Received: 6 January 2023 | Revised: 3 July 2023 | Accepted: 3 August 2023

DOI: 10.1002/jsp2.1276

RESEARCH ARTICLE

JOR Spine OPEN ACCESS

Deep learning-based detection and classification of lumbar disc herniation on magnetic resonance images

Weicong Zhang[1] | Ziyang Chen[1] | Zhihai Su[1] | Zhengyan Wang[2] | Jinjin Hai[2] | Chengjie Huang[1] | Yuhan Wang[1] | Bin Yan[2] | Hai Lu[1]

[1]Department of Spinal Surgery, The Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, Guangdong, China

[2]Henan Key Laboratory of Imaging and Intelligent Processing, PLA Strategic Support Force Information Engineering University, Zhengzhou, China

Correspondence
Hai Lu, Department of Spinal Surgery, The Fifth Affiliated Hospital of Sun Yat-sen University, 52 East Meihua Road, Xiangzhou District, Zhuhai 519000, Guangdong, China.
Email: lvhai@mail.sysu.edu.cn

Bin Yan, Henan Key Laboratory of Imaging and Intelligent Processing, PLA Strategic Support Force Information Engineering University, Zhengzhou 450000, China.
Email: tom.yan@gmail.com

Funding information
Joint Funding Scheme 2022 for Scientific Research Projects (FDCT-GDST Projects) by the Science and Technology Development Fund of Macau and the Department of Science and Technology of Guangdong Province, Grant/Award Number: 2022A0505020019; Zhuhai City Industry-University-Research Cooperation Project, Guangdong Province, China, Grant/Award Number: ZH22017002210017PWC; Zhuhai City Innovation and Innovation Team Project, Guangdong Province, China, Grant/Award Number: ZH0406190031PWC

Abstract

Background: The severity assessment of lumbar disc herniation (LDH) on MR images is crucial for selecting suitable surgical candidates. However, the interpretation of MR images is time-consuming and requires repetitive work. This study aims to develop and evaluate a deep learning-based diagnostic model for automated LDH detection and classification on lumbar axial T2-weighted MR images.

Methods: A total of 1115 patients were analyzed in this retrospective study; both a development dataset (1015 patients, 15 249 images) and an external test dataset (100 patients, 1273 images) were utilized. According to the Michigan State University (MSU) classification criterion, experts labeled all images with consensus, and the final labeled results were regarded as the reference standard. The automated diagnostic model comprised Faster R-CNN and ResNeXt101 as the detection and classification network, respectively. The deep learning-based diagnostic performance was evaluated by calculating mean intersection over union (IoU), accuracy, precision, sensitivity, specificity, F1 score, the area under the receiver operating characteristics curve (AUC), and intraclass correlation coefficient (ICC) with 95% confidence intervals (CIs).

Results: High detection consistency was obtained in the internal test dataset (mean IoU = 0.82, precision = 98.4%, sensitivity = 99.4%) and external test dataset (mean IoU = 0.70, precision = 96.3%, sensitivity = 97.8%). Overall accuracy for LDH classification was 87.70% (95% CI: 86.59%–88.86%) and 74.23% (95% CI: 71.83%–76.75%) in the internal and external test datasets, respectively. For internal testing, the proposed model achieved a high agreement in classification (ICC = 0.87, 95% CI: 0.86–0.88, $P < 0.001$), which was higher than that of external testing (ICC = 0.79, 95% CI: 0.76–0.81, $P < 0.001$). The AUC for model classification was 0.965 (95% CI: 0.962–0.968) and 0.916 (95% CI: 0.908–0.925) in the internal and external test datasets, respectively.

Weicong Zhang, Ziyang Chen, Zhihai Su, and Zhengyan Wang contributed significantly to this paper and are considered co-first authors.

Conclusions: The automated diagnostic model achieved high performance in detecting and classifying LDH and exhibited considerable consistency with experts' classification.

KEYWORDS
classification, deep learning, detection, lumbar disc herniation, magnetic resonance images

1 | INTRODUCTION

Lumbar disc herniation (LDH)-induced low back pain (LBP) is a significant global public health and socioeconomic concern.[1,2] Related studies reported that approximately 70%–80% of adults suffer from LBP, which considerably impacts their daily lives and work.[3] LDH patients are grouped into three categories based on their clinical features: Group I exhibits clear indications for operative intervention; Group II comprises patients who do not require surgery; and Group III comprises those with less clear surgical indications.[4] Magnetic resonance (MR) imaging, an essential diagnostic tool, can accurately assess LBP and measure the severity of LDH, which should influence the clinical decision of lumbar discectomy, especially in group III patients.[4,5]

However, MR image reading is highly dependent on the subjective judgments of observers, and interpreting these images is repetitive, time-consuming, and burdensome. Stacked images of each disc layer must be analyzed one after another.[6] Consequently, various automated diagnostic models have been explored in diagnosing lumbar diseases for potential applicability over the past decade. In 2010, Alomari et al.[7] developed three probabilistic Gaussian models according to Gibbs distribution for dichotomous classification of intervertebral disc degeneration. In 2016, He et al.[8] accurately and quickly identified the presence or absence of foramina stenosis using a synchronized superpixel representation model. However, due to the limited available data size, the aforementioned diagnostic models, which were based on conventional machine learning, require further validation.

