Received: 4 April 2024 | Revised: 15 July 2024 | Accepted: 18 August 2024

DOI: 10.1002/jsp2.70003

RESEARCH ARTICLE

JOR Spine OPEN ACCESS

Deep learning-based structure segmentation and intramuscular fat annotation on lumbar magnetic resonance imaging

Yefu Xu | Shijie Zheng | Qingyi Tian | Zhuoyan Kou | Wenqing Li | Xinhui Xie | Xiaotao Wu

Department of Spine Surgery, ZhongDa Hospital, School of Medicine, Southeast University, Nanjing, China

Correspondence
Xiaotao Wu, Department of Spine Surgery, Affiliated ZhongDa Hospital, School of Medicine, Southeast University, Nanjing, Jiangsu 210009, China.
Email: wuxiaotaospine@seu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 82372473

Abstract

Background: Lumbar disc herniation (LDH) is a prevalent cause of low back pain. LDH patients commonly experience paraspinal muscle atrophy and fatty infiltration (FI), which further exacerbates the symptoms of low back pain. Magnetic resonance imaging (MRI) is crucial for assessing paraspinal muscle condition. Our study aims to develop a dual-model for automated muscle segmentation and FI annotation on MRI, assisting clinicians evaluate LDH conditions comprehensively.

Methods: The study retrospectively collected data diagnosed with LDH from December 2020 to May 2022. The dataset was split into a 7:3 ratio for training and testing, with an external test set prepared to validate model generalizability. The model's performance was evaluated using average precision (AP), recall and F1 score. The consistency was assessed using the Dice similarity coefficient (DSC) and Cohen's Kappa. The mean absolute percentage error (MAPE) was calculated to assess the error of the model measurements of relative cross-sectional area (rCSA) and FI. Calculate the MAPE of FI measured by threshold algorithms to compare with the model.

Results: A total of 417 patients being evaluated, comprising 216 males and 201 females, with a mean age of 49 ± 15 years. In the internal test set, the muscle segmentation model achieved an overall DSC of 0.92 ± 0.10, recall of 92.60%, and AP of 0.98. The fat annotation model attained a recall of 91.30%, F1 Score of 0.82, and Cohen's Kappa of 0.76. However, there was a decrease on the external test set. For rCSA measurements, except for longissimus (10.89%), the MAPE of other muscles was less than 10%. When comparing the errors of FI for each paraspinal muscle, the MAPE of the model was lower than that of the threshold algorithm.

Conclusion: The models demonstrate outstanding performance, with lower error in FI measurement compared to thresholding algorithms.

KEYWORDS
deep learning, fatty infiltration, lumbar disc herniation, paraspinal muscles, segmentation

1 | INTRODUCTION

Lumbar disc herniation (LDH) is a common condition leading to low back pain.[1] Each year, a large number of individuals are affected by this condition, impacting both work and daily life.[2] Patients with LDH often experience accompanying atrophy of the paraspinal muscles and increased intramuscular fat content, which further exacerbates the symptoms of low back pain.[3–5] Research on LDH not only focuses on the herniated intervertebral disc itself but also on the condition of the paraspinal muscles.

Magnetic resonance imaging (MRI) provides clear visualization of the intervertebral disc, nerve roots, spinal cord, and paraspinal muscles.[6] It holds significant importance for both the diagnosis and treatment of LDH.[7] The muscles at the level of L2 to L5 in the lumbar region are the most abundant and serve as a focal area in many studies.[8,9] Assessing the condition of lumbar muscles on MRI involves measuring the area of muscle regions and the ratio of fat area to muscle area.[10] These procedures are time-consuming, and an efficient and accurate deep learning (DL) model can greatly assist in this regard.

Previous studies have reported segmentation models for lumbar structures.[11–13] However, these models do not encompass all regions comprehensively. The erector spinae have traditionally been regarded as a singular entity, yet they consist of the iliocostalis (IL) and longissimus (LO) components, which could be more accurately identified separately.[14] For the assessment of fat infiltration degree, the predominant method is the application of automated thresholding algorithm techniques.[15,16] While offering operational simplicity and speed, they may lack precision when quantifying fat content. Semantic segmentation techniques within DL enable simultaneous learning of both the morphological characteristics and signal intensity (SI) differences of regions of interest (ROIs),[17] making them particularly suitable for identifying fat regions within muscles.

Our study is focused on the development of a dual-model tailored for the fully automated segmentation of muscles and annotation of fat regions in lumbar spine MRI. This model aids clinical practitioners in swiftly and accurately delineating each structure within lumbar MRI scans, facilitating the calculation of muscle area and intramuscular fat ratio. Such capabilities are pivotal for the comprehensive assessment of LDH severity in clinical practice.

2 | METHOD

This research received approval from our institutional review board and complied with ethical regulations. Due to the retrospective nature of the study and minimal risks involved, patient informed consent was waived.

3 | DATASET PREPARATION

This study retrospectively collected patients diagnosed with LDH and admitted for treatment between December 2020 and May 2022. Inclusion criteria comprised: (1) The diagnosis of LDH confirmed; (2) The duration of low back pain and/or leg pain exceeding 6 months; (3) Undergoing lumbar spine MRI examination within 1 month prior to admission. Exclusion criteria included: (1) Previous lumbar spine surgery with internal fixation devices; (2) Unclear MRI images; (3) MRI performed with contrast enhancement. Cases meeting the selection criteria had their lumbar spine MRI images collected, along with general and clinical information during their inpatient stay (such as sex, age, disease duration, symptoms, the BMI, and chronic medical history).

The MRI scans selected axial T2-weighted images, encompassing the range from the second to the fifth lumbar vertebrae, with one image acquired per intervertebral space (three scans were performed for each intervertebral space, and the second scan, positioned at the center of the intervertebral disc, was chosen for analysis). Cases were randomly allocated into training/validation sets and internal testing sets at a 7:3 ratio. MRI scanning was conducted using a 1.5 T platform (General Electric, USA) with specific parameters: Echo time ranging from 120 to 125 ms, repetition time ranging from 3600 to 4000 ms, section thickness of 4.5 mm, matrix size of 512 × 512, and a field of view of 200 × 200 mm^2. The images were saved in DICOM format, standardized with uniform coding, and patient information was anonymized.

An independent medical institution collaborating with us collected a dataset of lumbar spine MRI scans from patients with LDH. In this dataset, samples were screened using the same inclusion and exclusion criteria as mentioned above. Among the screened samples, 50 samples were randomly selected as an external test set, with image extraction and processing methods consistent with the internal dataset.

4 | DATASET ANNOTATION

The training/validation set was annotated by three spinal surgeons. The annotated images did not contain any patient information. Initially, labeling for the lumbar vertebrae and muscles was performed using the open-source image annotation software LabelMe (version 3.16.2) in Python. The lumbar spine structures were delineated using polygonal annotation boxes, including: intervertebral discs (IVD), vertebral arch (VA), psoas major (PM), quadratus lumborum (QL), LO, IL, and multifidus (MF) (Figure S1). Subsequently, the pixel values of the images were standardized to a standard normal distribution using the SimpleITK (Python library). The ROIs for muscle categories were cropped from the images, and the intramuscular fat regions were labeled using 3D Slicer (Figure S2). For ensuring precise annotation, all images were manually labeled by raters, without employing threshold-based segmentation assistance. The paint mode with a 2.0 mm diameter was utilized, and the annotated results were saved in NRRD format.

