European Radiology
https://doi.org/10.1007/s00330-021-08034-1

MUSCULOSKELETAL

Evaluating subscapularis tendon tears on axillary lateral radiographs using deep learning

Yusuhn Kang[1] · Dongjun Choi[1] · Kyong Joon Lee[1] · Joo Han Oh[2] · Bo Ram Kim[1] · Joong Mo Ahn[1]

Received: 23 December 2020 / Revised: 6 April 2021 / Accepted: 30 April 2021

Abstract

Objective To develop a deep learning algorithm capable of evaluating subscapularis tendon (SSC) tears based on axillary lateral shoulder radiography.

Methods A total of 2,779 axillary lateral shoulder radiographs (performed between February 2010 and December 2018) and the patients' corresponding clinical information (age, sex, dominant side, history of trauma, and degree of pain) were used to develop the deep learning algorithm. The radiographs were labeled based on arthroscopic findings, with the output being the probability of an SSC tear exceeding 50% of the tendon's thickness. The algorithm's performance was evaluated by determining the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, negative predictive value (NPV), and negative likelihood ratio (LR-) at a predefined high-sensitivity cutoff point. Two different test sets were used, with radiographs obtained between January and December 2019; Test Set 1 used arthroscopic findings as the reference standard ($n = 340$), whereas Test Set 2 used MRI findings as the reference standard ($n = 627$).

Results The AUCs were 0.83 (95% confidence interval, 0.79–0.88) and 0.82 (95% confidence interval, 0.79–0.86) for Test Sets 1 and 2, respectively. At the high-sensitivity cutoff point, the sensitivity, NPV, and LR- were 91.4%, 90.4%, and 0.21 in Test Set 1, and 90.2%, 89.5%, and 0.21 in Test Set 2, respectively. Gradient-weighted Class Activation Mapping identified the subscapularis insertion site at the lesser tuberosity as the most sensitive region.

Conclusion Our deep learning algorithm is capable of assessing SSC tears based on changes at the lesser tuberosity on axillary lateral radiographs with moderate accuracy.

Key Points

- *We have developed a deep learning algorithm capable of assessing SSC tears based on changes at the lesser tuberosity on axillary lateral radiographs and previous clinical data with moderate accuracy.*
- *Our deep learning algorithm could be used as an objective method to initially assess SSC integrity and to identify those who would and would not benefit from further investigation or treatment.*

Keywords Deep learning · Rotator cuff tear · Subscapularis · Radiography

Abbreviations

AUC	Area under the receiver operating characteristic curve
CNN	Convolutional neural network
$Cutoff_{95\%}$	Cutoff point for an expected sensitivity of 95%
$Cutoff_{optimal}$	Optimal cutoff point determined by Youden's J statistic
DICOM	Digital Imaging and Communications in Medicine
Grad-CAM	Gradient-weighted Class Activation Mapping
LR	Positive likelihood ratio
LR-	Negative likelihood ratio
NPV	Negative predictive value

✉ Yusuhn Kang
yskang0114@gmail.com

✉ Kyong Joon Lee
kjoon31@gmail.com

[1] Department of Radiology, Seoul National University Bundang Hospital, 82 Gumi-ro, 173 Beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do 13620, South Korea

[2] Department of Orthopedic Surgery, Seoul National University Bundang Hospital, 82 Gumi-ro, 173 Beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do 13620, South Korea

Published online: 20 May 2021

PPV	Positive predictive value
ReLU	Rectified linear unit
ROC	Receiver operating characteristic
SSC	Subscapularis tendon
VAS	Visual analogue scale

Introduction

The subscapularis tendon (SSC) is the largest of the rotator cuffs that occupy the subscapular fossa, acting as a major internal rotator of the shoulder. The importance of the SSC is being increasingly emphasized, as it has an important function in maintaining anterior stability and shoulder coaptation [1, 2], through the formation of transverse force couples with the posterior parts of the rotator cuffs. SSC tears affect treatment and surgical approaches, and postoperative prognosis [3, 4].

The integrity of the SSC is typically assessed using various clinical tests, including the bear-hug, belly-press, and lift-off tests, and the internal rotation lag sign. Their diagnostic value in assessing SSC tears has been evaluated in various studies [5–8], which have demonstrated high specificity (85–100%), but limited sensitivity (12–52%) [6, 7]. Moreover, clinical examinations are subject to inter-examiner variability and may be less reliable when performed by non-orthopedic specialists or inexperienced physicians.

The initial radiographic examination of patients presenting with shoulder pain typically includes anteroposterior, axillary, and supraspinatus outlet or scapular Y views [9]. Among these, the axillary lateral view most readily depicts the lesser tuberosity, the attachment site of the SSC. Several studies with small sample sizes have indicated that radiographic changes at the lesser tubercle of the humerus may be associated with SSC lesions [9–13], although the results vary.

We hypothesized that by applying deep learning to the interpretation of radiographs, an objective assessment of SSC tears could be possible. Furthermore, by appropriately adjusting the cutoff points of the probability output of the algorithm, a high sensitivity could be achieved, allowing it to be used as an adjunct to clinical tests. Thus, the purpose of this study was to develop a deep learning algorithm capable of evaluating SSC tears based on axillary lateral shoulder radiographs and clinical information.

Materials and methods

The institutional review board of Seoul National University Bundang Hospital approved this study and waived the informed consent requirement.

