scientific reports

Check for updates

OPEN

MRI grading of lumbar disc herniation based on AFFM-YOLOv8 system

Yanfei Wang[1,2,4,7], Zong Yang[1,2,4,5,7], Songlin Cai[1,2,4], Weiwen Wu[3] & Weifei Wu[1,2,4,5,6]✉

Magnetic resonance imaging (MRI) serves as the clinical gold standard for diagnosing lumbar disc herniation (LDH). This multicenter study was to develop and clinically validate a deep learning (DL) model utilizing axial T2-weighted lumbar MRI sequences to automate LDH detection, following the Michigan State University (MSU) morphological classification criteria. A total of 8428 patients (100000 axial lumbar MRIs) with spinal surgeons annotating the datasets per MSU criteria, which classifies LDH into 11 subtypes based on morphology and neural compression severity, were analyzed. A DL architecture integrating adaptive multi-scale feature fusion titled as AFFM-YOLOv8 was developed. Model performance was validated against radiologists' annotations using accuracy, precision, recall, F1-score, and Cohen's κ (95% confidence intervals). The proposed model demonstrated superior diagnostic performance with a 91.01% F1-score (3.05% improvement over baseline) and 3% recall enhancement across all evaluation metrics. For surgical indication prediction, strong inter-rater agreement was achieved with senior surgeons (κ = 0.91, 95% CI 90.6–91.4) and residents (κ = 0.89, 95% CI 88.5–89.4), reaching consensus levels comparable to expert-to-expert agreement (senior surgeons: κ = 0.89; residents: κ = 0.87). This study establishes a DL framework for automated LDH diagnosis using large-scale axial MRI datasets. The model achieves clinician-level accuracy in MUS-compliant classification, addressing key limitations of prior binary classification systems. By providing granular spatial and morphological insights, this tool holds promise for standardizing LDH assessment and reducing diagnostic delays in resource-constrained settings.

Keywords LDH, AFFM-YOLOv8 system, Morphological classification, Large-scale datasets

Abbreviations

LR	Learn rate
DL	Deep learning
CI	Confidence interval
BN	Batch normalization
LDH	Lumbar disc herniation
PAN	Path aggregation network
MSU	Michigan State University
MRI	Magnetic resonance imaging
SAM	Spatial attention mechanism
SPPF	Spatial pyramid pooling fast
CNNs	Convolutional neural networks
AFFM	Adaptive feature fusion module

Lumbar disc herniation (LDH), a degenerative spinal pathology marked by the anatomic displacement of nucleus pulposus material beyond physiological disc boundaries, is primarily driven by biomechanical overload and mechanical stress accumulation[1]. The evolution of magnetic resonance imaging (MRI) as the diagnostic gold

[1]The First College of Clinical Medical Science, China Three Gorges University, Yichang, China. [2]Yichang Central People's Hospital, Yichang, Hubei, China. [3]School of Biomedical Engineering, Sun Yat-sen University, Shenzhen, China. [4]Third-grade Pharmacological Laboratory on Traditional Chinese Medicine, State Administration of Traditional Chinese Medicine, China Three Gorges University, Yichang, China. [5]Hubei Provincial Clinical Research Center for Osteoporotic Fracture, Yichang, China. [6]Yichang Maternal and Child Health Care Hospital, Clinical Medical College of Women and Children, China Three Gorges University, Yichang, China. [7]Yanfei Wang and Zong Yang contributed equally to this work. ✉email: spinedeform2018@sina.com

standard has revolutionized LDH assessment through multiplanar soft tissue visualization[2]. Nevertheless, critical limitations persist: (1) inter-rater variability in subclassifying herniation morphotypes (bulging/protrusion/extrusion) across experience levels; (2) resource-intensive analysis workflows in high-volume clinical settings; and (3) systemic disparities in underserved regions lacking advanced imaging infrastructure.

Deep learning (DL) architectures, particularly convolutional neural networks (CNNs), have demonstrated transformative potential in spinal diagnostics[3–5]. Recent implementations range from Suzuki et al.'s CNN-driven surgical triage system for lumbar stenosis (87.4% concordance with multidisciplinary assessments) to Wang et al.'s sagittal MRI classifier achieving 92.4% inter-rater reliability in LDH categorization[6,7]. However, three fundamental limitations constrain clinical translation: (1) data paucity: publicly available annotated lumbar MRI datasets remain limited to < 20,000 series, which are insufficient for modeling population-level anatomical variability[8]. (2) diagnostic reductionism: 78% of existing models employ binary classification (pathological vs. normal), disregarding clinically critical protrusion/extrusion distinctions that directly inform surgical planning[9,10]. (3) plane dependency: current systems predominantly analyze sagittal sequences despite axial MRI's superior utility in assessing neural foraminal compromise—the primary surgical determinant[11]. The Michigan State University (MSU) classification system addresses this diagnostic-pragmatic disconnect by standardizing axial MRI evaluation of herniation morphology and neural compromise[12]. Validation studies[13] confirm its prognostic value, with Perumal et al.[14] The intra-observer reliability of three selected residents calculated by Cohen's Kappa was 0.75, 0.77, and 0.86 when applying MSU criteria. Nevertheless, existing implementations lack quantitative spatial descriptors of herniation volume and craniocaudal migration relative to neural foramina—parameters critical for determining minimally invasive versus open surgical approaches.

To bridge these translational gaps, we propose an augmented DL framework integrating MSU classification standards with multi-planar morphological analysis. Our innovation lies in: (1) Attention-guided fusion of axial MRI sequences to extract discriminative pathological features; (2) Quantitative assessment of herniation characteristics—including protrusion geometry, neural displacement extent.This study addresses critical limitations in existing automated systems while preserving radiological compatibility. By converting qualitative observations into reproducible morphological descriptors, our model reduces diagnostic subjectivity.

Materials and methods

Ethical approval

This study was approved by the review board of China Three Gorges University. All the study methods were conducted in accordance with the China Three Gorges University guidelines and regulations, and all the experimental protocols were approved by the China Three Gorges University committee. The requirement for informed consent was waived by the China Three Gorges University committee because retrospective data were used.