The internal test set was annotated by a spinal expert with 26 years of clinical experience, serving as the ground truth labels. To facilitate comparison with the DL models, we also invited a spinal

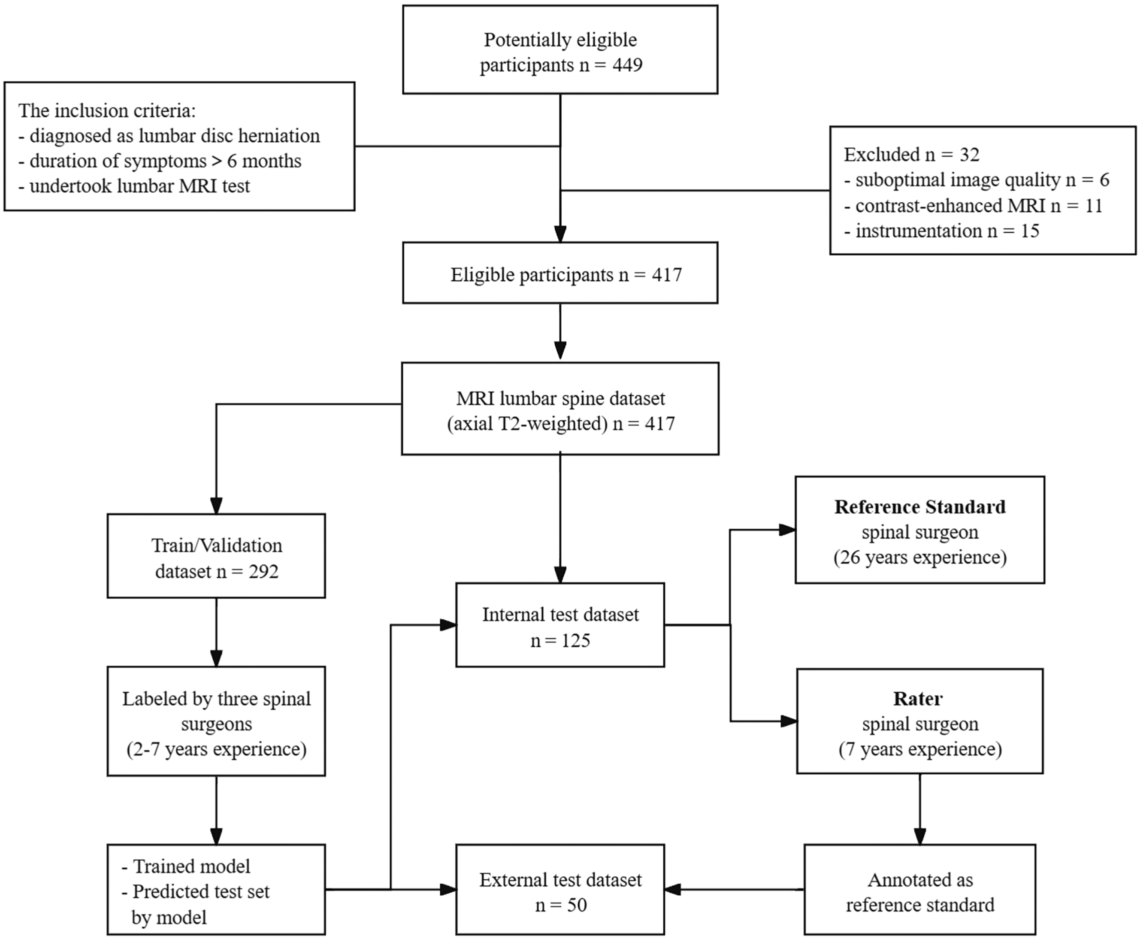
Potentially eligible participants n = 449
The inclusion criteria:
- diagnosed as lumbar disc herniation
- duration of symptoms > 6 months
- undertook lumbar MRI test
Excluded n = 32
- suboptimal image quality n = 6
- contrast-enhanced MRI n = 11
- instrumentation n = 15
Eligible participants n = 417
MRI lumbar spine dataset (axial T2-weighted) n = 417
Train/Validation dataset n = 292
Reference Standard spinal surgeon (26 years experience)
Internal test dataset n = 125
Labeled by three spinal surgeons (2-7 years experience)
Rater spinal surgeon (7 years experience)
- Trained model
- Predicted test set by model
External test dataset n = 50
Annotated as reference standard

FIGURE 1 Flowchart of the study design for the internal dataset and the external test set.

surgeon with 7 years of clinical experience to annotate the internal test set. Additionally, this surgeon served as the ground truth labels for annotating the external test set (Figure 1). The image annotation standards were consistent with those of the training set, and no pixel standardization was applied to the test set images.

5 | DL MODEL DEVELOPED

This study employed a dual-model architecture: Mask R-CNN[18] as the primary structure for lumbar structure recognition, and U-net[19] as the primary structure for intramuscular fat recognition.

The architecture of the Lumbar Structures Segmentation Model (Seg Model) includes a cascade region-based convolutional neural network[20] for instance segmentation, which employs Mask R-CNN as the backbone network combined with Feature Pyramid Network (FPN)[21] as an intermediate layer. The backbone network adopts the ResNet101[22] structure, consisting of four stages, and utilizes pre-trained model weights for initialization. The weights were sourced from the Open-MMLab model library (github.com/open-mmlab/mmdetection/tree/main/configs/cascade_rcnn). The FPN utilizes the feature maps of the backbone network to generate a multiscale feature pyramid. The region proposal network (RPN)[23] is responsible for generating candidate boxes for object detection, using the output of FPN as input, predicting the target scores and bounding box offsets of anchor boxes, and employing smooth L1 loss and cross-entropy loss for bounding box regression and target classification, respectively. The ROI head refines and predicts the class for each candidate box, employing three cascading stages, each with its own bounding box head. In the final cascading stage, a head measuring positional offsets is utilized based on the characteristics of this task. Overall, the model combines backbone network features with a multi-scale feature pyramid and utilizes a cascade approach for object detection and instance segmentation.

The Intramuscular Fat Infiltration Detection Model (FI Model) is based on a generative neural network with a U-Net structure, used for image semantic segmentation tasks. The U-Net structure consists of an encoder and a decoder. In this model, the encoder comprises a series of down-sample modules, each containing three consecutive convolutional layers followed by a max-pooling layer, for feature extraction and down-sampling of the input image. By gradually reducing spatial resolution, they transform the input image into a high-dimensional feature representation. The decoder consists of a series of up-sample modules, each containing a transposed convolutional layer and three consecutive convolutional layers, for gradually restoring the feature maps to the same size as the original input image. The up-sample modules gradually restore high-dimensional features through up-sampling operations and concatenate them with features from corresponding layers of the encoder, achieving refinement and precise segmentation of the image. The entire network structure aims to achieve hierarchical feature extraction and restoration of images through the encoder-decoder structure, thereby achieving precise image semantic segmentation. Initialization of the model excluded pre-trained weights; instead, initial weights were randomly assigned to the model's layers to encourage exploration of diverse feature representations during training.

The segmentation model comprises approximately 95.8 million parameters and was trained for a total duration of 49 h, using a batch size of 16 and an SGD optimizer with an initial learning rate of 0.02. The FI model, with approximately 50.2 million parameters, was trained for 15 h, using a batch size of 16 and an Adam optimizer with an initial learning rate of 0.001. Additionally, the FI model employed a custom Focal Loss function with $\gamma = 5$ to enhance its focus on crucial information.

