Skeletal Radiology (2021) 50:683–692
https://doi.org/10.1007/s00256-020-03599-2

SCIENTIFIC ARTICLE

Deep learning method for segmentation of rotator cuff muscles on MR images

Giovanna Medina[1] · Colleen G. Buckless[2] · Eamon Thomasson[2] · Luke S. Oh[1] · Martin Torriani[2]

Received: 15 June 2020 / Revised: 27 August 2020 / Accepted: 3 September 2020 / Published online: 16 September 2020

Abstract

Objective To develop and validate a deep convolutional neural network (CNN) method capable of (1) selecting a specific shoulder sagittal MR image (Y-view) and (2) automatically segmenting rotator cuff (RC) muscles on a Y-view. We hypothesized a CNN approach can accurately perform both tasks compared with manual reference standards.

Material and methods We created 2 models: model A for Y-view selection and model B for muscle segmentation. For model A, we manually selected shoulder sagittal T1 Y-views from 258 cases as ground truth to train a classification CNN (Keras/Tensorflow, Inception v3, 16 batch, 100 epochs, dropout 0.2, learning rate 0.001, RMSprop). A top-3 success rate evaluated model A on 100 internal and 50 external test cases. For model B, we manually segmented subscapularis, supraspinatus, and infraspinatus/teres minor on 1048 sagittal T1 Y-views. After histogram equalization and data augmentation, the model was trained from scratch (U-Net, 8 batch, 50 epochs, dropout 0.25, learning rate 0.0001, softmax). Dice (F1) score determined segmentation accuracy on 105 internal and 50 external test images.

Results Model A showed top-3 accuracy > 98% to select an appropriate Y-view. Model B produced accurate RC muscle segmentations with mean Dice scores > 0.93. Individual muscle Dice scores on internal/external datasets were as follows: subscapularis 0.96/0.93, supraspinatus 0.97/0.96, and infraspinatus/teres minor 0.97/0.95.

Conclusions Our results show overall accurate Y-view selection and automated RC muscle segmentation using a combination of deep CNN algorithms.

Keywords Shoulder · Muscles · Atrophy · MRI · Artificial intelligence · Segmentation · Rotator cuff

Introduction

Rotator cuff (RC) tendon tears are associated with varied degrees of muscle atrophy, manifested by decreased muscle bulk and fatty infiltration [1, 2]. Atrophy of RC musculature is linked to higher rates of repair failure and overall worse clinical outcomes [3–6]. MRI is the reference standard for imaging RC tendons for tears, severity of cuff abnormalities, and postoperative healing [7, 8]. Further, MRI is the preferred method to evaluate RC muscles, enabling quantification and longitudinal assessment of atrophy [2, 9, 10]. Fatty infiltration and degree of atrophy of the supraspinatus muscle have received most attention in studies correlating surgical decision-making and prognostic factors [5, 10, 11]. Importantly, tears of subscapularis and infraspinatus tendons may also occur, and atrophy of their muscles also carries important implications to functionality and postoperative outcome [4].

Although multiple approaches have been described for estimation of RC muscle atrophy, they are qualitative or semi-quantitative, with limitations in their reproducibility [1, 9, 10]. On the other hand, quantitative methods require accurate manual or semi-automated segmentation strategies that are time-consuming and may exhibit variability across operators [12–14]. Further, some of these techniques require water-fat separation sequences, which are not typically included in routine shoulder examinations [14, 15]. Consequently, their adoption for clinical management is limited, highlighting the need for reliable automated methodologies.

Giovanna Medina and Colleen G. Buckless contributed equally to this work.

✉ Martin Torriani
mtorriani@mgh.harvard.edu

[1] Department of Orthopedics, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

[2] Division of Musculoskeletal Imaging and Intervention, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street – YAW 6048, Boston, MA 02114, USA

Recently, automated segmentation of shoulder MRI images using deep learning techniques has been limited to supraspinatus muscle [16] and shoulder girdle muscles in birth-related brachial plexus palsy [17]. Previously, deep learning methods for segmentation of other muscles have also been reported on ultrasound [18] and MRI [19]. However, prior studies have not evaluated an automated workflow to select a specific shoulder image for segmentation and generate cross-sectional areas of multiple RC muscles. The purpose of our study was to develop deep convolutional neural networks (CNN) to identify a scapular Y-view (hereafter referred to as Y-view) from a routine sagittal T1-weighted shoulder MRI and another CNN to segment subscapularis, supraspinatus, and infraspinatus/teres minor muscles on a Y-view. We hypothesized that Y-view selection using the Inception v3 architecture [20] and multi-class segmentation using a modified U-net [21] would achieve high accuracy as compared with a reference standard of manual Y-view selection and manual muscle segmentation, respectively.

Materials and methods

Our study was IRB-approved and complied with Health Insurance Portability and Accountability Act (HIPAA) guidelines with exemption status for individual informed consent. MRI examinations obtained between October 2018 and January 2020 were collected retrospectively, regardless of indication. The shoulder MRIs were performed using 1.16-T, 1.5-T, and 3.0-T scanners (General Electric, Waukesha, WI, USA; Hitachi Medical Corporation, Tokyo, Japan; Siemens Healthcare, Erlangen, Germany; Philips Healthcare, Amsterdam, Netherlands) within our institution (hereafter referred to as "internal"). In addition, we separately collected shoulder MRIs performed at non-affiliated imaging facilities that were uploaded to our hospital's database for clinical consultation ("external").

Shoulder MRIs were obtained with the patient in supine position, head first, and using a dedicated shoulder coil. The field-of-view was adapted to the patient's body habitus. Only T1-weighted sagittal images were used for our study. They were prescribed parallel to the glenoid articular surface, with standard acquisition parameters: repetition time (TR) 400–775 ms, echo time (TE) 8–25 ms, field-of-view (FOV) 140–180 cm, number of excitations (NEX) 0.5–3, bandwidth 61–325 Hz, slice thickness 3–4.5 mm, and inter-slice gap 20–25% of slice thickness.