Dataset

From February 2010 to December 2019, 4,894 patients underwent radiography of the shoulder that included an axillary lateral view in addition to arthroscopic surgery. We excluded 1,024 patients with the following criteria: (a) age under 18 years (n = 55); (b) surgical reports following arthroscopic surgery for labral repair lacking detailed descriptions of rotator cuff status (n = 800); (c) prior history of surgery (n = 131); (d) diagnosis of infectious arthritis (n = 25); (e) the presence of fracture (n = 11) or tumorous condition (n = 2). We further excluded patients with an interval > 90 days between the radiographic examination and arthroscopic surgery (n = 737), and those with radiographs inadequate for evaluation (n = 14). We split the radiography dataset to generate a training set, a validation set and a test set; the data was first temporally split to generate a test set for external validation (n = 340, January 2019 to December 2019) (Test Set 1), and the remaining data was further split randomly into a training set (90%, n = 2,501, February 2010 to December 2018) and validation set (10%, n = 278, February 2010 to December 2018).

An additional test set (Test Set 2) was collected to assess the algorithm's performance in patients undergoing MRI for rotator cuff evaluations. Of 1,738 patients who underwent MRI of the shoulder joint between January 2019 and December 2019, 1,059 patients who underwent axillary lateral radiography within 90 days of MRI were included. Those with prior surgical history (n = 312), infectious arthritis (n = 2), MRI inadequate for evaluation (n = 18), radiographs inadequate for evaluation (n = 22), or missing clinical information (n = 78) were excluded. Among the 627 patients whose data were included in Test Set 2, 232 were also included in Test Set 1.

Clinical information

Patients' electronic medical records were retrospectively reviewed for clinical information, including age, traumatic history, whether the affected shoulder was of the dominant arm, and pain intensity assessed using the visual analogue scale (VAS). The results of the clinical tests performed to evaluate the SSCs were also collected, including the bear-hug, belly-press, and lift-off tests, and the internal rotation lag sign. The clinical tests were performed by a board-certified orthopedic surgeon in fellowship training within two weeks prior to surgery.

The clinical information pertaining to age, traumatic history, lateral dominance, and degree of pain were used for the development of the deep learning algorithm, whereas the results of the clinical tests were utilized only for comparing the diagnostic performance of the algorithm with that of established clinical tests.

Image acquisition

Shoulder radiography for suspected rotator cuff tears is performed with five radiographic views at our institution, including the true anteroposterior, abduction anteroposterior, 30° caudal tilt, supraspinatus outlet, and axillary lateral views. Among these, the axillary lateral view most readily depicts the lesser tuberosity, with minimal overlap with other bony structures; therefore, it was selected for SSC assessments and performed with the patient in a seated position with the arm abducted to the level of the shoulder and the elbow resting on the image detector. Patients were asked to lean laterally toward the table to ensure the center of the glenohumeral joint aligned with the center of the image receptor.

Digital Imaging and Communications in Medicine (DICOM) files of the radiographic images were downloaded from the picture archiving and communication system and anonymized for further use.

Shoulder MRI was performed using various 1.5-T and 3-T MRI systems, including 3-T MRI systems (Achieva, Ingenia, and Ingenia CX, Philips Healthcare) with dedicated shoulder receiver coils (8 channels; Sense Shoulder Coil, Philips Healthcare) for MR examinations performed at our hospital.

Image labeling

Radiographs of Test Set 1 were labeled based on the structured arthroscopic shoulder surgery reports, all of which were conducted by an orthopedic surgeon specializing in shoulder surgery (J.H.O.). The SSC was routinely evaluated using 30° and 70° arthroscopes. The tendon was probed to determine the presence of tears, which were classified as follows: (a) intact; (b) insignificant tear requiring only debridement, including partial-thickness tears involving < 50% of the tendon thickness; (c) significant tears requiring tendon repair, including partial-thickness tears > 50% of the tendon thickness and full-thickness tears. We converted the classification into two levels; label 0 indicated normal or insignificant rotator cuff abnormalities (including partial thickness tears involving < 50% of tendon thickness), and label 1 indicated significant rotator cuff abnormalities (partial-thickness tears > 50% of the tendon thickness and full-thickness tears). Concomitant supraspinatus/infraspinatus tears were also recorded.

For Test Set 2, radiographs were labeled based on the structured reports of the MRI examinations. The SSC was assessed and classified as follows: normal, tendinosis, low-grade partial tear (tear of ≤ 50% tendon thickness), high-grade partial tear (tear of > 50% tendon thickness), and full-thickness tear. The same dichotomized classification was applied to the MRI reports and used to label the radiographs. The presence of concomitant supraspinatus/infraspinatus tears was also recorded based on the MRI reports.

Image preprocessing

The pydicom library 1.2.0 was used to import DICOM files of radiographs. The axillary lateral radiographs were cropped to 42 × 56 mm^2 patches, with a manually defined line placed over the subscapularis insertion site at the lesser tuberosity, indicating the center point. The left shoulder images were horizontally flipped to match those of the right shoulder. The image patches were resized to 210 × 280 pixels using bilinear interpolation. For augmentation of the training data, the center point was altered randomly by applying horizontal and vertical shifts.

Deep learning algorithm

Axillary lateral radiographs and clinical information were trained simultaneously using a multimodal deep learning model [14]. We implemented the model using the TensorFlow library (version 1.12) and used the graphic processing unit (NVIDIA TITAN Xp) in the Linux operating system (Ubuntu 16.04) with NVIDIA CUDA/cuDNN (versions 7.5 and 7.3, respectively).