Considering the need for models to perform automatic recognition in clinical scenarios, we integrated the two models into a dual-model structure and added result visualization functionality. When test set images are inputted, the models automatically segment muscles, label fat tissues, calculate muscle area, and intramuscular fat ratio. Finally, it outputs the segmented image and the fat-labeled image (Figure 2). To evaluate the performance of the dual model, we divided the evaluation into three groups (Manual Segmentation & Threshold Algorithm, Manual Segmentation & FI Model, Seg Model & FI Model) compared to the standard group (Manual Segmentation & Manual Fat Annotation) (Figure 3).

Both models were implemented using PyTorch (version 1.8.0) and trained on NVIDIA GeForce RTX GPU devices using CUDA (version 11.1). The Seg Model was trained using the MM Detection API (version 2.25.0). Pre-trained model weights were obtained from the MM Detection Model Zoo.

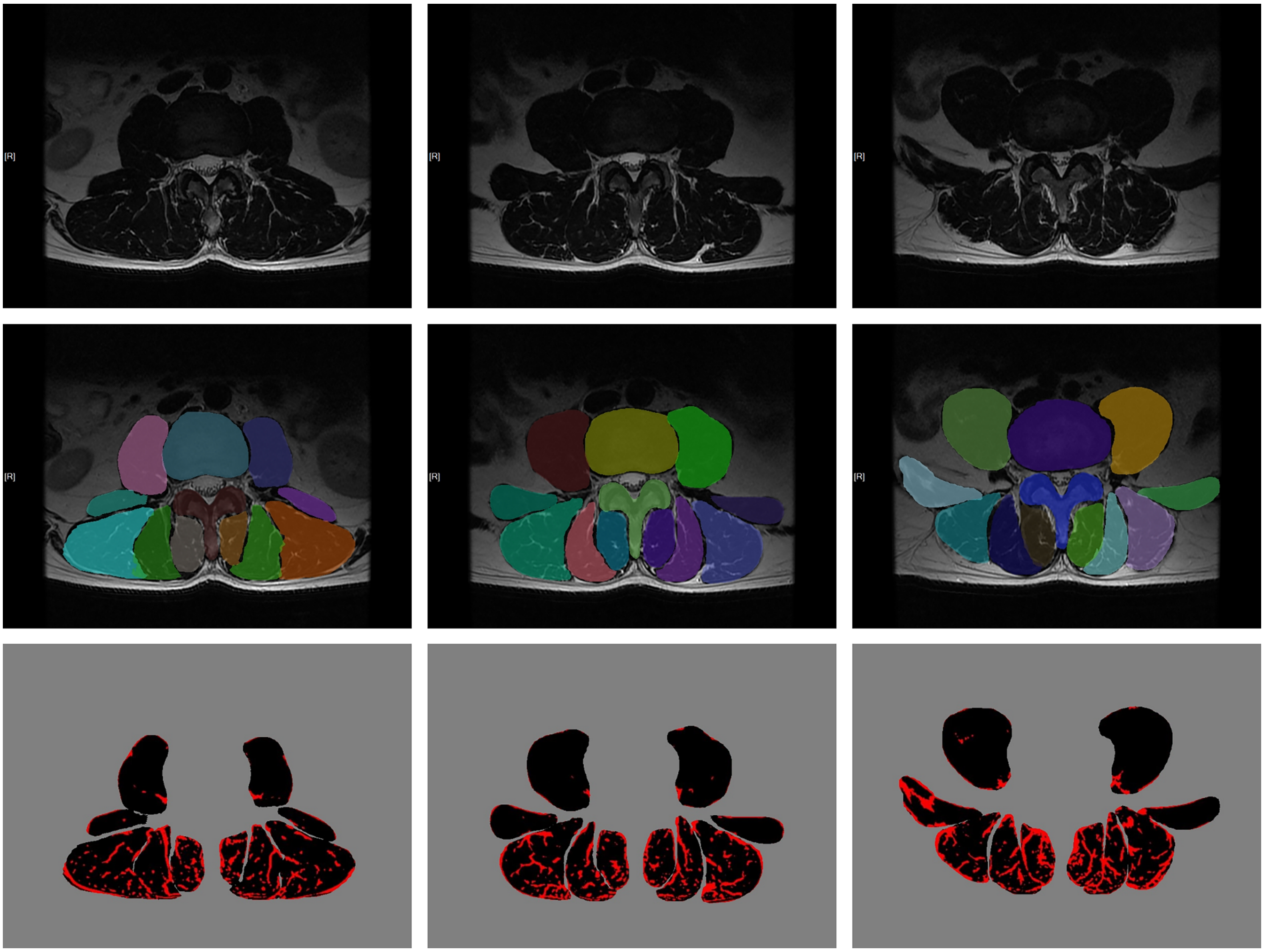

FIGURE 2 A 47-year-old male admitted with a 10-year history of low back pain, exacerbated with left lower limb pain for the past 2 months. From left to right are T2-weighted axial lumbar magnetic resonance imaging images at the L2/L3, L3/L4, and L4/L5 levels. From top to bottom are the original input images, lumbar structures segmentation model output images, and intramuscular fat infiltration detection model output images.

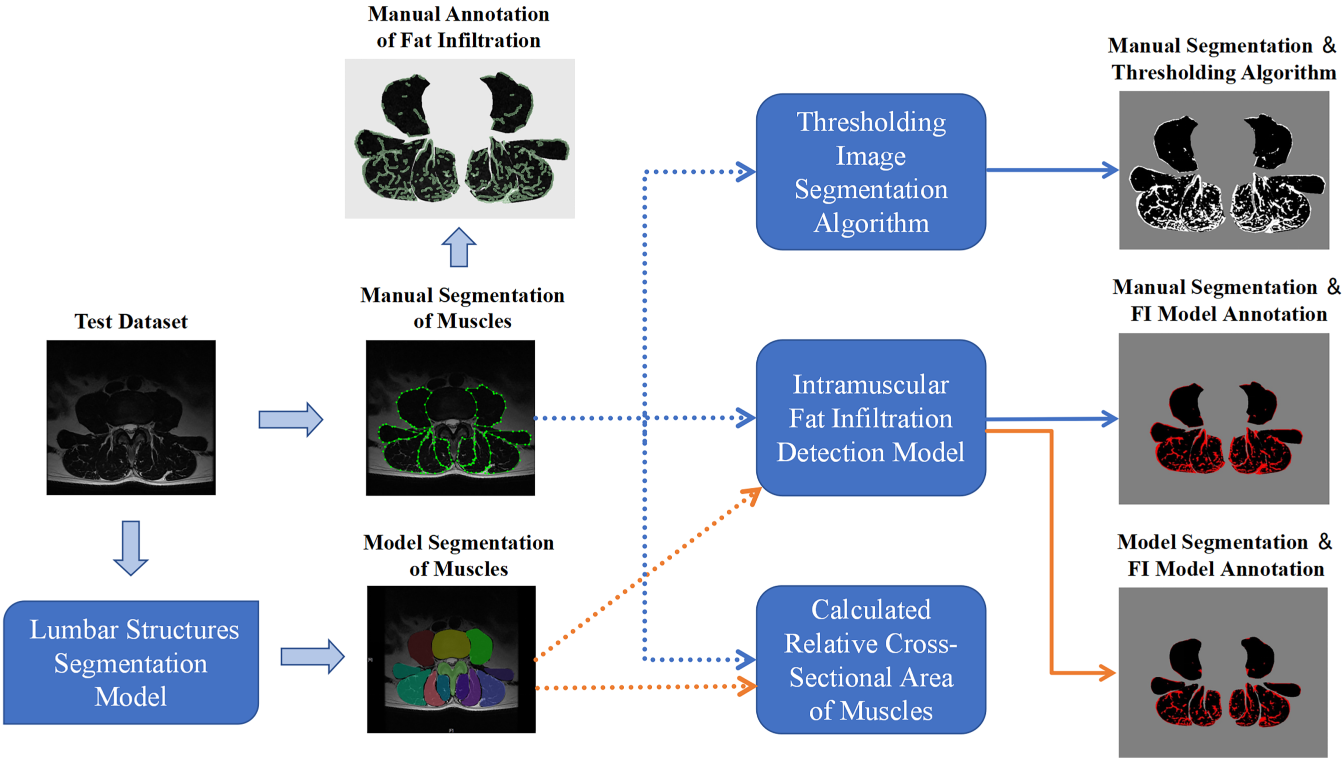
Manual Annotation of Fat Infiltration
Manual Segmentation of Muscles
Test Dataset
Lumbar Structures Segmentation Model
Model Segmentation of Muscles
Thresholding Image Segmentation Algorithm
Intramuscular Fat Infiltration Detection Model
Calculated Relative Cross-Sectional Area of Muscles
Manual Segmentation & Thresholding Algorithm
Manual Segmentation & FI Model Annotation
Model Segmentation & FI Model Annotation