We defined the Y-view as the most lateral image showing contact between scapular spine and posterior glenoid, forming a Y-letter shape [3, 14] (Fig. 1a). This image was used as it is recognizable, provides a representative cross section of RC muscles, and has been used in previous studies [3, 14, 16, 17].

No cases had intra-articular or intra-venous contrast injection. One-hundred and ninety scans were excluded due to the following: motion artifacts that severely degraded anatomic detail (N = 42), inadequate field-of-view for RC muscles (N = 97), and inadequate slice coverage of T1 sagittal images not including a proper Y-view (N = 51).

Two models were developed:

Model A (classifier), for Y-view selection; and,
Model B (segmentation), for RC muscle segmentation at a Y-view.

Model A (Y-view selection)

Ground truth labeling

T1-weighted sagittal shoulder MRI images were grouped into 3 anatomical zones in order to balance the classification task (Fig. 1b):

- *Zone 1*, from most lateral image to lateral acromioclavicular (AC) joint;
- *Zone 2*, from AC joint up to Y-view; and
- *Zone 3*, from Y-view to most medial image.

The AC joint and Y-view were selected as boundaries for each anatomical zone for being easily identifiable. These 3 zones were the ground truth labels for their respective images. An equal number of images in each zone was used to create model A.

To better characterize model A's cohort, two investigators classified Y-view images as either normal or pathologic, using the Goutallier grading system modified by Fuchs et al. [2]: *normal*, grade 1 images; *pathologic*, grades 2 and 3 (moderate and advanced fatty infiltration). Images were scored by consensus of two investigators with 6 years of medical image analysis and 10 years of clinical orthopedic experience. This data was not used for model development.

Training and testing

Contrast limited adaptive histogram equalization (CLAHE) was performed on all grayscale images and saved as Tag Image File Format (TIFF) files. For this classification task, we used the GoogLeNet Inception v3 CNN architecture, which was developed by Szegedy et al. [20]. Briefly, this architecture comprises 42 layers, incorporating three varieties of Inception modules that help reduce computation time relative to other architectures [20]. Input images were 299 × 299 pixels and 8-bit 3-channel grayscale. All images were normalized to the training dataset mean and standard deviation. Model A was trained using Python 3.6 (Python Software

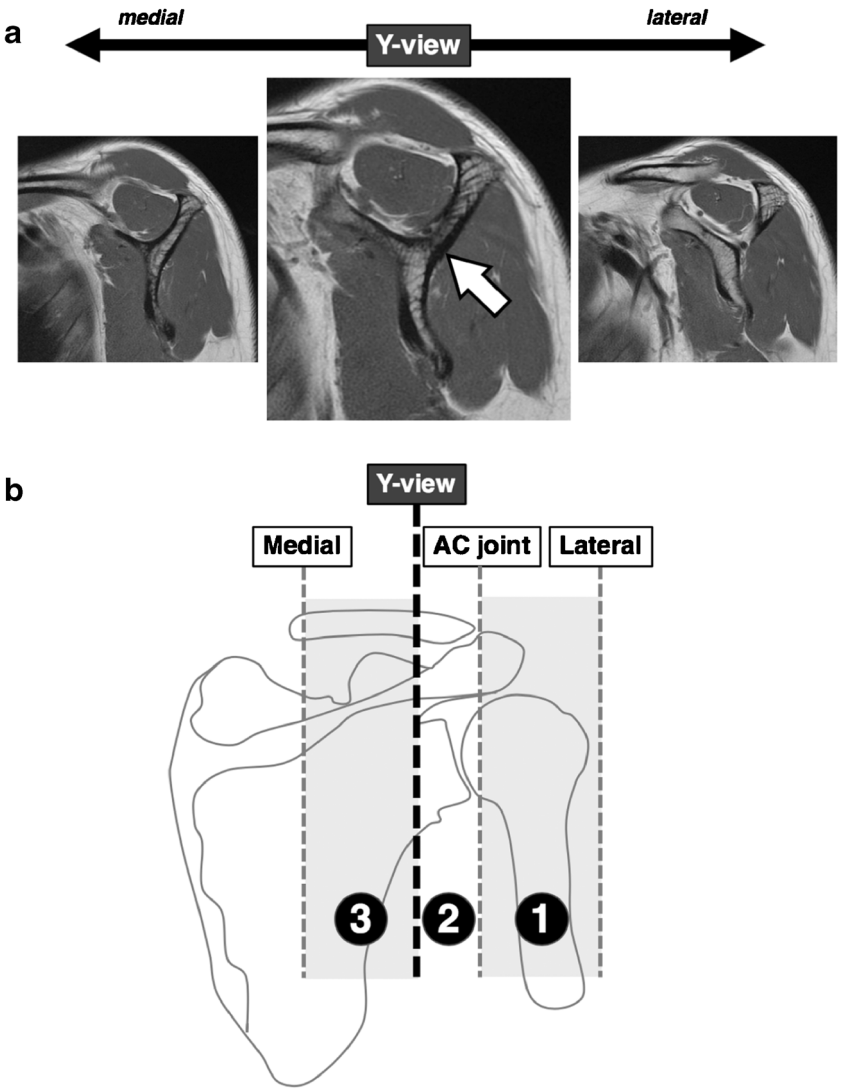

Fig. 1 **a** Definition of sagittal Y-view: we used the most lateral image that showed contact between scapular spine and posterior glenoid (arrow), forming a Y-letter shape (Refs. [3, 14]). **b** Grouping of sagittal images in 3 anatomic zones (1, 2, and 3) that served as ground truth labels for model A training and classification

Foundation, Beaverton, OR) and the Keras library (v2.2.4, https://keras.io) with Tensorflow 1.13.1 (Google, Mountain View, CA) backend [22]. The training dataset was split as 80% training and 20% validation images. Inline image augmentation was performed using the Keras built-in image generator, including rotation, magnification, cropping, horizontal flipping, and horizontal/vertical shifting. Batch size was 16 and we used the RMSprop optimizer (learning rate, 0.001; rho = 0.9). The model trained for 100 epochs on a Linux workstation (Ubuntu 14.04) with 4 NVIDIA Titan Xp Graphic Processing Units. We ran the training procedure in triplicate to produce 3 versions of model A and generate an average testing performance.