The multimodal deep learning model consisted of three components: an image component, a clinical information component, and a fusion component (Fig. 1). First, the image component was generated based on a dense block and a transition layer [15]. The input image was first passed through a 7 × 7 convolution layer and a 3 × 3 max pooling layer, with the four dense blocks and three transition layers stacked alternately. The dense block comprised a 1 × 1 convolutional layer, a 3 × 3 convolutional layer, and a concatenation operation. The transitional layer comprised a 1 × 1 convolutional layer and a 2 × 2 average pooling layer; squeeze-and-excitation attention was applied immediately before the pooling calculation [16, 17]. The convolutional layers in the dense block and transition layer were all applied in the same sequential order of batch normalization, rectified linear unit (ReLU) activation, and convolution operation. After the last dense block, the output of the image component was calculated through batch normalization and ReLU activation, and from the average pooling layer and fully connected layer. Next, the clinical information component comprised one fully connected layer with each input being one of four types of clinical information (age, traumatic history, whether the affected shoulder was of the dominant arm, and pain intensity assessed using the VAS). Finally, the fusion component consisted of a concatenation layer (for the output of the image and clinical information components) and a fully connected layer. After passing through this fully connected layer, a sigmoid function was applied to calculate the probability of an SSC tear (label 1).

The initial weights of the image component were obtained using a pre-training process to efficiently improve the model's performance [18], further details of which are described in the

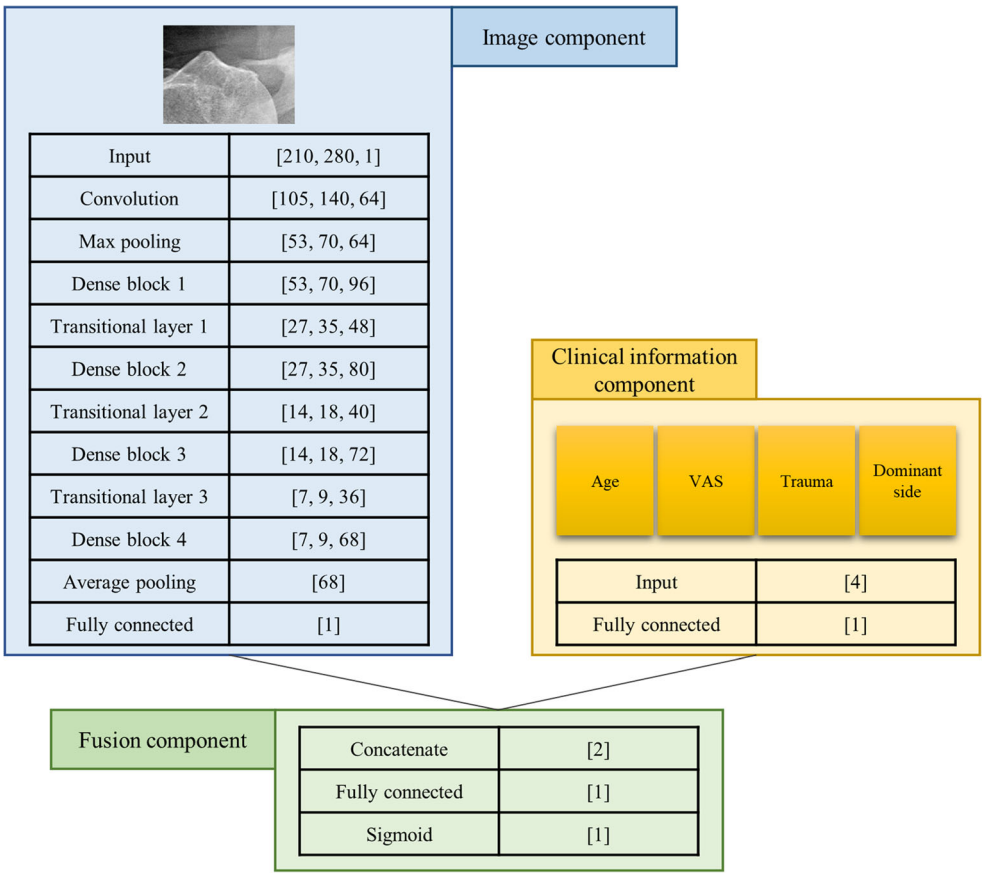

Fig. 1 A schematic diagram showing the training process of our deep learning model. The model consisted of three components, including an imaging component, a clinical information component, and a fusion component

Supplementary Appendix. The initial weights of the clinical information component were coefficients obtained by fitting a logistic regression model in the training set with clinical information as explanatory variables and SSC labels as a predictor. The initial weights of the fusion component were coefficients obtained by fitting a logistic regression model in the training set with outputs of the image and clinical information components as explanatory variables and SSC labels as a predictor.

The learning rate started at 0.0005 and decayed every 5,000 steps at a rate of 0.7. The mini-batch size was 8. Focal loss was the objective function [19], followed by minimization using the RMSprop optimizer [20]. The hyperparameters of focal loss, as well as alpha and gamma, were 0.3 and 2, respectively. L2 regularization was applied to the trainable weights to prevent overfitting.

The Gradient-weighted Class Activation Mapping (Grad-CAM) method was applied to identify the region in the images that was most predictive [21]. By linearly combining the layer just before the average pooling of the image component in our model and the gradient of the corresponding layer, the region in the image with the highest sensitivity was determined. ReLU activation was not applied to the linear combination to confirm both the positive and negative gradients. The high sensitivity regions were confirmed by distinguishing between regions with relatively high and low gradients.

