BMC Musculoskeletal Disorders

RESEARCH

Open Access

Vision transformer-based diagnosis of lumbar disc herniation with grad-CAM interpretability in CT imaging

Qingsong Chu[1,2†], Xingyu Wang[1,2†], Hao Lv[1,2†], Yao Zhou[1,2†] and Ting Jiang[1*]

Abstract

Background In this study, a computed tomography (CT)-vision transformer (ViT) framework for diagnosing lumbar disc herniation (LDH) was proposed for the first time by taking advantage of the multidirectional advantages of CT and a ViT.

Methods The proposed ViT model was trained and validated on a dataset consisting of 983 patients, including 2100 CT images. We compared the performance of the ViT model with that of several convolutional neural networks (CNNs), including ResNet18, ResNet50, LeNet, AlexNet, and VGG16, across two primary tasks: vertebra localization and disc abnormality classification.

Results The integration of a ViT with CT imaging allowed the constructed model to capture the complex spatial relationships and global dependencies within scans, outperforming CNN models and achieving accuracies of 97.13% and 93.63% in terms of vertebra localization and disc abnormality classification, respectively. The performance of the model was further validated via gradient-weighted class activation mapping (Grad-CAM), providing interpretable insights into the regions of the CT scans that contributed to the model predictions.

Conclusion This study demonstrated the potential of a ViT for diagnosing LDH using CT imaging. The results highlight the promising clinical applications of this approach, particularly for enhancing the diagnostic efficiency and transparency of medical AI systems.

Keywords CT, Deep learning, Diagnostic accuracy, Grad-CAM, LDH, Medical imaging, ViT

Background

LDH, which affects 1–2% of adults annually [1–3], results from nerve compression caused by intervertebral disc displacement. While MRI remains the diagnostic gold standard, CT imaging provides superior bone structure visualization and accessibility [4–8]. Research on low-dose CT denoising has further improved its clinical accessibility [9, 10].

AI, particularly deep learning models such as CNNs, has demonstrated success in medical image analysis tasks [11–13], including MRI-based LDH diagnosis [14, 15]. However, CNNs exhibit critical limitations in CT-based LDH analyses: (1) local receptive fields hinder the multi-vertebral relationship modelling process [16], (2) pooling operations degrade fine disc abnormalities [17], and (3) position sensitivity reduces cross-patient generalizability [18].

ViTs address these challenges through self-attention mechanisms [19–21], dynamically weighting interpatch

†Qingsong Chu, Xingyu Wang, Hao Lv and Yao Zhou contributed equally as co-first authors.

*Correspondence:
Ting Jiang
jiangting70@163.com
[1] The First Affiliated Hospital of Anhui University of Chinese Medicine, Hefei, China
[2] Anhui University of Chinese Medicine, Hefei, China

relationships. Unlike CNNs with fixed kernels [18, 19], ViTs focus on clinically relevant interfaces regardless of their spatial positions. Emerging applications validate the potential of ViTs in tasks such as dementia diagnosis via brain structural changes [22], brain tumour classification with cross-attention [23], emphysema subtype detection [24], and COVID- 19 severity grading [25].

The growing demand for interpretable AI in clinical practice [26, 27] contrasts with the current research trends. The current LDH research disproportionately focuses on MRI-CNN frameworks [28–30], neglecting the clinical advantages of CT. Moreover, the absence of interpretable CT-based LDH studies [28–30] and the architectural constraints of CNNs in spinal pathology analysis scenarios [31] underscore the need for innovative approaches.

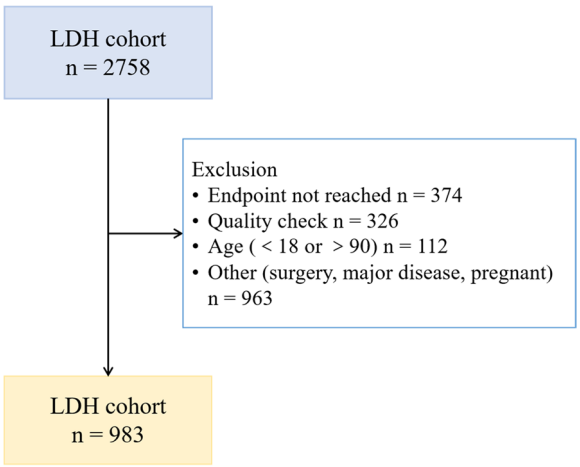

Fig. 1 Patient cohort of this study

Research contributions

- First CT-ViT Integration Method: Unlike the existing MRI-CNN framework, this study proposes an approach that integrates the strengths of CT in bony structure visualization with the global attention mechanism of a ViT for achieving improved LDH diagnoses.
- Multiscale Interpretation: We implement Grad-CAM across different transformer layers to provide interpretable disc pathology visualizations, offering new insights into LDH diagnoses.

Materials and methods

Patient cohort

The initial patient cohort was drawn from a pool of 2758 individuals presenting with clinical symptoms indicative of LDH at partner hospitals. After specific inclusion and exclusion criteria were applied, a total of 983 patients were selected for analysis. The inclusion criteria consisted of individuals aged 18 to 90 years who exhibited symptoms associated with LDH, such as low back pain, restricted lumbar mobility, or radicular symptoms. Patients were excluded if they had histories of lumbar surgeries, significant comorbidities, inadequate imaging quality, non-LDH diagnoses, ages outside the specified range, or pregnancy to minimize confounding factors (as shown in Fig. 1). The final patient group had an age range of 26 to 88 years, with a mean age of 48.3 ± 11.2 years and a male:female ratio of 1.2:1.

Dataset

The raw CT images in 2D format were subjected to a standardized preprocessing pipeline. First, all the CT images were normalized to [0, 1], preserving their diagnostically critical tissue contrast while standardizing the intensity distributions across the dataset. The images were then resized to 224 × 224 pixels using bicubic interpolation. To address the limited diversity of the data while respecting spinal structural consistency, we applied minimal data augmentation, which was constrained to horizontal rotation within ± 5° bounds.