Each version of model A was tested on sagittal T1-weighted series of internal and external MRI studies. Each full sagittal T1-weighted series was examined individually, in which its images were sequentially tested to predict anatomic zone assignment. The most lateral slice predicted as *zone 3* was considered the model's prediction for a Y-view. Additionally, one slice immediately lateral and one slice immediately medial to the predicted Y-view were added, yielding a 3-image prediction. The 3-image prediction was considered accurate if one of its images matched the ground truth Y-view (Fig. 2).

Model B (muscle segmentation)

Ground truth labeling

Manual segmentation was performed using the Horos DICOM viewer (version 6.5.2, www.horosproject.com) by a single operator with 10 years of clinical experience, with all images and segmentations inspected by a second investigator with 23 years of clinical experience. As shown in Fig. 3, examinations were segmented manually into 4 classes, as follows: (1) background pixels (all pixels outside RC muscles; black); (2) subscapularis (blue); (3) supraspinatus (red); (4) infraspinatus/teres minor (yellow).

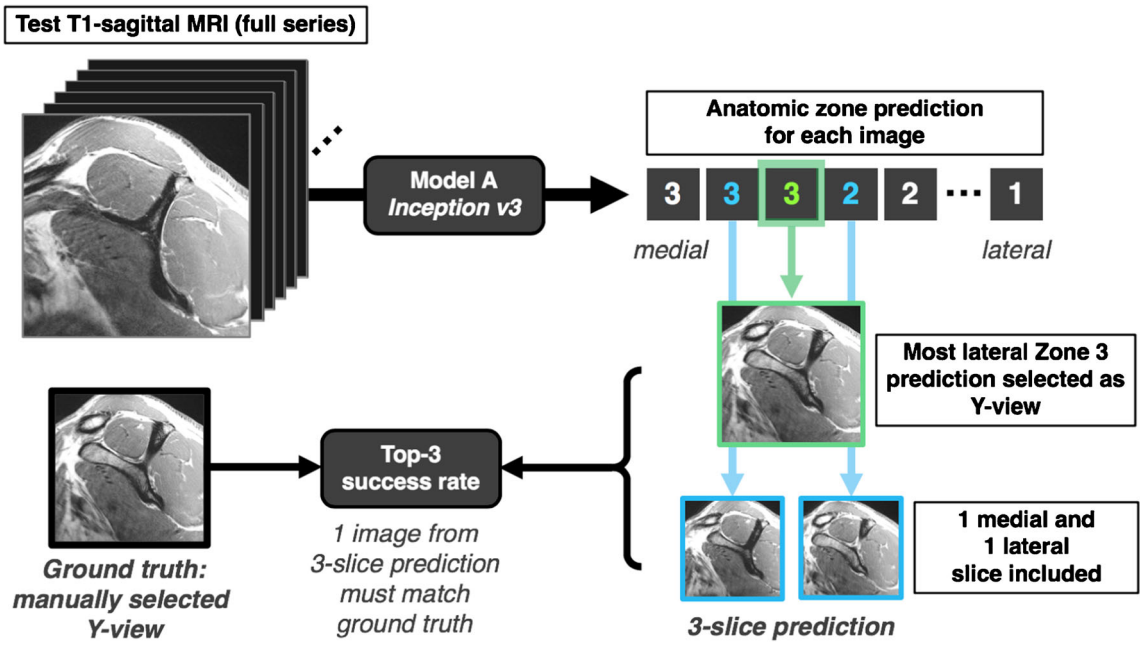

Fig. 2 Workflow for testing of model A. Images from a test T1 sagittal series were sequentially exposed to model A. The most lateral zone 3 prediction was considered the model's choice for a Y-view. One medial and one lateral adjacent images were combined to yield a 3-slice prediction, which was compared with the ground truth Y-view

Muscle segmentations comprised all pixels within the muscle boundary, including intramuscular fatty septae. If a muscle had severe fatty replacement, its fascial boundaries were traced, rather than delineating individual preserved fibers. The teres minor muscle was not individually segmented given its limited surgical importance and poorly defined boundaries with infraspinatus muscle. All segmented images were anonymized TIFF, with color ground truth masks being 8-bit 3-channel RGB. Corresponding grayscale images were 8-bit single-channel. All images were resized to 384 × 384 pixels.

To better characterize model B's cohort and understand potential biases, images were classified for muscle status as done for model A. This information was not used for model development.

Training and testing

The training dataset was split as 80% training and 20% validation images. CLAHE was performed on all grayscale images, which were then saved as JPEG files, followed by image augmentation of training dataset applying random rotation, horizontal flipping, cropping, and scaling, to achieve a total of N = 10,000. To further increase variability and improve generalizability, we applied Poisson noise to randomly selected 50% of augmented grayscale images. All images were normalized to the training dataset mean and standard deviation.

Model B employed a modified U-Net CNN architecture [21]. Briefly, images were input in our U-Net structure that consisted of five layers with four down-sampling steps

