Open Access Full Text Article

ORIGINAL RESEARCH

Deep-Learning-Based Radiomics to Predict Surgical Risk Factors for Lumbar Disc Herniation in Young Patients: A Multicenter Study

Zheng Fan[1], Tong Wu[1], Yang Wang[2], Zhuoru Jin[1], Tong Wang[3], Da Liu[1]

[1]Department of Orthopedics, Shengjing Hospital of China Medical University, Shenyang, People's Republic of China; [2]Department of Orthopedics, China Medical University Shenyang Fourth People's Hospital, Shenyang, People's Republic of China; [3]Department of Radiology, Shengjing Hospital of China Medical University, Shenyang, People's Republic of China

Correspondence: Da Liu, Department of Orthopedics, Shengjing Hospital of China Medical University, No. 36, Sanhao Street, Heping District, Shenyang, 110004, People's Republic of China, Email spinecmu@163.com

Objective: The aim of this study is to develop and validate a deep-learning radiomics model for predicting surgical risk factors for lumbar disc herniation (LDH) in young patients to assist clinicians in identifying surgical candidates, alleviating symptoms, and improving prognosis.

Methods: A retrospective analysis of patients from two medical centers was conducted. From sagittal and axial MR images, the regions of interest were handcrafted to extract radiomics features. Various machine-learning algorithms were employed and combined with clinical features, resulting in the development of a deep-learning radiomics nomogram (DLRN) to predict surgical risk factors for LDH in young adults. The efficacy of the different models and the clinical benefits of the model were compared.

Results: We derived six sets of features, including clinical features, radiomics features (Rad_SAG and Rad_AXI) and deep learning features (DL_SAG and DL_AXI) from sagittal and axial MR images, as well as fused deep-learning radiomics (DLR) features. The support vector machine(SVM) algorithm exhibited the best performance. The area under the curve (AUC) of DLR in the training and testing cohorts of 0.991 and 0.939, respectively, were significantly better than those of the models developed with radiomics (Rad_SAG=0.914 and 0.863, Rad_AXI=0.927 and 0.85) and deep-learning features(DL_SAG=0.959 and 0.818, DL_AXI=0.960 and 0.811). The AUC of DLRN coupled with clinical features(ODI, Pfirrmann grade, SLRT, MMFI, and MSU classification) were 0.994 and 0.941 in the training and testing cohorts, respectively. Analysis of the calibration and decision curves demonstrated good agreement between the predicted and observed outcomes, and the use of the DLRN to predict the need for surgical treatment of LDH demonstrated significant clinical benefits.

Conclusion: The DLRN established based on clinical and DLR features effectively predicts surgical risk factors for LDH in young adults, offering valuable insights for diagnosis and treatment.

Keywords: deep learning, radiomics, nomogram, lumbar disc herniation, surgical treatment, young adults

Introduction

Low back pain (LBP) is one of the most disabling musculoskeletal conditions globally.[1,2] Various factors contribute to LBP, with lumbar disc herniation (LDH) being the most common.[3] LDH can compress the spinal nerve roots, resulting in radiculopathy, which is often accompanied by severe lower limb pain, numbness, and diminished muscle strength.[4] Although most patients with LDH experience symptomatic relief after conservative treatment, some experience progressive neurological dysfunction, necessitating early surgical intervention to reduce short-term symptoms and improve long-term outcomes. LDH often affects working adults aged 20–40, and the burden on families and society caused by prolonged activity restrictions or disabilities is a significant issue.[5,6] The assessment of medical history and clinical signs for surgical risk, along with the interpretation of lumbar MRI findings, is a complex and labor-intensive process. With the

increasing number of young patients with LDH, there is an urgent need to devise methods for the swift and accurate identification of patients suitable for surgical treatment.

LDH diagnosis requires a comprehensive assessment of the patient's medical history, clinical signs, and imaging findings. Lumbar MRI plays a crucial role in evaluating the location, size, direction, and free nucleus of the disc, and provides essential information for surgical decision making. The Philips mDixon sequence is a rapid and noninvasive technique that can generate four images in a one-stop scan (water, fat, in-phase, and out-of-phase images), significantly reducing scanning time and simultaneously improving image quality through water–fat separation and B0 correction, and has been widely applied in clinical practice.[7]

Radiomics extracts numerous quantitative features from regions of interest (ROIs) within medical images and correlates them with clinical conclusions using machine learning (ML). Radiomics utilizes not only the shape, intensity (first-order), and texture (second-order) features of the original images, but also more advanced and abstract features such as wavelets.[8,9] Radiomics is widely employed in spinal disease diagnosis, including in the differential diagnosis of metastatic tumors and osteoporosis.[10]

Deep learning, which is a subset of artificial intelligence (AI) algorithms, offers techniques for extracting features from images to detect and classify objects. Neural networks are particularly well suited for medical imaging problems.[11] Moreover, ChatGPT demonstrates the potential to supplement and enhance neurosurgical practices.[12] In a deep-learning system, vertebrae and intervertebral discs can be detected by training a neural network, including feature extraction and image segmentation. The deep-learning radiomics nomogram (DLRN) is a graphical representation of a model that integrates radiomic and clinical features, enhancing the predictive efficiency for disease diagnosis. It can also incorporate deep-learning features autonomously learned from a convolutional neural network (CNN), thereby further improving the efficacy of the model.[13] Zhao et al[7] developed a deep-learning radiomics (DLR) model based on 222 radiomics features extracted from lumbar MR mDixon images for predicting vertebral osteoporosis; their results demonstrated the DLR model had excellent efficacy. Zhang et al[14] also reported on the efficacy of a model that combines CT-derived DLR features with traditional radiomics features in distinguishing acute and chronic vertebral compression fractures. The results showed that the DLRN model significantly outperformed traditional radiomics models.

The evaluation process for the indications for LDH surgery is complex. To date, the application of radiomics and deep-learning fusion technology to assess the surgical risk factors for LDH in young patients has not been explored. Therefore, in this study, the extraction and fusion of radiomics and deep-learning features from sagittal and axial lumbar MR mDxion sequences was combined with clinical features to develop a DLRN, followed by external validation to swiftly and precisely identify young adult patients suitable for surgical treatment.

Material and Methods
Patients
A retrospective cohort of 1066 patients aged 16–44 with LBP, treated at Shengjing Hospital of China Medical University from January 1, 2022, to January 1, 2024, was used as the training cohort, including 404 patients diagnosed with LDH who underwent surgery. Additionally, 191 patients with LBP treated at the China Medical University Shenyang Fourth People's Hospital were selected as the test cohort, with 47 patients diagnosed with LDH who underwent surgery. The inclusion criteria were as follows: (1) age 16–44 years and (2) availability of complete clinical data or ability to supplement by telephone follow-up. The exclusion criteria were (1) blurred lumbar MR (n = 28); (2) congenital diseases or spinal deformities (n = 22); (3) history of spinal fractures, tuberculosis, or metastatic spinal tumors (n = 21); and (4) previous spinal surgery (n = 9). The workflow of this study is illustrated in Figure 1. This study was approved by the Shengjing Hospital Ethics Committee (approval number KYCS2024040), and a waiver of informed consent was granted given the retrospective nature of the study and the minimal risk involved.

Dataset Collection
The clinical data of surgical inpatients and outpatients were extracted from the Neusoft Hospital Information Management System (HIS, version 5.0), and the integrity of the clinical information was ensured through telephone
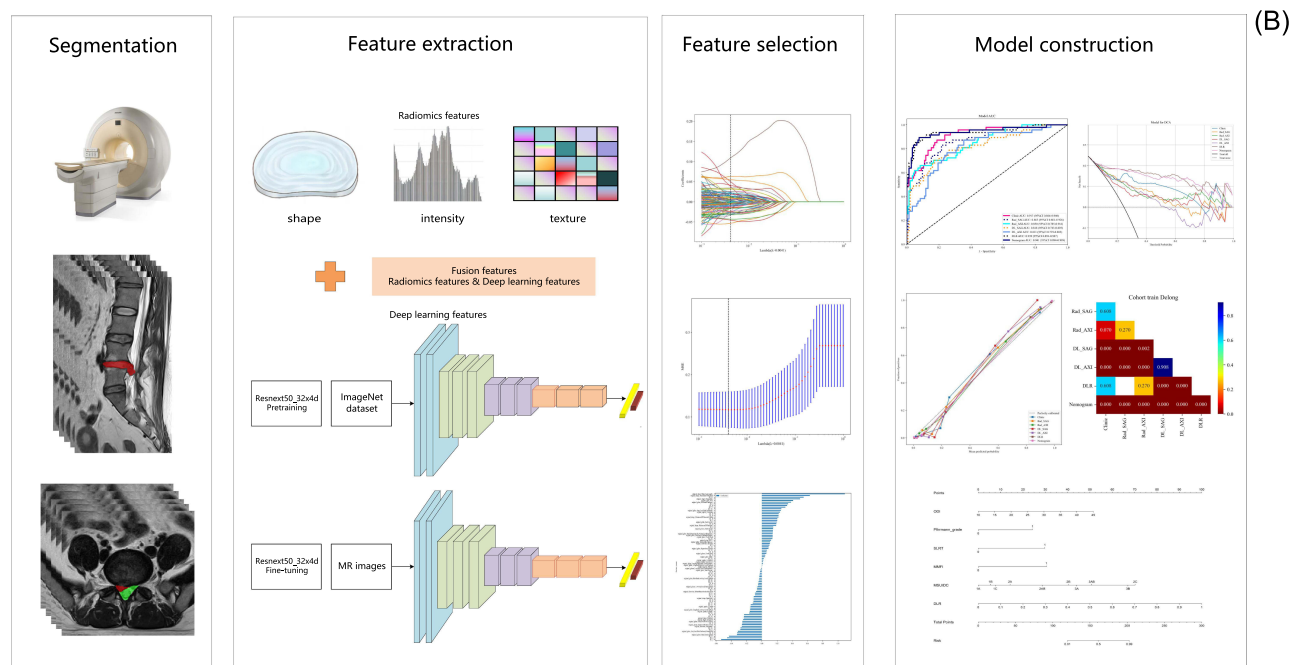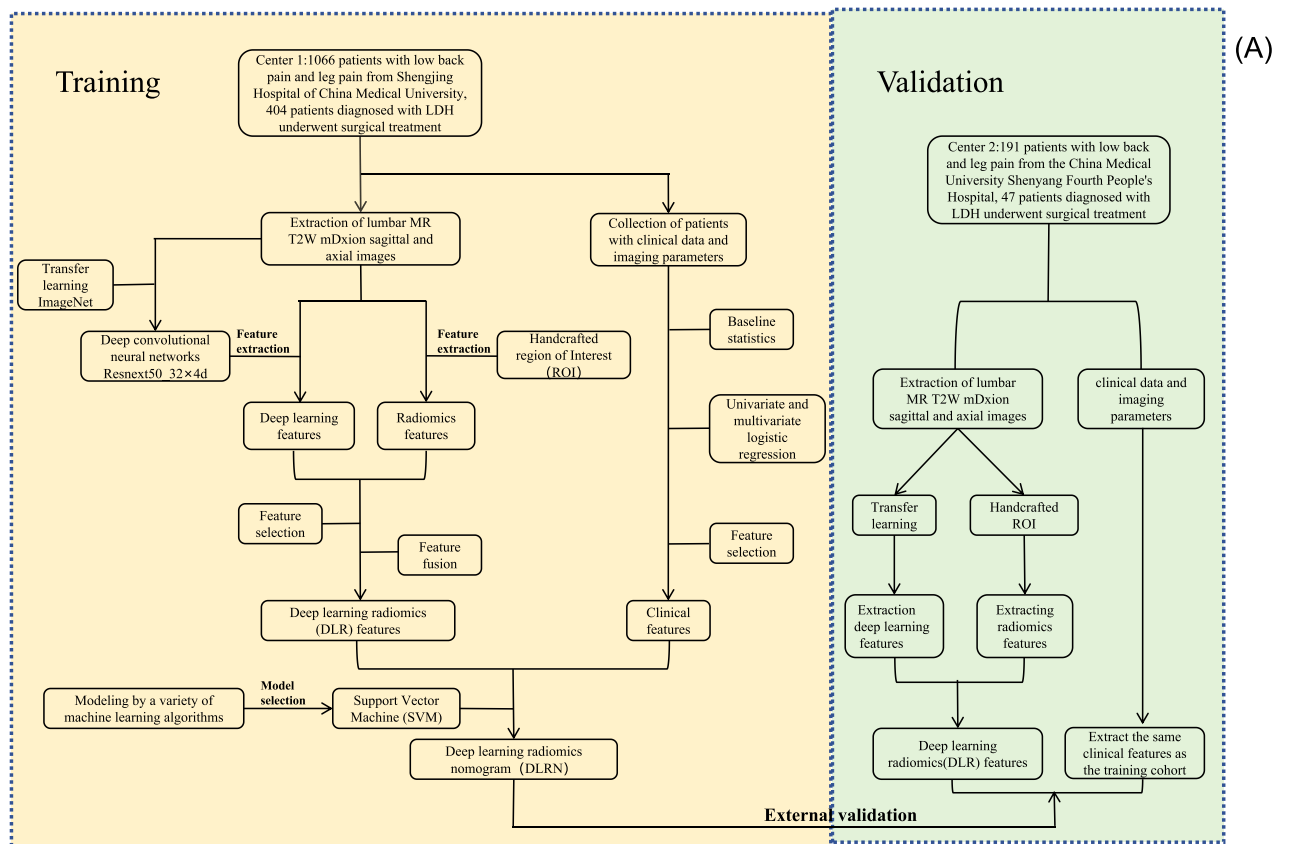
Figure 1 (**A**) Flowchart of this multicenter study and (**B**) workflow of deep learning radiomic nomogram (DLRN) modeling; MR: magnetic resonance.