With the advancement of graphics-processing-unit technology, deep learning (DL) methods have become prominent. Some intelligent diagnostic models that are based on DL algorithms have been developed, and researchers have observed that they are effective in diagnosing degenerative spinal diseases, such as spondylolisthesis,[9] fresh osteoporotic vertebral fractures,[10] lumbar spinal stenosis,[11] and LDH.[12] Notably, Jamaludin et al.[13] proposed an automated diagnostic system based on a convolutional neural network (CNN) to effectively grade lumbar degenerative diseases such as disc degeneration and central canal stenosis. However, many DL models for diagnosing lumbar disc degenerative diseases have been designed only for binary classification (disease-absent vs. disease-present).[12,14] In actual clinical practice, analyzing complex biomedical images requires a detailed description of LDH severity to guide surgical decision-making, especially for patients with unclear surgical indications. Although an automated LDH classification model has been developed to classify LDH into four different morphology categories, namely normal disc, bulging disc, protrusion, and extrusion,[15] surgeons are still subjected to uncertainty in the necessity and approach of operation without detailed information on disc herniation size.

Because the Michigan State University (MSU) classification is often considered an objective criterion for LDH surgical selection on lumbar axial MR images,[16] it has been widely applied in clinical practice.[17,18] To comprehensively describe LDH severity according to disc herniation size, this study aimed to develop an automated diagnostic model using the MSU classification. Thus, surgeons can accurately and rapidly make appropriate treatment options.

2 | MATERIALS AND METHODS

This study was approved by the Ethics Review Committees of two participating institutions: the Fifth Affiliated Hospital of Sun Yat-sen University and the Third Affiliated Hospital of Southern Medical University (No. 2021 K53-1). The waiver of informed consent was granted in light of the study's retrospective nature and the minimal risk involved.

2.1 | Dataset collection

Figure 1 depicts the data allocation and processing flowchart. A development dataset of 15 249 axial T2-weighted (T2W) MR images was collected from 1015 lumbar spine patients administrated at the Fifth Affiliated Hospital of Sun Yat-sen University (Zhuhai, Guangdong Province, China), from January 2015 to May 2019. Additionally, an external test dataset utilized for external validation comprised 1273 axial T2W MR images from 100 lumbar spine patients in the Third Affiliated Hospital of Southern Medical University (Guangzhou, Guangdong Province, China), from June 2016 to December 2017. The development dataset was randomly divided into an internal training set (12 012 images) and an internal test set (3237 images). Moreover, all datasets were anonymized and numbered.

The inclusion criteria for screening study participants were LBP patients who underwent routine lumbar spine MR scans.

The exclusion criteria were as follows:

1. inflammation, fractures, or deformities of the vertebral column;
2. a history of prior or concurrent malignancies;
3. previous spinal surgery and metallic implants at the corresponding level;
4. suboptimal image quality.

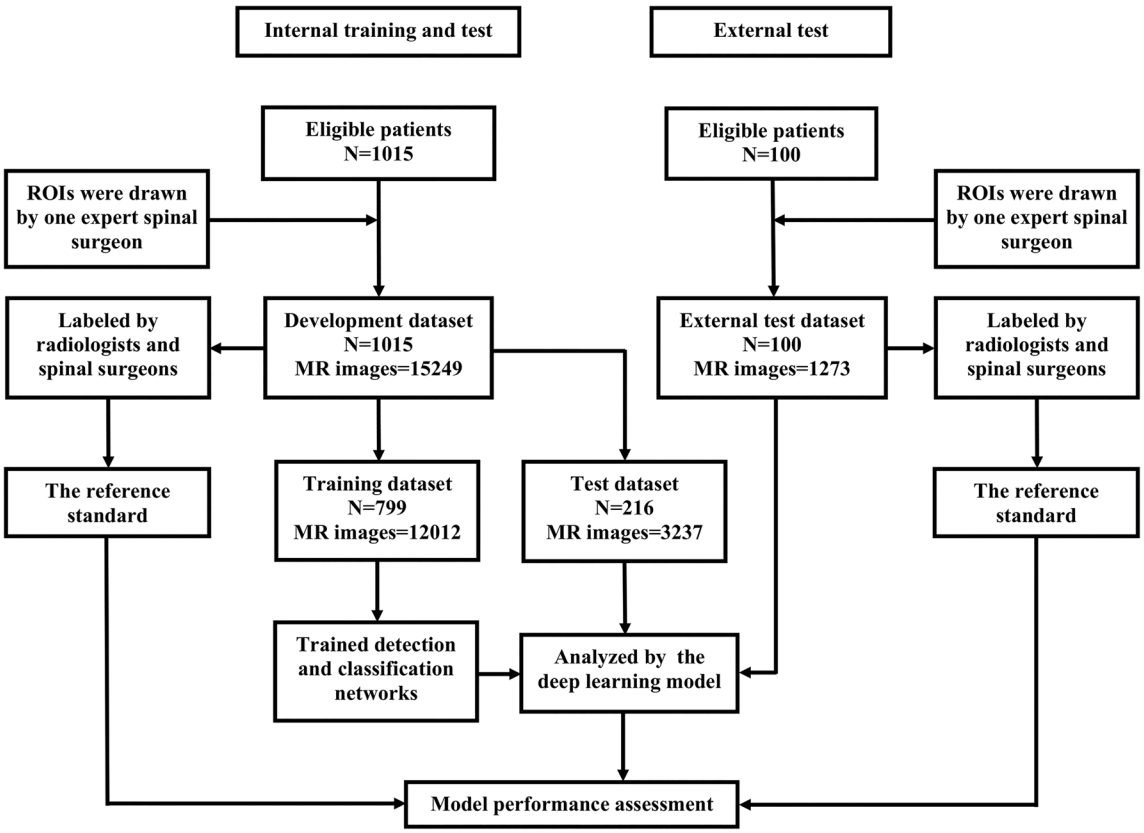

FIGURE 1 Data allocation and processing flowchart. MR, magnetic resonance; *N*, number of patients; ROIs, regions of interest.