Statistical analysis

The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR), and negative LR of the deep learning algorithm were calculated. Receiver operating characteristic (ROC) curves were constructed for the probability output of the deep learning algorithm, from which the area under the curves (AUCs) were calculated. The AUCs were calculated for the clinical information component alone, the image component alone, and the fusion component. Two cutoff points were determined based on the validation set and were applied to the test set; an optimal cutoff point determined by Youden's J statistic ($cutoff_{optimal}$) and a cutoff point for an expected sensitivity of 95% ($cutoff_{95\%}$).

To identify the factors associated with false negative test results, the demographic and clinical factors of the true and false negative groups were compared using Fisher's exact test. For continuous variables, two-tailed independent Student's *t* tests or Wilcoxon rank-sum tests were used to compare the two groups.

Table 1 Baseline characteristics

	Training set (n = 2501)	Validation set (n = 278)	Test set 1 (n = 340)	Test set 2 (n = 627)
Age (years)	60.7 ± 9.3	61.5 ± 9.7	61.8 ± 9.4	61.2 ± 11.0
Sex				
Female	1464 (58.5)	137 (49.3)	191 (56.2)	333 (53.1)
Male	1037 (41.5)	141 (50.7)	149 (43.8)	294 (46.9)
Dominant side				
Dominant side	1756 (70.2)	216 (77.7)	238 (70.0)	429 (68.4)
Non-dominant side	745 (29.8)	62 (22.3)	102 (30.0)	198 (31.6)
History of trauma				
Present	867 (34.7)	95 (34.2)	122 (35.9)	217 (34.6)
Absent	1634 (65.3)	183 (65.8)	218 (64.1)	410 (65.4)
Degree of pain (VAS)	6.21 ± 2.26	6.14 ± 2.22	5.67 ± 2.04	5.94 ± 2.09
SSC label				
Label 0	1939 (77.5)	216 (77.7)	224 (65.9)	403 (64.3)
SST/IST label 0	65 (3.4)	6 (2.8)	1 (0.4)	96 (23.8)
SST/IST label 1	1874 (96.6)	210 (97.2)	223 (99.6)	307 (76.2)
Label 1	562 (22.5)	62 (22.3)	116 (34.1)	224 (35.7)
SST/IST label 0	40 (7.1)	5 (8.1)	2 (1.7)	24 (10.7)
SST/IST label 1	522 (92.9)	57 (91.9)	114 (98.3)	200 (89.3)
Concomitant SST/IST lesion				
Label 0	105 (4.2)	11 (4.0)	3 (0.9)	120 (19.1)
Label 1	2396 (95.8)	267 (96.0)	337 (99.1)	507 (80.9)

Data are numbers of cases and percentages in the parentheses, unless otherwise specified

All statistical analysis was performed using R statistical software, version 3.6.1 (The R Foundation for Statistical Computing). p values < 0.05 were considered statistically significant.

Results

Demographic information, baseline characteristics, and the distribution of the SSC labels are summarized in Table 1.

The AUC of the clinical information component alone was 0.64 (95% confidence interval (CI), 0.58–0.70) for Test Set 1 and 0.68 (95% CI, 0.64–0.72) for Test Set 2. The AUC of the image component alone was 0.83 (95% CI, 0.78–0.87) and 0.81 (95% CI, 0.78–0.85) for Test Sets 1 and 2, respectively. Further details pertaining to the diagnostic performance of the clinical information component and the image component are included in the Supplementary Appendix. The algorithm based on the combination of the clinical information and image

Table 2 Diagnostic performance of the deep learning algorithm and clinical tests in patients undergoing arthroscopic surgery for rotator cuff pathology (Test Set 1)

	$Cutoff_{opt}$	$Cutoff_{95\%}$	Bear hug test	Belly press test	Lift off test	Internal rotation lag sign
Sensitivity	65.5% (76/116, 56.1–74.1%)	91.4% (106/116, 84.7–95.8%)	77.6% (90/116, 68.9–84.8%)	50.9% (59/116, 41.4–60.3%)	8.6% (10/116, 4.2–15.3%)	11.2% (13/116, 6.1–18.4%)
Specificity	86.6% (194/224, 81.4–90.8%)	42.0% (94/224, 35.4–48.7%)	80.8% (181/224, 75.0–85.7%)	93.8% (210/224, 89.7–96.5%)	99.6% (223/224, 97.5–100.0%)	99.6% (223/224, 97.5–100.0%)
Positive predictive value	71.7% (76/106, 62.1–80.0%)	44.9% (106/236, 38.5–51.5%)	67.7% (90/133, 59.0–75.5%)	80.8% (59/73, 69.9–89.1%)	90.9% (10/11, 58.7–99.8%)	92.9% (13/14, 66.1–99.8%)
Negative predictive value	82.9% (194/234, 77.5–87.5%)	90.4% (94/104, 83.0–95.3%)	87.4% (181/207, 82.1–91.6%)	78.7 (210/267, 73.2–83.4%)	67.8% (223/329, 62.4–72.8%)	68.4% (223/326, 63.1–73.4%)
Positive likelihood ratio	4.89 (3.42–7.00)	1.57 (1.39–1.78)	4.04 (3.04–5.38)	8.14 (4.75–13.93)	19.31 (2.50–149.01)	25.10 (3.32–189.53)
Negative likelihood ratio	0.40 (0.31–0.51)	0.21 (0.11–0.38)	0.28 (0.20–0.39)	0.52 (0.43–0.63)	0.92 (0.89–0.97)	0.89 (0.84–0.95)