follow-up. Image data were extracted using a picture archiving and communication system (PACS, version 5.5.0.20072). The two hospitals used the same Ingenia 3.0-T MR system (Philips Healthcare, Netherlands) with a spine coil. The MRI sequences are listed in Table 1.

ROI Segmentation

The lumbar MR images were imported into ITK-SNAP software (version 3.8.0, www.itksnap.org). Two senior radiologists and one experienced spine surgeon (with over 10 years of experience) separately outlined the intervertebral disc of the surgical segment on the sagittal plane, and the axial lumbar spinal canal included the intervertebral disc or nucleus pulposus tissue and dura mater. The non-surgical group was selected on the basis of the most evident LDH levels. Disagreements, if any, were resolved through discussion, and intra-and interobserver consistency was assessed using intraclass correlation coefficients (ICCs). Features with an ICC value greater than 0.75 were considered to have good consistency.

Radiomics and Deep-Learning Feature Extraction

The feature extraction in this study encompassed traditional handcrafted features, including geometric shape, intensity, and texture, along with deep-learning features that were transferred from the ImageNet database using CNNs.

All handcrafted features were extracted using Pyradiomics (http://pyradiomics.readthedocs.io) to implement an internal feature analysis program. The software processes both the original images and ROIs, categorizing the features into three groups: (I) geometric shape, (II) intensity, and (III) texture. Geometric shape features characterize the three-dimensional form of the ROI. The intensity features depict the primary statistical distribution of voxel intensities within the ROI. Texture features illustrate the spatial distribution of patterns or intensities in the second and higher orders. Various techniques have been utilized to extract texture features, including the gray-level co-occurrence matrix (GLCM), gray-level difference method (GLDM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLSZM), and neighborhood gray-tone difference matrix (NGTDM).

We used Resnext50_ 32 × 4d as the CNN framework to extract the deep-learning features. Initially, the model was pretrained on the ImageNet database. Subsequently, we extracted the ROIs, fed them into Resnext50_32×4d, and employed the model's penultimate average pooling layer for transfer learning, thereby extracting features.

Feature Selection and Fusion

We conducted the Mann–Whitney U-test and feature screening for all features, with $P < 0.05$. For features with high repeatability, Spearman correlation coefficient was also used, and one of the features with a correlation coefficient greater than 0.9 was retained. To retain the ability to depict features as much as possible, we used a greedy recursive deletion strategy for feature filtering; that is, the feature with the greatest redundancy in the current set was deleted each time.

The least absolute shrinkage and selection operator (LASSO) regression model was used for signature construction. Depending on the regulation weight λ, LASSO shrinks all regression coefficients towards zero and sets the coefficients of many irrelevant features exactly to zero. To find an optimal λ value, 10-fold cross validation with minimum criteria was

Table 1 Parameters of MRI mDxion Sequences Corresponding to the Patient

Parameters	T1W_mDixon_TSE Sagittal	T2W_mDixon_TSE Sagittal	T2W_mDixon_TSE Coronal	T2W_TSE_DRIVE
Flip Angle°	10	90	90	90
Repetition time (ms)	6.61	2500	3000	3112.5
Echo time (ms)	1.03	103	100	120
Section thickness/Intersection gap (mm)	4/2.0	4/4.4	4/4.8	4/4.4
Field of view (mm)	320	309	180	173
Matrix number	640×640	432×432	432×432	480×480
Number of excitation	2	2	2	1

employed, where the final value of λ yielded the minimum cross-validation error. The retained features with nonzero coefficients were used for regression model fitting and were combined into a radiomics signature. Subsequently, we obtained a radiomics score (rad-score) for each patient using a linear combination of the retained features weighted by their model coefficients. The Python scikit-learn package was used for the LASSO regression modeling.

Owing to the characteristics of the deep-learning dimension of 2048, we adopted principal component analysis (PCA) to balance and reduce its dimensions to 32, simultaneously improving the model's generalization and reducing the risk of overfitting. Subsequently, the radiomics features extracted from the ROI and deep-learning features selected by the model were initially fused. The Z-score method was used to standardize and calculate the mean and variance of each feature. Each feature column was transformed into a standard normal distribution by subtracting the mean, dividing it by the variance, and constructing the DLR features. Subsequently, feature selection was performed using the same process as for the radiomics features to achieve the optimal subset of fusion features.

The screening of clinical features began with baseline statistical analysis, employing univariate and multivariate logistic regression to identify significantly different variables and extracting features with $P < 0.05$. The feature selection process mirrored that of radiomics. To select the optimal clinical features, a receiver operating characteristic (ROC) curve was drawn and the top five features with the best area under the curve (AUC) values were included in the final model.

Model Construction and Validation

After feature screening and fusion, we obtained the Clinic, Rad_SAG, Rad_AXI, DL_SAG, DL_AXI, and DLR features, a total of 6 sets of features. These features were input into machine-learning models such as logistic regression (LR), support vector machine (SVM), k-nearest neighbor (KNN), decision trees, random forest (RF), extremely randomized trees, eXtreme gradient boosting (XGBoost), multi-layer perceptron (MLP), and light gradient boosting machine (LGBM), to construct the risk model. We compared the performance of the different models, and to prevent overfitting, a five-fold cross test was used to obtain the final DLR model. According to the performance of the model, the optimal rad-score was selected and fused with clinical features to construct a nomogram. The efficacy of the model was assessed by plotting the ROC curves and calculating the AUC, precision, sensitivity, and specificity. The DeLong test was employed to compare the performance of different models. Decision curve analysis (DCA) was used to evaluate clinical utility.

Statistical Analysis

Statistical analyses were conducted using SPSS (version 26.0; SPSS Inc., Armonk, NY, USA), R language (version 4.2.0), and GraphPad Prism V.8.2 (GraphPad Software Inc., San Diego, CA, USA). The Mann–Whitney U-test and Kruskal–Wallis test were used to compare non-normally distributed data, whereas the T-test or Fisher's exact test was used for normally distributed data. The X^2 test was employed for comparing counting data. Univariate and multivariate logistic regression analyses were conducted to determine the clinical parameters. ROC curves were generated to compare the predictive performance (including sensitivity and specificity), with the Yoden index set as the highest performance threshold. A nomogram combining radiomics and clinical features was developed. The ROC curves were plotted, calibration curves were used to evaluate calibration efficiency, and Hosmer–Lemeshow analytical fit was applied to assess its calibration capability. Statistical significance was defined as $P < 0.05$.

Results

Clinical Baseline Characteristics

The data of 1257 patients were included in this study; 1066 patient datasets were used for internal training, and 191 patient datasets were used for external testing. There was no significant difference between the two groups in gender, age, body mass index (BMI), occupation, anxiety score and LBP (all $P > 0.05$). Significant differences were found in visual analogue scale (VAS), Oswestry Disability Index (ODI), lower limb muscle strength, reflex, straight-leg raising test (SLRT), Michigan State University (MSU) classification of intervertebral disc, Pfirrmann grade, and fat infiltration of multifidus muscle (MMFI) (all $P < 0.05$); see Table 2. After univariate and multivariate logistic regression analyses were

Table 2 Clinical Characteristics and Lumbar MR Imaging Parameters (Baseline Analysis)

Variables	Training	Non-surgical	Surgical	P	Test	Non-surgical	Surgical	P
Sex				0.561				0.988
Male	528(49.53)	333(50.30)	195(48.27)		116(60.73)	88(61.11)	28(59.57)	
Female	538(50.47)	329(49.70)	209(51.73)		75(39.27)	56(38.89)	19(40.43)	
Age	33.23±7.38	33.09±6.99	33.45±7.99	0.156	33.10±6.90	33.49±6.84	31.91±7.02	0.196
BMI	23.59±3.58	23.56±3.26	23.62±4.05	0.924	23.37±3.25	23.11±2.77	24.17±4.35	0.33
Labor				0.872				0.078
Mild	626(58.72)	387(58.46)	239(59.16)		160(83.77)	125(86.81)	35(74.47)	
Heavy	440(41.28)	275(41.54)	165(40.84)		31(16.23)	19(13.19)	12(25.53)	
Smoke				0.953				0.314
No	924(86.68)	573(86.56)	351(86.88)		169(88.48)	125(86.81)	44(93.62)	
Yes	142(13.32)	89(13.44)	53(13.12)		22(11.52)	19(13.19)	3(6.38)	
Alcoholism				0.001				0.964
No	985(92.40)	626(94.56)	359(88.86)		169(88.48)	128(88.89)	41(87.23)	
Yes	81(7.60)	36(5.44)	45(11.14)		22(11.52)	16(11.11)	6(12.77)	
Duration				<0.001				0.01
<3 months	433(40.62)	296(44.71)	137(33.91)		105(54.97)	71(49.31)	34(72.34)	
>3 months	633(59.38)	366(55.29)	267(66.09)		86(45.03)	73(50.69)	13(27.66)	
Special diet				1.000				1.000
No	1065(99.91)	661(99.85)	404(100.00)		190(99.48)	143(99.31)	47(100.00)	
Yes	1(0.09)	1(0.15)	-		1(0.52)	1(0.69)	−	
Drug abuse				0.003				1.000
No	1054(98.87)	660(99.70)	394(97.52)		190(99.48)	143(99.31)	47(100.00)	
Yes	12(1.13)	2(0.30)	10(2.48)		1(0.52)	1(0.69)	−	
Education				0.02				<0.001
high	788(73.92)	506(76.44)	282(69.80)		152(79.58)	123(85.42)	29(61.70)	
low	278(26.08)	156(23.56)	122(30.20)		39(20.42)	21(14.58)	18(38.30)	
Income				0.235				<0.001
Above average	860(80.68)	542(81.87)	318(78.71)		149(78.01)	122(84.72)	27(57.45)	
Below average	206(19.32)	120(18.13)	86(21.29)		42(21.99)	22(15.28)	20(42.55)	
Diabetes				<0.001				0.324
No	1020(95.68)	647(97.73)	373(92.33)		185(96.86)	141(97.92)	44(93.62)	
Yes	46(4.32)	15(2.27)	31(7.67)		6(3.14)	3(2.08)	3(6.38)	
Hypertension				<0.001				0.303
No	1047(98.22)	659(99.55)	388(96.04)		188(98.43)	143(99.31)	45(95.74)	
Yes	19(1.78)	3(0.45)	16(3.96)		3(1.57)	1(0.69)	2(4.26)	
APSH				0.02				1.000
No	926(86.87)	588(88.82)	338(83.66)		177(92.67)	133(92.36)	44(93.62)	
Yes	140(13.13)	74(11.18)	66(16.34)		14(7.33)	11(7.64)	3(6.38)	
Depression score	40.78±9.70	40.08±10.31	41.92±8.50	0.012	39.92±10.58	39.28±11.17	41.87±8.32	0.085
Anxiety score	38.76±8.73	38.54±8.71	39.13±8.77	0.298	40.65±10.74	41.17±11.33	39.06±8.59	0.264
VAS	3.51±1.03	3.11±0.99	4.17±0.71	<0.001	3.32±1.03	3.03±0.97	4.23±0.56	<0.001
ODI	31.42±9.31	29.71±8.20	34.22±10.29	<0.001	30.03±8.57	28.30±7.64	35.34±9.16	<0.001
LLA(°)	28.41±10.47	28.05±10.79	29.01±9.89	0.143	25.21±9.99	24.34±9.81	27.87±10.17	0.035
LSA(°)	29.22±7.59	30.52±7.29	27.07±7.60	<0.001	28.94±7.30	30.06±7.25	25.49±6.35	<0.001
SSA(°)	42.96±7.92	43.43±7.49	42.18±8.53	0.012	42.30±7.44	43.19±7.00	39.57±8.14	0.004
PMMCSA(mm2)	1315.55±409.36	1284.72±399.56	1366.08±420.57	0.002	1320.55±398.45	1285.68±391.81	1427.40±403.78	0.062
QLMCSA(mm2)	495.96±152.68	474.41±141.49	531.27±163.58	<0.001	480.72±154.55	462.36±139.07	536.96±185.06	0.029
MMCSA(mm2)	685.43±177.92	685.98±170.76	684.52±189.29	0.95	671.31±168.05	665.51±158.58	689.09±195.06	0.327
ESMCSA(mm2)	1260.85±355.02	1260.85±345.07	1260.85±371.19	0.925	1270.05±362.59	1266.92±373.08	1279.62±331.96	0.536
L1/2SFTT(mm)	13.15±6.96	12.13±6.35	14.81±7.58	<0.001	11.98±6.30	10.95±4.89	15.16±8.71	<0.001