The annotations were performed by a multidisciplinary team including two board-certified radiologists (with 8 and 12 years of musculoskeletal imaging experience) and two orthopaedic surgeons (with 10 and 15 years of spinal disorder expertise). Each case was independently annotated by two specialists (one radiologist and one surgeon) to ensure both its imaging accuracy and clinical relevance. Discrepancies (observed in 2% of the cases) were resolved through blind adjudication by a third radiologist.

Table 1 presents a comprehensive overview of the distribution of the CT imaging data utilized for the thorough diagnosis of LDH. This includes localization, which identifies the intervertebral disc segments (e.g., L3-L4), and qualitative assessment, which categorizes the corresponding disc diagnoses into bulging, herniation, and normal findings. Figure 2 shows representative examples of the different types of images included in the dataset. To assess the generalization ability of the model, we collected an external test set from independent medical centres while maintaining the same preprocessing procedures as those applied to the training data.

Deep learning methods

The classification models used included a ViT, ResNet18, ResNet50, LeNet, AlexNet, and VGG16. We systematically compared the performance differences among these

Table 1 Distribution of the CT imaging data for patients with lumbar disc herniation based on location and qualitative assessment

CT Imaging Details	Number of training-validation images	Number of test images
Localization CT Images		
L3-L4	446	89
L4-L5	320	64
L5-S1	328	65
Total (Localization)	**1094**	**218**
Qualitative CT Images		
Bulging	358	71
Herniation	261	52
Normal	387	77
Total (Qualitative)	**1006**	**200**
Grand Total	**2100**	**418**

models in terms of vertebra localization, disc herniation classification, and integrated diagnostic tasks. The outline of the study is shown in Fig. 3.

ViT model

A ViT adapts the transformer architecture for image recognition, enhancing its feature modelling capabilities for complex medical imaging tasks. As shown in Fig. 4, a ViT first divides the input image into fixed-size patches $x = \{x_1, x_2, \ldots, x_N\}$. Each patch is linearly projected into high-dimensional embeddings via a learnable matrix E, forming the initial embeddings.

$$\mathbf{z}_0 = [x_1E; x_2E; \ldots; x_NE] + E_{\text{pos}} \tag{1}$$

The positional encoding matrix E_{pos} preserves the spatial topology of anatomical structures in CT images, which is crucial for analysing pathological correlations across vertebral bodies.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

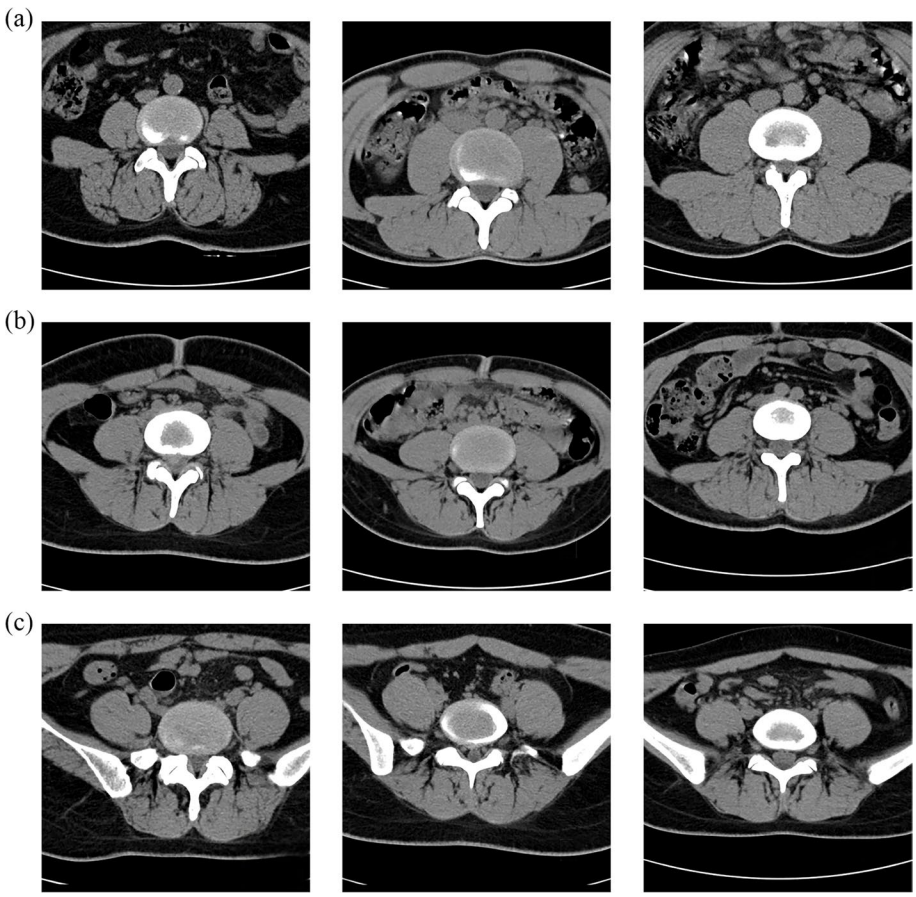

Fig. 2 **a** L3 - 4; **b** L4 - 5; **c** L5-S1 (from left to right: bulging, herniation and normal)

