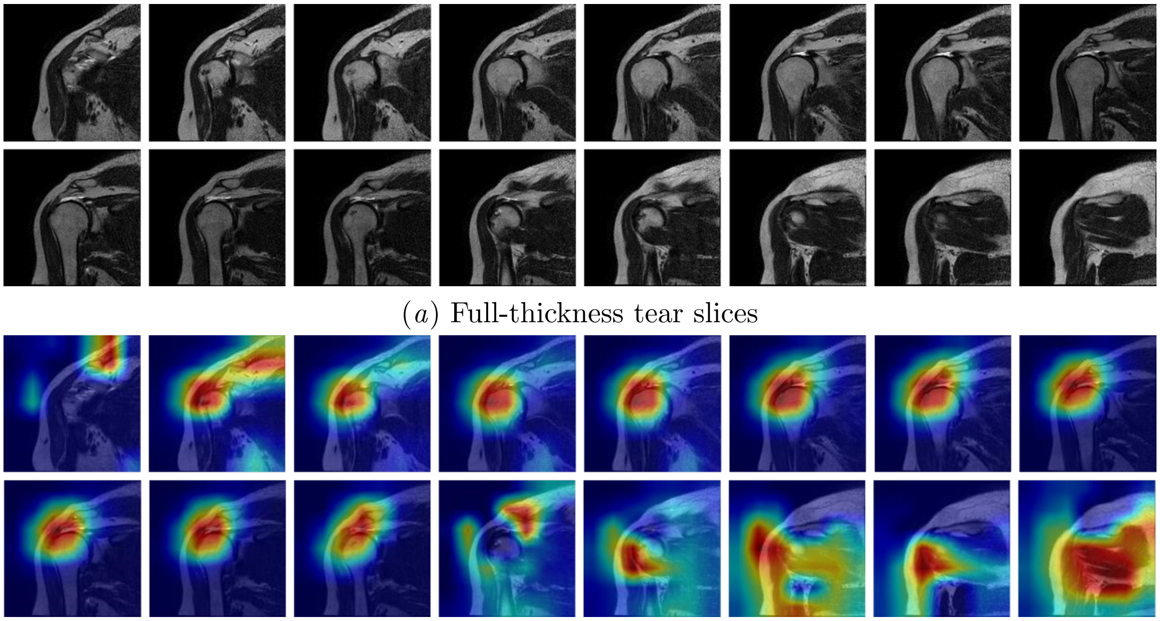
Figure 6: Grad-CAM results for normal case slices.



(a) Partial-thickness tear slices

(b) Grad-CAM output

Figure 7: Grad-CAM results for partial-thickness tear slices.



(*a*) Full-thickness tear slices

(*b*) Grad-CAM output

Figure 8: Grad-CAM results for full-thickness tear slices.

values in the context of brain image analysis, this approach is not able to extract effective features from the given MRI scans.

In terms of model robustness, our approach outperforms all baselines in terms of all metrics. In particular, our approach is able to obtain the highest m-AUC score (98%) and, as shown in Figure 4a, the highest specificity value, thus demonstrating a high ability to correctly identify patients who are not suffering from RCTs.

Lastly, to gain insight into the impact of class imbalance, we plotted the PR curves for two representative models in Figure 5. Compared to the PR curve of XGBoost, the PR curve of our approach is better for partial-thickness tears (by 38%). Furthermore, the confusion matrix presented in Figure 4b shows that our approach obtains an accuracy of about 38% for the partial-thickness tear class in the presence of a high class imbalance. However, it is also clear that improving the effectiveness of the proposed approach in the presence of a high class imbalance remains a future work item.

5. Discussion

The most important finding of our study is that, for computer-aided diagnosis of RCTs in 3-D MRI scans, the use of a CNN with a weighted linear combination layer is able to produce a higher diagnostic accuracy than human annotators. In addition, the approach proposed in this study is able to make a distinction between low-grade partial-thickness RCTs (with a thickness of less than 50%), high-grade partial-thickness RCTs (with a thickness of more than 50%), and full-thickness RCTs. Moreover, our approach is able to localize the pathological lesions related to RCTs, such as the footprint at the musculotendinous junction, as highlighted in Figure 6, Figure 7, and Figure 8 using Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017). Based on the outcome of this study, it should be possible to develop a segmentation technique that is able to automatically detect rotator cuff lesions. Furthermore, it may be helpful to construct a model for 3-D reconstruction of the shape of an RCT.