FIGURE 3 The process for grouped comparison of measurements: Relative cross-sectional area and fat infiltration of paraspinal muscles. FI model: Intramuscular fat infiltration detection model.

5.1 | Thresholding image segmentation algorithm

Initially, utilizing ROIs delineated based on the reference standard, an in-depth analysis of the SI of each pixel within is conducted, aiming to derive an optimal threshold that minimizes the intra-class variance of pixel SIs. Subsequently, employing OpenCV, the primary signal thresholds for each muscle are computed, with subsequent statistical analysis of the thresholds' mean and standard deviation. For individual muscles, should the primary threshold exceed the range of ±1.96 standard deviations, the mean threshold derived from the remaining nine muscles is employed as its definitive threshold; conversely, if falling within the ±1.96 standard deviation range, the primary threshold is retained. Ultimately, pixels with signal intensities below the definitive threshold are classified as muscle tissue, while those surpassing it are designated as fat tissue. The "imshow" function (Python library) is utilized to visually render the binary masked image, thus facilitating result visualization.

5.2 | Clinical measurements

To measure parameters using ROIs, including: the total cross-sectional area (tCSA), fat cross-sectional area (fCSA), and fat infiltration (FI). tCSA represents the area of each ROI, fCSA denotes the area of fat tissue within the ROI, and FI is defined as the ratio of fCSA to tCSA. To mitigate the impact of height, weight, and body type on paraspinal muscle parameters, we computed the relative cross-sectional area (rCSA): defined as the ratio of paraspinal muscle tCSA to intervertebral disc tCSA. The DL model performed separate segmentation of muscles and CSA computation for both the left and right sides.

5.3 | Statistical analyses

To assess the performance of the model on test set, we evaluated the Intersection over Union (IoU) of predicted masks and ground truth masks, with a threshold set at 0.75. Recall and average precision (AP) were utilized to evaluate the Seg Model's identification efficacy, and Dice similarity coefficient (DSC) were utilized to report the consistency between raters. For the FI Model, Cohen's Kappa coefficient[24,25] was employed to assess the consistency between the model's predictions and ground truth, along with the calculation of Recall, Specificity, Precision, F1 Score, and the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve to evaluate its performance. The levels of agreement defined by Cohen's Kappa are as follows[26]: Less than 0: poor; 0–0.2: slight; 0.21–0.4: fair; 0.41–0.6: moderate; 0.61–0.80: substantial; and 0.81–1: almost-perfect agreement.

The rCSA and FI derived from the model's predictions need to be quantitatively compared with the standard group, utilizing the mean squared error (MSE) and mean absolute percentage error (MAPE). These statistical metrics are computed using the Scikit-learn (Python 3.9).

6 | RESULTS

Among 449 patients, cases with a history of lumbar spine surgery and internal fixation devices (*n* = 15), suboptimal image quality (*n* = 6), or contrast-enhanced MRI scans (*n* = 11) were excluded. A total of 417 patients were assessed, involving 1251 slices utilized for analysis. The cohort comprised 216 males and 201 females, aged between 18 and 89 years, with a mean age of (49 ± 15) years. They were randomly divided into 292 patients (70%) for training and validation, and 125 patients (30%) for internal testing. Within the internal testing set, there were 64 males and 61 females, with ages ranging from 19 to 88 years and a mean age of (48 ± 16) years (Table 1).

6.1 | Model performance evaluation

The performance of the Seg Model on the test set is shown in Table 2. In the internal test set, the model demonstrated excellent recognition performance, with an overall DSC of 0.92 ± 0.10, a recall of 92.60%, and an AP of 0.98. Its performance is very close to RATER (DSC 0.93 ± 0.08, recall 93.29%). The DSC of IL and MF reached 0.91 ± 0.06 and 0.91 ± 0.08, respectively, which are higher than RATER's results (0.90 ± 0.10, 0.88 ± 0.08). The recognition performance of VA (DSC 0.90 ± 0.05, recall 93.02%) is worse than RATER (0.95 ± 0.06, 95.73%). The Precision-Recall curves for each category are shown in Figure S3.

In the external test set, the overall DSC of the model reached 0.91 ± 0.09, with a recall of 90.77%, which is slightly lower than that of the internal test set. The AP is 0.98, equal to the internal test set (Figure 4). LO showed a recall rate of 73.46% and an AP of 0.80, significantly lower than the internal test set. The metrics for the remaining categories show slight decreases compared to the internal test set, and the Precision-Recall curves for each category are shown in Figure S3.

TABLE 1 Patient demographic and clinical characteristics.

Characteristics	All datasets (*n* = 417)	Test dataset (*n* = 125)
Age (years)[a]	49 ± 15 (18–89)	48 ± 16 (19–88)
Sex (male/female)	216/201	64/61
BMI (kg/m^2)[a]	24.20 ± 3.57	25.03 ± 4.38
Symptom duration (months)[a]	10.50 ± 11.92	12.25 ± 14.03

[a] Data are means ± standard deviations, with ranges in parentheses.

The performance of the FI Model is shown in Table 3. In the internal test set, the model demonstrates outstanding recognition efficacy: Recall 91.30%, Specificity 90.80%, F1 Score 0.82, and the consistency with the reference standard is substantial (Cohen's Kappa 0.76). In the external test set, the model's efficacy significantly decreases, with precision experiencing the most noticeable decline (74.27 internally and 57.67 externally), and the consistency also decreases significantly (Cohen's Kappa 0.61). In the ROC curve (Figure 5), it can be observed that the overall performance of the model is satisfactory, with the AUC of the internal test set (0.97) higher than that of the external AUC (0.93).

6.2 | Measurement of paraspinal muscles rCSA

The measurement results of the Seg Model are compared with the reference standard, and the rCSA of each muscle at each level are recorded in Table 4. As the lumbar levels descend, the areas of IL and LO gradually decrease, while those of PM and MF gradually increase. In the internal test set, the model's measurements are very close to the standard results: the MAPE for all muscles is less than 10%, except LO (10.89%); the MSE for PM, QL, and MF (0.0019, 0.0010, 0.0019) is small, while that for IL and LO (0.0048, 0.0030) is slightly larger. In the external test set, the model's predicted values are relatively close to the standard results; there is a slight increase in MSE and MAPE for each category compared to the internal test set, with the MAPE for LO (12.10%) and MF (11.03%) exceeding 10%.