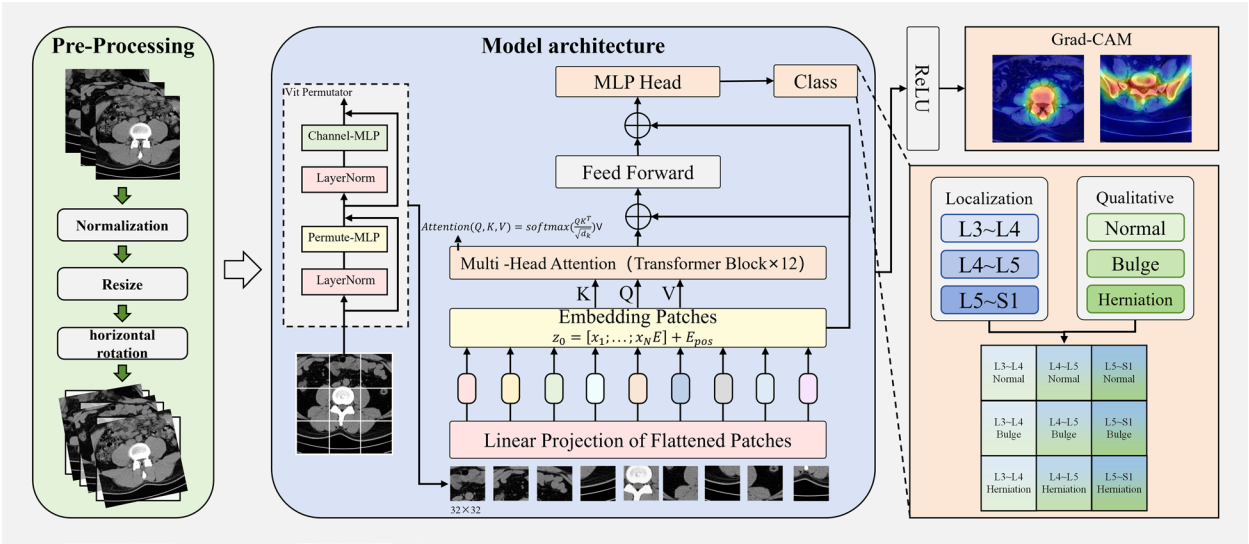

Fig. 3 Schematic outline of the study

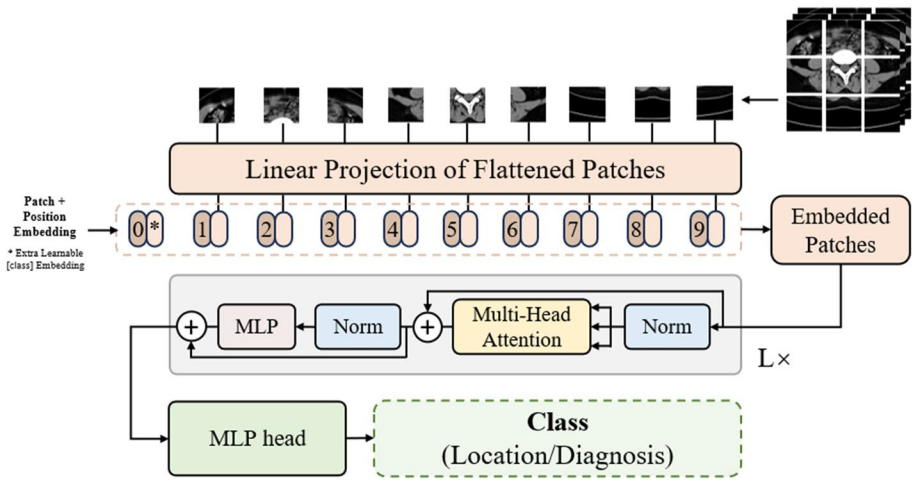

Fig. 4 Architecture of a VIT

The core innovation lies in the multihead self-attention mechanism, which establishes global dependencies across different image regions through dynamic attention weights between the $Q = z_l W^Q$, $K = z_l W^K$, and $V = z_l W^V$ matrices. The 12-layer transformer architecture integrates residual connections and 12 attention heads for performing multiscale feature extraction, with each layer containing self-attention sublayers and MLP blocks. This design enables the ViT to exhibit enhanced spatial modelling capabilities when processing complex structures. Additionally, the positional encoding scheme of the ViT effectively preserves the axial spatial continuity within CT series, which is a crucial feature for multiplanar reconstruction analyses.

We applied a fivefold cross-validation framework for model training and evaluation purposes. Through a grid search, hyperparameter optimization was performed over the parameter space. The final selected hyperparameters are presented in Table 2.

Grad-CAM visualization

Grad-CAM is a gradient-based visualization technique that identifies the critical regions influencing classification decisions by quantifying the sensitivity of the target class scores to the feature maps of a neural network. The core principle involves computing the gradients of the target class score with respect to the final convolutional

Table 2 Parameters of the models

Training parameters	Value
Patch size	32
Epochs	100
dim	1536
depth	12
heads	12
Mlp-dim	2048
Optimizer	Adam
Learning rate	0.0003
Loss function	Categorical Cross-Entropy

(or transformer-generated) feature maps, followed by implementing channelwise weighting through global average pooling. These weights are combined with the corresponding feature maps, normalized via ReLU activation, and superimposed on the original image to generate a heatmap highlighting the decision-critical regions.

In CNNs, the feature maps generated through convolutional kernels maintain spatial consistency with the original image, where weights reflect the importance of localized patterns. To interpret model decisions, we employed a Grad-CAM variant adapted for ViTs to address their lack of convolutional inductive biases [26]. For a ViT, gradients are computed from the final feature maps of the transformer block, capturing global dependencies via self-attention mechanisms. Specifically, the input CT image is processed through the ViT, and the gradients of the target class scores are globally averaged with respect to the features output by the transformer to generate weights. These weights are combined with the feature maps, and this is followed by ReLU activation to suppress negative values, producing a heatmap that highlights the regions that are critical for classification.

Statistical analysis

To evaluate the classification performance of the tested models, various statistical analysis methods were applied, including the concordance index (C-index), receiver operating characteristic (ROC) curve, confusion matrix, accuracy, F1 score, sensitivity (recall), and precision. The following formulas define the different evaluation metrics:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Results

Performance of the tested localization models

Cross-validation results

The ViT model achieved the highest validation accuracy (97.03%) in vertebra localization, outperforming the CNN-based models, with detailed metrics provided in Supplementary Figure S1 and Supplementary Table S1. The performance stability of the ViT was further confirmed through a best-validation analysis (Supplementary Figure S2).

Testing results

The ViT exhibited strong performance across different spinal regions, with maximal precision in L3-L4 (0.98) and L5-S1 (0.98), coupled with high recall rates of 0.97 and 0.95, respectively. For the L4-L5 region, it maintained a recall of 0.98 and an F1 score of 0.96, demonstrating robust diagnostic consistency. A comparative analysis revealed that ResNet50 ranked second in accuracy (89.11%), whereas LeNet attained the lowest performance (61.63%).