All patients underwent lumbar MR scans on a 3.0-T platform (Magnetom Verio; Siemens, Germany) using T2W turbo spin echo sequences. Scanning parameters differed between development and external test datasets (echo time: 94–120 ms, repetition time: 3500–3775 ms, slice thickness: 4–4.5 mm, field of view: 153 × 153 mm^2, bandwidth: 250 kHz). All MR images were stored as Digital Imaging and Communications in Medicine (DICOM) files for subsequent analyses.

2.2 | Dataset labeling and reference standard

A spinal surgeon interested in reading lumbar MR images drew bounding boxes manually to segment regions of interest (ROIs) at all images using the MRIcro software. All LDH images were subsequently interpreted by multiple radiology experts and senior spinal surgeons; all experts reached a consensus as per the standard MSU classification criterion.[16] The aforementioned diagnostic results obtained from image analysis were regarded as the reference standard. The manual analysis of an axial lumbar MR image consumed approximately 2.3–6.5 s.

The LDH size of the MSU classification is reported in three precise increments, described as Grades 1, 2, and 3. All grades are measured by a herniated disc position relative to one intra-facet line drawn transversely across the lumbar canal, starting and ending at the medial edges of the left and right facet joint articulations.[16] This measurement requires only a single T2W axial MR image cut corresponding to the maximal herniation. We divided LDH into four grades based on the MSU classification, including the normal disc. Grade 0 denotes no disc herniation; Grade 1 denotes disc herniation that extends up to (less than) 50% of the distance from the non-herniated posterior aspect of the disc to the intra-facet line; Grade 2 denotes disc herniation that extends more than 50% of the aforementioned distance; and Grade 3 denotes the herniation that extends completely beyond the intra-facet line. Supporting Information indicate detailed information on the size grading of the classification utilized herein.

2.3 | DL model construction

We constructed a DL model by selecting Faster R-CNN as the detection network and ResNeXt101 as the classification network, respectively. First, Faster R-CNN located the effective image area for judging LDH. Subsequently, the ResNeXt101 classified LDH using the effective area after data augmentation. Figure 2 illustrates the detailed flowchart of LDH detection and classification. The entire process consumed only 20 ms on average to automatically analyze an axial lumbar MR image.

2.3.1 | Detection network architecture

We trained a Faster R-CNN network of two modules to detect the lumbar disc region from the MR images. The region proposal network (RPN), a fully convolutional network, served as the first module to

FIGURE 2 Lumbar disc herniation detection and classification flowchart. MR, magnetic resonance; ROI, region of interest.

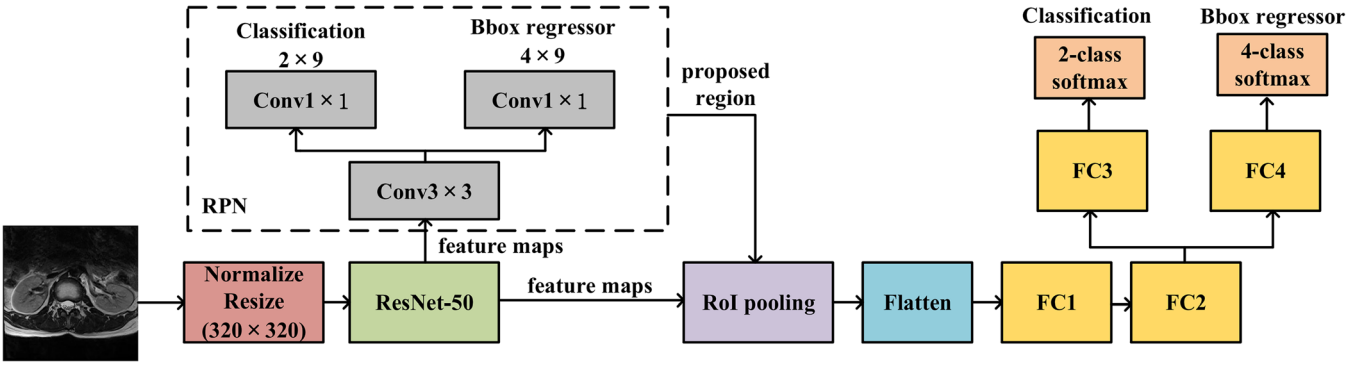

FIGURE 3 Faster R-CNN network for lumbar region detection. Bbox, bounding box; Conv, convolutional kernel; FC, fully connected layer; RoI, region of interest; RPN, region proposal network.