Data are percentages with the nominator/denominator and/or 95% confidence interval in the parentheses

Table 3 Diagnostic performance of the deep learning algorithm in patients undergoing shoulder MRI for suspected rotator cuff pathology (Test set 2)

	Test set 2	
	$Cutoff_{opt}$	$Cutoff_{95\%}$
Sensitivity	65.6% (147/224, 59.0–71.8%)	90.2% (202/224, 85.5–93.7%)
Specificity	85.4% (344/403, 81.5–88.7%)	46.4% (187/403, 41.5–51.4%)
Positive predictive value	71.4% (147/206, 64.7–77.4%)	48.3% (202/418, 43.4–53.2%)
Negative predictive value	81.7% (344/421, 77.7–85.3%)	89.5% (187/209, 84.5–93.3%)
Positive likelihood ratio	4.48 (3.48–5.78)	1.68 (1.52–1.86)
Negative likelihood ratio	0.40 (0.33–0.48)	0.21 (0.14–0.32)

Data are percentages with the nominator/denominator and/or 95% confidence interval in the parentheses

components (fusion component) was used for subsequent analysis. The diagnostic performance of the algorithm using Test Sets 1 and 2 is summarized in Tables 2 and 3, respectively.

For Test Set 1, the AUC of the deep learning algorithm was 0.83 (95% CI, 0.79–0.88). At the $cutoff_{95\%}$, the sensitivity and specificity were 91.4% (106/116, 84.7–95.8%) and 42.0% (94/224, 35.4–48.7%), respectively. The algorithm could rule out a significant SSC tear with an NPV of 90.4% (94/104, 83.0–95.3%) and a negative LR of 0.21 (0.11–0.38). With the $cutoff_{optimal}$, the algorithm exhibited sensitivity of 65.6% (76/116; 95% CI, 56.1–74.1%), specificity of 86.6% (194/224, 81.4–90.8%), positive LR of 4.89 (3.42–7.00), and negative LR of 0.40 (0.31–0.51). The sensitivity and specificity of the clinical tests ranged from 8.6 to 77.6% and from 80.8 to 99.6%, respectively.

The deep learning algorithm exhibited similar diagnostic performance with Test Set 2. The AUC was 0.82 (95% CI,

Table 4 Clinical factors associated with false negative test results in Test Set 1 when $cutoff_{95\%}$ is applied

	Total negative cases (n = 104)	True negative (n = 94)	False negative (n = 10)	p value
Age				
< 60 years	62.5% (65/104, 52.5–71.8%)	64.9% (61/94, 54.4–74.5%)	40.0% (4/10, 12.2–73.8%)	0.17
≥ 60 years	37.5% (39/104, 28.2–47.5%)	35.1% (33/94, 25.5–45.6%)	60.0% (6/10, 26.2–87.8%)	
Sex				
Female	61.5% (64/104, 51.5–70.9%)	62.8% (59/94, 52.2–72.5%)	50.0% (5/10, 18.7–81.3%)	0.50
Male	38.5% (40/104, 29.1–48.5%)	37.2% (35/94, 27.5–47.8%)	50.0% (5/10, 18.7–81.3%)	
Dominancy				
Dominant side	66.3% (69/104, 56.4–75.3%)	67.0% (63/94, 56.6–76.4%)	60.0% (6/10, 26.2–87.8%)	0.73
Non-dominant side	33.7% (35/104, 24.7–43.6%)	33.0% (31/94, 23.6–43.4%)	40.0% (4/10, 12.2–73.8%)	
Trauma				
Present	28.8% (30/104, 20.4–38.6%)	28.7% (27/94, 19.9–39.0%)	30.0% (3/10, 6.7–65.2%)	> 0.99
Absent	71.2% (74/104, 61.4–79.6%)	71.3% (67/94, 61.0–80.1%)	70.0% (7/10, 34.8–93.3%)	
Degree of pain				
Mild	15.4% (16/104, 9.1–23.8%)	14.9% (14/94, 8.4–23.7%)	20.0% (2/10, 2.5–55.6%)	0.16
Moderate	61.5% (64/104, 51.5–70.9%)	59.6% (56/94, 49.0–69.6%)	80.0% (8/10, 44.4–97.5%)	
Severe	23.1% (24/104, 15.4–32.4%)	25.5% (24/94, 17.1–35.6%)	0.0% (0/10, 0.0–30.8%)	
Concomitant SST/IST tear				
Present	99.0% (103/104, 94.8–100.0%)	98.9% (93/94, 94.2–100.0%)	100.0% (10/10, 69.2–100.0%)	> 0.99
Absent	1.0% (1/104, 0.0–5.2%)	1.1% (1/94, 0.0–5.8%)	0.0% (0/10, 0.0–30.8%)	

Data are percentages with the nominator/denominator and/or 95% confidence interval in the parentheses