As visualized in Fig. 5, the confusion matrices provided granular insights into class-specific prediction patterns. These results aligned with those of the macroaveraged ROC analysis (Fig. 6), where the AUC values of the ViT across all the spinal regions exceeded those of the competing models. The complete metric distributions, including both macroaveraged and class imbalance-adjusted weighted averages, are systematically catalogued in Table 3.

Performance of qualitative models

Cross-validation results

The ViT demonstrated the most stable qualitative classification performance (validation loss: 0.0251). LeNet exhibited significant overfitting tendencies (training accuracy: 97.67% vs. validation: 79.97%). Detailed cross-validation metrics are provided in Supplementary Table S2 and Figures S3-S4.

Testing results

In the qualitative assessment of the disc herniation classification results, the ViT still attained satisfactory performance, with a macroaveraged precision of 0.92 and an overall accuracy of 93.63% (Table 4).

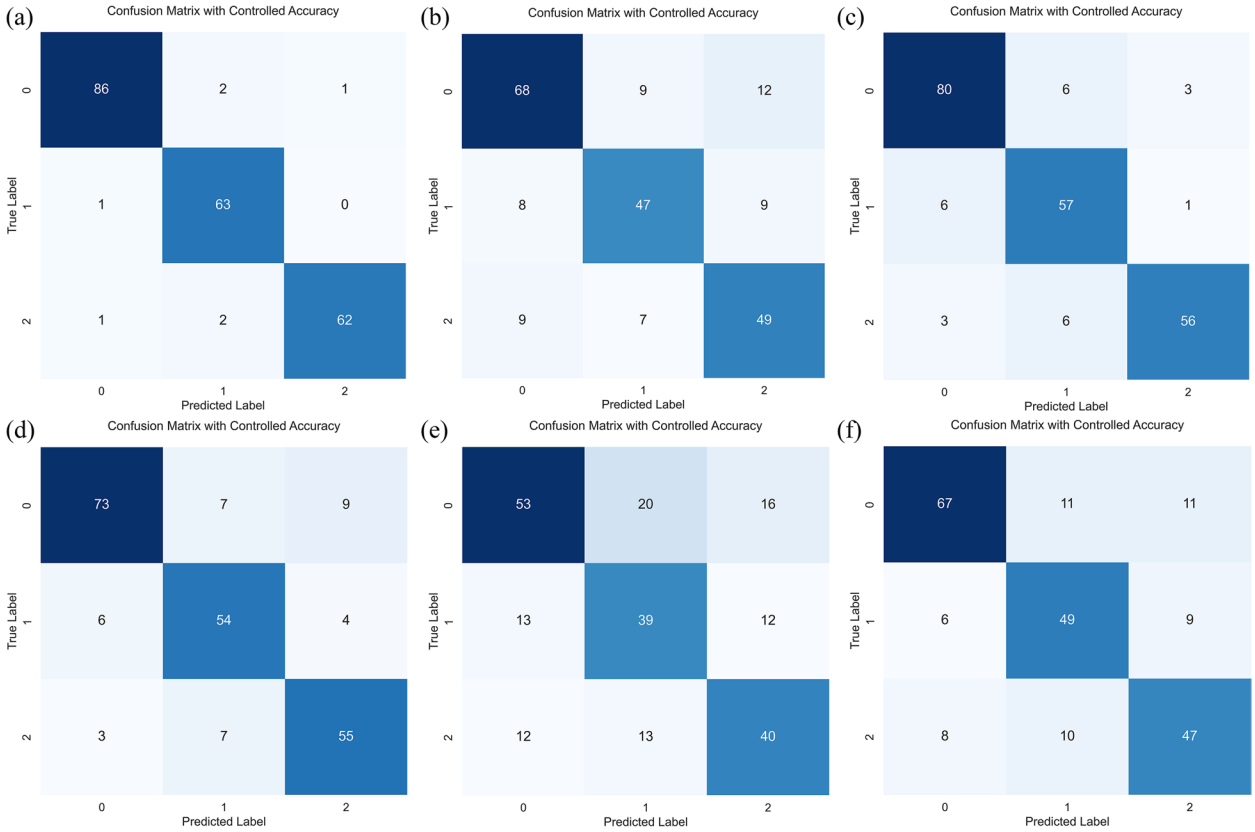

Fig. 5 Confusion matrices: **a** ViT; **b** VGG16; **c** ResNet50; **d** ResNet18; **e** LeNet; **f** AlexNet; classes 0, 1, and 2 are L3-L4, L4-L5, and L5-S1, respectively