6.3 | Intramuscular FI measurement

We compared three measurement methods with the standard group (Table 5). According to the measurement results of the standard

TABLE 2 The lumbar structures segmentation model performance.

	Rater (internal dataset)		Model (internal dataset)			Model (external dataset)		
	DSC	Recall (C75, %)	DSC	Recall (C75, %)	AP (C75)	DSC	Recall (C75, %)	AP (C75)
IVD	0.97 ± 0.05	100.00	0.97 ± 0.03	100.00	1.00	0.97 ± 0.02	100.00	1.00
PM	0.94 ± 0.07	99.55	0.95 ± 0.06	99.42	1.00	0.95 ± 0.05	99.62	1.00
QL	0.93 ± 0.09	95.09	0.92 ± 0.10	94.57	0.99	0.91 ± 0.09	93.46	0.98
IL	0.90 ± 0.10	92.49	0.91 ± 0.06	95.16	0.97	0.91 ± 0.06	93.46	0.97
LO	0.89 ± 0.09	79.64	0.88 ± 0.09	77.90	0.86	0.88 ± 0.10	73.46	0.80
MF	0.88 ± 0.08	86.98	0.91 ± 0.08	92.05	0.97	0.90 ± 0.08	88.46	0.95
VA	0.95 ± 0.06	95.73	0.90 ± 0.05	93.02	0.97	0.90 ± 0.05	92.31	0.96
Total	0.93 ± 0.08	93.29	0.92 ± 0.10	92.60	0.98	0.91 ± 0.09	90.77	0.98

Abbreviations: AP, average precision; C75, The intersection over union threshold was 0.75; DSC, Dice Similarity Coefficient; IL, iliocostalis; IVD, intervertebral disc; LO, longissimus; MF, multifidus; PM, psoas major; QL, quadratus lumborum; VA, vertebral arch.

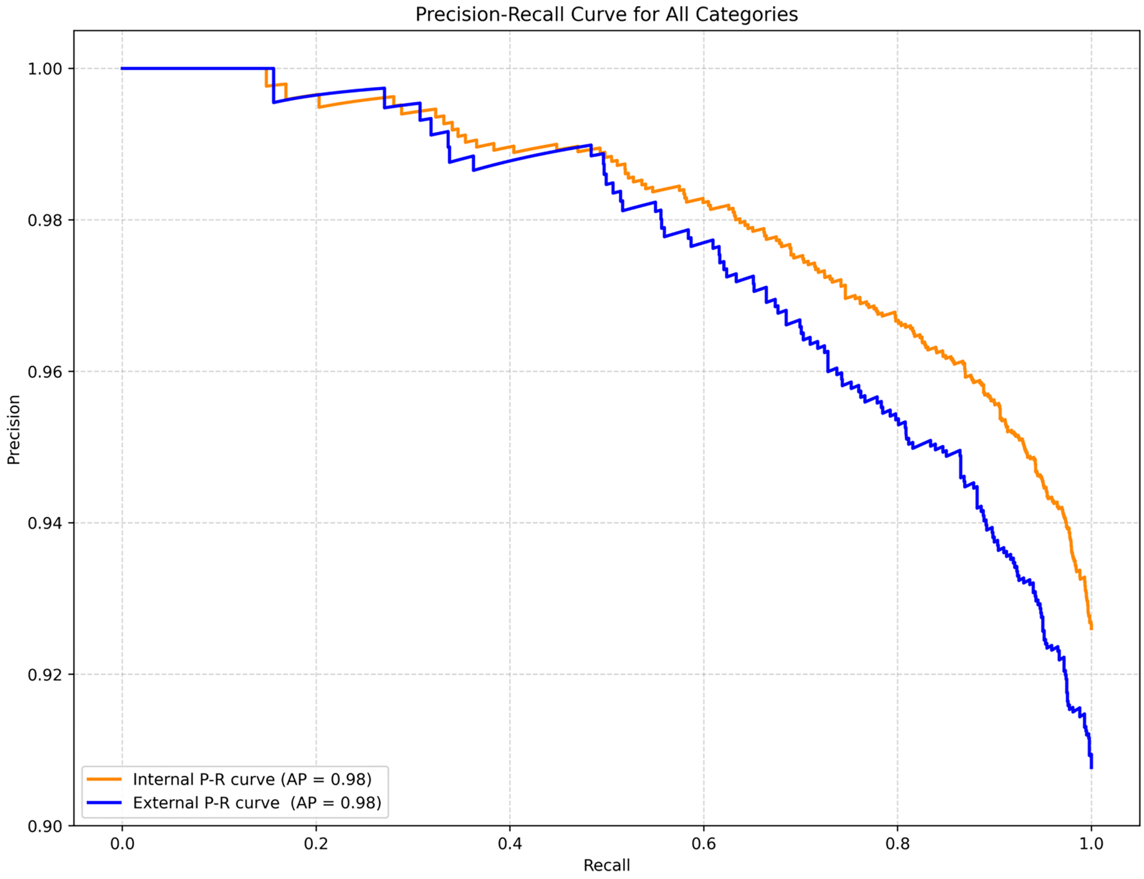

FIGURE 4 The Precision-Recall Curve for lumbar structures segmentation model performance on internal and external test datasets. AP, Average precision.

TABLE 3 The intramuscular fat infiltration detection model performance.

	Recall	Specificity	Precision	F1 score	Cohen's Kappa
Internal dataset	91.30	90.80	74.27	0.82	0.76
External dataset	89.42	84.08	57.67	0.70	0.61

group, the muscle with the highest FI is MF: 0.30 ± 0.06 in the internal test set and 0.28 ± 0.04 in the external test set; and the least is QL (0.16 ± 0.04 internally, 0.15 ± 0.04 externally). Among the three groups, the measurement results of FI Model group are closest to the standard group, with errors significantly lower than the other two groups. The measurement errors of Seg & FI Model group and thresholding algorithm group are similar, with slightly better performance in Seg & FI Model group.

In the internal test set, the measurement results of FI Model closely approximate the standard, with the highest MAPE for QL (35.22%) and the lowest for IL (22.39%). In the Seg & FI Model group, the results for PM (30.62%), IL (27.96%), and MF (30.87%) are relatively good, while QL (42.14%) and LO (44.18%) have slightly larger MAPEs. The MAPE of thresholding algorithm group is significantly higher compared to the other two groups.

In the external test set, the results of FI Model group are closest to the standard group, while the performance of Seg & FI Model group and thresholding algorithm group is essentially the same. Furthermore, the MAPE of both FI Model group and Seg & FI Model group increases significantly compared to the internal test set, while there is no significant change in thresholding algorithm group.

7 | DISCUSSION

Changes in paraspinal muscles can significantly impact the therapeutic efficacy of lumbar diseases.[27] Studies have identified impaired function of paraspinal muscles can affect lumbar alignment and compromise spinal biomechanics, thereby increasing the risk of intervertebral disc injury.[28,29] In addition, research has reported that paraspinal muscle atrophy and significant intramuscular fat infiltration exist in patients with low back pain.[30] Clearly, evaluating the condition of paraspinal muscle is essential before formulating treatment strategies for LDH.