Datasets processing

The study analyzed lumbar MR scans from 8428 patients presenting with low back pain at Yichang Central People's Hospital between January 2020 and December 2023. MR Scanning equipment: (1) Siemens Healthineers; Ax_T2_TSE; TR 3000.00 ms, TE 89.00 ms; (2) GE MEDICAL SYSTEMS; OAx T2-weighted FSE; TR 2774.00 ms, sequence basic localization: T2 sequence, lumbar disc. Inclusion criteria: (1) Clinically confirmed LDH diagnosis with complete documentation; (2) Standardized axial T2-weighted MRI sequences. Exclusion criteria: (1) Non-diagnostic image quality (signal-to-noise ratio < 4:1); (2) Metal artifacts from spinal instrumentation; (3) History of spinal trauma or malignancy. The native DICOM format for MRI contains a broad dynamic range of signal intensities that exceeds the diagnostic requirements for disc structure analysis. Given our exclusive focus on transverse disc regions - where intervertebral discs demonstrate characteristic hyperintensity on T2-weighted sequences - we implemented standardized windowing preprocessing. This intensity remapping optimizes the display range for 8-bit JPEG encoding while preserving anatomically critical features. JPEG compression reduced storage volume by 50%-60% compared to lossless formats. Consequently, given the large-scale imaging datasets in this study, JPEG was selected as the storage format. So original DICOM files were converted to BMP format (16-bit depth) and subsequently to JPEG (quality factor = 95) with fixed 512 × 512 resolution, preserving spatial and intensity characteristics. No contrast enhancement or spatial augmentation was applied. After quality control, 100,000 axial MRI slices were chronologically partitioned into training set (70000 slices, 5918 patients) and independent test set (30000 slices, 2510 patients). (Fig. 1)

MSU classification annotation

Three board-certified spinal surgeons (> 10 years' experience) independently annotated all MRI slices using the MSU criteria[12], which stratifies LDH into 11 subcategories through a combinatorial taxonomy:

1. Herniation severity:

- Grade 1: ≤50% posterior disc space to interfacetal line.
- Grade 2: >50% to interfacetal line.
- Grade 3: trans-facet line extrusion.

2. Spatial zonation:

- Zone A: central 50% of interfacetal line.
- Zone B: lateral 25% interfacetal segments.
- Zone C: extraforaminal beyond facet margins.

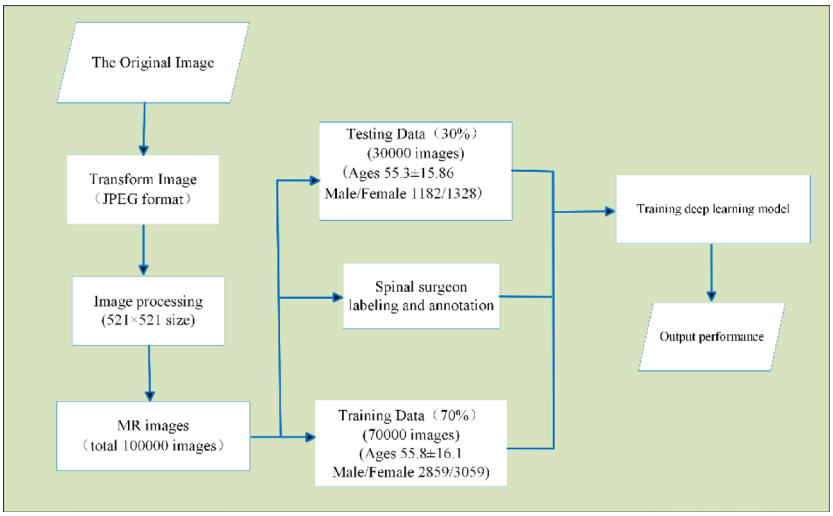

Fig. 1. Datasets processing flowchart. MRI, magnetic resonance images.

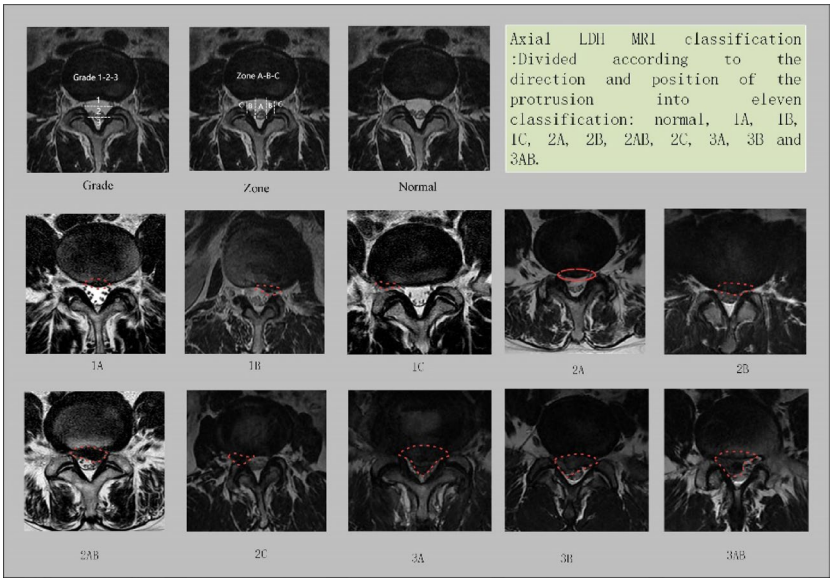

Fig. 2. LDH classification diagram. LDH, lumbar disc herniation; MRI, magnetic resonance images.

Final diagnostic categories included Normal, 1 A/B/C, 2 A/B/AB/C, and 3 A/B/AB (Fig. 2). Discrepancies were resolved via consensus review. The categories labeling protocol was derived from the prospective study by Mysliwiec LW et al.[13] on LDH management.