(*Continued*)

Table 2 (Continued).

Variables	Training	Non-surgical	Surgical	P	Test	Non-surgical	Surgical	P
LBP				0.444				0.371
No	298(27.95)	191(28.85)	107(26.49)		27(14.14)	18(12.50)	9(19.15)	
Yes	768(72.05)	471(71.15)	297(73.51)		164(85.86)	126(87.50)	38(80.85)	
Sciatica				<0.001				<0.001
No	597(56.00)	343(51.81)	254(62.87)		169(88.48)	144(100.00)	25(53.19)	
Yes	469(44.00)	319(48.19)	150(37.13)		22(11.52)	–	22(46.81)	
Conservative treatment				0.786				0.783
No	598(56.10)	374(56.50)	224(55.45)		158(82.72)	118(81.94)	40(85.11)	
Yes	468(43.90)	288(43.50)	180(44.55)		33(17.28)	26(18.06)	7(14.89)	
Superficial sensation				0.045				<0.001
No	1012(94.93)	621(93.81)	391(96.78)		173(90.58)	137(95.14)	36(76.60)	
Yes	54(5.07)	41(6.19)	13(3.22)		18(9.42)	7(4.86)	11(23.40)	
Reflex				<0.001				0.481
No	746(69.98)	510(77.04)	236(58.42)		155(81.15)	119(82.64)	36(76.60)	
Yes	320(30.02)	152(22.96)	168(41.58)		36(18.85)	25(17.36)	11(23.40)	
Weakened Strength				<0.001				<0.001
No	862(80.86)	597(90.18)	265(65.59)		169(88.48)	144(100.00)	25(53.19)	
Yes	204(19.14)	65(9.82)	139(34.41)		22(11.52)	–	22(46.81)	
SLRT				<0.001				<0.001
No	1022(95.87)	649(98.04)	373(92.33)		182(95.29)	142(98.61)	40(85.11)	
Yes	44(4.13)	13(1.96)	31(7.67)		9(4.71)	2(1.39)	7(14.89)	
MSU classification				<0.001				<0.001
1A	137(12.76)	136(20.54)	–		25(13.09)	25(17.36)	–	
1B	62(5.82)	59(8.91)	3(0.74)		9(4.71)	9(6.25)	–	
1C	22(2.06)	21(3.17)	1(0.25)		7(3.66)	6(4.17)	1(2.13)	
2A	260(24.39)	236(35.65)	24(5.94)		54(28.27)	52(36.11)	2(4.26)	
2B	220(20.17)	121(17.98)	99(23.76)		49(25.65)	34(23.61)	15(31.91)	
2C	8(0.75)	2(0.30)	6(1.49)		0(0.00)	–	–	
3A	86(8.07)	31(4.68)	55(13.61)		14(7.33)	8(5.56)	6(12.77)	
3B	141(13.23)	12(1.81)	129(31.93)		16(8.38)	–	16(34.04)	
2AB	72(6.75)	43(6.50)	29(7.18)		11(5.76)	10(6.94)	1(2.13)	
3AB	58(5.44)	2(0.30)	56(13.86)		6(3.14)	–	6(12.77)	
Pfirrmann grade				<0.001				0.011
Mild	716(67.17)	500(75.53)	216(53.47)		139(72.77)	112(77.78)	27(57.45)	
Severe	350(32.83)	162(24.47)	188(46.53)		52(27.23)	32(22.22)	20(42.55)	
Modic change				<0.001				<0.001
No	897(84.15)	588(88.82)	309(76.49)		160(83.77)	132(91.67)	28(59.57)	
Yes	169(15.85)	74(11.18)	95(23.51)		31(16.23)	12(8.33)	19(40.43)	
Annulus fibrosus tear				0.001				<0.001
No	723(67.82)	425(64.20)	298(73.76)		133(69.63)	115(79.86)	18(38.30)	
Yes	343(32.18)	237(35.80)	106(26.24)		58(30.37)	29(20.14)	29(61.70)	
Lumbodorsal myofascitis				0.943				0.075
No	939(88.09)	584(88.22)	355(87.87)		187(97.91)	143(99.31)	44(93.62)	
Yes	127(11.91)	78(11.78)	49(12.13)		4(2.09)	1(0.69)	3(6.38)	
PMMFI				1.000				1.000
Mild	1065(99.91)	661(99.85)	404(100.00)		191(100.00)	144(100.00)	47(100.00)	
Severe	1(0.09)	1(0.15)	–		0(0.00)	–	–	
QLMFI				0.059				0.347
Mild	1029(96.53)	645(97.43)	384(95.05)		185(96.86)	138(95.83)	47(100.00)	
Severe	37(3.47)	17(2.57)	20(4.95)		6(3.14)	6(4.17)	–	

(*Continued*)

Table 2 (Continued).

Variables	Training	Non-surgical	Surgical	P	Test	Non-surgical	Surgical	P
ESMFI				<0.001				0.129
Mild	961(90.15)	622(93.96)	339(83.91)		165(86.39)	128(88.89)	37(78.72)	
Severe	105(9.85)	40(6.04)	65(16.09)		26(13.61)	16(11.11)	10(21.28)	
MMFI				<0.001				0.044
Mild	922(86.49)	598(90.33)	324(80.20)		180(94.24)	139(96.53)	41(87.23)	
Severe	144(13.51)	64(9.67)	80(19.80)		11(5.76)	5(3.47)	6(12.77)	

Abbreviations: BMI, body Mass Index; APSP, history of abdominal or pelvic surgery; VAS, visual analogue scale; ODI, Oswestry disability index; LLA, lumbar lordosis angle; LSA, lumbosacral angle; SSA, sacrum slant angle; PMMCSA, cross sectional area of the psoas major muscle; QLMCSA, cross sectional area of quadratus lumbosae muscle; MMCSA, cross sectional area of multifidus muscle; ESMCSA, cross sectional area of erector spinae muscle; PMMFI, fat infiltration in the psoas major muscle; QLMFI, fat infiltration in the quadratus lumbosae muscle; MMFI, fat infiltration in the multifidus muscle; ESMFI, fat infiltration in the erector spinae muscle; L1/2SFTT, Subcutaneous fat tissue thickness at L1-2 level; MSU classification, Michigan State University (MSU) classification of intervertebral disc; SLRT, straight leg raising test.

performed, some factors that may cause bias due to the small sample size, such as diabetes, hypertension, adverse drug use, alcohol abuse, and education, were excluded. The remaining factors are drawn to ROC, and the top five features with the highest AUC values (ODI, Pfirrmann grade, SLRT, MMFI, and MSU classification) were included in the final clinical model (Table 3 and Figure 2).

Table 3 Univariate and Multivariate Analysis of Clinical Features and Lumbar MR Imaging Parameters

Variables	Univariate Analysis		Multivariate Analysis	
	OR (95% CI)	P	OR (95% CI)	P
L1/2SFTT	1.057(1.038–1.077)	<0.001	1.042 (1.016–1.069)	0.001
ESMFI	2.307(1.619–3.3)	<0.001	1.674 (1.025–2.738)	0.039
MMFI	2.982(1.977–4.548)	<0.001	3.114 (1.778–5.533)	<0.001
QLMFI	1.976(1.023–3.861)	0.043		
PMMFI	0(NA-3.165)	0.97		
ESMCSA	1 (0.998–1.002)	<0.001	1 (0.999–1.002)	0.266
MMCSA	1(0.999–1.001)	0.896		
QLMCSA	1.002(1.002–1.003)	<0.001	1.002 (1.001–1.003)	<0.001
PMMCSA	1 (1–1.001)	0.002		
Lumbodorsal myofascitis	1.033(0.703–1.507)	0.866		
Annulus fibrosus tear	0.638(0.484–0.836)	0.001	0.508 (0.342–0.746)	0.001
Modic change	2.443(1.752–3.418)	<0.001	1.388 (0.882–2.188)	0.156
Pfirrmann grade	2.686(2.065–3.501)	<0.001	1.986 (1.381–2.864)	<0.001
MSU classification	1.003(1.001–1.006)	<0.001	1.007 (1.005–1.009)	<0.001
LSA	0.98(0.965–0.996)	0.012	1.033 (1.004–1.063)	0.025
LLA	0.939(0.922–0.955)	<0.001	0.918 (0.889–0.946)	<0.001
SSA	1.009(0.997–1.021)	0.144		
Weakened Strength	4.818(3.484–6.722)	<0.001	3.773 (2.422–5.945)	<0.001
SLRT	2.388(1.828–3.126)	<0.001	3.322 (2.251–4.943)	<0.001
Reflex	4.149(2.192–8.302)	<0.001	7.266 (3.192–17.46)	<0.001
Superficial sensation	0.504(0.257–0.926)	0.035		
Conservative treatment	1.044(0.813–1.339)	0.738		
ODI	1.056(1.041–1.072)	<0.001	1.030 (1.010–1.050)	0.003
VAS	3.591(3.031–4.295)	<0.001	3.477 (2.842–4.303)	<0.001
Sciatica	0.635(0.492–0.817)	<0.001	0.438 (0.301–0.634)	<0.001
Duration	1.576(1.221–2.04)	0.001	1.126 (0.811–1.562)	0.551

(Continued)

Table 3 (Continued).