The ground truth for the raw RCT MRI dataset was created by a single orthopedic surgeon who has more than 10 years of experience as a shoulder specialist, with reviewing taking place two times, using an interval of two weeks in-between. Moreover, the RCT diagnoses were confirmed through MRI diagnosis by a musculoskeletal radiologist. In this context, it is worth mentioning that it may be more accurate to confirm the RCT diagnoses by arthroscopic findings. However, patients with low-grade partial-thickness RCTs were treated conservatively, and as such, verification was done through MRI diagnosis by a radiologist.

The dataset used in this study comes with a skewed distribution, with the number of normal MRI scans being much higher than the number of MRI scans having partial- or full-thickness tears. Indeed, the raw RCT MRI dataset was obtained in a university hospital that uses MRI as a screening test. Due to many refractory patients who failed conservative treatments, there may be many normal rotator cuff patients, compared to the number of RCT patients. Moreover, patients with Bankart lesions, SLAP tears, or recurrent dislocation were classified as normal control patients if the rotator cuff did not have any pathological lesions, leading to a further increase in the number of normal patients. However, the bias introduced by the skewed distribution may be minimal, given that this study did not focus on the size or shape of RCTs.

Limitations In this study, a comparison was made using testing set between the diagnostic values obtained by orthopedic residents and the diagnostic values obtained by the newly proposed computational model, and where the orthopedic residents cannot be considered experts yet. This is different from other studies, which typically compare the diagnostic values between radiologists and computer-aided models. Therefore, the diagnostic values obtained by the human annotators in our study were relatively lower than the ones presented in other studies (Dinnes et al., 2003; Lenza et al., 2013; Liu et al., 2020). This hints at the difficulty of diagnosing RCTs in 3-D MRI scans when someone is not specialized for the shoulder. As such, we believe that the predictive model proposed in this study could be used as a supplementary screening tool to overcome diagnostic difficulties in a more general clinical setting. Finally, we would like to point out that our model for RCT diagnosis has not been validated yet by making use of an external RCT dataset, given that such a dataset is currently not publicly available.

Future Work In future research, we plan to improve the diagnostic accuracy of the proposed approach, particularly focusing on addressing class imbalance issues. Furthermore, we plan to develop a model that gives insight into where a tear can be found, as well as its approximate shape and size, so to enable more effective shoulder surgery.

Acknowledgments

The research efforts described in this paper were funded by Ghent University, Ghent University Global Campus, Flanders Innovation & Entrepreneurship (VLAIO), the Fund for Scientific Research-Flanders (FWO-Flanders), the European Union, and the Korean government (MEST) (No. 2019R1F1A1063714).

References

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Beth G Ashinsky, Mustapha Bouhrara, Christopher E Coletta, Benoit Lehallier, Kenneth L Urish, Ping-Chang Lin, Ilya G Goldberg, and Richard G Spencer. Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative. *Journal of Orthopaedic Research*, 35 (10):2243–2250, 2017.

Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS medicine*, 15(11):e1002699, 2018.

Pierre-Henri Conze, Sylvain Brochard, Valérie Burdin, Frances T Sheehan, and Christelle Pons. Healthy versus pathological learning transferability in shoulder muscle MRI segmentation using deep convolutional encoder-decoders. *arXiv preprint arXiv:1901.01620*, 2019.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

J Dinnes, E Loveman, L McIntyre, and N Waugh. The effectiveness of diagnostic tests for the assessment of shoulder pain due to soft tissue disorders: a systematic review. In *NIHR Health Technology Assessment programme: Executive Summaries*. NIHR Journals Library, 2003.