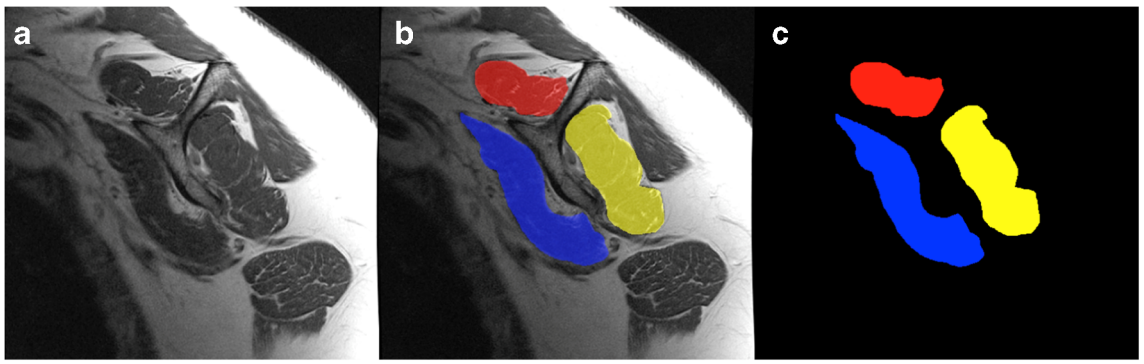

Fig. 3 Example of manual ground truth segmentation. The sagittal T1-weighted at the scapular Y-view (**a**) is manually segmented into multiple muscles (**b**) resulting in a mask (**c**) with four classes: background pixels (all pixels outside RC muscles; *black*), subscapularis (*blue*), supraspinatus (*red*), and infraspinatus/teres minor (*yellow*)

followed by four up-sampling steps. Each step consisted of two successive 3 × 3 padded convolutions, and in the down-sizing steps, a dropout of 0.25 was applied. This was followed by a rectified linear unit (ReLU) activation function and a max-pooling operation with a 2 × 2 pixel kernel size. The up-sampling operations were performed using a 2 × 2 transposed convolution followed by a 3 × 3 filter size convolution, after which the output concatenates with the corresponding decoding step. The final layer consisted of a 1 × 1 convolution followed by a sigmoid function, resulting in an output pixelwise prediction score for each class.

Model B was trained in using the Python/Keras/Tensorflow stack as previously described. During training, the 4 classes were adjusted for imbalances by weighting prevalence, which penalized predictions of classes with highest number of pixels (e.g., background). Batch size was 8 and we used the Adadelta optimizer (learning rate, 0.0001). The model trained for 25 epochs with early stopping enabled. Multi-class Dice loss was used as cost function. Training was performed on a Linux workstation (Ubuntu 14.04) with 4 NVIDIA Titan Xp Graphic Processing Units. We ran the training procedure in triplicate to produce 3 versions of model B and generate an average testing performance.

Model B was tested on internal and external Y-view images to output predictions in 4 classes (background, subscapularis, supraspinatus, and infraspinatus/teres minor) that were compared with manual segmentations using Dice (F1) score [23].

Statistical analysis

Descriptive statistics are reported in terms of percentages and means ± standard deviations (SD). For model A, a top-3 success rate was used to evaluate performance. The top-3 success rate was determined by comparing the manually selected ground truth Y-view to the 3-image prediction. The 3-image prediction was considered accurate if one of its images matched the ground truth Y-view. For model B, the Dice (F1) score was used to assess similarity between the manual segmentations and the CNN predicted segmentations [23]. A Dice score of 1.00 is a perfect similarity. We also obtained mean precision (positive predictive value) and mean recall (sensitivity) for model B tests.

Results

Model A

Model A was trained on 258 scans (N = 4320 images) from patients with mean age 56.2 ± 14.3 years. Each of 3 shoulder anatomic zones was represented by an equal number of images (N = 1440 images per zone). Model A was tested on 100 internal scans (N = 3197 images; mean age, 56.0 ± 15.0 years) and separately on 50 external scans (N = 1205 images; mean age, 55.0 ± 17.2 years). Cohort characteristics regarding RC muscle status in training and test datasets for model A are outlined in Table 1. Overall, the subscapularis muscle was more frequently normal, followed by supraspinatus and infraspinatus/teres minor.

Training took 1 h 20 min per run (training was run 3 times). Mean top-3 success rates to detect a proper Y-view were 98.7 ± 1.0% (internal test dataset) and 99.7 ± 1.0% (external). The few errors observed were due to the predicted Y-view being 2 slices apart from the manually determined Y-view. Mean top-1 success rates to detect the singular ground truth Y-view were 80.0 ± 3.0% (internal) and 91.0 ± 3.0% (external). On our workstation, detecting a Y-view took 1.8 s per test scan (each scan contained a full T1 sagittal series).

Table 1 Cohort characteristics: rotator cuff muscle status in training, internal testing, and external testing datasets

		Subscapularis	Supraspinatus	Infraspinatus
Model A				
Training (258 scans)	Normal	239 (92.6%)	229 (88.8%)	199 (77.1%)
	Pathologic	19 (7.4%)	29 (11.2%)	59 (22.9%)
Internal testing (100 scans)	Normal	90 (90%)	87 (87%)	82 (82%)
	Pathologic	10 (10%)	13 (13%)	18 (18%)
Model B				
Training (943 scans)	Normal	860 (91.2%)	818 (86.7%)	717 (76.0%)
	Pathologic	83 (8.8%)	125 (13.2%)	226 (24.0%)
Internal testing (105 scans)	Normal	96 (91.4%)	89 (84.8%)	77 (73.3%)
	Pathologic	9 (8.6%)	16 (15.3%)	28 (26.7%)
Model A and model B				
External testing (50 scans)	Normal	44 (88%)	40 (80%)	36 (72%)
	Pathologic	6 (12%)	10 (20%)	14 (28%)

Table 2 Mean Dice, precision, and recall scores for model B segmentation on internal and external test datasets. Values are mean ± SD from testing on models generated in 3 distinct runs

	Background	Subscapularis	Supraspinatus	Infraspinatus
Internal test dataset (105 images)				
Dice (F1 score)	0.994 ± 0.001	0.957 ± 0.001	0.965 ± 0.001	0.968 ± 0.001
Precision	0.994 ± 0.001	0.957 ± 0.001	0.971 ± 0.001	0.967 ± 0.002
Recall	0.994 ± 0.001	0.959 ± 0.002	0.961 ± 0.001	0.969 ± 0.001
External test dataset (50 images)				
Dice (F1 score)	0.989 ± 0.001	0.933 ± 0.050	0.964 ± 0.022	0.951 ± 0.033
Precision	0.985 ± 0.001	0.963 ± 0.019	0.975 ± 0.009	0.959 ± 0.026
Recall	0.994 ± 0.001	0.919 ± 0.066	0.957 ± 0.033	0.947 ± 0.041

Model B

A total of 1048 scans, from which one Y-view image was used per scan, were collected from 1030 patients (mean age of 56.1 ± 14.5 years). The images were divided into 90% training (N = 943) and 10% test (N = 105) datasets. Cohort characteristics regarding RC muscle status were similar to model A (i.e., larger proportion of normal subscapularis muscle) (Table 2). External test cases for model B were the same used to test model A.