Variables	Univariate Analysis		Multivariate Analysis	
	OR (95% CI)	P	OR (95% CI)	P
LBP	1.126(0.854–1.489)	0.404		
Anxiety score	1.008(0.994–1.022)	0.28		
Depression score	1.02(1.007–1.033)	0.003	1.018 (1.000–1.038)	0.046
Alcoholism	2.18(1.383–3.46)	0.001		
Smoke	0.972(0.672–1.396)	0.879		
APSH	1.552(1.083–2.219)	0.016		
Income	1.221(0.895–1.663)	0.205		
Education	1.403(1.062–1.852)	0.017		
Drug abuse	8.376(2.194–54.67)	0.006		
Special diet	0(NA-3.165)	0.97		
Hypertension	9.058(2.995–39.12)	<0.001		
Diabetes	3.585(1.942–6.906)	<0.001		
Labor	0.972(0.755–1.249)	0.822		

Abbreviations: BMI, body Mass Index; APSP, history of abdominal or pelvic surgery; VAS, visual analogue scale; ODI, Oswestry disability index; LLA, lumbar lordosis angle; LSA, lumbosacral angle; SSA, sacrum slant angle; PMMCSA, cross sectional area of the psoas major muscle; QLMCSA, cross sectional area of quadratus lumbosae muscle; MMCSA, cross sectional area of multifidus muscle; ESMCSA, cross sectional area of erector spinae muscle; PMMFI, fat infiltration in the psoas major muscle; QLMFI, fat infiltration in the quadratus lumbosae muscle; MMFI, fat infiltration in the multifidus muscle; ESMFI, fat infiltration in the erector spinae muscle; L1/2SFTT, Subcutaneous fat tissue thickness at L1-2 level; MSU classification, Michigan State University (MSU) classification of intervertebral disc; SLRT, straight leg raising test.

Features Selection

A total of 107 radiomics features were initially extracted from the T2W_mDixon_TSE sagittal images. After screening by LASSO and 10-fold cross testing, the penalty coefficient (λ = 0.0015) was determined, and 35 features were obtained. Similarly, 95 radiomics features were initially extracted from the T2W_mDixon_TSE axial images. Using the same method, the penalty coefficient was determined (λ = 0.0063), and 18 features were obtained.

By using the Resnext50_32x4d model architecture, the network was initially pretrained on the ImageNet database and then fed with the maximum-level ROIs, resulting in deep-learning features with 2040 dimensions. After PCA reduced the dimensions to 32, using LASSO and 10-fold cross-validation testing, the sagittal penalty coefficient (λ = 0.0020) and axial penalty coefficient (λ = 0.0036) were determined, and finally 31 sagittal and 30 axial deep-learning features were extracted (Figure 3 and Supplementary Figure 1).

DLR was obtained by fusion of these features after LASSO and a 10-fold cross-validation test, with a penalty coefficient λ of 0.0041, yielding 114 feature parameters (Figure 4 and Supplementary Figure 2).

Predictive Performance of the Models

In this study, multiple features were incorporated into various machine-learning models. External testing results show that, in Rad_SAG, the three models with best performance were LR (AUC = 0.879), MLP (AUC = 0.874), and SVM (AUC = 0.863), and the three models with the best performance in Rad_AXI were SVM (AUC = 0.85), XGBoost (AUC = 0.842) and MLP (AUC = 0.833). SVM (AUC = 0.818), LR (AUC = 0.8), MLP (AUC = 0.792) were the three models with the best performance in DL_SAG, and MLP (AUC = 0.856), LR (AUC = 0.831) and SVM (AUC = 0.811) were the three models with the best performance in DL_AXI. The three models with the best performance in DLR were SVM (AUC = 0.939), LR (AUC = 0.932) and MLP (AUC = 0.918). The three best-performing clinical models were LightBGM (AUC = 0.922), XGBoost (AUC = 0.914), and SVM (AUC = 0.904); see Supplementary Figure 3 and Supplementary Tables 1–7. According to these results, no algorithm performed the best among all the models; therefore, the top three results were counted and voted according to the number of times. Finally, we chose SVM as the framework
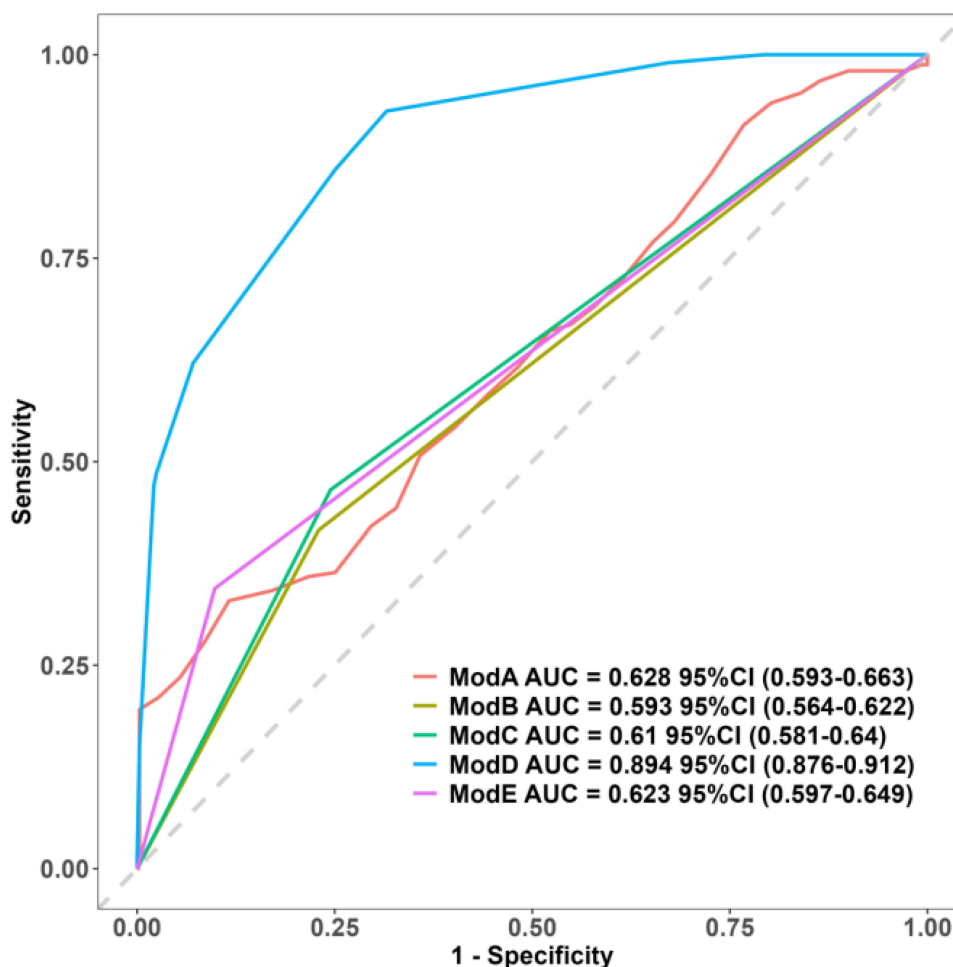
Figure 2 ROC curves for the clinical characteristics. The final selected 5 features had the following AUC values: ModA was ODI (AUC=0.628), ModB was SLRT (AUC=0.593), ModC was Pfirrmann grading (AUC=0.61), ModD was MSU classification (AUC=0.894), and ModE was MMFI (AUC=0.623).

for model construction. The efficiency of the nomogram in the test cohort was 0.941 (95% CI 0.894–0.989), which was good and avoided the bias caused by the confusing algorithm (Figures 5 and 6 and Table 4).

The SVM was judged the optimal model for radiomics. Among all the models, DLR had the best performance (Figure 7 and Supplementary Figure 3), with an AUC of 0.991 (95% CI 0.985–0.997) in the training cohort and 0.939 (95% CI 0.891–0.987) in the test cohort. The nomogram constructed by the fusion of clinical features had an AUC of 0.994 (95% CI 0.989–0.999) in the training cohort and 0.941 (95% CI 0.894–0.989) in the test cohort. Analysis using the DeLong test found statistical differences between the radiomics, DL, and DLR models vs nomogram in the training and test cohort ($P < 0.05$). In the training cohort, the difference between the clinical model and the nomogram was statistically significant ($P < 0.05$), but not in the test cohort ($P = 0.338$) (Figure 8). The calibration curves showed close agreement between the predicted and observed values in the training and test cohorts, and the Hosmer–Lemeshow test ($P > 0.05$) showed good adaptability (Figure 9). DCA showed that the nomogram had significant clinical benefits in predicting probabilities over a wide range, and its efficacy was higher than that of the other models (Figure 10).

Nomogram Construction

The five selected clinical features (ODI, Pfirrmann grade, SLRT, MMFI, and MSU classification) and DLR were combined to construct a DLRN, which more intuitively demonstrated the efficacy of the model and was conducive to individualized risk prediction and clinical promotion (Figure 11).
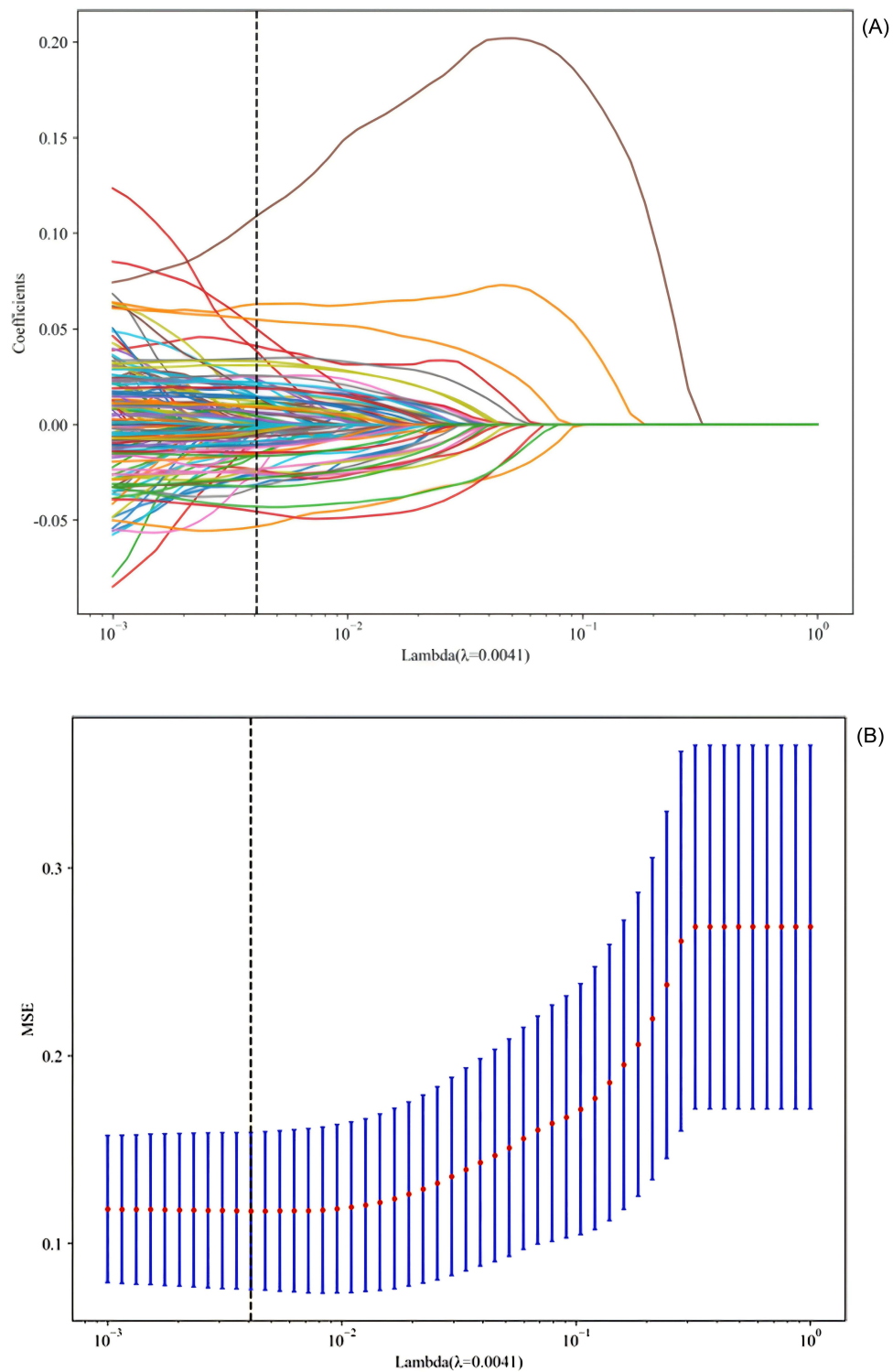
Figure 3 (**A**) Coefficient distribution and (**B**) mean standard error of 10-fold cross validation in the LASSO feature selection, also showing the optimal penalty coefficient λ of 0.0041 for deep-learning radiomics (DLR).