Our developed lumbar structures segmentation model can automatically identify various paraspinal muscles, while the intramuscular fat

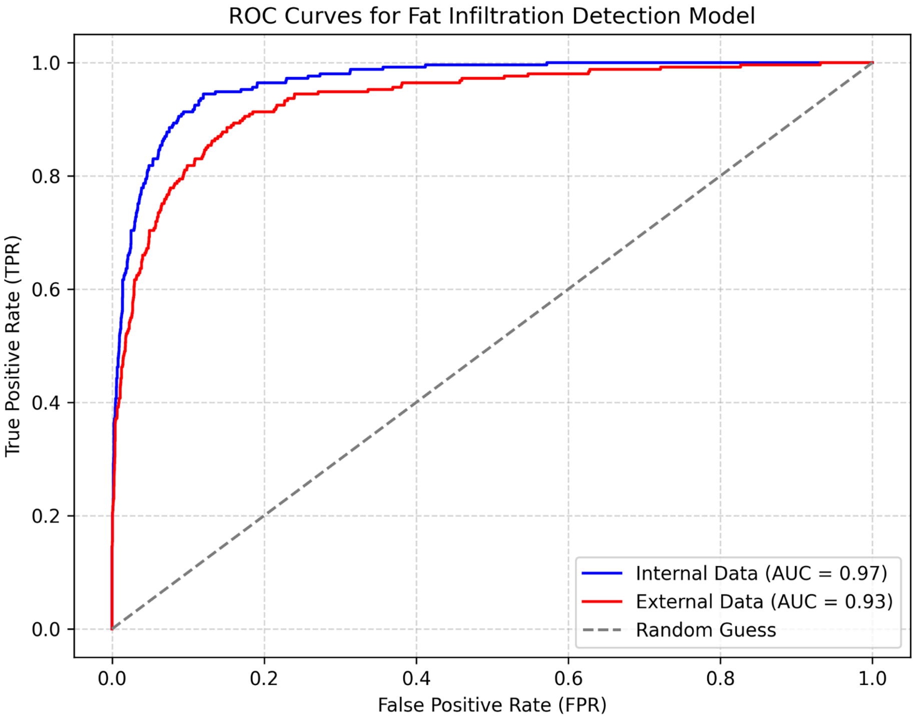

FIGURE 5 The receiver operating characteristic (ROC) curve for intramuscular fat infiltration detection model performance on internal and external test datasets. AUC, Area under the curve.

TABLE 4 The comparison of relative cross-sectional area (rCSA) of paraspinal muscles: standard measurement and model prediction.

	Standard rCSA			Seg Model rCSA				
	L2/L3	L3/L4	L4/L5	L2/L3	L3/L4	L4/L5	MSE of total	MAPE of total (%)
Internal data								
PM	0.80 ± 0.17	0.88 ± 0.17	0.95 ± 0.18	0.78 ± 0.16	0.86 ± 0.17	0.94 ± 0.16	0.0019	5.44
QL	0.27 ± 0.09	0.29 ± 0.09	0.32 ± 0.10	0.26 ± 0.09	0.28 ± 0.09	0.31 ± 0.08	0.0010	8.42
IL	0.75 ± 0.16	0.70 ± 0.16	0.55 ± 0.14	0.73 ± 0.15	0.67 ± 0.14	0.52 ± 0.13	0.0048	7.72
LO	0.48 ± 0.10	0.39 ± 0.09	0.34 ± 0.10	0.45 ± 0.09	0.37 ± 0.09	0.35 ± 0.08	0.0030	10.89
MF	0.23 ± 0.07	0.32 ± 0.08	0.52 ± 0.08	0.22 ± 0.06	0.29 ± 0.06	0.49 ± 0.07	0.0019	9.98
External data								
PM	0.79 ± 0.18	0.86 ± 0.16	0.97 ± 0.17	0.78 ± 0.17	0.84 ± 0.16	0.96 ± 0.16	0.0027	5.87
QL	0.27 ± 0.09	0.29 ± 0.10	0.33 ± 0.10	0.26 ± 0.08	0.28 ± 0.08	0.32 ± 0.09	0.0014	9.08
IL	0.76 ± 0.17	0.70 ± 0.15	0.58 ± 0.14	0.74 ± 0.14	0.64 ± 0.14	0.58 ± 0.13	0.0056	8.16
LO	0.50 ± 0.09	0.41 ± 0.09	0.37 ± 0.10	0.47 ± 0.08	0.40 ± 0.08	0.37 ± 0.10	0.0036	12.10
MF	0.23 ± 0.07	0.33 ± 0.07	0.53 ± 0.08	0.20 ± 0.06	0.30 ± 0.06	0.49 ± 0.07	0.0025	11.03

Abbreviations: IL, iliocostalis; IVD, intervertebral disc; LO, longissimus; MAPE, mean absolute percentage error; MF, multifidus; MSE, mean squared error; PM, psoas major; QL, quadratus lumborum; Seg Model: lumbar structures segmentation model.

infiltration detection model can mark the fat within these muscles. Both models have demonstrated excellent performance and reliable generalizability on internal and external test sets. Moreover, the consistency between the model's identification results and the annotations by spine specialists is remarkably high. These models are capable of assisting doctors in performing rapid and automated evaluations of the paraspinal muscles.

Numerous studies[31–33] have manually measured the rCSA of paraspinal muscles in patients with lumbar degeneration, which aligns with the measurement results of the DL model in this study. Manual measurements are time-consuming, making the development of a fully automated segmentation model essential. Shen et al.[11] reported their designed Spine Explorer model, which segmented structures in lumbar MRI. The Spine Explorer model recognition was limited to images of the L4-L5 intervertebral space. Our model recognition covered the mid-lumbar spine (L2-L5) region. Measurements across multiple intervertebral spaces tend to yield more reliable results.[34] Additionally, the test set used in our study had more samples and included an external test set to evaluate the model's generalizability. Li et al.[35] developed a model capable of segmenting the erector spinae and multifidus in lumbar MRI,

TABLE 5 Comparison of fat infiltration measurement results from three techniques with reference standard.

	Standard	FI model		Thresholding algorithm		Seg model and FI model	
	Fat infiltration	Fat infiltration	MAPE (%)	Fat infiltration	MAPE (%)	Fat infiltration	MAPE (%)
Internal dataset							
PM	0.19 ± 0.05	0.23 ± 0.09	24.05	0.27 ± 0.12	49.97	0.24 ± 0.10	30.62
QL	0.16 ± 0.04	0.21 ± 0.08	35.22	0.24 ± 0.11	57.26	0.22 ± 0.08	42.14
IL	0.25 ± 0.05	0.30 ± 0.07	22.39	0.35 ± 0.12	47.55	0.30 ± 0.09	27.96
LO	0.24 ± 0.05	0.31 ± 0.10	33.19	0.37 ± 0.11	62.42	0.34 ± 0.10	44.18
MF	0.30 ± 0.06	0.36 ± 0.09	28.65	0.43 ± 0.10	50.79	0.38 ± 0.10	30.87
External dataset							
PM	0.18 ± 0.05	0.25 ± 0.09	44.79	0.27 ± 0.10	55.81	0.26 ± 0.10	54.97
QL	0.15 ± 0.04	0.22 ± 0.10	50.69	0.22 ± 0.10	55.76	0.23 ± 0.10	57.06
IL	0.23 ± 0.05	0.30 ± 0.08	35.15	0.34 ± 0.12	54.83	0.32 ± 0.09	45.46
LO	0.23 ± 0.06	0.32 ± 0.12	45.24	0.36 ± 0.13	65.98	0.33 ± 0.13	49.42
MF	0.28 ± 0.04	0.37 ± 0.10	38.62	0.40 ± 0.13	49.12	0.37 ± 0.11	42.79

Abbreviations: FI Model, intramuscular fat infiltration detection model; IL, iliocostalis; LO, longissimus; MAPE, mean absolute percentage error; MF, multifidus; PM, psoas major; QL, quadratus lumborum; Seg Model, lumbar structures segmentation model.

demonstrating excellent performance. In contrast, our model, while maintaining robust performance, was able to identify a greater number of lumbar structures. Wang et al.[36] developed a model for three-dimensional morphological recognition of erector spinae, intervertebral discs, neural roots, and spinal cord at the L4-L5 level. The model proposed by Hess et al.[37] can recognize psoas major, multifidus, erector spinae, and quadratus lumborum in the axial plane, as well as vertebrae and intervertebral discs in the sagittal plane.