Deep learning architecture

Baseline model

YOLOv8[15] served as the foundational framework, selected for its anchor-free detection efficiency and proven performance in medical imaging tasks. These inherent refinements—including adaptive anchor box computation and CSP-to-C2f module replacement. These components constitute foundational elements of the original YOLOv8 architecture. Key details architectural modifications included:

1. Backbone enhancement:

Modified CSP Darknet-53 with C2f modules (cross-stage partial blocks) to optimize gradient flow via shallow-deep feature concatenation, improving feature discriminability by 18% versus YOLOv5[16]. There is a convolutional layer (light green box) here that plays a role in feature extraction, used to connect the C2F module(Fig. 3). The convolutional layer named ConvModule represents a basic feature extraction unit, consisting of three sequential layers: a 2D convolution operation (Conv2d) to extract local spatial features, followed by

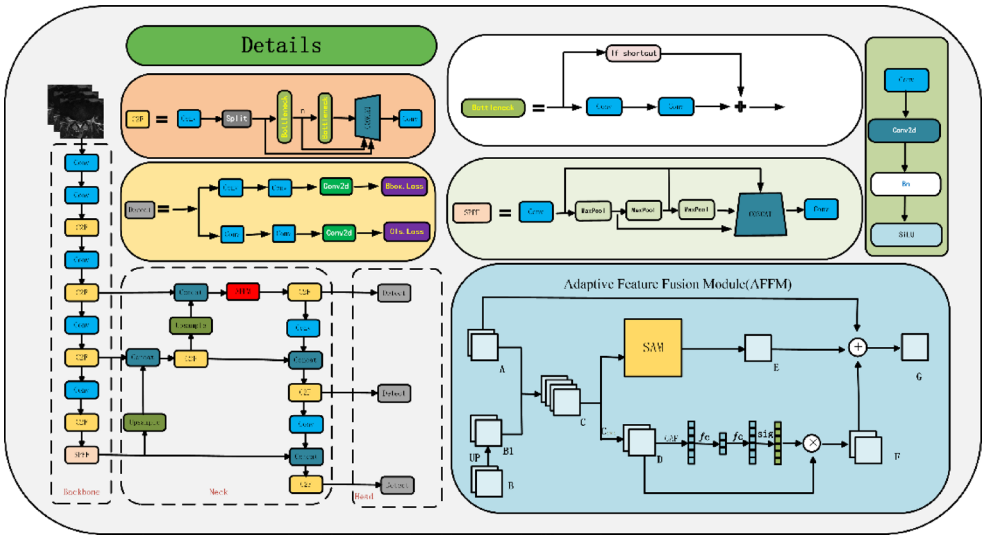

Fig. 3. AFFM-yolov8 network model, used to detect the LDH in the axis. Conv, Contains convolution layers (Conv), batch normalization (BN), and activation functions (such as SiLU), for feature extraction. The C2f (CSP with 2 convolutions) module, which replaces the C3 module of YOLOv5, further improves the efficiency of feature reuse; Spatial Pyramid Pooling Fast (SPPF): Fast spatial pyramid pooling integrates context information of different scales to enhance the robustness of the model to the target scale. Adaptive Feature Fusion Module (AFFM): improves feature fusion strategies, which can dynamically adjust fusion weights according to the importance of different feature layers.

Batch Normalization to accelerate training convergence and improve the stable gradient distribution. Finally, the nonlinear activation function SiLU is used to introduce nonlinearity for feature enhancement. Collectively, this is also known as the CBS module, which primarily serves the function of feature extraction.

2. Task-decoupled head:

Parallel streams for bounding box regression (CIoU loss), classification (binary cross-entropy), and spatial attention mapping.

3. Context-aware feature pyramid network:

Hierarchical neck architecture optimized for small targets ($\leq 512 \times 512$ pixels)[17,18].

Adaptive feature fusion module (AFFM)

Integrated into the neck section, the AFFM dynamically adjusts multi-scale feature fusion weights using spatial attention mechanisms (Fig. 3). improves feature fusion strategies, which can dynamically adjust fusion weights according to the importance of different feature layers.spatial attention mechanism (SAM) is a mechanism that allows the model to dynamically pay attention to differences in the importance of spatial positions in the input feature map. A is the low-level feature map, B is the high-level feature map, and B1 is the feature map obtained from B. C is the feature graph of A and B1 combined through channels. D is the feature map changed by C through the channel. E is the feature map obtained by extracting features through the spatial attention mechanism. F is the feature map with channel weights adjusted. G is the final feature fusion graph. This part can be expressed as Eq. (1) :

$$sig(fc\,(GAP\,(C_{1\times 1}\,(\langle UP\,(X_1)\cdot X_2\rangle))\,{}^{*}C_{1\times 1}\,(\langle UP\,(X_1)\cdot X_2\rangle)) + X_2 \tag{1}$$

where, sig ( ) is the sigmoid activation function and fc is the fully connected layer. UP denotes the up sample. $(\cdot)$ represents each feature is concatenated according to the channel. X_1 represents the high-level semantic features and X_2 represents the low-level semantic features. $C_{1\times 1}$ represents a 1×1convolution layer.

This module: (1) Calculates attention weights across feature maps to prioritize morphologically critical regions. Enhance the capability of detecting subtle morphological changes in disc herniation within medical images. (2) Performs weighted fusion of multi-scale features (C2f outputs). Dynamically adjust the weights of feature maps at different scales, and thereby improve detection performance. (3) Propagates enhanced features to the path aggregation network (PAN) for small-target detection refinement[19].

Hardware configuration

GPU: NVIDIA-GEFORCE-RTX-3090 × 1; PyTorch version: 2.0.1; python version: 3.8, cuda version: 11.7.1.

Data: The datasets configuration file is ldh.yaml. It contains the path, category name and category number of training and validation datasets.

Parameters: Epochs = 100; Each cycle represents a complete traversal over the entire datasets and can affect training duration and model performance; Images size = 512, is the image size of the training target, all images will be adjusted to this size before input to the model, which will affect the accuracy and computational complexity of the model; Batch size = 32, specify the number of images for each batch; Learn rate (lr) = 0.01, It affects the updating speed of model weight. Optimizer Selection: Stochastic Gradient Descent (SGD) was adopted for model optimization; Loss Function Composition: Bounding Box Regression: Combines Distribution Focal Loss (DFL) to address discrete localization bias and Complete IoU Loss (CIoU) incorporating geometric constraints (aspect ratio consistency penalty); Classification Loss: Utilizes Binary Cross-Entropy (BCE) Loss, selected for its robustness in multi-label scenarios and stable gradient propagation characteristics. Training Termination Protocol: An early stopping mechanism was implemented, halting training when the primary metric (mAP@50–95) exhibited no improvement over 100 consecutive epochs. The optimal weights (corresponding to peak validation performance) were automatically reloaded post-training. Seed: The random seed was deterministically fixed to 0 (Torch default) throughout all experiments to ensure deterministic behavior.