Manual segmentations were accomplished in approximately 5–10 min per image. Training took 1 h per run (training was performed 3 times). Overall, mean muscle segmentation Dice scores for internal and external test datasets were > 0.93 and are outlined in Table 2. Figure 4 shows examples of accurate CNN muscle segmentations. Although overall accuracy was high on internal and external test datasets, minor prediction errors were seen especially along the inferior contour of the subscapularis, where teres major/latissimus dorsi and axillary artery produce challenging boundaries (Fig. 5). There was only one instance—from a total of 155 test cases—in which the model misclassified a larger muscle area (Fig. 6). On our workstation, each automated segmentation was accomplished in 0.02 s per test image.

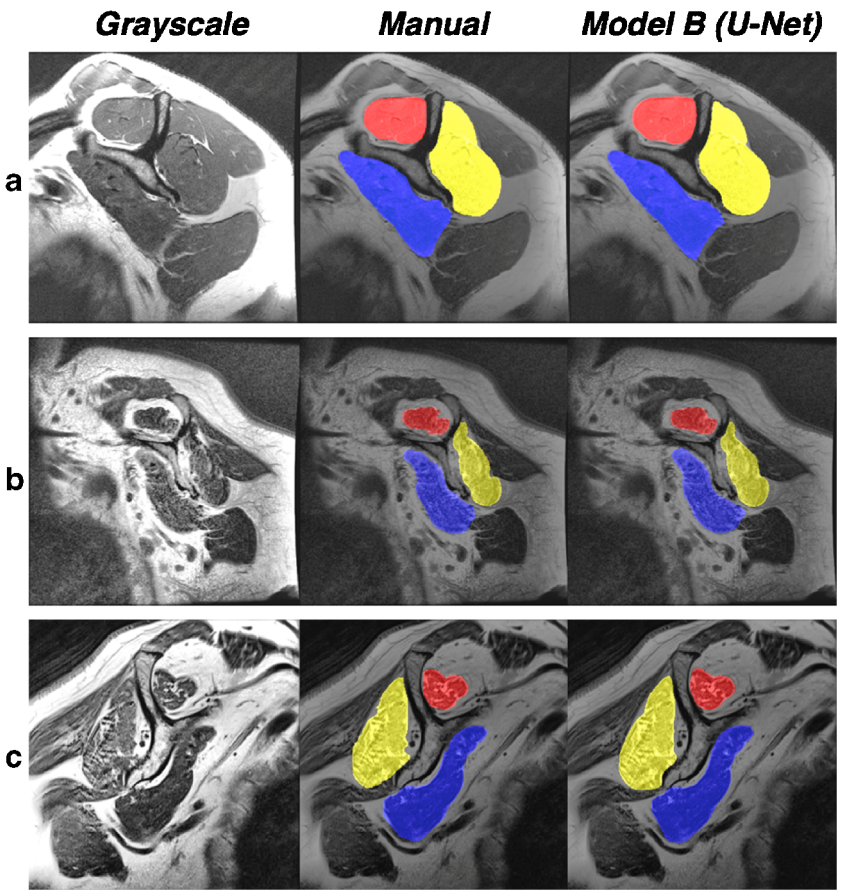

Fig. 4 Examples of accurate muscle segmentation using model B, with each row containing test images from different patients, with normal rotator cuff muscle appearance (**a**), and varied degrees of muscle atrophy and fatty infiltration (**b**, **c**). In grayscale images from all 3 cases, note the challenging boundaries between infraspinatus and teres minor. Manual, manual tracing; Model B (U-Net), model prediction by CNN