Discussion

The aim of this study was the innovative use of lumbar MR T2W_mDxion sagittal and axial images with separately handcrafted ROIs to extract traditional radiomic features. The Resnext50_32x4d model architecture was pre-trained in the ImageNet database and used to extract deep-learning features, after which LASSO and 10-fold cross tests were used
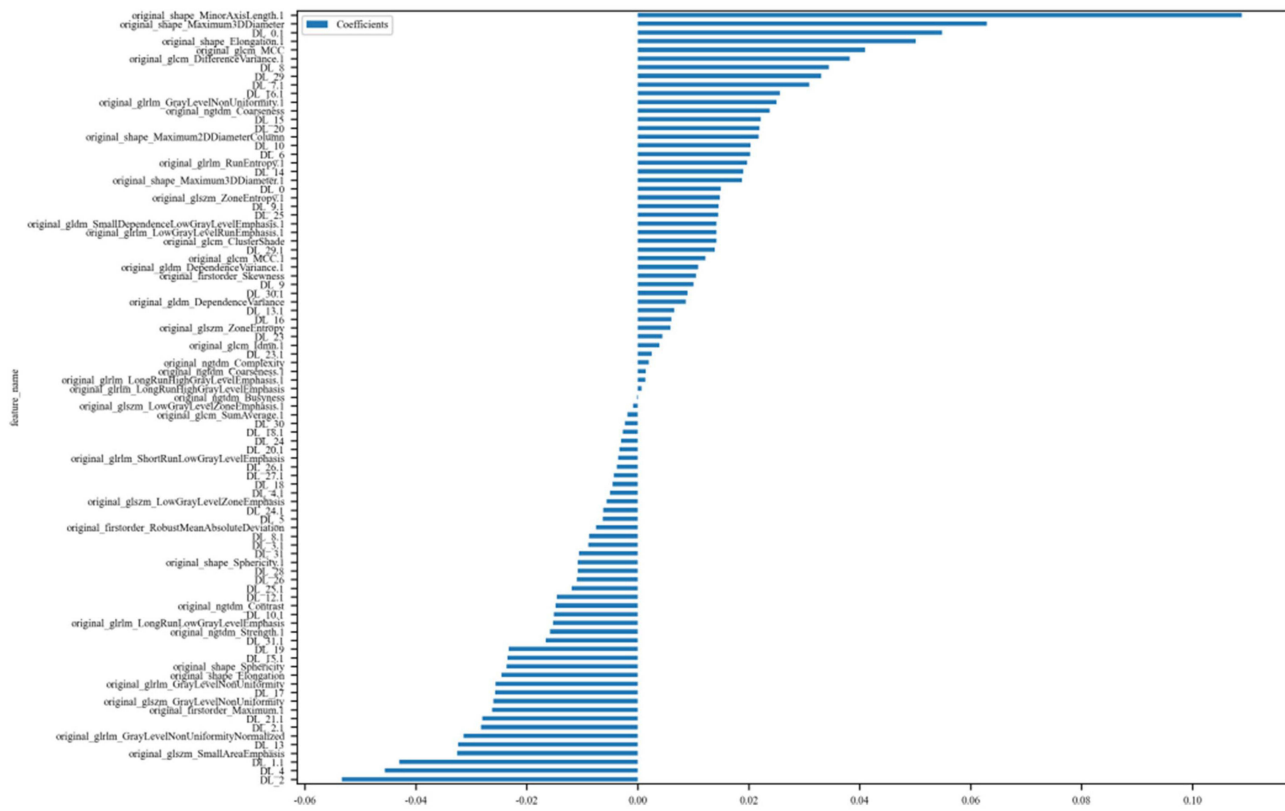
Figure 4 Histogram of the deep-learning radiomics (DLR) feature importance score.

for feature screening, multiple machine-learning algorithms were used for modeling, and external tests were conducted simultaneously to verify the model efficiency and build the DLRN. In this study, nomogram showed excellent predictive performance and good calibration of surgical risk factors for LDH in young adults, with an AUC of 0.994 (95% CI 0.989–0.999) in the training cohort and 0.941 (95% CI 0.894–0.989) in the test cohort; DCA showed better clinical benefits.

Currently, the use of artificial intelligence (AI) in lumbar disease diagnosis mainly includes automatic positioning and sizing of the disc and positioning and measurement of the spinal canal. Zheng et al[15] used BianqueNet to automatically segment the vertebra and disc and predict the degree of disc degeneration with an accuracy of 92%. Zhou et al[16] applied a CNN and used a transfer-learning method to automatically locate lumbar vertebrae from L1 to S1. Its advantage was that it did not use labeled MR images for training, and the algorithm achieved 98.6% accuracy. Hallinan et al[17] developed a deep-learning model using Resnet101, and found that it was reliable for diagnosing central canal stenosis (AUC = 0.82), lateral recess stenosis (AUC = 0.72), and foraminal stenosis (AUC = 0.75). Won et al[18] reported no significant difference between the grading of spinal stenosis using a deep-learning model and that by spinal surgeons. However, these studies mainly focused on imaging, lacked clinical signs, and could not predict the risk factors for lumbar diseases.

Studies have also shown that AI can be helpful in predicting LDH prognosis. Harada et al[19] developed a risk prediction model for preventing postoperative lumbar disc recurrence by using XGBoost, and finally included seven factors such as ODI, symptom duration, and BMI (AUC = 0.72); they found that young patients with LDH have a higher risk of postoperative recurrence, which is similar to the findings of Abdu et al, who published a study on the effect of surgery on recurrent lumbar discs.[20] Saravi et al[4] extracted the preoperative radiomics features of lumbar MR and used SVM, XGBoost, and random trees combined with clinical features to predict the postoperative effect of LDH; the results showed that the accuracy was higher than that of the clinical prediction model alone. Nevertheless, these studies mainly focused on clinical signs. Here, we used radiomics-DL fusion feature modeling and included the patients' clinical characteristics to predict the surgical risk of young patients with LDH. We found that DLR had better prediction
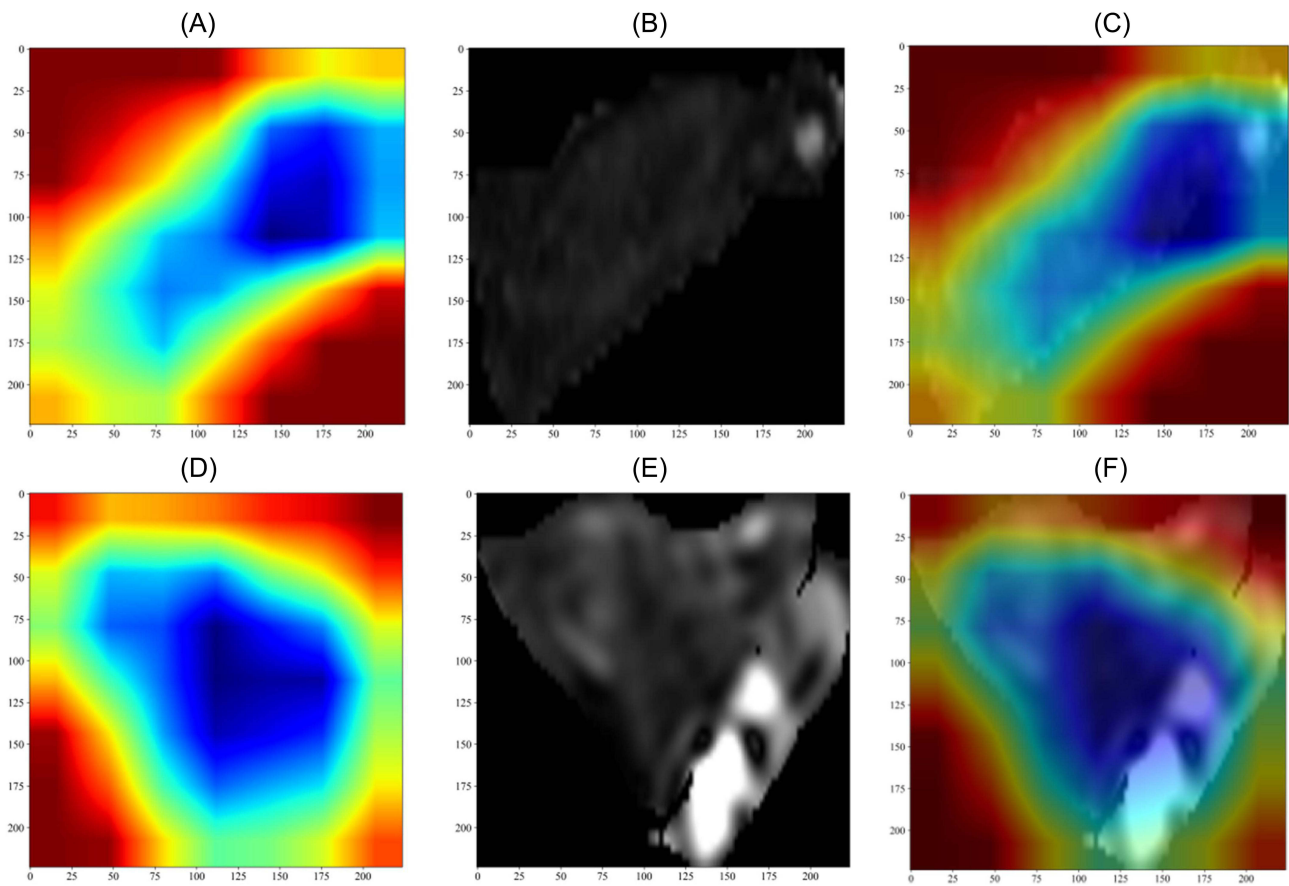
Figure 5 Visualization of deep-learning models: (**A**) DL_SAG and (**D**) DL_AXI showed the local entropy image; (**B**) DL_SAG and (**E**) DL_AXI showed the ROI of lumbar MR, (**B**) was the intervertebral disc, (**E**) including the intervertebral disc protruding into the spinal canal and the dura in the spinal canal; (**C**) DL_SAG and (**F**) DL_AXI were the results of the cluster showing that the model had a good fusion effect.
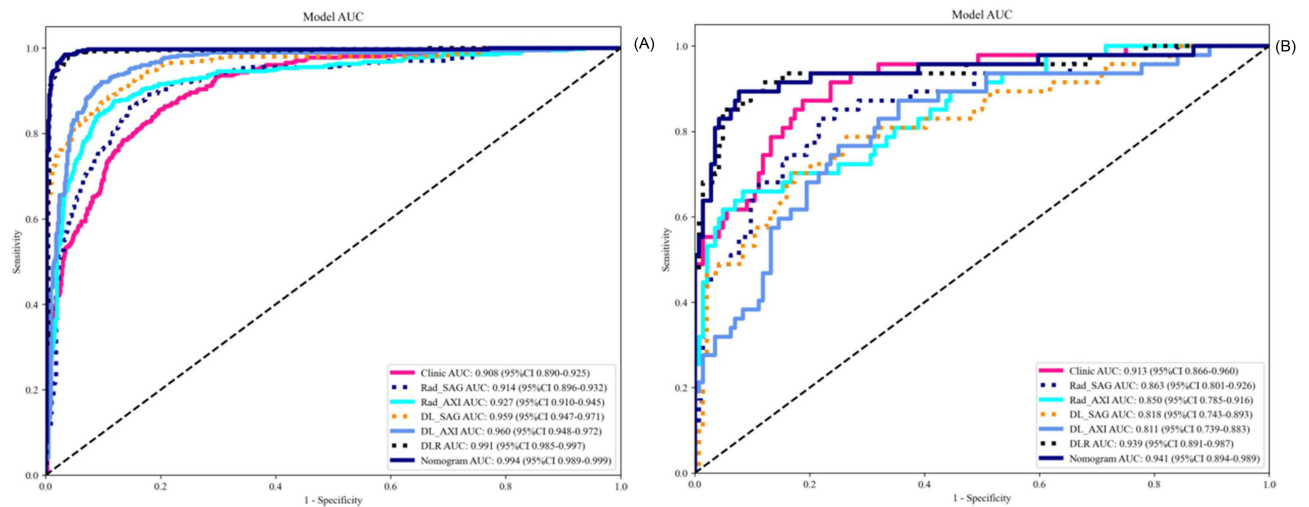

Figure 6 Comparison of AUC values for each group of features under the SVM algorithm in the (**A**) training and (**B**) test cohort; in both cohorts, the AUC of DLR was higher than any other radiomics feature models, and DLRN was the optimal model.