Ravikumar Gurusamy and Vijayan Subramaniam. A machine learning approach for MRI brain tumor classification. *Computers, Materials & Continua*, 53(2):91–108, 2017.

J Christoph Katthagen, Gabriella Bucci, Gilbert Moatshe, Dimitri S Tahal, and Peter J Millett. Improved outcomes with arthroscopic repair of partial-thickness rotator cuff tears: a systematic review. *Knee Surgery, Sports Traumatology, Arthroscopy*, 26(1):113–124, 2018.

Joo Young Kim, Kyunghan Ro, Sungmin You, Bo Rum Nam, Sunhyun Yook, Hee Seol Park, Jae Chul Yoo, Eunkyoung Park, Kyeongwon Cho, Baek Hwan Cho, et al. Development of an automatic muscle atrophy measuring algorithm to calculate the ratio of supraspinatus in supraspinous fossa using deep learning. *Computer methods and programs in biomedicine*, 182:105063, 2019.

Yang-Soo Kim, Sung-Eun Kim, Sung-Ho Bae, Hyo-Jin Lee, Won-Hee Jee, and Chang Kyun Park. Tear progression of symptomatic full-thickness and partial-thickness rotator cuff tears as measured by repeated MRI. *Knee Surgery, Sports Traumatology, Arthroscopy*, 25(7):2073–2080, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Mario Lenza, Rachelle Buchbinder, Yemisi Takwoingi, Renea V Johnston, Nigel CA Hanchard, and Flavio Faloppa. Magnetic resonance imaging, magnetic resonance arthrography and ultrasonography for assessing rotator cuff tears in people with shoulder pain for whom surgery is being considered. *Cochrane Database of Systematic Reviews*, (9), 2013.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Fang Liu, Zhaoye Zhou, Alexey Samsonov, Donna Blankenbaker, Will Larison, Andrew Kanarek, Kevin Lian, Shivkumar Kambhampati, and Richard Kijowski. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. *Radiology*, 289(1):160–169, 2018.

Fanxiao Liu, Jinlei Dong, Wun-Jer Shen, Qinglin Kang, Dongsheng Zhou, and Fei Xiong. Detecting Rotator Cuff Tears: A Network Meta-analysis of 144 Diagnostic Studies. *Orthopaedic Journal of Sports Medicine*, 8(2):2325967119900356, 2020.

Yunpeng Liu, Renfang Wang, Ran Jin, Dechao Sun, Huixia Xu, and Chen Dong. Shoulder Joint Image Segmentation Based on Joint Convolutional Neural Networks. In *Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence*, pages 236–241, 2019.

Deepak Ranjan Nayak, Ratnakar Dash, and Banshidhar Majhi. Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests. *Neurocomputing*, 177:188–197, 2016.

Leonardo Osti, Matteo Buda, Mattia Andreotti, Raffaella Osti, Leo Massari, and Nicola Maffulli. Transtendon repair in partial articular supraspinatus tendon tear. *British medical bulletin*, 123(1):19–34, 2017.

V Roblot, Y Giret, M Bou Antoun, C Morillot, X Chassin, A Cotten, J Zerbib, and L Fournier. Artificial intelligence to diagnose meniscus tears on MRI. *Diagnostic and interventional imaging*, 100(4):243–249, 2019.

Robert E Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Lior Shamir, Nikita Orlov, D Mark Eckley, Tomasz Macura, Josiah Johnston, and Ilya G Goldberg. Wndchrm–an open source utility for biological image analysis. *Source code for biology and medicine*, 3(1):13, 2008.

Gururaj Sharma, Sudarshan Bhandary, Ganesh Khandige, and Utkarsh Kabra. MR imaging of rotator cuff tears: correlation with arthroscopy. *Journal of clinical and diagnostic research: JCDR*, 11(5):TC24, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Khalid Usman and Kashif Rajpoot. Brain tumor classification from multi-modality MRI using wavelets and machine learning. *Pattern Analysis and Applications*, 20(3):871–881, 2017.

Bing Wang and Tuan D Pham. MRI-based age prediction using hidden Markov models. *Journal of neuroscience methods*, 199(1):140–145, 2011.