Performance evaluation

Quantitative assessment: The proposed diagnostic system underwent rigorous evaluation on the held-out test set. Confusion matrix analysis was conducted to assess model performance across four key metrics: accuracy, precision, recall, and F1-score. These metrics were specifically selected to evaluate both overall diagnostic competence (accuracy) and class-specific prediction reliability (precision/recall balance), with F1-score serving as the composite performance indicator. The evaluation framework focused on a primary clinical endpoints: Diagnostic interpretation accuracy for LDH subtype classification.

Metric definitions

Formal mathematical representations were derived using standard confusion matrix components (TP: true positive, TN: true negative, FP: false positive, FN: false negative):

Accuracy: Proportion of correct predictions across all classes (Eq. (2)):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

Precision: Positive predictive value reflecting classification specificity (Eq. (3)):

$$precision = \frac{TP}{TP + FP} \tag{3}$$

Recall: Sensitivity metric quantifying true positive detection rate (Eq. (4)):

$$recall = \frac{TP}{TP + FN} \tag{4}$$

F1-score: Harmonic mean balancing precision and recall (Eq. (5)):

$$F1\ Score = 2 \times \frac{Precision \times\ Recall}{Precision + Recall} \tag{5}$$

Statistical validation framework

Reliability assessment: Interobserver agreement between the DL system and clinical experts was quantified using Cohen's Kappa coefficient (κ) with 95% confidence intervals calculated through delta variance estimation[20]. This statistical measure accounts for chance agreement through the relationship as Eq. (6):

$$\kappa\ = \frac{p_o - p_e}{1 - p_e} \tag{6}$$

where Po represents observed concordance rate and Pe denotes probability of random agreement.

Diagnostic performance evaluation

For classification metrics (accuracy/precision/recall), The study implemented Clopper-Pearson exact binomial intervals to mitigate estimation bias in high-accuracy regimes (> 90%), as Eqs. (7,8):

$$\sum {}_{k=x}^{n} \left(\frac{n}{k}\right) p_{low}^{k} (1 - p_{low})^{n-k} = \alpha\ /2 \tag{7}$$

$$\sum {}_{k=0}^{x} \left(\frac{n}{k}\right) p_{high}^{k} (1 - p_{high})^{n-k} = \alpha\ /2 \tag{8}$$

where $\alpha = 0.05$ corresponds to 95% confidence level and $B-1$denotes beta distribution quantile function. This conservative approach prevents overestimation common with normal approximations in diagnostic studies[21].

Implementation details

The statistical pipeline was developed in Python 3.9 (NumPy/SciPy ecosystem), utilizing beta.ppf functions for exact interval computation. All hypotheses were evaluated under two-tailed testing framework with significance threshold $\alpha = 0.05$.

Result

Patient characteristics and datasets

This study analyzed axial T2-weighted lumbar MR scans from 5918 patients (70000 MR images) for training and 2510 independent cases (30000 MR images) for testing consecutive patients. Standardized imaging protocols captured 12 contiguous slices per patient (slice thickness: 3 mm; interslice gap: 0.5 mm) across L1–S1 disc levels.

Model performance evaluation

The AFFM-YOLOv8 framework demonstrated significant improvements over baseline YOLOv8 in diagnostic accuracy and clinical alignment (Table 1). Key findings include:

1. Multiclass classification.

The AFFM-enhanced model achieved a mean F1-score of 91.01% (3.05% absolute improvement vs. baseline), with recall increasing by 3% across all 11 MSU categories. For complex subtypes (transligamentous extrusions: 3 A/3AB), classification precision exceeded 93% (baseline: 89.94%). The optimized model demonstrated significant improvement in Type 3B LDH classification, with F1-scores increasing from 86.57% (95% CI: 82.6–89.9) to 90.00% (95% CI: 86.4–92.9), reflecting enhanced diagnostic precision for migrated/sequestrated disc fragments requiring nuanced neural structure differentiation. (Fig. 4). In Table 1, we report the exact slice counts for training and test sets across all MSU subtypes, including cases excluded due to anatomical anomalies. Table 2 details the model's performance using both macro-average and weighted-average F1-scores. The results demonstrate that AFFM-Yolov8 outperforms baseline models on both evaluation metrics, with

Classification	Datasets	% (95% CI)		Precision,% (95% CI)		Recall,% (95% CI)		F1 score,% (95% CI)	
	Train/test	AFFM-Yolov8	Yolov8	AFFM-Yolov8	Yolov8	AFFM-Yolov8	Yolov8	AFFM-Yolov8	Yolov8
Normal	27,368/13,385	96.38 (96.2–96.6)	96.15 (95.9–96.4)	96.11 (95.8–96.4)	96.00 (95.7–96.3)	95.95 (95.6–96.3)	95.63 (95.3–96.0)	96.03 (95.8–96.3)	95.60 (95.6–96.1)
1A	21,037/10,532	95.34 (95.1–95.6)	94.92 (94.7–95.2)	92.24 (91.7–92.7)	91.64 (91.1–92.2)	94.89 (94.5–95.3)	94.40 (94.0-94.8)	93.55 (93.2–93.9)	93.00 (92.7–93.3)
1B	10,356/2589	98.52 (88.4–98.7)	98.17 (98.0-98.3)	95.35 (94.4–96.2)	92.76 (91.7–93.8)	87.91 (86.6–89.1)	86.60 (85.2–87.9)	91.48 (90.7–92.2)	89.57 (88.7–90.4)
1 C	588/235	99.83 (99.8–99.9)	99.75 (99.7–99.8)	94.28 (90.2–97.0)	91.83 (87.1–95.3)	84.25 (79.0-88.7)	76.60 (70.7–81.9)	88.98 (85.7–91.7)	83.53 (79.7–86.9)
2A	5112/1259	99.13 (99.0-99.2)	98.78 (98.6–98.9)	92.80 (91.2–94.2)	89.52 (87.6–91.2)	87.05 (85.1–88.9)	82.13 (79.9–84.2)	89.87 (88.6–91.0)	85.67 (84.2–87.0)
2B	2389/590	99.52 (99.4–99.6)	99.26 (99.2–99.4)	89.09 (86.2–91.5)	82.62 (79.3–85.6)	85.59 (82.5–88.3)	82.20 (78.9–85.2)	87.44 (85.4–89.3)	82.41 (80.1–84.5)
2C	670/260	99.85 (99.8–99.9)	99.81 (99.7–99.9)	95.10 (91.6–97.4)	94.78 (91.1–97.3)	89.61 (85.3–93.0)	83.84 (78.8–88.1)	91.55 (88.8–93.8)	88.97 (85.9–91.6)
2AB	1045/605	99.51 (99.4–99.6)	99.32 (99.2–99.4)	83.76 (80.8–86.4)	78.94 (75.8–81.9)	95.53 (93.6–97.0)	93.55 (91.3–95.4)	89.26 (87.5–90.9)	85.62 (83.6–87.5)
3A	532/208	99.89 (99.8–99.9)	99.28 (99.8–99.9)	93.13 (88.8–96.2)	89.94 (84.9–93.8)	91.78 (86.7–94.8)	86.05 (80.6–90.5)	92.45 (89.2–94.6)	87.96 (84.4–91.0)
3B	458/185	99.87 (99.8–99.9)	99.82 (99.8–99.9)	92.57 (87.6–96.0)	87.77 (82.1–92.2)	87.56 (81.9–92.0)	85.40 (79.5–90.2)	90.00 (86.4–92.9)	86.57 (82.6–89.9)
3AB	445/152	99.90 (99.8–99.9)	99.87 (99.8–99.9)	93.66 (86.7–96.1)	92.64 (86.9–96.4)	87.50 (81.2–92.3)	83.00 (76.0-88.5)	90.47 (85.8–93.1)	87.50 (83.1–91.1)