Table 4 Performance of Each Model in the Training and Test Cohorts Under the SVM Algorithm

Model	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Threshold	Task
Clinic	0.825	0.906	0.8885–0.9238	0.822	0.826	0.743	0.884	0.283	train
Clinic	0.801	0.904	0.8549–0.9538	0.809	0.799	0.567	0.927	0.278	test
Rad_SAG	0.856	0.914	0.8956–0.9324	0.859	0.853	0.782	0.908	0.357	train
Rad_SAG	0.791	0.863	0.8006–0.9255	0.809	0.785	0.551	0.926	0.444	test
Rad_AXI	0.878	0.927	0.9100–0.9446	0.874	0.881	0.817	0.92	0.367	train
Rad_AXI	0.848	0.85	0.7853–0.9156	0.638	0.917	0.714	0.886	0.646	test
DL_SAG	0.888	0.959	0.9469–0.9710	0.884	0.891	0.832	0.926	0.355	train
DL_SAG	0.791	0.818	0.7431–0.8928	0.681	0.826	0.561	0.888	0.475	test
DL_AXI	0.909	0.96	0.9478–0.9720	0.918	0.903	0.853	0.948	0.378	train
DL_AXI	0.696	0.811	0.7393–0.8828	0.851	0.646	0.44	0.93	0.272	test
DLR	0.967	0.991	0.9851–0.9968	0.975	0.962	0.94	0.985	0.366	train
DLR	0.921	0.939	0.8914–0.9874	0.83	0.951	0.848	0.945	0.624	test
Nomogram	0.973	0.994	0.9894–0.9990	0.983	0.967	0.947	0.989	0.329	train
Nomogram	0.911	0.941	0.8936–0.9885	0.872	0.924	0.788	0.957	0.57	test

performance than the above single models, and DLRN combined with clinical characteristics was more convincing in predicting efficiency.

In this study, after a voting process was conducted, SVM was ultimately chosen to build the model, and a nomogram with an AUC of 0.941 in the test cohort showed good performance. Gaonkar et al[21] also used an SVM to segment the lumbar canal area on axial lumbar MRI images, and the Deep-U-Net model was used to segment the disc in the sagittal plane, which can assist in the diagnosis of spinal stenosis. Hashia et al[22] reported a texture feature (GLRLM) based on sagittal lumbar MR and modeling by SVM to distinguish disc herniation, with an accuracy of 0.833. In summary, SVM
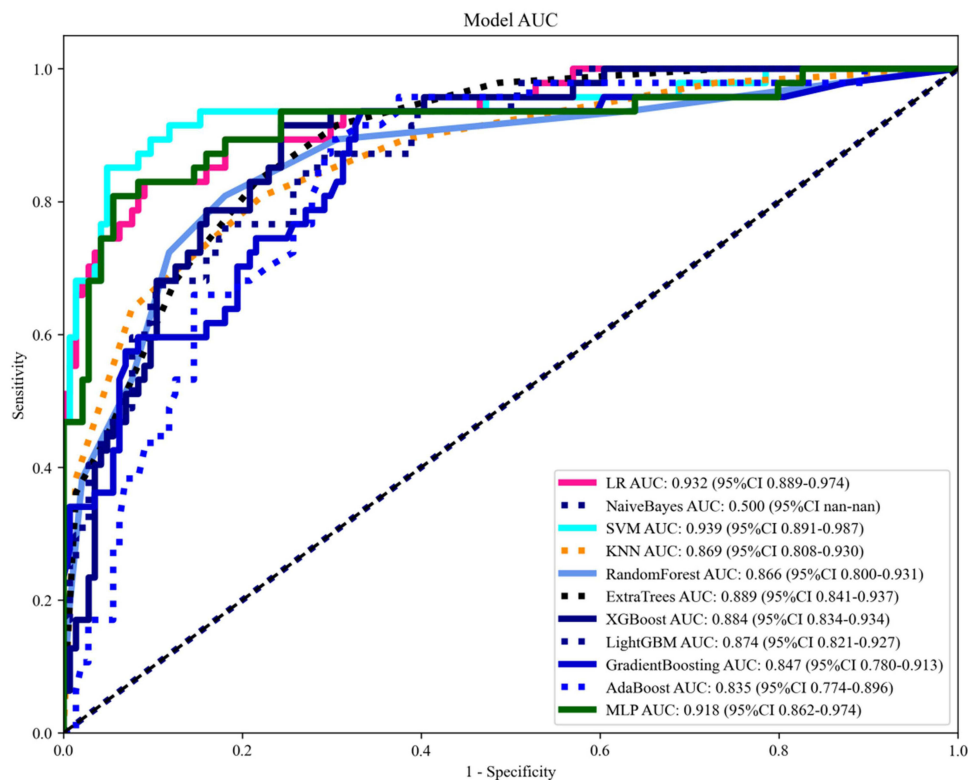

Figure 7 ROC curves of DLR features for different deep-learning algorithms in the test cohort; the results showed the SVM was the optimal algorithm.
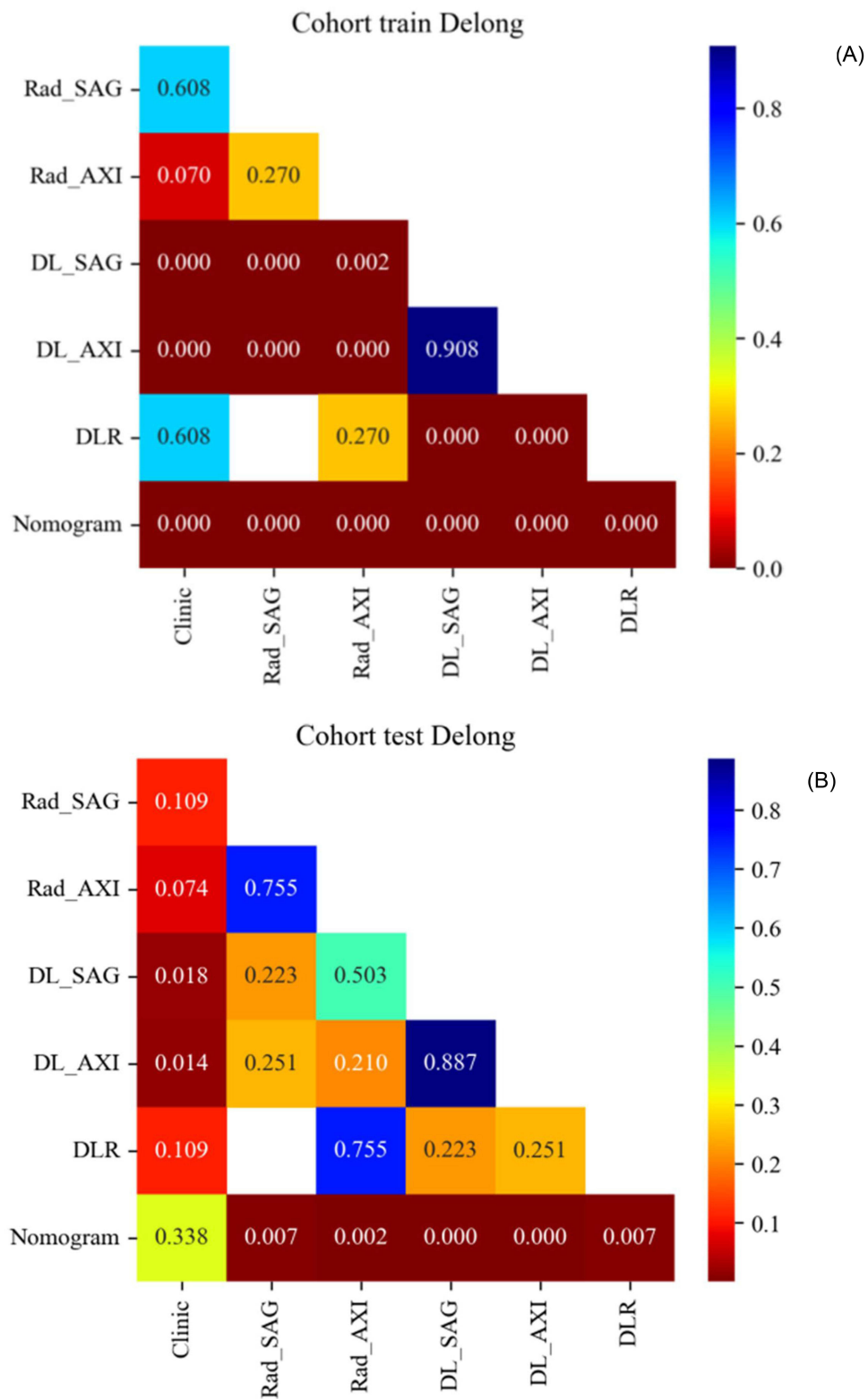
Figure 8 Delong test for the (**A**) training and (**B**) testing cohorts; (**A**) revealed that the nomogram outperformed models constructed with other features ($P < 0.05$), and (**B**)indicated that the nomogram surpassed models constructed with radiomics features, deep learning, or DLR ($P < 0.05$), yet its performance enhancement was not notably distinct when compared to clinical models ($P = 0.338$).
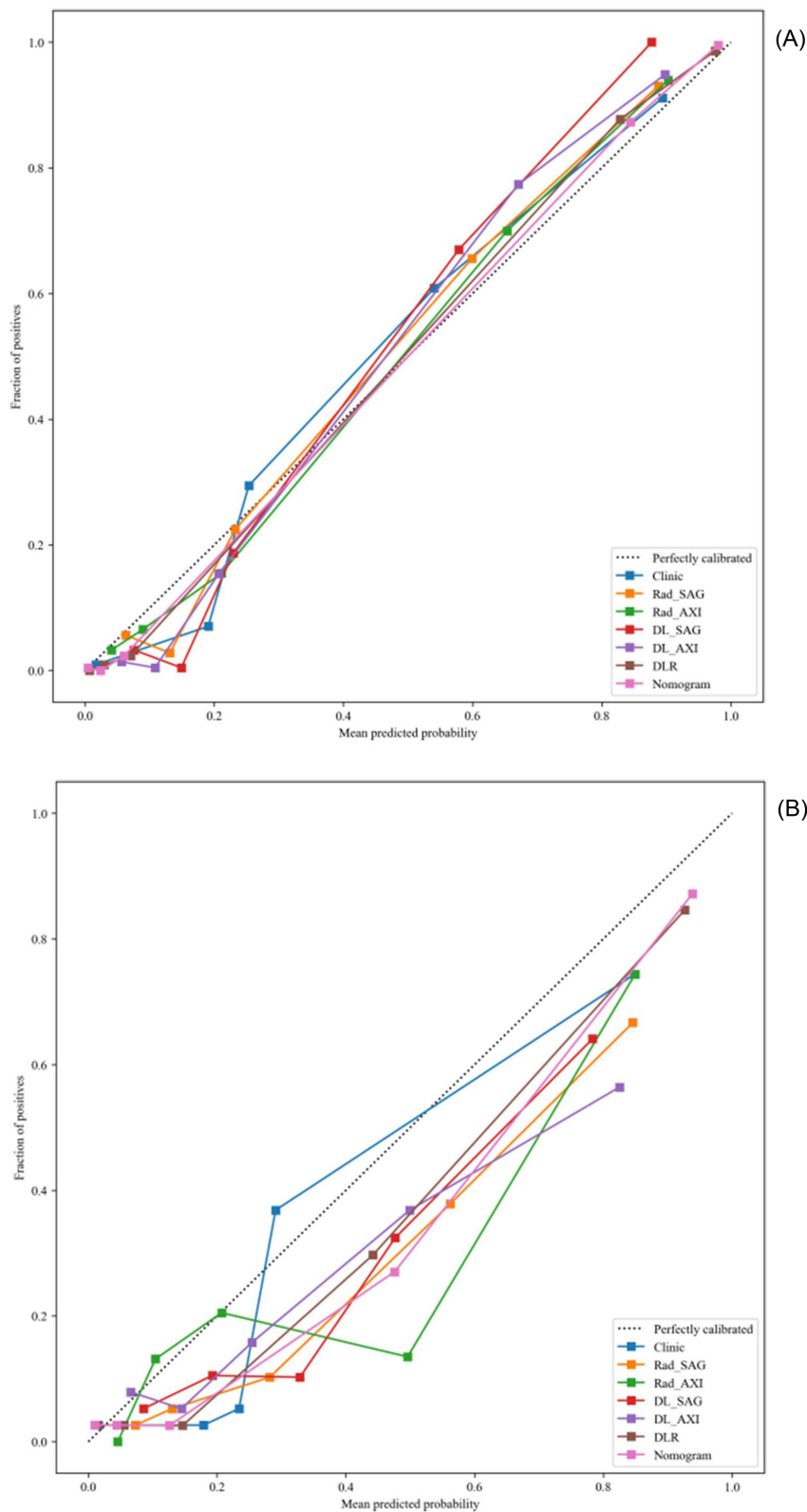
(A)