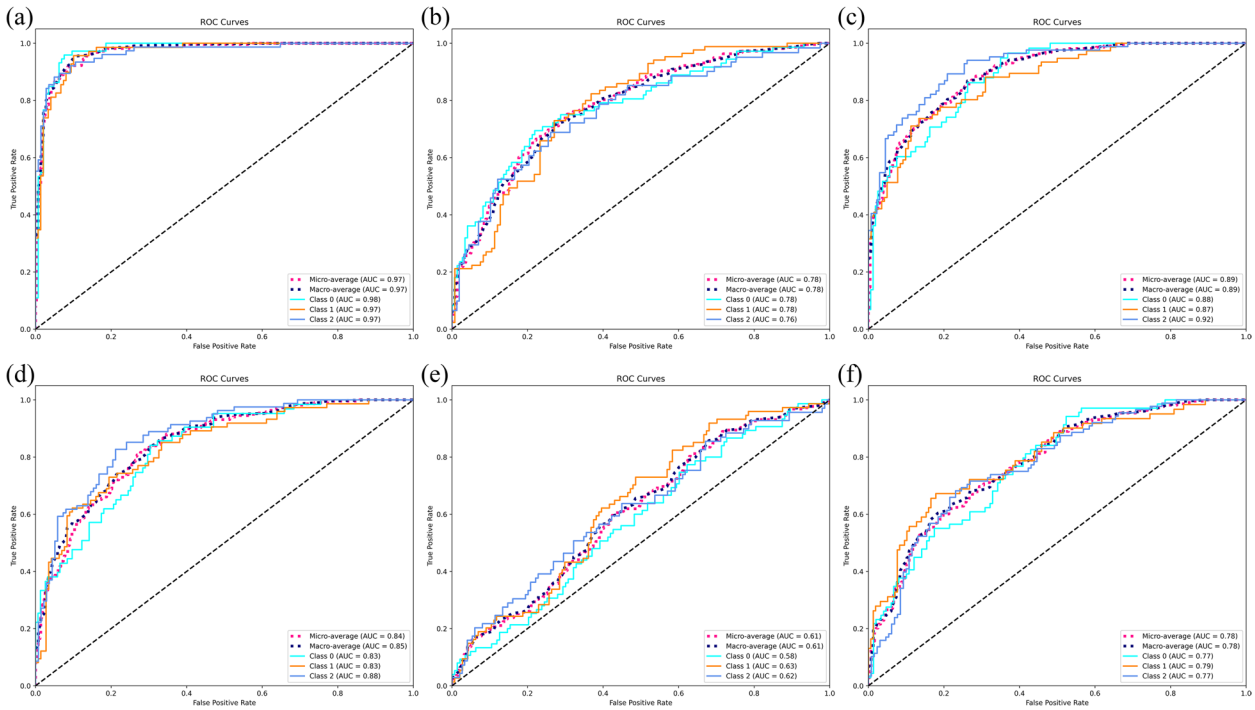

Fig. 6 AUC curves: **a** ViT; **b** VGG16; **c** ResNet50; **d** ResNet18; **e** LeNet; **f** AlexNet; classes 0, 1, and 2 are L3-L4, L4-L5, and L5-S1, respectively

Table 3 Classification results of the location models

	ViT	ResNet18	ResNet50	LeNet	AlexNet	VGG16
Precision						
L3~L4	0.98	0.89	0.90	0.68	0.83	0.80
L4~L5	0.94	0.79	0.83	0.54	0.70	0.75
L5~S1	0.98	0.81	0.93	0.59	0.70	0.70
macro avg	0.97	0.83	0.89	0.60	0.74	0.75
weighted avg	0.97	0.84	0.89	0.61	0.75	0.75
Recall						
L3~L4	0.97	0.82	0.90	0.60	0.75	0.76
L4~L5	0.98	0.84	0.89	0.61	0.77	0.73
L5~S1	0.95	0.85	0.86	0.62	0.72	0.75
macro avg	0.97	0.84	0.88	0.61	0.75	0.75
weighted avg	0.97	0.83	0.89	0.61	0.75	0.75
F1 score						
L3~L4	0.97	0.85	0.90	0.63	0.79	0.78
L4~L5	0.96	0.82	0.86	0.57	0.73	0.74
L5~S1	0.97	0.83	0.90	0.60	0.71	0.73
macro avg	0.97	0.83	0.88	0.60	0.74	0.75
weighted avg	0.97	0.84	0.89	0.61	0.75	0.75
Inference time (s)	106.65	135.32	129.99	39.97	47.05	153.09
Accuracy	97.13%	84.65%	89.11%	61.63%	75.17%	75.89%

Table 4 Classification results of the qualitative models

	ViT	ResNet18	ResNet50	LeNet	AlexNet	VGG16
Precision						
Bulging	0.94	0.85	0.85	0.60	0.72	0.72
Herniation	0.94	0.78	0.76	0.49	0.66	0.66
Normal	0.89	0.83	0.88	0.61	0.79	0.79
macro avg	0.92	0.82	0.83	0.57	0.72	0.72
weighted avg	0.92	0.83	0.84	0.57	0.73	0.73
Recall						
Bulging	0.93	0.82	0.85	0.56	0.72	0.77
Herniation	0.90	0.83	0.81	0.56	0.73	0.77
Normal	0.92	0.83	0.84	0.58	0.73	0.75
macro avg	0.92	0.82	0.83	0.57	0.73	0.77
weighted avg	0.92	0.83	0.84	0.57	0.73	0.77
F1 score						
Bulging	0.94	0.83	0.85	0.58	0.72	0.77
Herniation	0.92	0.80	0.79	0.52	0.69	0.75
Normal	0.90	0.83	0.86	0.60	0.76	0.77
macro avg	0.92	0.82	0.83	0.57	0.72	0.76
weighted avg	0.92	0.83	0.84	0.57	0.73	0.77
Inference time (s)	98.57	124.07	119.83	42.04	44.77	4000.00
Accuracy	93.63%	83.63%	83.89%	57.26%	74.11%	76.73%

While ResNet50 achieved comparable macro-F1 scores (0.83), its accuracy (83.89%) lagged behind that of the ViT by nearly 10 percentage points. Notably, the ViT maintained a relatively fast speed (98.57 s) among the tested deep learning architectures.

At the class-specific level, the ViT achieved robust performance in terms of herniation detection (precision: 0.94; recall: 0.90) and normal disc identification (precision: 0.89; recall: 0.92). The produced confusion matrices (Fig. 7) and ROC curves (Fig. 8) quantitatively validated these findings.

Comprehensive localization and qualitative results

We combined the results produced by the ViT model in the localization and quantitative tasks to obtain a comprehensive diagnostic result for lumbar disc herniation using the ViT model. The accuracy is demonstrated using a confusion matrix, as shown in Fig. 9. The ViT algorithm attained high diagnostic accuracy across all stages, as indicated by the confusion matrix results.

Grad-Cam visual explanation of the results based on the ViT model

The Grad-CAM heatmaps (Fig. 10) reveal that the ViT model focused on the diagnostically critical regions in the CT scans, aligning with the radiologists' anatomical and pathological reasoning processes. We conducted a hierarchical analysis of these heatmaps across two levels, vertebra–disc localization and herniation classification, and compared the attention patterns of the model with those of established clinical workflows.