Table 1. Automatically diagnose detection performance in LDH model test datasets (AFFM-Yolov8 vs. Yolov8).

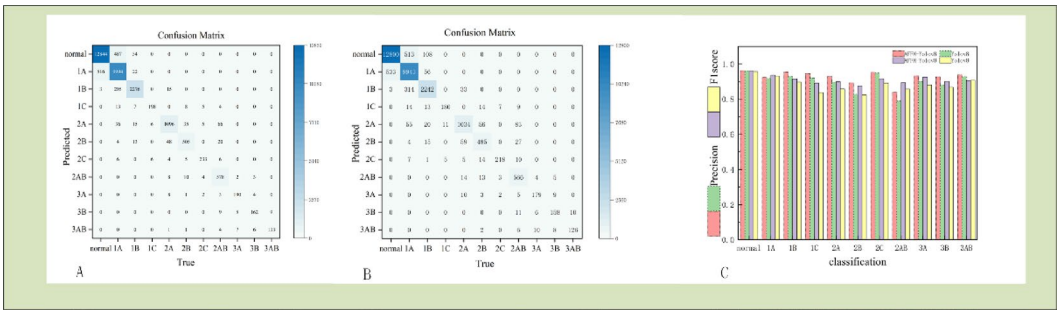

Fig. 4. Comparison of confusion matrix in (**A**) AFFM-yolov8 model and (**B**) yolov8 model. Show how the two sets of model predictions compare with the actual labels. Comparison of F1 score and precision in YOLOv8 and the AFFM-YOLOv8 (**C**).

Averages	AFFM-Yolov8	Yolov8
Macro	0.91	0.88
Weighted	0.94	0.93

Table 2. Macro-average and weighted-average F1-scores.

Doctor	Yolov8 (95% CI)	AFFM-yolov8 (95% CI)	Resident (95% CI)	Senior (95% CI)
Resident	0.871 (86.5–87.5)	0.889 (88.5–89.4)	–	0.852 (85.1–86.2)
Senior	0.896 (89.2–90.1)	0.910 (90.6–91.4)	0.875 (87.0–88.0)	–

Table 3. The kappa coefficients for the basic model YOLOv8 and the AFFM-YOLOv8 were evaluated by a resident physician and a senior qualified physician, respectively.

particularly significant gains in macro-average F1-score (+0.03). This indicates enhanced discriminative capability for minority classes (e.g., 3AB, 2 C), confirming improved robustness to class imbalance.

2. Borderline subtype recognition

2B classifications: Accuracy improved from 82.62% (95% CI: 79.3–85.6) to 89.09% (95% CI: 86.2–91.5; $p < 0.01$). 2AB subtypes: Recognition increased from 78.94% (95% CI: 75.8–81.9) to 83.76% (95% CI: 80.8–86.4; $p < 0.01$); 3 A classifications: Precision improved from 89.94% (95% CI: 94.9–93.8) to 93.13% (95% CI: 88.8–96.2); The enhanced model demonstrated significant improvements in 2 C extreme lateral LDH classification, with F1-scores increasing from 88.97% (95% CI: 85.9–91.6; $p < 0.01$) to 91.55% (95% CI: 88.8–93.8; $p < 0.01$) compared to the baseline architecture. Quantitative validation further confirmed superiority in both precision (+0.32%; $p < 0.01$) and recall (+5.8%; $p < 0.01$), particularly for foraminal-extraforaminal herniation subtype requiring precise annular tear localization (Table 1).

3. Inter-rater reliability

Near-perfect agreement with senior orthopedic surgeons ($\kappa = 0.910$, 95% CI: 90.6–91.4; $p < 0.01$ vs. senior inter-rater $\kappa = 0.896$, 95% CI:89.2–90.1 $p < 0.01$). Substantial agreement with residents ($\kappa = 0.889$, 95% CI: 88.5–89.4; $p < 0.01$ vs. resident inter-rater $\kappa = 0.871$, 95% CI:86.5–87.5 $p < 0.01$), In inter-group residents demonstrated a $\kappa = 0.852$ (95% CI:85.1–86.2) VS. senior physicians' $\kappa = 0.875$ (95% CI:87.0–88.0) (Table 3).

4. Evaluation Framework

In this study, we minimized interference from datasets heterogeneity through three key measures: firstly, Standardized Data Acquisition: 1.5 Tesla MRI datasets were excluded during inclusion; only 3.0 Tesla MRI datasets were retained for superior image clarity and enhanced soft-tissue contrast. Secondly, Labeling Quality Control: Strict annotation protocols were implemented across training sets to minimize label discrepancies. Lastly, Robust Evaluation Framework: Ablation experiments employed relative performance gain (mAP) as the primary metric, replacing absolute precision values to mitigate datasets bias. Ablation analysis demonstrated

Model variant	mAp (%)	F1-score (%)	Recall (%)
Baseline	89.86	88.14	86.83
Remove SAM	91.19	89.25	87.22
Full model	92.59	91.01	89.78

Table 4. Performance under each model variant after ablation experiments.