(B)

Figure 9 Calibration curves of different models in the (**A**) training and (**B**) test cohorts; the curves demonstrated close agreement between model predictions and actual observations. Nomogram exhibited the best performance, with *P* > 0.05 obtained from the Hosmer–Lemeshow test. The horizontal axis represents the predicted probability, and the vertical axis represents the actual probability. The diagonal dotted line in the graph signifies perfect alignment between predicted and actual probabilities under ideal conditions.
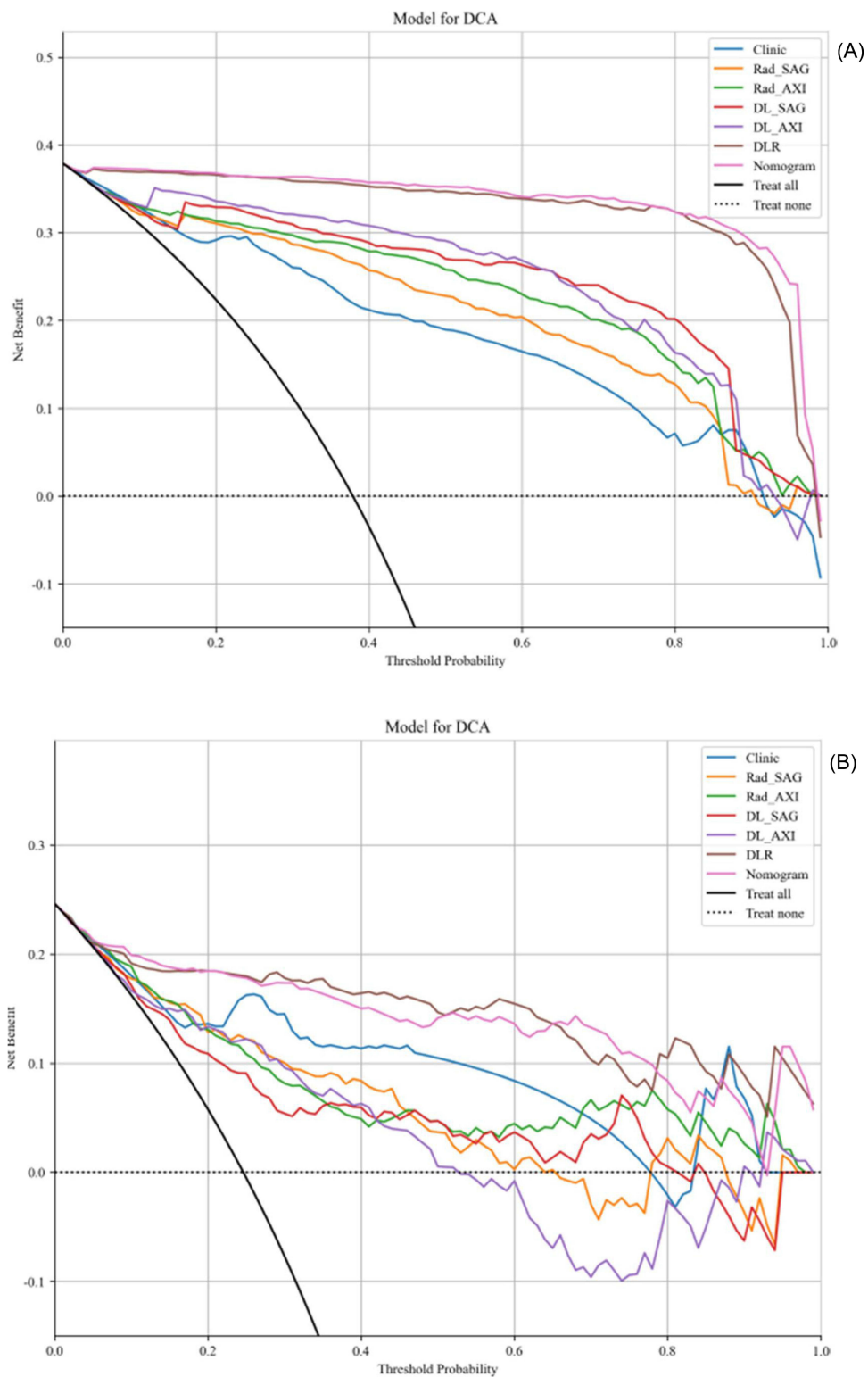
Figure 10 DCA in the (**A**) training and (**B**) test cohorts; the X-axis represents threshold probability, and the Y-axis represents net benefit. The black line represents all positive assumptions, and the dashed line represents negative assumptions. The results indicate that the nomogram achieved significant clinical benefit in both the training and testing cohorts; furthermore, when compared to clinical features or DLR, the threshold probability range of the nomogram surpassed that of other features.
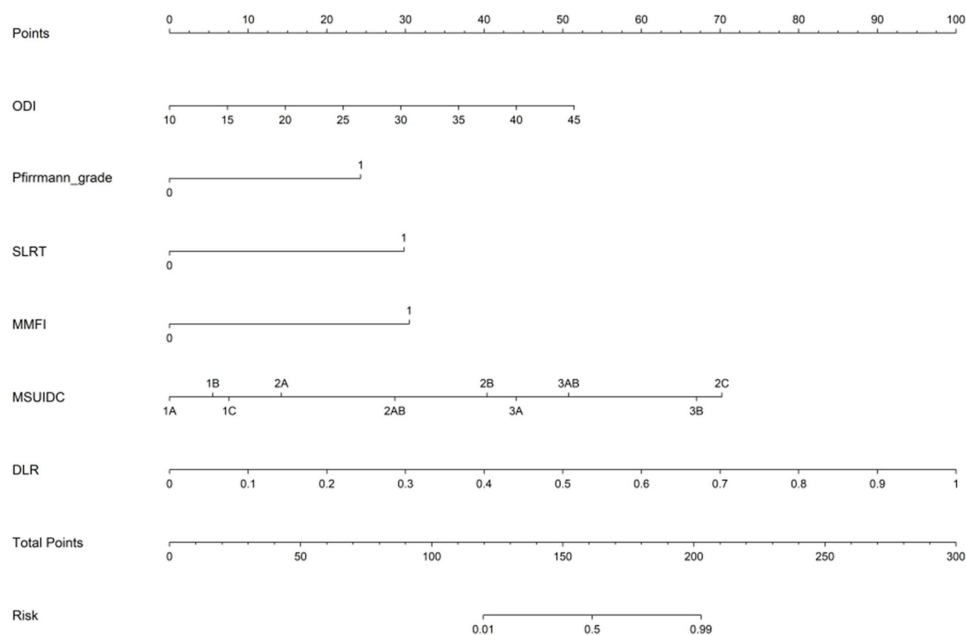
Figure 11 Constructing the deep learning radiomics nomogram: In the nomogram, a vertical line was drawn on the point axis to obtain the individual points corresponding to each model under different values. The scores of all features were added together to obtain the total points of the patient, and then a vertical line was drawn downward at the position of the total score to finally obtain the prediction probability.

has been widely used in the diagnosis of spine diseases and has achieved good performance. We used ResNext50_32x4d to pretrain the model in the ImageNet database and extract DL features. ImageNet, the world's largest image database with excellent generalizability, has been used extensively across clinical domains. For instance, Zhang et al[23] employed three DL models (ResNet50, ResNet101, and DenseNet121), pre-training on ImageNet to develop a model for predicting sacroiliac arthritis. The test cohort achieved an AUC of 0.91, demonstrating superior performance. ResNext merges the Inception philosophy with ResNet, inheriting the strengths of both. It innovatively introduces the cardinality dimension, surpassing the classification accuracy of ResNet, and addresses the complexity inherent in Inception's structural design. It enhances accuracy while maintaining model complexity, requires fewer hyperparameters for migration, and exhibits strong scalability.[24]

Our results showed that clinical factors including ODI, Pfirrmann grade, MSU classification, SLRT, and MMFI objectively reflect the LDH severity. ODI can comprehensively assess the patient's pain, life self-care ability, and other factors; and VAS is less effective than ODI.[25] The Pfirrmann grade is widely used to clinically assess disc degeneration. This system categorizes the disc degeneration process into five levels, with Grades I–III indicating mild degeneration and normal disc height and Grades IV and V indicating severe degeneration and collapse of the disc height. Diverse research has employed radiomics or DL to automate the identification of disc degeneration, emphasizing the critical role of the Pfirrmann grade in disc disorders; the MSU classification of discs on axial MRI images is regarded a crucial indicator for surgical indications. This method divides disc herniation (degree and size) into categories 1-2-3, and the protrusion site (location) into A-B-C regions. Surgical validation showed 98% reliability with postoperative ODI improvement, indicating the widespread clinical applicability of this method.[26] Zhang et al[27] categorized the MSU classification into grades 0, I, II, and III by employing faster R-CNN localization to identify the effective LDH area. Subsequently, a prediction model was developed using ResNext101, which demonstrated effective performance. The straight-leg raise and strengthening tests, particularly the contralateral versions, are critical indicators of LDH severity. For patients with progressive neurological dysfunction, early surgery is advised,[28] with greater functional recovery than that in the non-surgical group; six months may serve as a threshold.[29,30] The paraspinal muscles play a key role in maintaining the stability and dynamic regulation of the spine, with the multifidus muscle located on the innermost side having a prominent effect.[31] Current research indicates that MMFI is associated with LBP in young patients.[32,33] Additionally,

studies have demonstrated a correlation between the MMFI and lumbar disc degeneration (LDD). Faur et al[34] assessed the degree of LDD at L4/5 and L5/S1 by analyzing lumbar MRI from 35 patients with CLBP and found a strong correlation between LDD and MMFI. In this study, we qualitatively analyzed MMFI using the Goutallier grading system, a widely accepted standard for evaluating fat infiltration in paraspinal muscles. The criteria were as follows: 0 points for almost no fat in the muscle, 1 point for fat stripes, 2 points for less fat than muscle, 3 points for roughly equal amounts of muscle and fat, and 4 points for more fat than muscle tissue.[35,36] We further classified 0–2 points as mild and 3–4 points as severe. The results confirmed that the MMFI grade in patients undergoing surgery for LDH was higher than that in the non-surgery group, which also indicated that the MMFI was correlated with the severity of LDH. In summary, we included ODI, which represents patients' subjective feelings; SLRT, which represents patients' clinical signs; lumbar MR parameters such as Pfirrmann grade, MSU classification, and MMFI; and combined DLR to construct a nomogram.