For spinal segment identification (Fig. 10a–c), the model prioritized bony landmarks analogously to the radiologists' manual localization strategies. For L3–L4 localization (Fig. 10a), the heatmap highlights the vertebral endplates and pedicle morphology, which are key features that radiologists use to distinguish lumbar levels. For L5–S1 (Fig. 10c), the model focused on the sacral ala and facet joint orientations, which is consistent with the clinical protocols for differentiating lumbosacral transitions. Crucially, spinous process angulation—which is a primary radiologic marker for axial slice alignment—received strong attention weights (mean Grad-CAM intensity: 0.82 ± 0.05). This anatomical alignment

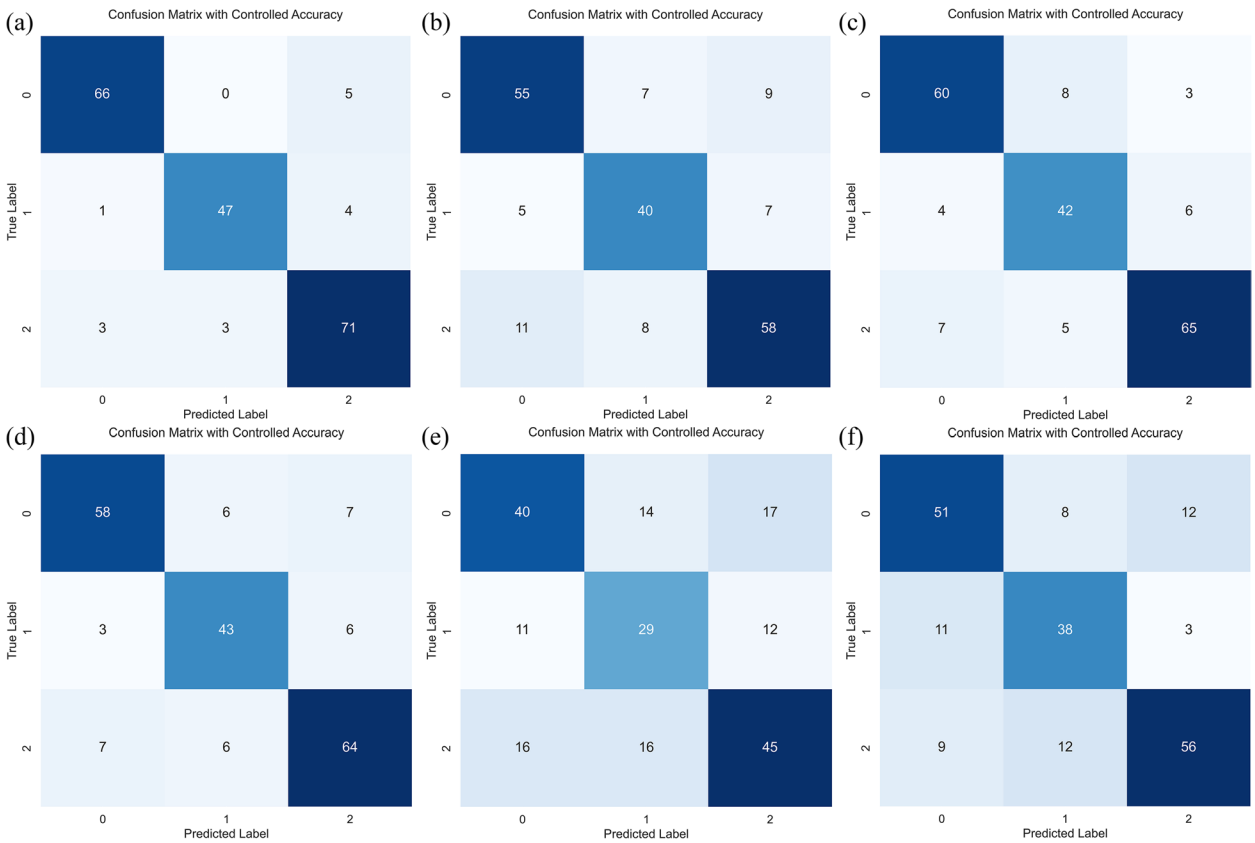

Fig. 7 Confusion matrices: **a** ViT; **b** VGG16; **c** ResNet50; **d** ResNet18; **e** LeNet; **f** AlexNet; classes 0, 1, and 2 are bulging, herniation, and normal, respectively

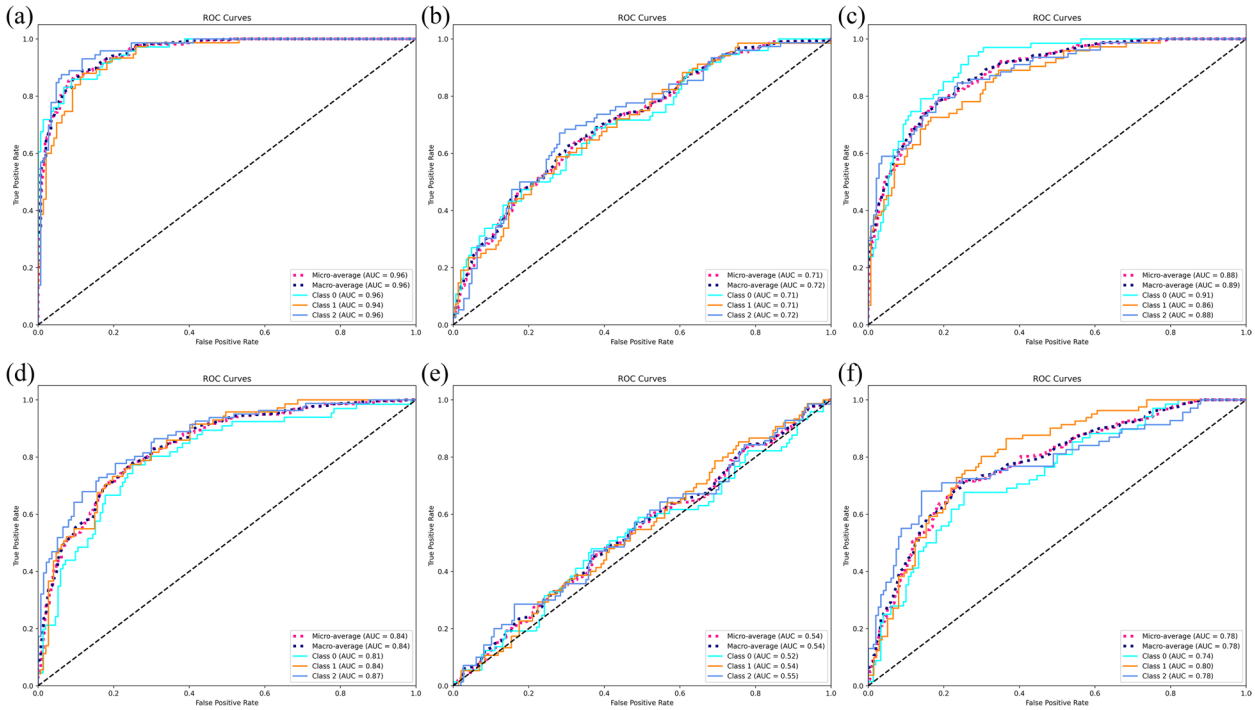

Fig. 8 AUC curves: **a** ViT; **b** VGG16; **c** ResNet50; **d** ResNet18; **e** LeNet; **f** AlexNet; classes 0, 1, and 2 are bulging, herniation, and normal, respectively