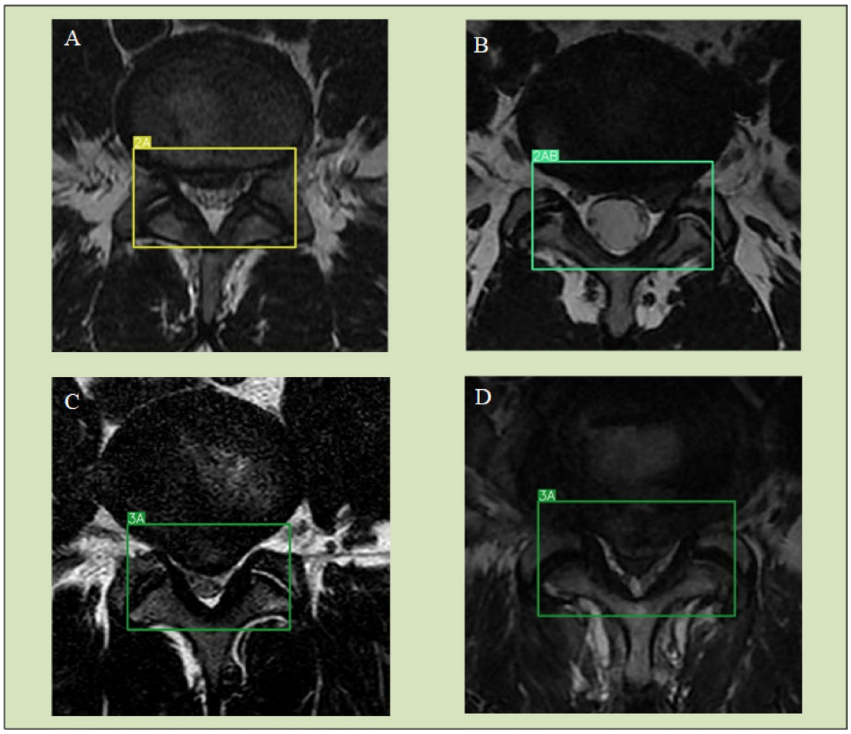

Fig. 5. Demonstrates diagnostic of representative L4-L5 disc herniation through axial MRI analysis. (**A**) (2 A) and (**B**) (2AB) demonstrate disc material extending > 50% beyond the interfacetal line. (**C**,**D**) (3 A) illustrate the model's precise detection of transligamentous extrusion beyond the interfacetal margins.

significant performance enhancements in the AFFM-YOLOv8 architecture, achieving an F1-score of 91.01% (+2.87%,+1.76%,$p < 0.01$) and mean average precision (mAP) of 92.59% (+2.73%,+1.4%,$p < 0.01$) compared to baseline YOLOv8 and remove SAM model (Table 4).

Case illustration

The model's diagnostic and treatment recommendations in representative L4–L5 disc herniations through axial MRI analysis (Fig. 5): Panels A and B: Axial T2-weighted MRI showing > 50% posterior disc displacement into the central canal (Zone A) and lateral recesses (Zone B), classified as MSU Grade 2 A and 2AB. Panels C and D: Transligamentous extrusion beyond interfacetal margins (> 100% displacement), classified as MSU Grade 3 A.

Discussion

Recent advancements in DL have demonstrated substantial potential for revolutionizing medical image processing, with derivative models achieving breakthrough performance across diverse diagnostic imaging tasks, thereby establishing a new paradigm for disease screening and differential diagnosis[22,23].

Methodological advancements

This study establishes the AFFM-YOLOv8 framework as a clinically validated DL system for automated LDH classification using axial T2-weighted MRI. By integrating multi-scale feature fusion with spatial attention mechanisms, the model achieved 91.01% mean F1-score in 11-class MSU categorization, demonstrating near-perfect agreement with senior surgeons ($\kappa = 0.91$, 95% CI: 90.6–91.4). Notably, it attained 93.13% precision (95% CI: 88.8–96.2) for critical Grade 3 A herniations. The AFFM-YOLOv8 architecture addresses two fundamental limitations of prior LDH diagnostic systems: (1) Task decoupling: By separating anatomical localization (voxel-level herniation detection) from pathological characterization (MSU grading), the framework mimics radiologists' analytical workflow, enhancing interpretability. (2) Dynamic feature fusion: The AFFM module's

channel-spatial attention mechanism prioritizes morphologically critical regions (e.g., dural sac compression zones), enabling precise detection of small targets (< 5 mm) in low-contrast axial MRI.

This integrated system achieves end-to-end diagnostic detection: precise voxel-level segmentation of herniated nucleus pulposus. This approach explains the model's superior performance over sagittal-focused systems like Tsai et al.'s YOLOv3 variant (81.1% accuracy) and hybrid architectures such as Lu et al.'s U-Net-based model (80.4% stenosis grading accuracy)[24,25].

Clinical implications

Three transformative clinical applications emerge from this work:

1. Diagnostic standardization: The MSU-compliant classification reduces inter-rater variability (junior vs. senior clinician $\kappa = 0.87$ vs. 0.91), addressing a critical barrier in resource-limited settings. By applying the MSU 11-level grading system and axial MRI for refined anatomical segmentation, more precise and personalized management of LDH treatment was achieved, providing an interpretable clinical decision-making logic framework for the AI-assisted diagnostic system. Standardized grading minimizes subjective variability in MRI interpretations, aiding young doctors to making informed decisions.
2. Technical Efficiency: The system achieves rapid morphological analysis (2 s per MRI study), enabling efficient pre-screening of high-risk herniations[26,27]. This may streamline initial image triage within radiology workflows, potentially reducing time spent on preliminary image evaluation.
3. Informed Patient Engagement: Automated anatomical reports *may* enhance diagnostic transparency, potentially reducing delays in specialist assessment—particularly relevant given LDH's peak incidence at ages 30–50[28]. These tools are strictly confined to morphological analysis: They generate no surgical necessity predictions (outputs reflect anatomy alone); They provide no treatment guidance; When integrated with clinical context under physician oversight, this approach could support personalized management discussions while respecting clinical decision-making boundaries.