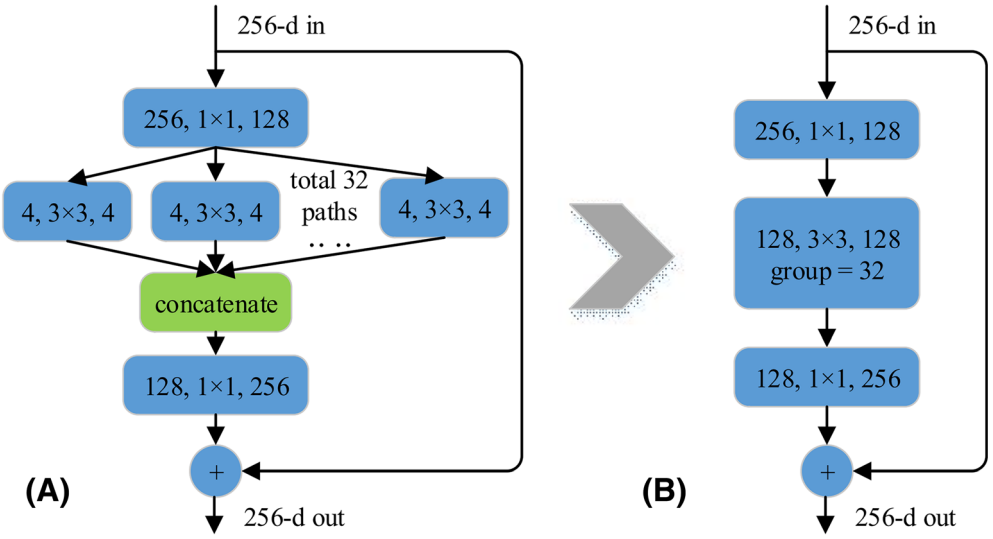

FIGURE 4 ResNeXt block. (A) A block is equivalent to B, implemented as grouped convolutions. (B) A block of ResNeXt network.

generate each image's regional proposals, which were subsequently fed into the second module. The second module extracted features from the region proposals using a Fast R-CNN detector. These features were further separately utilized as inputs of a 2-class softmax classifier to detect lumbar disc regions and of a 4-class softmax classifier to improve bounding box coordinates. The architecture of the Faster R-CNN approach is illustrated in Figure 3. The core idea entailed sharing the same convolutional layers for RPN and the Fast R-CNN detector. Therefore, the lumbar axial MR images passed through the CNN only once to produce and refine region proposals. Figure 3 indicates that the images should be resized to a standard size of before inputting the lumbar axial MR images into the Faster R-CNN network. For more technical details, please refer to the original Faster R-CNN study.[19]

2.3.2 | Classification network architecture

ResNeXt101, the classification network, was enhanced based on ResNet.[20,21] For the proposed ResNeXt classification network, we set the neuron number of the last fully connected layer to 4, the same as the LDH category. The structure diagram of the ResNeXt network block is illustrated in Figure 4. First, to decrease the data's feature dimension while also reducing the number of model parameters, a convolution layer with a 1×1 convolution kernel was utilized. Subsequently, we utilized a group convolution with a group number of 32 and a convolution kernel of 3×3 for feature extraction. Furthermore, a convolution layer with a 1×1 convolution kernel was applied to increase the feature dimension. Finally, the output of 1×1 the convolution layer was added to the original input of the block to obtain the final result. The ResNeXt input data

were the Faster R-CNN detected lumbar ROIs. Subsequently, we unified the image size to 96 × 160 (the maximum size of lumbar ROIs) using mirror filling because the input image size needed to be the same. Meanwhile, we standardized each image's pixel values to improve the model's generalization ability. Random translation and rotation of images were performed for data augmentation.

2.4 | Statistical analysis

The performance of the automated diagnostic model was evaluated on two test datasets, namely internal and external. Based on intersection over union (IoU), we assessed ROIs detection for all bounding boxes with a 0.5 threshold. Mean IoU, precision, and sensitivity were applied to evaluate detection performance. After constructing the confusion matrix, we measured the classification performance of the diagnostic model by calculating accuracy, precision, sensitivity, specificity, F1 score, the area under the receiver operating characteristics curve (AUC), and intraclass correlation coefficient (ICC) with their 95% confidence intervals (CI).[22] ICC defined levels of agreement as follows: a 0–0.4 value is poor; 0.41–0.75 is moderate; 0.76–0.9 is good; and 0.91–1 is excellent. A bilateral $P < 0.05$ was considered statistically significant. Python was used to perform analyses.

3 | RESULTS

3.1 | Patient characteristics and reference standard in datasets

The development dataset comprised 1015 patients of 49 ± 14 (age range: 13–88) years. The external test dataset consisted of 100 patients with an average age of 53 ± 16 (age range: 17–84) years. Table 1 lists the patient characteristics for both datasets. Table 2 provides an overview of standard reference grading for different datasets. Predominantly, all datasets were biased toward Grades 0 and 1. In the development dataset, Grades 2 and 3 accounted for 12.6% (1508 of 12 012) and 11.6% (377 of 3237) in the training and test sets, respectively. Regarding the external test dataset, 17.6% (224 of 1273) of labels were classified as Grades 2 and 3.

TABLE 1 Patient characteristics of development and external test datasets.

Characteristics	Development dataset ($n = 1015$)	External test dataset ($n = 100$)
Age (years)	49 ± 14 (13–88)	53 ± 16 (17–84)
Women	492 (48.5)	51 (51.0)
Men	523 (51.5)	49 (49.0)

Note: Data are expressed as *n* (%), or mean ± standard deviation (range).