ensured that the spatial reasoning of the model matched the interpretative frameworks of humans.

Discussion

The heatmaps similarly show the ability of the ViT model to align its diagnostic reasoning process with clinical criteria across different disc classification scenarios. For normal discs (Fig. 10d), the model emphasized symmetric anatomical integrity, focusing on preserved annulus fibrosus boundaries and uniform disc heights, whereas the minimal activation at the posterior margin reflects the absence of a containment loss, mirroring radiologists' exclusion of herniation through an evaluation of posterior longitudinal ligament continuity. For the cases classified as bulging discs (Fig. 10e), the heatmaps highlight circumferential annular expansion (60–80% disc circumference) with maintained posterior symmetry, which is consistent with the radiologic threshold for diagnosing bulges (> 50% circumferential displacement). For herniated discs (Fig. 10f), the diagnostic logic of the model shifted markedly, with asymmetric attention given to posterolateral disc margins that directly corresponded to radiologists' localization results for containment breaches. Strong activation values at sites with epidural fat obliteration and nerve root displacement replicated manual assessments of neural compression severity, whereas the concurrent attention paid to endplate irregularities suggests that the model recognized degenerative precursors that were predisposed to herniation, aligning with radiologists' holistic evaluation of structural and contextual biomarkers. This layered activation pattern not only validates the clinical relevance of the model but also quantifies its decision hierarchy—from anatomical preservation in normal cases to nuanced pathologic differentiation in abnormalities—thereby bridging AI interpretability with radiologic expertise.

Interestingly, Fig. 11 presents an unexpected observation, where the Grad-CAM heatmap highlights the erector spinae muscle, which is a region that clinicians typically do not prioritize when directly assessing LDH. This raises an important question about the interpretability and focus of the model: is this a divergence from clinically significant regions, or could the model be identifying subtle patterns associated with disc herniation that are not easily discerned by the human eye?

On the one hand, this could represent a deviation from the current clinical focus areas, suggesting that the model may not be entirely aligned with the primary regions of interest for diagnosing LDH. On the other hand, this divergence might also highlight the ability of the model to capture subtle yet meaningful associations that are not always evident in standard clinical assessments. The fact that the erector spinae muscle is highlighted could indicate that the model detected potential correlations between muscle morphology or tension and the presence of disc herniation. There is evidence in the literature that

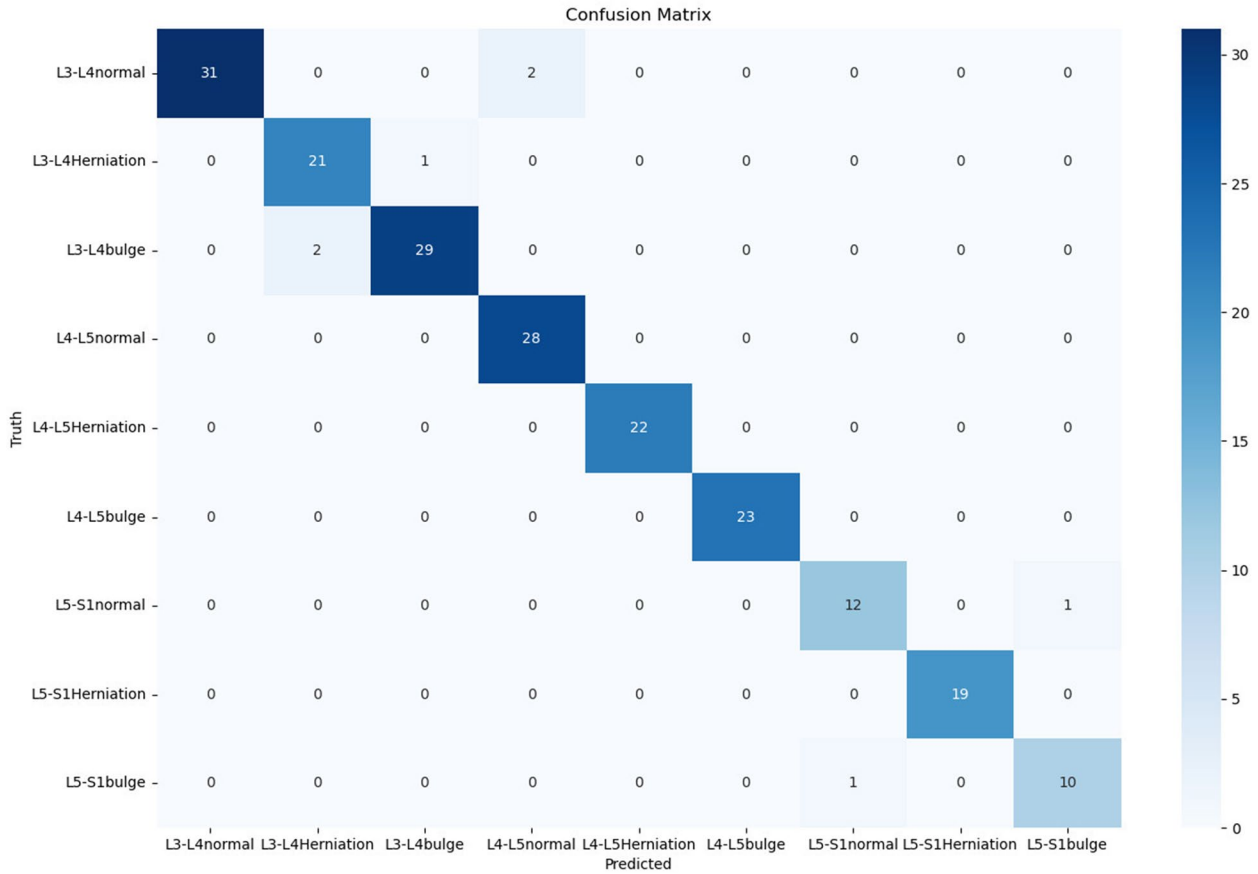

Fig. 9 Combined confusion matrix for the localization and quantitative ViT models