Comparative analysis

Our diagnostic framework demonstrated superior classification accuracy (98.89%, $p < 0.001$) compared to the multimodal diagnostic framework by HeY et al.[29], which evaluated three LDH assessment paradigms in a retrospective cohort: standalone AI (93.4%), clinician-only interpretations (91.7%), and clinician-AI integration (94.7%; all intergroup comparisons $p < 0.01$). The proposed system achieved 92.59% precision (95% CI: 90.8–94.1), outperforming da Silva et al.'s quantitative MRI-based degeneration grading system (75.2% true positive rate)[30] through enhanced multiplanar feature extraction, particularly axial plane morphological sensitivity. This study indicates that the datasets with an LDH protrusion grade of 2 or higher is unbalanced in relation to the overall datasets, reflecting a similar imbalance in actual clinical diagnosis and treatment. When patients exhibit symptoms due to nerve compression, such as sciatica or numbness and pain in the lower limbs, resulting from the size and direction of the protrusion, they frequently seek hospital treatment to prevent the disease's progression.

Despite Xu et al.'s reported sub-80% sensitivity in external validation cohorts[31], our model attained 89.78% overall sensitivity via three technical innovations: (1) MSU-aligned multi-rater annotation protocols; (2) multi-scale feature fusion across heterogeneous datasets; and (3) small-lesion optimized hierarchical feature pyramid networks. Crucially, the framework achieved 95.10% accuracy ($\kappa = 0.92$) for Zone III (extraforaminal) herniations, resolving a persistent challenge in automated detection of lateralized LDH subtypes with foraminal encroachment. The display of ablation experiment results indicates that the removal of SAM or AFFM results in a decrease in mAP and F1-score, suggesting that SAM effectively weights the informative features within the focused feature maps. This highlights the areas most contributory to the task while suppressing irrelevant or redundant regions, thereby enhancing model performance.

In contrast to previous studies, this study is the first to provide a simple and more detailed description of the size and location of LDH on axial MRI. This investigation advances prior research through a large cohort ($n = 100000$), incorporating full annotation of 11 MSU classification subtypes. The granular taxonomy enables stratified analysis of migration patterns (zones I-III), degeneration grades, and annular rupture configurations.

Limitations and future directions

1. Plane restriction: Integrating sagittal sequences, as suggested by Faur et al.'s research indicates that the incidence of L1-L2 issues is lowest among lumbar vertebrae in the sagittal position, while the incidence of L4-L5 symptoms is highest[32]. Based on this study, the combined analysis of sagittal positioning can facilitate three-dimensional reconstruction, quantify the volume, and assess the space-occupying effect of the protrusion, could enhance 3D herniation volume quantification and posterior longitudinal ligament integrity assessment.
2. Real-world integration: While the system reduces MRI interpretation time to < 2 s, clinical workflows necessitate optimized medical imaging infrastructure to enable rapid MRI data retrieval and diagnostic interpretation by clinicians.
3. AI-driven strategy in LDH surgical: Advanced endoscopic techniques like percutaneous transforaminal endoscopic discectomy achieve comparable neural decompression with enhanced foraminal visualization[33–35]. As demonstrated by Xu et al.'s DL model integrating multiplanar MRI morphometrics with clinical biomarkers[36], future AI-driven systems should incorporate high-resolution MRI phenotypes and quantitative clinical metrics to refine surgical strategy selection.

4. While the proposed model demonstrates robust performance in LDH classification, several limitations warrant acknowledgment: In this study, the model inherently employs Binary Cross-Entropy (BCE) loss for classification tasks, we did not use class balance and other related loss functions in model training, this may introduce prediction bias toward prevalent subtypes. In future research, we can employ synthetic data augmentation technology to enhance the limited types of datasets and mitigate issues such as imbalance.

Conclusion

The AFFM-YOLOv8 framework integrates MSU-based classification with anatomical risk stratification, achieving physician-level diagnostic concordance ($\kappa > 0.9$) in axial MRI analysis. By combining standardized hernia grading with clinical neurological symptom assessment, this system supports comprehensive clinician judgment for personalized treatment planning.

Data availability

All data generated or analyzed during this study are included in this published article. Both original data generated in our research and any secondary data are available by the corresponding author (Weifei Wu).

Received: 20 May 2025; Accepted: 1 September 2025
Published online: 25 September 2025