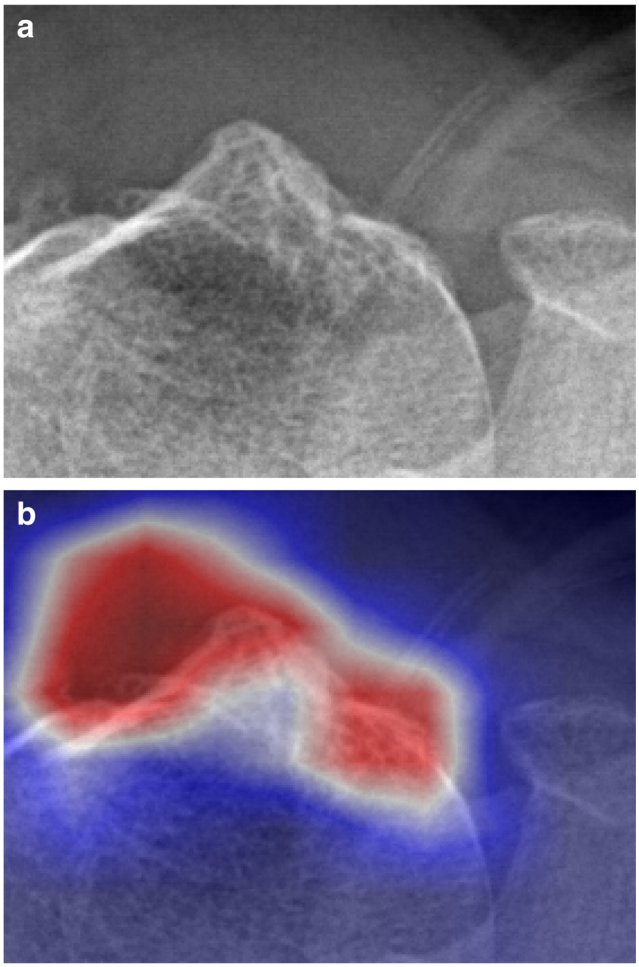

Fig. 2 A representative case for label 1 (significant subscapularis tendon tear). The axillary lateral radiograph (**a**) shows cortical irregularity of the lesser tuberosity, which corresponds to the region of highest intensity on the Gradient-weighted Class Activation Map (**b**)

0.79–0.86). At the cutoff$_{95\%}$, the sensitivity and specificity were 90.2% (202/224, 85.5–93.7%) and 46.4% (187/403, 41.5–51.4%), respectively. Significant SSC tears could be ruled out with a NPV of 89.5% (187/209, 84.5–93.3%) and a negative LR of 0.21 (0.14–0.32).

The results of subgroup analysis of cases with negative test results are shown in Table 4. In the subgroup analysis of cases with negative test results, false negatives were more frequent in patients ≥ 60 years old and in those with moderate pain. However, there was no statistically significant association between patient age, sex, dominance of the affected arm, traumatic history, pain intensity, or concomitant SST/IST tears and false negative results.

The Grad-CAM method was used to qualitatively assess the deep learning algorithm's performance. The region with the highest sensitivity in the axillary lateral view radiographs was the lesser tuberosity, which is the attachment site of the SSC (Figs. 2 and 3).

Discussion

Our deep learning algorithm identified SSC tears based on axillary lateral radiographs with moderate accuracy (AUC 0.83) in patients undergoing arthroscopic surgery. With a high-sensitivity cutoff, it achieved a sensitivity of 91.4% and a negative LR of 0.21.

Our deep learning algorithm exhibited a negative predictive value of 90.4% and a negative LR of 0.21 at a cutoff point with an expected sensitivity of 95%. Diagnostic tests provide strong evidence when the negative LR is below 0.2, small but sometimes important diagnostic evidence at 0.2–0.5, and small, rarely important evidence at 0.5–1 [9]. Although our algorithm did not achieve a negative LR below 0.2, it was lower than those of single clinical tests commonly used for initial SSC assessments, which have negative predictive values of 67.8–87.4% and negative LRs of 0.28–0.92. A recent systematic review showed that by combining the bear-hug and belly-press tests in parallel, a negative LR of 0.21 is achievable [22]. Our deep learning algorithm could enable the

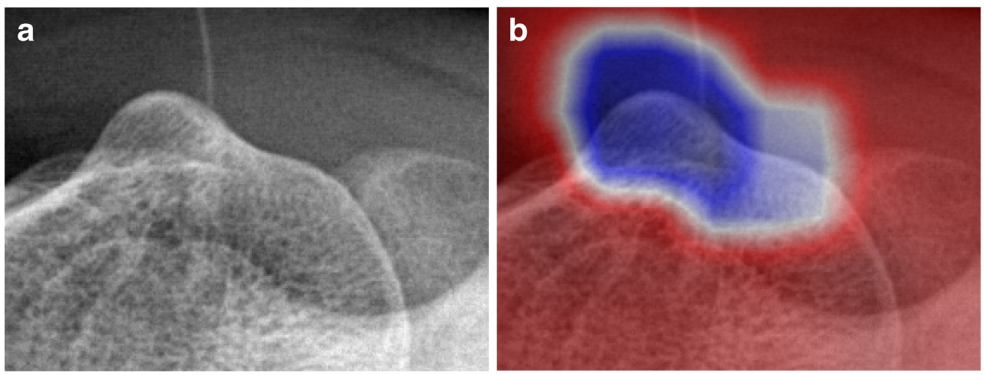

Fig. 3 A representative case for label 0 (normal or insignificant subscapularis tendon tear). The axillary lateral radiograph (**a**) shows the absence of abnormal cortical changes at the lesser tuberosity, which corresponds to the region of highest intensity on the Gradient-weighted Class Activation Map (**b**)

use of axillary lateral radiographs for objectively assessing SSC tears and to identify patients for whom further imaging would be unnecessary, while minimizing the large inter-observer variability of clinical tests.