MR mDixon is a technique that utilizes the chemical shift effect to perform water-fat separation through in-phase and anti-phase imaging. This enables quantitative fat analysis, which has demonstrated exceptional efficacy.[37] Currently, the use of mDixon is expanding in the assessment of lumbar paraspinal muscle fat infiltration and sarcopenia. Khil et al[35] reported the use of mDixon to evaluate lumbar and dorsal muscle fat infiltration in asymptomatic volunteers and yielded promising results. Chen et al[38] reported that fat infiltration in the multifidus muscle increases with age. mDixonTSE enabled T1-, T2-, and PD-weighted imaging, offering precise water-fat separation technology that was particularly effective in situations involving difficult field homogenization, irregular anatomical structures, or metal implants. Additionally, it produces superior fat suppression images, especially suited for intricately structured areas such as the spine, thereby facilitating the provision of highly accurate imaging parameters that are helpful in model feature extraction and the improvement of model efficiency.

Limitations: 1. Despite being a multicenter trial, the target population exhibited regional characteristics and bias, necessitating future testing with large samples from multiple centers and ethnicities. 2. The focus of this study was young adults; conversely, elderly patients primarily suffer from lumbar spinal stenosis, which has a more complex etiology than LDH and could be a focus of future research. 3. Factors such as smoking, basic diseases, adverse medication history, income, and education level did not contribute to the final clinical features, owing to the limited representation of young patients in this study. Prospective multicenter studies are recommended to enhance the efficacy of the model.

Conclusion

The results of this study confirm that the DLR prediction model surpasses standalone clinical models in assessing surgical indications in young patients with LDH. Additionally, a comprehensive nomogram was developed to aid clinicians in better managing surgical indications, thereby guiding patients towards early treatment, symptom relief, and improved long-term prognosis.

Data Sharing Statement

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Ethics Approval and Consent to Participate

This study was approved by the Shengjing Hospital Ethics Committee (approval number KYCS2024040), in conformity to the Declaration of Helsinki, and a waiver of informed consent was granted given the retrospective nature of the study and the minimal risk involved.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

Disclosure

The authors report no conflicts of interest in this work.

References

1. Buchbinder R, van Tulder M, Öberg B, et al. Low back pain: a call for action. *Lancet*. 2018;391(10137):2384–2388. doi:10.1016/S0140-6736(18)30488-4

2. Angarita-Fonseca A, Pagé MG, Meloto CB, et al. The Canadian version of the National Institutes of Health minimum dataset for chronic low back pain research: reference values from the Quebec low back pain study. *Pain*. 2023;164(2):325–335. doi:10.1097/j.pain.0000000000002703

3. Özüdoğru A, Canlı M, Ceylan İ, Kuzu Ş, Alkan H, Karaçay BÇ. Five times sit-to-stand test in people with non-specific chronic low back pain-a cross-sectional test-retest reliability study. *Ir J Med Sci*. 2023;192(4):1903–1908. doi:10.1007/s11845-022-03223-3

4. Saravi B, Zink A, Ülkümen S, et al. Clinical and radiomics feature-based outcome analysis in lumbar disc herniation surgery. *BMC Musculoskelet Disord*. 2023;24(1):791. doi:10.1186/s12891-023-06911-y

5. Arts MP, Kuršumović A, Miller LE, et al. Comparison of treatments for lumbar disc herniation: systematic review with network meta-analysis. *Medicine (Baltimore)*. 2019;98(7):e14410. doi:10.1097/MD.0000000000014410

6. Matsuyama Y, Chiba K, Iwata H, Seo T, Toyama Y. A multicenter, randomized, double-blind, dose-finding study of condoliase in patients with lumbar disc herniation. *J Neurosurg Spine*. 2018;28(5):499–511. doi:10.3171/2017.7.SPINE161327

7. Zhao Y, Zhao T, Chen S, et al. Fully automated radiomic screening pipeline for osteoporosis and abnormal bone density with a deep learning-based segmentation using a short lumbar mDixon sequence. *Quant Imaging Med Surg*. 2022;12(2):1198–1213. doi:10.21037/qims-21-587

8. Song Y, Zhang J, Zhang YD, et al. FeAture Explorer (FAE): a tool for developing and comparing radiomics models. *PLoS One*. 2020;15(8):e0237587. doi:10.1371/journal.pone.0237587

9. Mayerhoefer ME, Materka A, Langs G, et al. Introduction to Radiomics. *J Nucl Med*. 2020;61(4):488–495. doi:10.2967/jnumed.118.222893

10. Song MX, Yang H, Yang HQ, Li SS, Qin J, Xiao Q. MR imaging radiomics analysis based on lumbar soft tissue to evaluate lumbar fascia changes in patients with low back pain. *Acad Radiol*. 2023;30(11):2450–2457. doi:10.1016/j.acra.2023.02.038

11. Kocaman H, Yıldırım H, Gökşen A, Arman GM. An investigation of machine learning algorithms for prediction of lumbar disc herniation. *Med Biol Eng Comput*. 2023;61(10):2785–2795. doi:10.1007/s11517-023-02888-x

12. Bečulić H, Begagić E, Skomorac R, Mašović A, Selimović E, Pojskić M. ChatGPT's contributions to the evolution of neurosurgical practice and education: a systematic review of benefits, concerns and limitations. *Med Glas*. 2024. doi:10.17392/1661-23

13. Dong D, Fang MJ, Tang L, et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. *Ann Oncol*. 2020;31(7):912–920. doi:10.1016/j.annonc.2020.04.003

14. Zhang J, Liu J, Liang Z, et al. Differentiation of acute and chronic vertebral compression fractures using conventional CT based on deep transfer learning features and hand-crafted radiomics features. *BMC Musculoskelet Disord*. 2023;24(1):165. doi:10.1186/s12891-023-06281-5

15. Zheng HD, Sun YL, Kong DW, et al. Deep learning-based high-accuracy quantitation for lumbar intervertebral disc degeneration from MRI. *Nat Commun*. 2022;13(1):841. doi:10.1038/s41467-022-28387-5

16. Zhou Y, Liu Y, Chen Q, Gu G, Sui X. Automatic lumbar MRI detection and identification based on deep learning. *J Digit Imaging*. 2019;32(3):513–520. doi:10.1007/s10278-018-0130-7

17. Hallinan JTPD, Zhu L, Yang K, et al. Deep learning model for automated detection and classification of central canal, lateral recess, and neural foraminal stenosis at lumbar spine. *MRI Radiology*. 2021;300(1):130–138. doi:10.1148/radiol.2021204289

18. Won D, Lee HJ, Lee SJ, Park SH. Spinal stenosis grading in magnetic resonance imaging using deep convolutional neural networks. *Spine*. 2020;45(12):804–812. doi:10.1097/BRS.0000000000003377

19. Harada GK, Siyaji ZK, Mallow GM, et al. Artificial intelligence predicts disk re-herniation following lumbar microdiscectomy: development of the "RAD risk profile. *Eur Spine J*. 2021;30(8):2167–2175. doi:10.1007/s00586-021-06866-5

20. Abdu RW, Abdu WA, Pearson AM, Zhao W, Lurie JD, Weinstein JN. Reoperation for recurrent intervertebral disc herniation in the spine patient outcomes research trial: analysis of rate, risk factors, and outcome. *Spine*. 2017;42(14):1106–1114. doi:10.1097/BRS.0000000000002088

21. Gaonkar B, Villaroman D, Beckett J, et al. Quantitative analysis of spinal canal areas in the lumbar spine. an imaging informatics and machine learning study. *AJNR Am J Neuroradiol*. 2019;40(9):1586–1591. doi:10.3174/ajnr.A6174

22. Hashia B, Mir AH. Texture features' based classification of MR images of normal and herniated intervertebral discs. *Multimedia Tools Appl*. 2020;79(21):15171–15190. doi:10.1007/s11042-018-7011-4

23. Zhang K, Liu C, Pan J, et al. Use of MRI-based deep learning radiomics to diagnose sacroiliitis related to axial spondyloarthritis. Eur. *J Radiol*. 2024;172:111347. doi:10.1016/j.ejrad.2024.111347

24. Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. *IEEE*. 2016. doi:10.1109/CVPR.2017.634

25. Hara S, Andresen H, Solheim O, et al. Effect of spinal cord burst stimulation vs placebo stimulation on disability in patients with chronic radicular pain after lumbar spine surgery: a randomized clinical trial. *JAMA*. 2022;328(15):1506–1514. doi:10.1001/jama.2022.18231

26. 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(5):501–509. doi:10.1111/papr.12878

27. Zhang W, Chen Z, Su Z, et al. Deep learning-based detection and classification of lumbar disc herniation on magnetic resonance images. *JOR Spine*. 2023;6(3):e1276. doi:10.1002/jsp2.1276

28. Costa F, Oertel J, Zileli M, Restelli F, Zygourakis CC, Sharif S. Role of surgery in primary lumbar disk herniation: WFNS spine committee recommendations. *World Neurosurg X*. 2024;22:100276. doi:10.1016/j.wnsx.2024.100276

29. Petr O, Glodny B, Brawanski K, et al. Immediate versus delayed surgical treatment of lumbar disc herniation for acute motor deficits: the impact of surgical timing on functional outcome. *Spine*. 2019;44(7):454–463. doi:10.1097/BRS.0000000000002295

30. Siccoli A, Staartjes VE, de Wispelaere MP, Schröder ML, de Wispelaere MP. Association of time to surgery with leg pain after lumbar discectomy: is delayed surgery detrimental? *J Neurosurg Spine*. 2019;32(2):160–167. doi:10.3171/2019.8.SPINE19613

31. Bailey JF, Fields AJ, Ballatori A, et al. The relationship between endplate pathology and patient-reported symptoms for chronic low back pain depends on lumbar paraspinal muscle quality. *Spine*. 2019;44(14):1010–1017. doi:10.1097/BRS.0000000000003035

32. Özcan-Ekşi EE, Ekşi MŞ, Turgut VU, Canbolat Ç, Pamir MN. Reciprocal relationship between multifidus and psoas at L4-L5 level in women with low back pain. *Br J Neurosurg*. 2021;35(2):220–228. doi:10.1080/02688697.2020.1783434

33. Fan Z, Wang T, Wang Y, Zhou Z, Wu T, Liu D. Risk factors in patients with low back pain under 40 years old: quantitative analysis based on computed tomography and magnetic resonance imaging mDIXON-quant. *J Pain Res*. 2023;16:3417–3431. doi:10.2147/JPR.S426488

34. Faur C, Patrascu JM, Haragus H, Anglitoiu B. Correlation between multifidus fatty atrophy and lumbar disc degeneration in low back pain. *BMC Musculoskelet Disord*. 2019;20(1):414. doi:10.1186/s12891-019-2786-7

35. Khil EK, Choi JA, Hwang E, Sidek S, Choi I. Paraspinal back muscles in asymptomatic volunteers: quantitative and qualitative analysis using computed tomography (CT) and magnetic resonance imaging (MRI). BMC. *Musculoskelet Disord*. 2020;21(1):403. doi:10.1186/s12891-020-03432-w

36. Hu X, Feng Z, Shen H, et al. New MR-based measures for the evaluation of age-related lumbar paraspinal muscle degeneration. Eur Spine. *Eur Spine J*. 2021;30(9):2577–2585. doi:10.1007/s00586-021-06811-6

37. Kukuk GM, Hittatiya K, Sprinkart AM, et al. Comparison between modified Dixon MRI techniques, MR spectroscopic relaxometry, and different histologic quantification methods in the assessment of hepatic steatosis. *Eur Radiol*. 2015;25(10):2869–2879. doi:10.1007/s00330-015-3703-6

38. Chen P, Zhou Z, Sun L, et al. Quantitative multi-parameter assessment of age- and gender-related variation of back extensor muscles in healthy adults using Dixon MR imaging. *Eur Radiol*. 2024;34(1):69–79. doi:10.1007/s00330-023-09954-w