Our model improved upon previous approaches by distinguishing the erector spinae into two targets: the IL and LO. IL primarily supports spine extension and lateral flexion, while LO contributes not only to these movements but also plays a role in head and neck movement control. Their differing activation patterns and susceptibility to atrophy emphasize the need for separate assessment. This segmentation enhanced precision in muscle function evaluation and supports targeted treatment strategies.

Initial studies[38–40] employed visual grading for the assessment of fat infiltration. This method relies on subjective evaluation and lacks quantitative analysis. Subsequent research[11,15] proposed utilizing automatic thresholding algorithm for assessment. This technique is user-friendly and enables quantitative analysis of fat infiltration levels. However, the performance of thresholding technique is limited and is effective for simple differences in SI. Thresholding technique performs poorly when dealing with complex structures and blurred boundaries, especially when there is uneven SI distribution or partial volume effects. In comparison, DL models exhibits better robustness to variations in noise and image quality. Models can automatically learn complex features, identifying textures and high-level semantic information in images. Our research demonstrated that DL models outperform thresholding techniques. For Fat Infiltration Measurement of each muscle, the MAPE of DL models is significantly lower than that of the thresholding technique.

Mask R-CNN excels in instance segmentation tasks, effectively capturing the morphological characteristics and positional information of targets, making it highly suitable for lumbar structure segmentation tasks. U-net has significant advantages in semantic segmentation. Its encoder-decoder structure is adept at capturing detailed information in images, making it suitable for pixel-level recognition tasks. This characteristic enables U-net to perform exceptionally well in fat identification tasks.

Our research found that FI model's performance on external test set was subpar compared to its performance on internal test set, while the segmentation model's performance was comparable in both sets. This situation is related to the characteristics of how the models capture features, as mentioned earlier. The segmentation model relies on the morphology and location of targets for identification, which makes it less susceptible to variations in image quality. In contrast, fat identification focuses more on pixel-level features in images, which requires higher image quality.

The MR machines for the internal and external test sets come from different manufacturers, and different parameters were used during scanning. This resulted in variations in image quality. Comparing the performance on both test sets can reflect model's generalizability and test whether the model is applicable to any type of MRI machine. Our study showed that the segmentation model performed reliably under various conditions, indicating its potential for widespread clinical application. On the other hand, the FI model's performance relied on high image quality, which may affect its broader adoption.

Additionally, we observed that in fat infiltration measurement, the model's MAPE was relatively high, and the FI measurements for each category were higher than the standard. We analyzed the possible reasons as follows: The rater only labeled the fat within the muscle, while fat outside muscle or in the muscle gaps was not marked. However, the model sometimes mislabels fat outside the muscles, resulting in FI measurements higher than the standard, thereby increasing the MAPE. For such complex situations, the model's judgment ability still lags behind manual assessment. Nevertheless, compared to previous FI measurement methods, the DL model significantly reduces the error. Relying on its automatic and rapid characteristics, the model can provide a preliminary muscle quality assessment for patients undergoing MRI, thereby improving the efficiency of clinical workflow.

MR Dixon sequence and MR fingerprinting are novel methods for measuring FI, with high accuracy and stability. However, these techniques are not routinely performed in patients with LDH, and adding these extra examinations would increase the economic burden on patients. In contrast, DL models can perform evaluations using T1 or T2 sequences from conventional lumbar MRI scans. DL models can automatically process large volumes of images without manual intervention, offering advantages in speed and cost-efficiency.

To meet the demands of clinical settings, we integrated two models through an intermediate module and added visualization functionality, achieving fully automated output results. After inputting the image, the system can automatically output segmented muscle images, annotated fat images, as well as the measured values of muscle rCSA and intramuscular FI. This system has significant potential for application in medical imaging platforms. Before doctors diagnose MRI scans, the DL model can automatically label images, calculate rCSA and FI, and conduct preliminary assessments. This capability enhances diagnostic efficiency, and reduces human error.

For this study, there are still some limitations. Firstly, the recognition area of the DL model is from L2 to L5 segments, which does not cover the entire lumbar region. Given the richness and representativeness of muscles in the L2–L5 segments, previous studies have commonly chosen this area to assess paraspinal muscles.[41,42] Thus, our DL model evaluated paraspinal muscles based on the L2–L5 segments is reliable. Additionally, the performance of FI Model on external test sets showed a significant gap compared to internal test sets. We believe that enriching the database by including multicenter MRI images in the training set can address this issue. However, increasing the sample size implies the involvement of more doctors in the annotation process, which poses another challenge. Lastly, due to limited available time, the expert completed annotations only for the internal test set. Since there was only one rater for the external test set, we assessed the inter-rater annotation repeatability by having the rater independently annotate the dataset twice, with a 2-week interval between annotations.

8 | CONCLUSION

The Seg Model and FI Model demonstrate outstanding performance, with lower error in FI measurement compared to thresholding algorithms. The model can automatically identify images and output results, providing effective assistance to physicians in assessing the condition of patients' paraspinal muscles.

AUTHOR CONTRIBUTIONS

Yefu Xu and Shijie Zheng conceptualized the study, designed the research, and drafted the manuscript for intellectual content. Shijie Zheng and Qingyi Tian trained the deep-learning model, conducted data analysis, and evaluated its performance. Wenqing Li and Zhuoyan Kou were responsible for data collection and literature review. Yefu Xu developed deep-learning algorithms. Xiaotao Wu and Xinhui Xie edited and reviewed the manuscript, ensuring its integrity. All authors contributed significantly to the study and approved the final version of the manuscript for publication.

ACKNOWLEDGMENTS

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work was supported by the National Natural Science Foundation of China (82372473).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Our code and example images are available on supplementary documents.