3.2 | ROIs detection

IoU measures the overlap between detected and reference standard labels (bounding boxes). Concerning mean IoU, the proposed model detection achieved 0.82 and 0.70 for the internal and external test datasets, respectively. At a 0.5 IoU matching threshold, the model correctly recognized the ROIs with excellent detection performance. Bounding boxes generated by model detection displayed 98.4% precision and 99.4% sensitivity in the internal test dataset. Similar results were found in the external test dataset with 96.3% precision and 97.8% sensitivity.

3.3 | Internal test dataset classification

Figure 5 indicates the relevant statistics of the four MSU classification grades. Table 3 provides detailed results pertaining to the internal test dataset's accuracy, precision, sensitivity, specificity, F1 score, and ICC. Generally, predictive accuracy levels for Grades 0, 1, 2, and 3 were 92.86% (95% CI: 91.99%–93.77%, 3006 of 3237), 88.60% (95% CI: 87.52%–89.72%, 2868 of 3237), 94.66% (95% CI: 93.88%–95.44%, 3064 of 3237), and 99.30% (95% CI: 98.98%–99.59%, 3214 of 3237). Furthermore, the F1 score of the four grades was 92.98% (95% CI: 92.09%–93.90%), 84.64% (95% CI: 83.11%–86.24%), 76.20% (95% CI: 72.78%–79.55%), and 56.60% (95% CI: 38.97%–72.41%) in sequence. A good consistency (ICC: 0.87, 95% CI: 0.86–0.88, $P < 0.001$) was observed between the model classification and the reference standard. For the internal test, the model classification AUC was 0.965 (95% CI: 0.962–0.968).

3.4 | External test dataset classification

Table 4 includes the external test dataset's accuracy, precision, sensitivity, specificity, F1 score, and ICC details. The predictive accuracy of the four grades was 87.90% (95% CI: 86.04%–89.77%, 1119 of 1273), 78.00% (95% CI: 75.69%–80.43%, 993 of 1273), 86.33%

TABLE 2 Reference standard classification of the LDH at axial lumbar T2W MR images, *n* (%).

LDH severity	Internal training dataset	Internal test dataset	External test dataset
Grade 0	5925 (49.3)	1647 (50.9)	550 (43.2)
Grade 1	4579 (38.1)	1213 (37.5)	499 (39.2)
Grade 2	1374 (11.5)	347 (10.7)	196 (15.4)
Grade 3	134 (1.1)	30 (0.9)	28 (2.2)
Total	12 012 (100)	3237 (100)	1273 (100)

Abbreviations: LDH, lumbar disc herniation; MR, magnetic resonance.

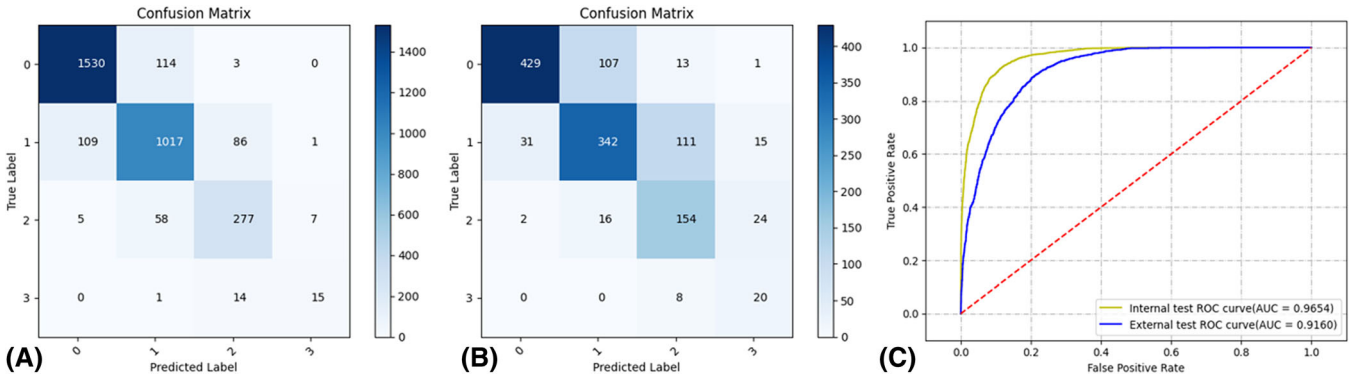

FIGURE 5 Confusion matrices and receiver operating characteristics (ROC) curves of lumbar disc herniation (LDH) automated classification. Confusion matrices of LDH grading in the internal test dataset (A) and external test dataset (B). ROC curves of LDH grading (C).

TABLE 3 Classification performance of the LDH diagnostic model in the internal test dataset.

Predicted grading	Accuracy, % (95% CI)	Precision, % (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	F1 Score, % (95% CI)	ICC (95% CI)
LDH	87.70 (86.59–88.86)	87.70 (86.59–88.86)	87.70 (86.59–88.86)	95.90 (95.53–96.29)	87.70 (86.59–88.86)	0.87 (0.86–0.88)
Grade 0	92.86 (91.99–93.77)	93.07 (91.81–94.37)	92.90 (91.67–94.14)	92.83 (91.54–94.18)	92.98 (92.09–93.90)	
Grade 1	88.60 (87.52–89.72)	85.46 (83.43–87.49)	83.84 (81.76–86.06)	91.45 (90.22–92.67)	84.64 (83.11–86.24)	
Grade 2	94.66 (93.88–95.44)	72.89 (68.55–77.27)	79.83 (75.52–84.01)	96.44 (95.76–97.13)	76.20 (72.78–79.55)	
Grade 3	99.30 (98.98–99.59)	65.22 (43.76–86.60)	50.00 (31.22–67.51)	99.75 (99.57–99.94)	56.60 (38.97–72.41)	

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient; LDH, lumbar disc herniation.