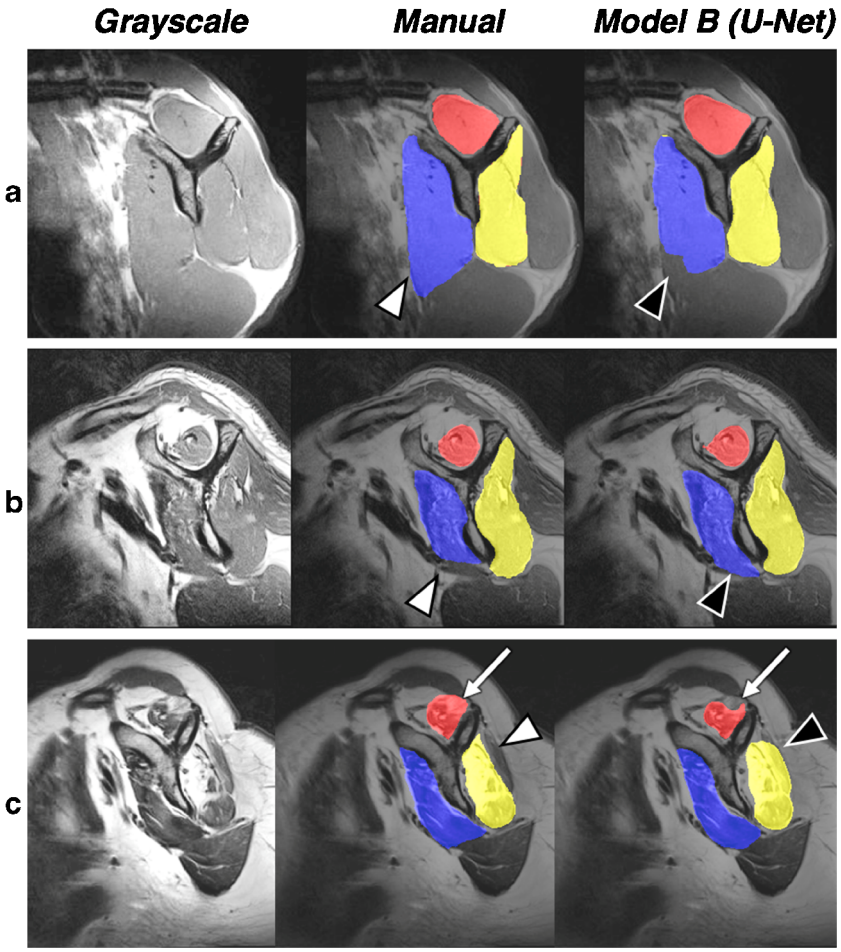

Fig. 5 Prediction errors on test images from 3 different subjects (one per row). Errors at caudal contour of subscapularis muscle (arrowheads), due to challenging muscle boundaries with teres major/latissimus dorsi (**a**) and axillary artery (**b**). **c** Underestimation of supraspinatus segmentation (arrow) and overestimation of infraspinatus/teres minor (white arrowhead), with model misclassifying portion of trapezius (black arrowhead). Manual, manual tracing; Model B (U-Net), model prediction by CNN

Discussion

The findings of our study are twofold: (1) a CNN classification method is able to accurately select an appropriate shoulder Y-view and (2) another U-Net-based CNN is able to accurately segment multiple RC muscles at that level. Importantly, our results show the feasibility of these methods in a large cohort of randomly selected shoulder MRIs obtained both inside and outside our institution.

RC muscle atrophy affects the repairability of tendons, with greater muscle atrophy predisposing higher rates of retear and unfavorable outcomes [3–6]. Prior qualitative [1, 9] and quantitative [10] studies have graded fatty infiltration and atrophy of RC muscles. A commonly used qualitative system is the Goutallier classification [1], originally described on non-contrast shoulder CT scans and later adapted for MRI [2] using Y-view T1-weighted images, yielding wide interobserver [2, 24–27] and intra-observer [24–27] reliability measures. Subsequently, the tangent sign was introduced as a binary qualitative assessment of supraspinatus muscle atrophy [9], and Thomazeau et al. [10] proposed an occupation ratio to determine supraspinatus muscle atrophy. Taken together, although these methods provide insight into RC muscle status, their limitations include subjective prescription of anatomic landmarks and manual tracings, which are time-consuming and present variability across operators [12–14]. These factors represent drawbacks for broader implementation of quantitative RC muscle measures in clinical care.

In our study, we focused on the Y-view for its familiar bony landmarks, good representation of RC muscle status, and frequent use in RC muscle atrophy studies [3, 14]. Automatic slice selection methods have been previously described to identify anatomical landmarks using atlas-based approaches and deep learning [28–30]. For musculoskeletal applications, Zhou et al. [31] had success using a CNN to select a knee sagittal slice for anterior cruciate ligament tear classification with an accuracy of 0.98. Our results are novel in presenting accurate methods for Y-view selection that can be the initial step in a workflow for automated RC muscle segmentation at that level.

Our automated segmentation of RC muscles showed an accuracy comparable or better to other deep learning methodologies. For example, Kim et al. [16] manually selected a Y-view on 240 patients and found Dice scores of 0.95 for supraspinatus muscle and 0.97 for supraspinatus fossa.

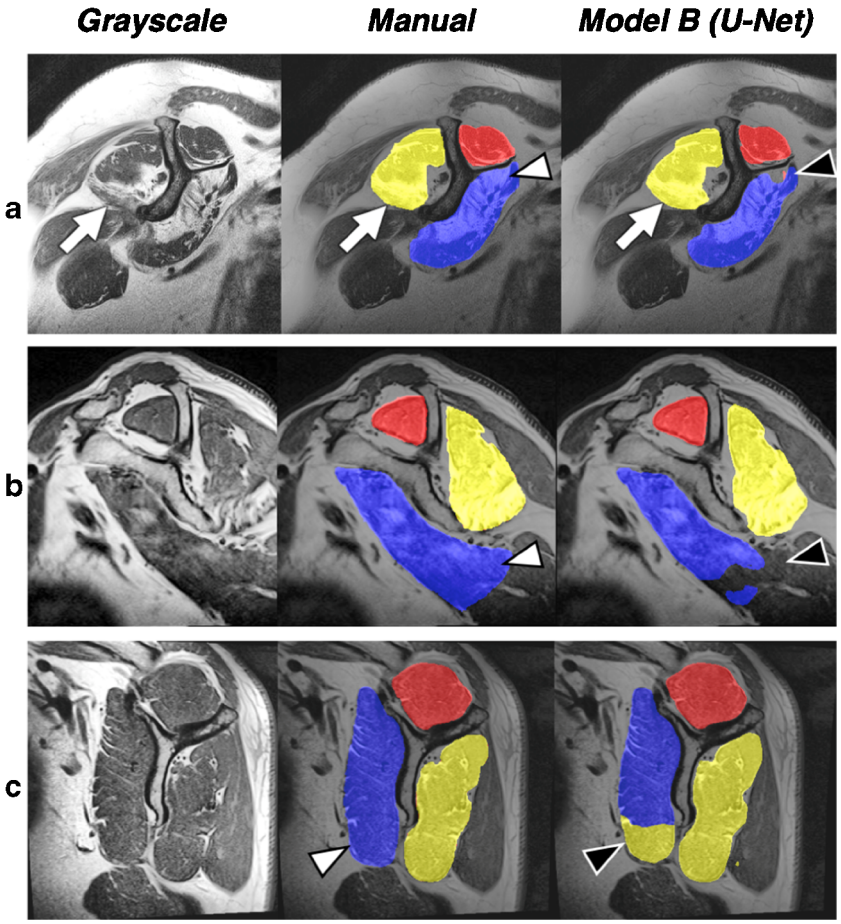

Fig. 6 Prediction errors on test images from 3 different subjects (one subject per row). **a** Although most atrophied subscapularis (white arrowhead) was correctly predicted by the model, its cranial portion was missed and a stray incorrect prediction is noted in this area (red pixels, black arrowhead). Note the correct inclusion of atrophied teres minor (white arrows) in infraspinatus/teres minor segmentation (yellow). **b** Underestimation of subscapularis segmentation (arrowheads) in area of pulsation artifact from axillary artery. **c** Rare example of model uncertainty assigning an area of subscapularis as belonging to infraspinatus/teres minor class (arrowheads). Manual, manual tracing; Model B (U-Net), model prediction by CNN

Similarly, Conze et al. [17] used a CNN to obtain volumetric segmentations from 24 pediatric healthy and pathological shoulder exams finding Dice scores of 0.71, 0.83, and 0.82 for supraspinatus, subscapularis, and infraspinatus, respectively. They also noted an improved performance when images of pathological shoulders were combined with images of unaffected shoulders [17].