ORCID

Yefu Xu https://orcid.org/0000-0003-0794-2116

REFERENCES

1. Zhang AS, Xu A, Ansari K, et al. Lumbar disc herniation: diagnosis and management. *Am J Med*. 2023;136(7):645-651.
2. Veresciagina K, Ambrozaitis KV, Spakauskas B. Health-related quality-of-life assessment in patients with low back pain using SF-36 questionnaire. *Medicina (Kaunas)*. 2007;43(8):607-613.
3. Xiao Y, Fortin M, Ahn J, Rivaz H, Peters TM, Battié MC. Statistical morphological analysis reveals characteristic paraspinal muscle asymmetry in unilateral lumbar disc herniation. *Sci Rep*. 2021;11(1):15576.
4. Ranger TA, Cicuttini FM, Jensen TS, et al. Are the size and composition of the paraspinal muscles associated with low back pain? A systematic review. *Spine J*. 2017;17(11):1729-1748.
5. Hildebrandt M, Fankhauser G, Meichtry A, Luomajoki H. Correlation between lumbar dysfunction and fat infiltration in lumbar multifidus muscles in patients with low back pain. *BMC Musculoskelet Disord*. 2017;18(1):12.
6. Brinjikji W, Luetmer PH, Comstock B, et al. Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. *AJNR Am J Neuroradiol*. 2015;36(4):811-816.
7. Oliveira CB, Maher CG, Pinto RZ, et al. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. *Eur Spine J*. 2018;27(11):2791-2803.
8. Park MS, Moon SH, Kim TH, et al. Paraspinal muscles of patients with lumbar diseases. *J Neurol Surg A Cent Eur Neurosurg*. 2018;79(4):323-329.
9. Stevens S, Agten A, Timmermans A, Vandenabeele F. Unilateral changes of the multifidus in persons with lumbar disc herniation: a systematic review and meta-analysis. *Spine J*. 2020;20(10):1573-1585. doi:10.1016/j.spinee.2020.04.007

10. Fortin M, Lazáry À, Varga PP, McCall I, Battié MC. Paraspinal muscle asymmetry and fat infiltration in patients with symptomatic disc herniation. *Eur Spine J*. 2016;25(5):1452-1459. doi:10.1007/s00586-016-4503-7
11. Shen H, Huang J, Zheng Q, et al. A deep-learning-based, fully automated program to segment and quantify major spinal components on axial lumbar spine magnetic resonance images. *Phys Ther*. 2021; 101(6):pzab041.
12. Vitale J, Sconfienza LM, Galbusera F. Cross-sectional area and fat infiltration of the lumbar spine muscles in patients with back disorders: a deep learning-based big data analysis. *Eur Spine J*. 2024;33(1):1-10.
13. Niemeyer F, Zanker A, Jonas R, Tao Y, Galbusera F, Wilke HJ. An externally validated deep learning model for the accurate segmentation of the lumbar paravertebral muscles. *Eur Spine J*. 2022;31(8): 2156-2164. doi:10.1007/s00586-022-07320-w
14. Dickx N, Cagnie B, Achten E, Vandemaele P, Parlevliet T, Danneels L. Differentiation between deep and superficial fibers of the lumbar multifidus by magnetic resonance imaging. *Eur Spine J*. 2010;19(1):122-128.
15. Fortin M, Omidyeganeh M, Battié MC, Ahmad O, Rivaz H. Evaluation of an automated thresholding algorithm for the quantification of paraspinal muscle composition from MRI images. *Biomed Eng Online*. 2017;16(1):61. doi:10.1186/s12938-017-0350-y
16. Zhang R, He A, Xia W, et al. Deep learning-based fully automated segmentation of regional muscle volume and spatial intermuscular fat using CT. *Acad Radiol*. 2023;30(10):2280-2289.
17. Zhou S, Nie D, Adeli E, et al. Semantic instance segmentation with discriminative deep supervision for medical images. *Med Image Anal*. 2022;82:102626.
18. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*; 2020;42(2):386-397.
19. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A. (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, Springer; 2015;9351:234-241.
20. Cai Z, Vasconcelos N. Cascade r-cnn: delving into high quality object detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT; 2018:6154-6162.
21. Lin T, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, HI; 2017: 936-944.
22. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, NV; 2016:770-778.
23. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(6):1137-1149.
24. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37-46. doi:10.1177/001316446002000104
25. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70:213-220.
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
27. Stanuszek A, Jędrzejek A, Gancarczyk-Urlik E, et al. Preoperative paraspinal and psoas major muscle atrophy and paraspinal muscle fatty degeneration as factors influencing the results of surgical treatment of lumbar disc disease. *Arch Orthop Trauma Surg*. 2022;142(7): 1375-1384. doi:10.1007/s00402-021-03754-x
28. Tamai K, Chen J, Stone M, et al. The evaluation of lumbar paraspinal muscle quantity and quality using the Goutallier classification and lumbar indentation value. *Eur Spine J*. 2018;27:1005-1012.
29. Bailey JF, Miller SL, Khieu K, et al. From the international space station to the clinic: how prolonged unloading may disrupt lumbar spine stability. *Spine J*. 2018;18:7-14.
30. Seyedhoseinpoor T, Taghipour M, Dadgoo M, et al. Alteration of lumbar muscle morphology and composition in relation to low back pain: a systematic review and meta-analysis. *Spine J*. 2022;22(4):660-676.
31. Wang W, Guo Y, Li W, Chen Z. The difference of paraspinal muscle between patients with lumbar spinal stenosis and normal middle-aged and elderly people, studying by propensity score matching. *Front Endocrinol*. 2022;13:1080033.
32. Ranger TA, Cicuttini FM, Jensen TS, Heritier S, Urquhart DM. Paraspinal muscle cross-sectional area predicts low back disability but not pain intensity. *Spine J*. 2019;19(5):862-868.
33. Goubert D, De Pauw R, Meeus M, et al. Lumbar muscle structure and function in chronic versus recurrent low back pain: a cross-sectional study. *Spine J*. 2017;17(9):1285-1296.
34. Urrutia J, Besa P, Lobos D, Andia M, Arrieta C, Uribe S. Is a single-level measurement of paraspinal muscle fat infiltration and cross-sectional area representative of the entire lumbar spine? *Skeletal Radiol*. 2018;47:939-945.
35. Li H, Luo H, Liu Y. Paraspinal muscle segmentation based on deep neural network. *Sensors*. 2019;19(12):2650.
36. Wang M, Su Z, Liu Z, et al. Deep learning-based automated magnetic resonance image segmentation of the lumbar structure and its adjacent structures at the L4/5 level. *Bioengineering*. 2023;10(8):963.
37. Hess M, Allaire B, Gao KT, et al. Deep learning for multi-tissue segmentation and fully automatic personalized biomechanical models from BACPAC clinical lumbar spine MRI. *Pain Med*. 2023;24(Suppl 1): S139-S148.
38. Meyer DC, Wieser K, Farshad M, Gerber C. Retraction of supraspinatus muscle and tendon as predictors of success of rotator cuff repair. *Am J Sports Med*. 2012;40(10):2242-2247. doi:10.1177/0363546512457587
39. Wall LB, Teefey SA, Middleton WD, 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.
40. Kang CH, Shin MJ, Kim SM, Lee SH, Lee CS. MRI of paraspinal muscles in lumbar degenerative kyphosis patients and control patients with chronic low back pain. *Clin Radiol*. 2007;62(5):479-486. doi:10.1016/j.crad.2006.12.002
41. Crawford RJ, Elliott JM, Volken T. Change in fatty infiltration of lumbar multifidus, erector spinae, and psoas muscles in asymptomatic adults of Asian or Caucasian ethnicities. *Eur Spine J*. 2017;26:3059-3067.
42. Hebert JJ, Kjaer P, Fritz JM, Walker BF. The relationship of lumbar multifidus muscle morphology to previous, current, and future low back pain: a 9-year population-based prospective cohort study. *Spine*. 2014;39(17):1417-1425.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Xu Y, Zheng S, Tian Q, et al. Deep learning-based structure segmentation and intramuscular fat annotation on lumbar magnetic resonance imaging. *JOR Spine*. 2024;7(3):e70003. doi:10.1002/jsp2.70003