TABLE 4 Classification performance of the LDH diagnostic model in the external test dataset.

Predicted grading	Accuracy, % (95% CI)	Precision, % (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	F1 Score, % (95% CI)	ICC (95% CI)
LDH	74.23 (71.83–76.75)	74.23 (71.83–76.75)	74.23 (71.83–76.75)	91.41 (90.61–92.25)	74.23 (71.83–76.75)	0.79 (0.76–0.81)
Grade 0	87.90 (86.04–89.77)	92.86 (90.56–95.21)	78.00 (74.47–81.56)	95.44 (93.90–96.99)	84.78 (82.39–87.19)	
Grade 1	78.00 (75.69–80.43)	73.55 (69.36–77.86)	68.54 (64.46–72.92)	84.11 (81.45–86.77)	70.95 (67.62–74.46)	
Grade 2	86.33 (84.43–88.36)	53.58 (48.01–59.73)	78.57 (72.62–84.53)	87.74 (85.73–89.89)	63.90 (58.81–68.93)	
Grade 3	96.23 (95.15–97.30)	33.33 (21.05–45.83)	71.43 (53.66–89.22)	96.79 (95.77–97.80)	45.45 (31.91–58.59)	

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient; LDH, lumbar disc herniation.

(95% CI: 84.43%–88.36%, 1099 of 1273), and 96.23% (95% CI: 95.15%–97.30%, 1225 of 1273) in sequence. In addition, the F1 score of the four grades was 84.78% (95% CI: 82.39%–87.19%), 70.95% (95% CI: 67.62%–74.46%), 63.90% (95% CI: 58.81%–68.93%), and 45.45% (95% CI: 31.91%–58.59%) in sequence. A good overall agreement in the external test dataset was shown with a 0.79 ICC value (95% CI: 0.76–0.81; $P < 0.001$), which was relatively lower than that in the internal test dataset. The model classification AUC was 0.916 (95% CI: 0.908–0.925) for the external test.

4 | DISCUSSION

The MR scan plays a crucial role in assessing LDH and in the definite diagnosis of LBP.[4,5,23] However, analyzing MR images can be repetitive, intensive, and time-consuming.[6] Therefore, an intelligent diagnostic model that can reliably grade LDH severity is imperative. Herein, we developed an automated diagnostic model based on the MSU classification criterion to detect and classify LDH on axial lumbar MR images.[16] The results demonstrated that the detection and classification networks, which represented two components of the automated diagnostic model, performed effectively. Bounding boxes generated by the detection network exhibited satisfactory consistency with reference labels in the internal test dataset (precision: 98.4%, sensitivity: 99.4%) and external test dataset (precision: 96.3%, sensitivity: 97.8%). Regarding classification, the overall LDH grading accuracy was 87.70% (95% CI: 86.59%–88.86%) in the internal test dataset and 74.23% (95% CI: 71.83%–76.75%) in the external test dataset, respectively. The model achieved a strong classification agreement with radiology experts and spinal surgeons both for the

internal test (ICC: 0.87, 95% CI: 0.86–0.88, $P < 0.001$) and external test (ICC: 0.79, 95% CI: 0.76–0.81, $P < 0.001$).

Various classification methods are available for LDH diagnosis. Nevertheless, only a few methods could provide a simple, universal description of disc herniation size and location that enabled surgeons to accurately and objectively assess disease severity and optimize surgical decisions. Fardon and Milette.[24] divided disc herniation into protrusion, extrusion, and sequestration. Nevertheless, because it is difficult to identify the exact size and location of disc herniation without reference to the MR images, this classification exhibited a significant disadvantage. Wiltse et al.[25] and Mysliwiec et al.[16] proposed some classification methods to overcome the aforementioned issue. Wiltse et al.[25] classified disc herniation into distinct zones (medial-lateral, axial plane) and levels (caudocranial, sagittal plane) according to anatomic location. However, the disc herniation size was ranked as either normal, mild, moderate, moderately severe, or severe; this subjective ranking fails to objectively define and consider the individual anatomic differences. In contrast, the MSU classification developed by Mysliwiec et al.[16] determined LDH size by measuring the relative position between the herniation and an intrafacet reference line; thus, it is straightforward and objective. A herniation size can be classified as 1-2-3. Several clinical studies have confirmed that in regard to selecting appropriate LDH patients who urgently require surgery, the MSU classification is more consistent with that of surgeons.[17,18] Patients with size-2 or size-3 lesions may account for a higher proportion of satisfactory surgical outcomes when using the MSU classification to guide surgical options.[16] To our knowledge, this is the first LDH diagnostic criterion closely linked with severity evaluation and corresponding clinical decision-making.

Due to the development of artificial intelligence, automated LDH diagnosis has become a reality. The automated diagnostic models initially focused on binary classification, either disease-absent or disease-present.[12,14] Alomari et al.[14] developed an LDH diagnostic model that was based on a binary Bayesian classifier to automatically categorize each disc as normal or herniated on lumbar MR images. Their model achieved an average accuracy of 92.5% over 65 cases. Recently, DL has exhibited prospective utility in diagnostic radiology.[26,27] Using deep-learning algorithms, a computerized diagnostic model constructed by Tsai et al.[12] could diagnose disc herniations with an average precision of 92.4%. Although these previous models displayed high-quality results, they failed to realize that LDH multiclass classification critically impacts surgical selection.