The overall accuracy of identifying SSC tears with our deep learning algorithm was moderate, with an AUC of 0.83. In the study by Kim et al [23], a deep learning algorithm assessing significant tear in the supraspinatus/infraspinatus complex was developed based on radiographs, and the algorithm showed an AUC of 0.91. Several factors may have attributed to the lower accuracy of the deep learning algorithm in assessing SSC tears compared to the supraspinatus/infraspinatus complex. The study by Kim et al utilized three different radiographic views (true anteroposterior, caudal 30° tilt, and supraspinatus outlet view), whereas the assessment of the SSC tears in the current study was based on only a single radiographic view. Although arthroscopy is accepted as the gold standard in diagnosing subscapular tendon tears, the insertion of the subscapularis tendon at the lesser tuberosity is incompletely visualized on arthroscopy, and lesions may be left undetected [24]. The limited accuracy of the arthroscopy in diagnosing subscapularis may have caused noisy labels in the training data leading to a degradation in the performance of the deep learning algorithm.

Our algorithm was qualitatively evaluated via the Grad-CAM method, revealing the lesser tubercle of the humerus as the optimal region for predicting the probable likelihood of SSC tears. This was consistent with previous studies showing an association between bone changes in the lesser tuberosity and SSC tears [10–13, 25–27]. In a study by Studler et al [9], isolated radiographic findings of lesser tuberosity cysts and cortical irregularity showed a sensitivity of 21–51% and a specificity 65–87% in diagnosing SSC tears when interpreted by radiologists. Several studies have also shown an association between lesser tuberosity cysts detected via MRI and SSC tears [10, 12, 13]. Our Grad-CAM results support the fact that the performance of our deep learning algorithm was not achieved by chance and that the probability output was based on sound imaging evidence.

Herein, the arthroscopic findings were used as the reference standard for diagnosing SSC tears. The accuracy of MRI in diagnosing SSC abnormalities has long been debated, and it is known that the accuracy in diagnosing subscapularis tendon tear is lower than that of overall rotator cuff tears [28–32]. In a recent meta-analysis by Malavolta et al [28], the sensitivity and specificity of MRI in diagnosing overall SSC tears were 0.68 and 0.90, respectively. We opted to use arthroscopic findings of the SSC as the reference standard, as it is a more reliable evaluation method, although our test set, composed of patients undergoing arthroscopic surgery, was not free from verification and spectrum biases. The prevalence of SSC tears would be higher in patients undergoing arthroscopic surgery compared to a clinical context in which shoulder radiographs are performed as an initial imaging assessment. The higher prevalence could have resulted in the overestimation of the diagnostic performance [33]. To partially compensate for this bias, we introduced a second test set collected from patients undergoing MRI for rotator cuff evaluations, demonstrating that the diagnostic performance of our algorithm was comparable in this second test set. To further examine the diagnostic performance of our algorithm, however, a prospective evaluation of the algorithm is warranted.

In the subgroup analysis of cases with negative test results, false negatives were more frequent in patients over 60 years old than in those under 60, and in those with moderate pain compared to those with milder pain. This indicates that negative test results should be interpreted with caution in patients of advanced age and moderate pain; however, no statistically significant association was found between clinical factors and false negativity in our study.

The study had several limitations. Firstly, the data were retrospectively collected in a single institution, and the possibility of verification bias limits the generalizability of our results. A prospective study is needed to further evaluate our deep learning algorithm's performance in the initial assessment of patients with shoulder pain. Secondly, the clinical test results we included were not all performed at the initial presentation. In most cases, the examiner had prior knowledge of the MRI results, and the clinical test results were recorded after the decision to repair the rotator cuff had been made. Thirdly, arthroscopy has limited use in evaluating the SSC, as interstitial tears may evade detection.

In conclusion, our deep learning algorithm is capable of assessing SSC tears based on changes at the lesser tuberosity on axillary lateral radiographs with moderate accuracy. Our algorithm could be used as an objective method to initially assess SSC integrity, although studies are needed to further evaluate the generalizability and its clinical applicability.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-08034-1.

Acknowledgements The authors sincerely thank Jeongmin Choi for her contribution in data collection.

Funding This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1F1A1060126) and grant from the SNUBH Research Fund (grant no.14-2020-038).

Declarations

Guarantor The scientific guarantor of this publication is Yusuhn Kang.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors (Dongjun Choi) has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap Some study subjects or cohorts have been previously reported in Kim Y, Choi D, Lee KJ et al (2020) Ruling out rotator cuff tear in shoulder radiograph series using deep learning: redefining the role of conventional radiograph. Eur Radiol 30:2843-2852.