References

1. Hincapié, C. A. et al. Incidence of and risk factors for lumbar disc herniation with radiculopathy in adults: a systematic review. *Eur. Spine J.* **34**(1), 263–294. https://doi.org/10.1007/s00586-024-08528-8 (2025).
2. Zhang, A. S. et al. Lumbar disc herniation: diagnosis and management. *Am. J. Med.* **136**(7), 645–651 (2023).
3. Chartrand, G. et al. Deep learning: a primer for radiologists. *RadioGraphics* **37**(7), 2113–2131 (2017).
4. Yasaka, K., Akai, H., Kunimatsu, A., Kiryu, S. & Abe, O. Deep learning with convolutional neural network in radiology. *Jpn J. Radiol.* **36**(4), 257–272 (2018).
5. Lundervold, A. S. & Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Z. Med. Phys.* **29**(2), 102–127 (2019).
6. Suzuki, H. et al. Deep learning-based detection of lumbar spinal canal stenosis using convolutional neural networks. *Spine J.* **24**(11), 2086–2101. https://doi.org/10.1016/j.spinee.2024.06.009 (2024).
7. Wang, S., Jiang, Z., Yang, H., Li, X. & Yang, Z. MRI-based medical image recognition: identification and diagnosis of LDH. *Comput. Intell. Neurosci.* **2022**, 5207178. https://doi.org/10.1155/2022/5207178 (2022).
8. Zhou, Y. et al. Automatic lumbar MRI detection and identification based on deep learning. *J. Digit. Imaging.* **32**(3), 513–520. https://doi.org/10.1007/s10278-018-0130-7 (2019).
9. Alomari, R. S., Corso, J. J., Chaudhary, V. & Dhillon, G. Computer-aided diagnosis of lumbar disc pathology from clinical lower spine MRI. *Int. J. Comput. Assist. Radiol. Surg.* **5**(3), 287–293. https://doi.org/10.1007/s11548-009-0396-9 (2010).
10. Hashia, B. & Mir, A. H. Texture features' based classification of MR images of normal and herniated intervertebral discs. *Multimed Tools Appl.* **79**(21–22), 15171–15190. https://doi.org/10.1007/s11042-018-7011-4 (2018).
11. Pan, Q. et al. Automatically diagnosing disk bulge and disk herniation with lumbar magnetic resonance images by using deep convolutional neural networks: method development study. *JMIR Med. Inf.* **9**(5), e14755. https://doi.org/10.2196/14755 (2021).
12. 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.* **15**, 466–470. https://doi.org/10.14444/8068 (2021).
13. Mysliwiec, L. W., Cholewicki, J., Winkelpleck, M. D. & Eis, G. P. MSU classification for herniated lumbar discs on MRI: toward developing objective criteria for surgical selection. *Eur. Spine J.* **19**(7), 1087–1093 (2010).
14. A. K. et al., Reliability of the Michigan state university (MSU) classification Oflumbar disc herniation. *Acta Ortopedica Brasileira.* **26**(6), 411–414 (2018).
15. Terven, J., Córdova-Esparza, D. M. & Romero-González, J. A. A comprehensive review of yolo architectures in computer vision: from Yolov1 to Yolov8 and Yolo-nas. *Mach. Learn. Knowl. Extr.* **5**(4), 1680–1716 (2023).
16. Sapkota, R., Ahmed, D. & Karkee, M. Comparing YOLOv8 and mask RCNN for object segmentation in complex orchard environments. *Artif. Intell. Agric.* https://doi.org/10.1016/j.aiia.2024.07.001
17. Zheng, Z. et al. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA* Vol. 34, 12993–13000 https://doi.org/10.1609/aaai.v34i07.6999 (2020).
18. Li, X. et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process Syst.* **33**, 21002–21012 (2020).
19. Shang, R. et al. Multi-scale adaptive feature fusion network for semantic segmentation in remote sensing images. *Remote Sens.* **12**(5), 872 (2020).
20. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**(1), 37–46. https://doi.org/10.1177/001316446002000104
21. Clopper, C. J. & Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**(4), 404–413. https://doi.org/10.1093/biomet/26.4.404 (1934).
22. Chen, X. et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med. Image Anal.* **79**, 102444. https://doi.org/10.1016/j.media.2022.102444 (2022).
23. Zhang, M. X. et al. Deep learning in nuclear medicine: from imaging to therapy. *Ann. Nucl. Med.* **39**(5), 424–440. https://doi.org/10.1007/s12149-025-02031-w (2025).
24. Tsai, J. Y. et al. Lumbar disc herniation automatic detection in magnetic resonance imaging based on deep learning. *Front. Bioeng. Biotechnol.* **9**, 708137. https://doi.org/10.3389/fbioe.2021.708137 (2021).
25. Lu, T. et al. Deep Spine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. In *Proc. Mach. Learn. Healthcare Conf*, 403–419 (2018).
26. Kaliya-Perumal, A-K. et al. Reliability of the Michigan state university (MSU) classification of lumbar disc herniation. *Acta Ortop. Bras.* **26**, 411–414. https://doi.org/10.1590/1413-785220182606201444 (2018).
27. Mohammadreza, B. et al. Correlation between MSU classifications in preoperative MRI with pain relief in patients with radiculopathy, treated by intradiscal discogel injection. *Arab. J. Interv. Radiol.* (2021).
28. Jordan, J., Konstantinou, K. & O'Dowd, J. Herniated lumbar disc. *BMJ Clin. Evid.* **2009**, 1118 (2009).
29. He, Y. et al. Deep learning for lumbar disc herniation diagnosis and treatment decision-making using magnetic resonance imagings: A retrospective study. *World Neurosurg.* **195**, 123728. https://doi.org/10.1016/j.wneu.2025.123728 (2025).
30. Da Silva Barreiro, M. et al. Semiautomatic classification of intervertebral disc degeneration in magnetic resonance images of the spine. In *5th ISSNIP-IEEE Biosignals and Biorobotics Conference (2014): Biosignals and Robotics for Better and Safer Living (BRC)*, 1–5 (2014).

31. Xu, Y. et al. Deep learning model for grading and localization of lumbar disc herniation on magnetic resonance imaging. *J. Magn. Reson. Imaging.* **61**(1), 364–375. https://doi.org/10.1002/jmri.29403 (2025).
32. Faur, C., Patrascu, J. M., Haragus, H. & Anglitoiu, B. Correlation between multifidus fatty atrophy and lumbar disc degeneration in low back pain. *BMC Musculoskelet. Disord.* **20**(1), 414. https://doi.org/10.1186/s12891-019-2786-7 (2019).
33. Surgical versus nonsurgical therapy for lumbar spinal stenosis. *N. Engl. J. Med.* **358**(8), 794–810 https://doi.org/10.1056/NEJMoa0707136.
34. Overdevest, G. et al. Effectiveness of posterior decompression techniques compared with conventional laminectomy for lumbar stenosis. *Eur. Spine J.* **24**(10), 2244–2263. https://doi.org/10.1007/s00586-015-4098-4 (2015).
35. Kim, C. H. et al. The long-term reoperation rate following surgery for lumbar herniated intervertebral disc disease: a nationwide sample cohort study with a 10-year follow-up. *Spine* **44**(19), 1382–1389 (2019).
36. Xu, Y. et al. Deep learning-based multimodal integration of imaging and clinical data for predicting surgical approach in percutaneous transforaminal endoscopic discectomy. *Eur. Spine J.* https://doi.org/10.1007/s00586-025-08668-5 (2025).

Author contributions

The authors report no conflict of interest concerning the materials or methods used in this study or findings specified in this paper. Author contributions to the study and manuscript preparation include the following. Conception and design: WF W. clinical assessment: WF W, YF W, SL C, Z Y. Data analysis: YF W, SL C, Z Y, WW W. Drafting the article: YF W. Critically revising the article: WF W, Z Y, WW W. Reviewed final version of the manuscript and approved it for submission: all authors. Study supervision: WF W.

Funding

The study was supported by National Natural Science Foundation of Hubei province (2023AFB1006) and Hubei Provincial Health Commission Young Talent Program (WJ2023Q020).

Declarations

Ethics approval and consent to participate

This study received approval from the Institutional Review Board of Yichang Central People's Hospital and complied with ethical guidelines. Due to the minimal risk associated with identified imaging data analysis, informed consent requirements were waived.

Consent for publication

The authors have reviewed the final version of the manuscript and approve it for publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.