The MSU classification could objectively divide LDH into several size-based typologies and help the patient be selected for single-level discectomy.[16] Notably, this classification suggests no surgery for patients with size-1 lesions, whereas patients with size-2 or size-3 lesions may benefit from micro discectomies.[16] Relying on the MSU classification, we proposed a novel, automated model that utilized axial lumbar MR images to group LDH into four grades; thus, we optimized severity assessment and the selection of suitable operative candidates. The proposed model achieved 87.7% and 74.2% average accuracy for internal and external test datasets, respectively, which indicated its superiority over other automated diagnostic models with multiclass classification (67.1%–81.2% for neural foraminal stenosis; and 70.6%–86.9% for central canal stenosis).[28–30] We observed 74.2%–84.2% accuracy in LDH grading without automatic detection in our prior research using a ResNet50 framework.[30] By comparison, this study further constructed a fully automated detection and classification workflow for grading LDH, and it obtained more optimal accuracy. In addition, we found the trained model consumed only 20 ms on average to automatically complete this workflow in test datasets. The model was much faster than manual analysis, which required approximately 2.3–6.5 s. This time range of manual analysis is roughly similar to that needed to interpret one MR image (3–4 s).[31] Because the model exhibited satisfactory classification results in internal and external test datasets collected from two hospitals, we believe it is a promising tool to be generalized across various medical institutes. Next, we will improve the deep-learning algorithm to further enhance our model's robustness and generalization. A more extensive multicenter longitudinal study will further validate the clinical effectiveness and reliability of this automated diagnostic model in our subsequent research. Beyond the preoperative application of the intelligent model, we can utilize the model to postoperatively regrade the herniation at different time points and to evaluate the surgical outcome by comparing herniation changes before and after surgery. Thus, the model will play a crucial role in longitudinal follow-up.

There are some limitations in our study. First, this retrospective study may be biased by potential selection. Further prospective studies should be strictly conducted to confirm the clinical feasibility and effectiveness of this LDH automated diagnostic model. Second, LDH Grades 2 and 3 exhibited a relatively low diagnostic precision, which is probably attributed to data class imbalance. In future studies, we will collect higher-quality image data, especially from high-grade LDH patients, to further enhance diagnostic precision. Third, we utilized only axial T2W MR images for DL. We would extract more informative features for automatic image analysis by integrating multiple MR sequences and sagittal images with this model. In addition, all bounding boxes were drawn by one expert spinal surgeon. Through a consensus labeling approach, more accurate ROIs can be obtained. Finally, we will optimize this diagnostic model with the latest deep-learning algorithms to explore more lumbar diseases, such as lumbar spinal stenosis, spondylolisthesis, and vertebral fracture.

5 | CONCLUSIONS

In this study, we developed an automated diagnostic model that can accurately and rapidly classify LDH on axial lumbar T2W MR images using deep-learning algorithms. The model effectively detected and graded LDH based on the MSU classification criterion. Moreover, it achieved a high consistency with specialists in classifying LDH, which can objectively enable spinal surgeons to evaluate disease severity and select appropriate treatment options.

AUTHOR CONTRIBUTIONS

Weicong Zhang and Zhihai Su conceptualized and designed the study, and drafted the manuscript for intellectual content. Ziyang Chen and

Zhengyan Wang constructed the deep-learning model and analyzed the data. Jinjin Hai conducted statistical data analyses and evaluated the performance of the deep-learning model. Chengjie Huang and Yuhan Wang collected the data and conducted a literature search. Bin Yan developed deep-learning algorithms. Hai Lu and Bin Yan edited and reviewed the manuscript, and took responsibility for the integrity of the work. All authors contributed to the study and approved the manuscript for publication.

ACKNOWLEDGMENTS

This study was supported by Zhuhai City Innovation and Innovation Team Project, Guangdong Province, China (ZH0406190031PWC), Zhuhai City Industry-University-Research Cooperation Project, Guangdong Province, China (ZH220170022100017PWC), and Joint Funding Scheme 2022 for Scientific Research Projects (FDCT-GDST Projects) by the Science and Technology Development Fund of Macau and the Department of Science and Technology of Guangdong Province (2022A0505020019).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Hai Lu https://orcid.org/0000-0002-4212-3282