Methodology
- Retrospective
- Experimental
- Performed at one institution

References

1. Richards DP, Burkhart SS, Lo IK (2003) Subscapularis tears: arthroscopic repair techniques. Orthop Clin North Am 34:485–498
2. Denard PJ, Burkhart SS (2013) Arthroscopic recognition and repair of the torn subscapularis tendon. Arthrosc Tech 2:e373–e379
3. Lee SH, Nam DJ, Kim SJ, Kim JW (2017) Comparison of clinical and structural outcomes by subscapularis tendon status in massive rotator cuff tear. Am J Sports Med 45:2555–2562
4. Kim SJ, Jung M, Lee JH, Park JH, Chun YM (2015) Arthroscopic repair of a significant (> 50%) partial-thickness subscapularis tear concomitant with a full-thickness supraspinatus tear: technical considerations for subscapularis repair (transtendon technique versus tear completion). J Shoulder Elbow Surg 24:875–881
5. Barth J, Audebert S, Toussaint B et al (2012) Diagnosis of subscapularis tendon tears: are available diagnostic tests pertinent for a positive diagnosis? Orthop Traumatol Surg Res 98:S178–S185
6. Kappe T, Sgroi M, Reichel H, Daexle M (2018) Diagnostic performance of clinical tests for subscapularis tendon tears. Knee Surg Sports Traumatol Arthrosc 26:176–181
7. Yoon JP, Chung SW, Kim SH, Oh JH (2013) Diagnostic value of four clinical tests for the evaluation of subscapularis integrity. J Shoulder Elbow Surg 22:1186–1192
8. Bartsch M, Greiner S, Haas NP, Scheibel M (2010) Diagnostic values of clinical tests for subscapularis lesions. Knee Surg Sports Traumatol Arthrosc 18:1712–1717
9. Studler U, Pfirrmann CW, Jost B, Rousson V, Hodler J, Zanetti M (2008) Abnormalities of the lesser tuberosity on radiography and MRI: association with subscapularis tendon lesions. AJR Am J Roentgenol 191:100–106
10. Celikyay F, Yuksekkaya R, Deniz C, Inal S, Gokce E, Acu B (2015) Locations of lesser tuberosity cysts and their association with subscapularis, supraspinatus, and long head of the biceps tendon disorders. Acta Radiol 56:1494–1500
11. Wissman RD, Kapur S, Akers J, Crimmins J, Ying J, Laor T (2009) Cysts within and adjacent to the lesser tuberosity and their association with rotator cuff abnormalities. AJR Am J Roentgenol 193:1603–1606
12. Wissman RD, Ingalls J, Hendry D, Gorman D, Kenter K (2012) Cysts within and adjacent to the lesser tuberosity: correlation with shoulder arthroscopy. Skeletal Radiol 41:1105–1110
13. Cetinkaya M, Oner AY, Ataoglu MB, Ozer M, Ayanoglu T, Kanatli U (2017) Lesser tuberosity cysts and their relationship with subscapularis tears and subcoracoid impingement. J Orthop Sci 22:63–68
14. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP (2020) Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med 3:136
15. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 2261–2269
16. Li X, Shen X, Zhou Y, Wang X, Li T-Q (2020) Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet). PLoS One 15:e0232127
17. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: CVPR. IEEE Computer Society, pp 7132–7141. https://doi.org/10.1109/CVPR.2018.00745
18. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: Ghahramani Z, Welling M, Cortes C, et al. (eds) NIPS. pp 3320–3328
19. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: ICCV. IEEE Computer Society, pp 2999–3007
20. Hinton G, Srivastava N, Swersky K (2012) Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Cited on 14(8)
21. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2015) Grad-cam: visual explanations from deep networks via gradient-based Localization. In: Proc. IEEEInt. Conf. Computer Vision (ICCV), pp 618–626
22. Dakkak A, Krill MK, Krill ML, Nwachukwu B, McCormick F (2020) Evidence-based physical examination for the diagnosis of subscapularis tears: a systematic review. Sports Health. https://doi.org/10.1177/1941738120936232:1941738120936232
23. Kim Y, Choi D, Lee KJ et al (2020) Ruling out rotator cuff tear in shoulder radiograph series using deep learning: redefining the role of conventional radiograph. Eur Radiol 30:2843–2852
24. Wright JM, Heavrin B, Hawkins RJ, Noonan T (2001) Arthroscopic visualization of the subscapularis tendon. Arthroscopy 17:677–684
25. Shah SH, Small KM, Sinz NJ, Higgins LD (2016) Morphology of the lesser tuberosity and intertubercular groove in patients with arthroscopically confirmed subscapularis and biceps tendon pathology. Arthroscopy 32:968–975
26. Cho JH, Han KJ, Lee DH, Chung NS, Park DY (2015) Pit above the lesser tuberosity in axial view radiography. Knee Surg Sports Traumatol Arthrosc 23:370–375
27. Lee JH, Rhyou IH, Ahn KB (2020) Prediction of the anterior shoulder pain source by detecting indirect signs for partial articular subscapularis tendon tears through conventional magnetic resonance imaging. Knee Surg Sports Traumatol Arthrosc. https://doi.org/10.1007/s00167-020-06259-z
28. Malavolta EA, Assuncao JH, Gracitelli MEC, Yen TK, Bordalo-Rodrigues M, Ferreira Neto AA (2019) Accuracy of magnetic resonance imaging (MRI) for subscapularis tear: a systematic review and meta-analysis of diagnostic studies. Arch Orthop Trauma Surg 139:659–667
29. Malavolta EA, Assuncao JH, Guglielmetti CL et al (2016) Accuracy of preoperative MRI in the diagnosis of subscapularis tears. Arch Orthop Trauma Surg 136:1425–1430
30. Furukawa R, Morihara T, Arai Y et al (2014) Diagnostic accuracy of magnetic resonance imaging for subscapularis tendon tears using radial-slice magnetic resonance images. J Shoulder Elbow Surg 23:e283–e290

31. Foad A, Wijdicks CA (2012) The accuracy of magnetic resonance imaging and magnetic resonance arthrogram versus arthroscopy in the diagnosis of subscapularis tendon injury. Arthroscopy 28:636–641
32. Adams CR, Schoolfield JD, Burkhart SS (2010) Accuracy of preoperative magnetic resonance imaging in predicting a subscapularis tendon tear based on arthroscopy. Arthroscopy 26:1427–1433
33. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Group Q-S (2013) A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 66:1093–1104