In our study, both models were trained and tested on datasets containing a variety of RC muscle conditions (i.e., normal, moderate, and severe atrophy). Although our accuracy and short analysis time per image for model B are promising, areas of over- and underestimation were seen. Some minor errors occurred most commonly at muscle boundaries with adjacent fat planes and likely represent low-impact quantitative issues. More prominent errors were noted along the inferior contour of subscapularis (adjacent to axillary vessels) and inferior contour of infraspinatus/teres minor. Despite high overall and per-muscle Dice scores, strategies to improve these errors should include expanding training datasets with more cases containing confounding features in those areas. Inclusion of a larger variety of supraspinatus atrophy states may also benefit segmentation performance. As noted by Kim et al. [16], a possible explanation for lower supraspinatus muscle Dice score is due to variations in cross-sectional area caused by supraspinatus tendon tears and atrophy.

Strengths of our study include successful slice selection using a classification algorithm and demonstration of accurate automated RC muscle segmentations on routine T1 sagittal MRIs from a large and varied cohort. This simple yet robust technique has not been previously described and yielded excellent results. Importantly, both our models were tested on datasets from studies obtained outside our institution, rendering comparable accuracies. The size of our training and testing dataset is another advantage as compared with prior studies [16, 17]. Further, although Conze et al. [17] examined a pediatric population, our work investigated a wider range of adult shoulder MRI images.

An important focus of our study was to separately validate methods for proper Y-view selection (model A) and accurate muscle boundary determination (model B). We did not test the performance of an integrated pipeline across both models; therefore, our top-3 performance for model A should not imply passing a 3-image dataset to model B would be used in such a workflow. An integrated pipeline, currently in development by our group, requires specific procedures and modifications to training and testing datasets that were beyond the scope of the current study. Our algorithm was not designed to

quantify the degree of atrophy in each muscle, which will require an additional stage of thresholding muscle vs. fat pixels within each segmentation. This desirable feature will be the subject of future development, which, however, relies first on robust and reliable localization of muscle boundaries, which was the key effort in our study. Our manual tracing also included fatty septae and fat replacement within the boundaries of each cross-sectional area, with the future expectation of separating muscle from fat pixels using dedicated methods. In principle, a similar procedure could be adopted to more accurately separate the infraspinatus from teres minor. Altogether, such developments may allow prompt determination of rotator cuff muscle cross-sectional area in clinical workstations, which could automatically provide overlays on specific images and data on dictation platforms.

Limitations of our study include model B being trained on a single standardized sagittal image. Volumetric (3D) muscle quantification using a CNN approach has been demonstrated in prior studies [17]. The use of 3D measures of RC muscle volume produces more accurate measures, which however require multiple slice segmentation and longer imaging time to cover the entire shoulder girdle, which is rarely accomplished in clinical practice [13]. Furthermore, previous studies have found that a single slice is appropriate for clinical assessment of fatty infiltration [13]. Another limitation is a relatively lower proportion of severely atrophic RC muscles in our datasets. Although this reflected the typical patient population at our imaging/clinical services, future work will develop models on datasets with higher degrees of fatty infiltration. For this reason, performance of our method could vary in a cohort with higher prevalence of severe RC muscle atrophy.

In conclusion, we demonstrate novel and accurate methods to select a Y-view image and segment multiple RC muscles using a combination of CNN models. Our work extends prior studies examining a larger and diverse cohort of patients. By offering automated and reliable muscle area quantification, our methods have potential use in surgical outcomes research and clinical assessment of RC pathology.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical approval All procedures performed in studies involving human participants were carried out in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Informed consent Informed consent was waived for individual participants included in the study. The study was approved by the local institutional review board (IRB) and HIPAA compliant.