REFERENCES

1. Knezevic NN, Candido KD, Vlaeyen JWS, Van Zundert J, Cohen SP. Low back pain. *Lancet*. 2021;*398*:78-92.
2. Amin RM, Andrade NS, Neuman BJ. Lumbar disc herniation. *Curr Rev Musculoskelet Med*. 2017;*10*:507-516.
3. Paolucci T, Attanasi C, Cecchini W, Marazzi A, Capobianco SV, Santilli V. Chronic low back pain and postural rehabilitation exercise: a literature review. *J Pain Res*. 2018;*12*:95-107.
4. Weber H. Lumbar disc herniation. A controlled, prospective study with ten years of observation. *Spine (Phila Pa 1976)*. 1983;*8*:131-140.
5. Qaseem A, Wilt TJ, McLean RM, et al. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. *Ann Intern Med*. 2017;*166*:514-530.
6. Markotić V, Pojužina T, Radančević D, Miljko M, Pokrajčić V. The radiologist workload increase; where is the limit?: mini review and case study. *Psychiatr Danub*. 2021;*33*:768-770.
7. Alomari RS, Corso JJ, Chaudhary V, Dhillon G. Computer-aided diagnosis of lumbar disc pathology from clinical lower spine MRI. *Int J Comput Assist Radiol Surg*. 2010;*5*:287-293.
8. He X, Yin Y, Sharma M, Brahm G, Mercado A, Li S. Automated diagnosis of neural foraminal stenosis using synchronized superpixels representation. *MICCAI*. 2016;*9901*:335-343.
9. Cai Y, Leung S, Warrington J, Pandey S, Shmuilovich O, Li S. Direct spondylolisthesis identification and measurement in MR/CT using detectors trained by articulated parameterized spine model. *SPIE Proc*. 2017;*10133*:1013319.
10. Yabu A, Hoshino M, Tabuchi H, et al. Using artificial intelligence to diagnose fresh osteoporotic vertebral fractures on magnetic resonance images. *Spine J*. 2021;*21*:1652-1658.
11. Lewandrowskl KU, Muraleedharan N, Eddy SA, et al. Feasibility of deep learning algorithms for reporting in routine spine magnetic resonance imaging. *Int J Spine Surg*. 2020;*14*:S86-S97.
12. Tsai JY, Hung IY, Guo YL, et al. Lumbar disc herniation automatic detection in magnetic resonance imaging based on deep learning. *Front Bioeng Biotechnol*. 2021;*9*:708137.
13. Jamaludin A, Kadir T, Zisserman A. SpineNet: automated classification and evidence visualization in spinal MRIs. *Med Image Anal*. 2017;*41*:63-73.
14. Alomari RS, Corso JJ, Chaudhary V, Dhillon G. Toward a clinical lumbar CAD: herniation diagnosis. *Int J Comput Assist Radiol Surg*. 2011;*6*:119-126.
15. Lehnen NC, Haase R, Faber J, et al. Detection of degenerative changes on mr images of the lumbar spine with a convolutional neural network: a feasibility study. *Diagnostics (Basel)*. 2021;*11*:902.
16. Mysliwiec LW, Cholewicki J, Winkelpleck MD, Eis GP. MSU classification for herniated lumbar discs on MRI: toward developing objective criteria for surgical selection. *Eur Spine J*. 2010;*19*:1087-1093.
17. Beyaz SG, Ülgen AM, Kaya B, et al. A novel combination technique: three points of epiduroscopic laser neural decompression and percutaneous laser disc decompression with the Ho:YAG laser in an MSU classification 3AB herniated disc. *Pain Pract*. 2020;*20*:501-509.
18. d'Ercole M, Innocenzi G, Ricciardi F, Bistazzoni S. Prognostic value of Michigan State University (msu) classification for lumbar disc herniation: is it suitable for surgical selection? *Int J Spine Surg*. 2021;*15*:466-470.
19. 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*:1137-1149.
20. He K, Zhang X, Ren S, Sun J. *Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE; 2016:770-778.
21. Xie S, Girshick R, Dollár P, Tu Z, He K. *Aggregated residual transformations for deep neural networks. IEEE conference on computer vision and pattern recognition*. IEEE; 2017:5987-5995.
22. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;*86*:420-428.
23. Kaliya-Perumal AK, Luo CA, Yeh YC, Tsai YF, Chen MJ, Tsai TT. Reliability of the Michigan State University (MSU) classification of lumbar disc herniation. *Acta Ortop Bras*. 2018;*26*:411-414.
24. Fardon DF, Milette PC. Combined task forces of the North American Spine Society, American Society of Spine Radiology, and American Society of Neuroradiology. Nomenclature and classification of lumbar disc pathology. Recommendations of the combined task forces of the North American Spine Society, American Society of Spine Radiology, and American Society of Neuroradiology. *Spine (Phila PA 1976)*. 2001;*26*:E93-E113.
25. Wiltse LL, Berger PE, McCulloch JA. A system for reporting the size and location of lesions in the spine. *Spine (Phila Pa)*. 1976;*1997*(22):1534-1537.
26. Moawad AW, Fuentes DT, ElBanan MG, et al. Artificial intelligence in diagnostic radiology: where do we stand, challenges, and opportunities. *J Comput Assist Tomogr*. 2022;*46*:78-90.
27. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging*. 2019;*49*:939-954.
28. Lu JT, Pedemonte S, Bizzo B, et al. Deep spine: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. Machine learning for healthcare conference (MLHC). *PMLR*. 2018;*85*:403-419.
29. Won D, Lee HJ, Lee SJ, Park SH. Spinal stenosis grading in magnetic resonance imaging using deep convolutional neural networks. *Spine (Phila Pa)*. 1976;*2020*(45):804-812.
30. Su ZH, Liu J, Yang MS, et al. Automatic grading of disc herniation, central canal stenosis and nerve roots compression in lumbar magnetic resonance image diagnosis. *Front Endocrinol (Lausanne)*. 2022;*13*:890371.

31. McDonald RJ, Schwartz KM, Eckel LJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad Radiol*. 2015;*22*:1191-1198.

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: Zhang, W., Chen, Z., Su, Z., Wang, Z., Hai, J., Huang, C., Wang, Y., Yan, B., & Lu, H. (2023). Deep learning-based detection and classification of lumbar disc herniation on magnetic resonance images. *JOR Spine*, *6*(3), e1276. https://doi.org/10.1002/jsp2.1276