References

1. Goutallier D, Postel J-M, Bernageau J, Lavau L, Voisin M-C. Fatty Muscle Degeneration in cuff ruptures: pre- and postoperative evaluation by CT scan. Clin Orthop. 1994;304:78–83.
2. Fuchs B, Weishaupt D, Zanetti M, Hodler J, Gerber C. Fatty degeneration of the muscles of the rotator cuff: assessment by computed tomography versus magnetic resonance imaging. J Shoulder Elb Surg. 1999;8(6):599–605.
3. Mellado JM, Calmet J, Olona M, Esteve C, Camins A, Pérez Del Palomar L, et al. Surgically repaired massive rotator cuff tears: MRI of tendon integrity, muscle fatty degeneration, and muscle atrophy correlated with intraoperative and clinical findings. AJR Am J Roentgenol. 2005;184(5):1456–63.
4. Jung W, Lee S, Hoon KS. The natural course of and risk factors for tear progression in conservatively treated full-thickness rotator cuff tears. J Shoulder Elb Surg. 2020;29(6):1168–1176.
5. Wieser K, Joshy J, Filli L, Kriechling P, Sutter R, Fürnstahl P, et al. Changes of supraspinatus muscle volume and fat fraction after successful or failed arthroscopic rotator cuff repair. Am J Sports Med. 2019;47(13):3080–8.
6. Kijowski R, Thurlow P, Blankenbaker D, Liu F, McGuine T, Li G, et al. Preoperative MRI shoulder findings associated with clinical outcome 1 year after rotator cuff repair. Radiology. 2019;291(3):722–9.
7. Iannotti JP, Zlatkin MB, Esterhai JL, Kressel HY, Dalinka MK, Spindler KP. Magnetic resonance imaging of the shoulder. Sensitivity, specificity, and predictive value. J Bone Joint Surg Am. 1991;73(1):17–29.
8. Lee SC, Williams D, Endo Y. The repaired rotator cuff: MRI and ultrasound evaluation. Curr Rev Musculoskelet Med. 2018;11(1):92–101.
9. Zanetti M, Gerber C, Hodler J. Quantitative assessment of the muscles of the rotator cuff with magnetic resonance imaging. Investig Radiol. 1998;33(3):163–70.
10. Thomazeau H, Rolland Y, Lucas C, Duval J-M, Langlais F. Atrophy of the supraspinatus belly. Assessment by MRI in 55 patients with rotator cuff pathology. Acta Orthop Scand. 1996;67(3):264–8.
11. Di Benedetto P, Beltrame A, Cicuto C, Battistella C, Gisonni R, Cainero V, et al. Rotator cuff tears reparability index based on preoperative MRI: our experience. Acta Bio Medica Atenei Parm. 2019;90(1-S):36–46.
12. Lehtinen J, Tingart M, Apreleva M, Zurakowski D, Palmer W, Warner J. Practical assessment of rotator cuff muscle volumes using shoulder MRI. Acta Orthop Scand. 2003;74(6):722–9.
13. Lee Y-B, Yang C-J, Li CZ, Zhuan Z, Kwon S-C, Noh K-C. Can a single sagittal magnetic resonance imaging slice represent whole fatty infiltration in chronic rotator cuff tears at the supraspinatus? Clin Orthop Surg. 2018;10(1):55.
14. Davis DL, Kesler T, Gilotra MN, Almardawi R, Hasan SA, Gullapalli RP, et al. Quantification of shoulder muscle intramuscular fatty infiltration on T1-weighted MRI: a viable alternative to the Goutallier classification system. Skelet Radiol. 2019;48(4):535–41.
15. Matsumura N, Oguro S, Okuda S, Jinzaki M, Matsumoto M, Nakamura M, et al. Quantitative assessment of fatty infiltration and muscle volume of the rotator cuff muscles using 3-dimensional 2-point Dixon magnetic resonance imaging. J Shoulder Elb Surg. 2017;26(10):e309–18.
16. Kim JY, Ro K, You S, Nam BR, Yook S, Park HS, et al. Development of an automatic muscle atrophy measuring algorithm to calculate the ratio of supraspinatus in supraspinous fossa using deep learning. Comput Methods Prog Biomed. 2019;182:105063.
17. Conze P-H, Brochard S, Burdin V, Sheehan FT, Pons C. Healthy versus pathological learning transferability in shoulder muscle MRI

segmentation using deep convolutional encoder-decoders. ArXiv190101620 Cs [Internet]. 2019 Feb 27 [cited 2020 Mar 23]; Available from: http://arxiv.org/abs/1901.01620.
18. Wang Y-W, Lee C-C, Lo C-M. Supraspinatus segmentation from shoulder ultrasound images using a multilayer self-shrinking snake. IEEE Access. 2019;7:146724–31.
19. Sezer A, Sezer HB. Capsule network-based classification of rotator cuff pathologies from MRI. Comput Electr Eng. 2019;80:106480.
20. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception architecture for computer vision. ArXiv151200567 Cs [Internet]. 2015 Dec 11 [cited 2020 Mar 31]; Available from: http://arxiv.org/abs/1512.00567.
21. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. ArXiv150504597 Cs [Internet]. 2015 May 18 [cited 2020 Mar 23]; Available from: http://arxiv.org/abs/1505.04597.
22. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. ArXiv160508695 Cs [Internet]. 2016 May 31 [cited 2020 Mar 23]; Available from: https://arxiv.org/abs/1605.08695.
23. Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297–302.
24. Oh JH, Kim SH, Choi J-A, Kim Y, Oh CH. Reliability of the grading system for fatty degeneration of rotator cuff muscles. Clin Orthop. 2010;468(6):1558–64.
25. Slabaugh MA, Friel NA, Karas V, Romeo AA, Verma NN, Cole BJ. Interobserver and intraobserver reliability of the Goutallier classification using magnetic resonance imaging: proposal of a simplified classification system to increase reliability. Am J Sports Med. 2012;40(8):1728–34.
26. Wall LB, Teefey SA, Middleton WD, Dahiya N, Steger-May K, Kim HM, et al. Diagnostic performance and reliability of ultrasonography for fatty degeneration of the rotator cuff muscles. J Bone Joint Surg Am. 2012;94(12):e83.
27. Schiefer M, Mendonça R, Magnanini MM, Fontenelle C, Pires Carvalho AC, Almeida M, et al. Intraobserver and interobserver agreement of Goutallier classification applied to magnetic resonance images. J Shoulder Elb Surg. 2015;24(8):1314–21.
28. Chang S-C, Lee Y-W, Lai Y-C, Tiu C-M, Wang H-K, Chiou H-J, et al. Automatic slice selection and diagnosis of breast strain elastography. Med Phys. 2014;41(10):102902.
29. Paknezhad M, Marchesseau S, Brown MS. Automatic basal slice detection for cardiac analysis. J Med Imaging [Internet]. 2016 Jul [cited 2020 Mar 31];3(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5028419/
30. Kanavati F, Islam S, Aboagye EO, Rockall A. Automatic L3 slice detection in 3D CT images using fully-convolutional networks. ArXiv181109244 Cs [Internet]. 2018 Nov 22 [cited 2020 Apr 2]; Available from: http://arxiv.org/abs/1811.09244
31. Zhou Z, Zhao G, Kijowski R, Liu F. Deep convolutional neural network for segmentation of knee joint anatomy. Magn Reson Med. 2018;80(6):2759–70.