the erector spinae muscle may play a role in the biomechanical changes associated with LDH, and the focus of the model on this area might reflect an advanced ability to identify patterns beyond those traditionally assessed by clinicians. If this is the case, the resulting heatmap could provide valuable insights into previously underexplored diagnostic markers, offering a fresh perspective concerning the pathology of LDH [32].

This study presents promising results regarding the use of ViTs for the diagnosis of LDH using CT imaging. However, several limitations must be addressed to enhance the clinical applicability and generalizability of the findings.

First, despite a rigorous curation process, the relatively limited size (983 patients) and single-centre origin of our dataset may limit its generalizability to diverse populations and imaging protocols. One key challenge lies in the limited availability of large, annotated CT datasets that are specific to LDH. Expanding the dataset size will be critical for improving the generalizability of the model and ensuring its robustness across diverse populations. Future work should focus on the creation of larger, publicly available annotated datasets that can serve as valuable resources for researchers and clinicians alike. Additionally, testing the model on multicentre datasets will help further assess its performance across various imaging conditions and patient demographics, strengthening its generalizability.

Another major consideration is the integration of models into clinical workflows. While ViTs show promise, the real-world adoption of such models faces significant barriers. One challenge concerns the regulatory hurdles associated with deploying medical AI systems in clinical environments. Achieving regulatory approval requires rigorous validations through clinical trials to ensure the safety, reliability, and efficacy of the applied model. The model should be seamlessly integrated into routine diagnostic practice to provide real-time decision support, ensuring minimal disruption to existing workflows. Evaluating the real-world impact of the model on the resulting diagnostic accuracy and the efficiency of care is essential for gauging its clinical utility.

The interpretability of AI models remains a crucial factor in their adoption, particularly in health care.

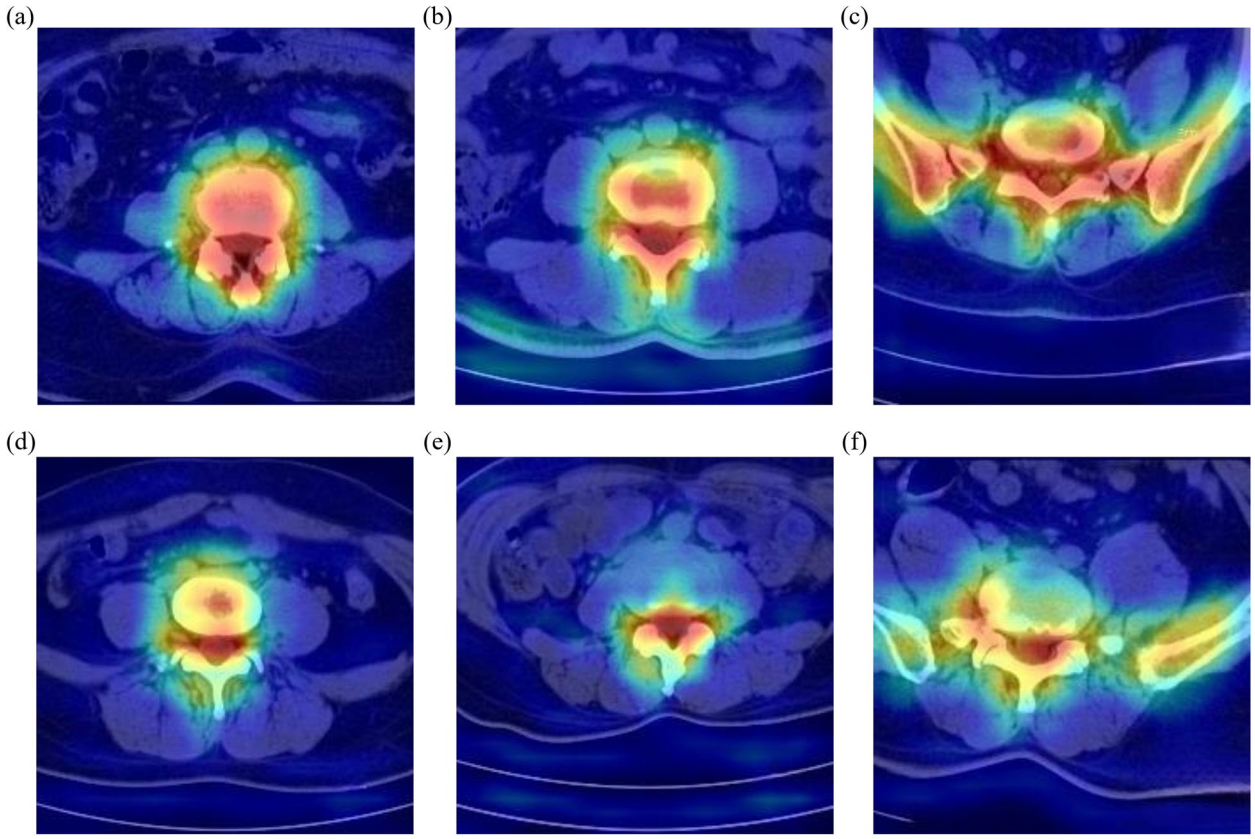

Fig. 10 Grad-CAM visualizations produced for predicting a real case of LDH: **a** location model (predicted class: L3-L4), **b** location model (predicted class: L4-L5), **c** location model (predicted class: L5-S1), **d** classification model (predicted class: normal), **e** classification model (predicted class: bulge), and **f** classification model (predicted class: herniation)

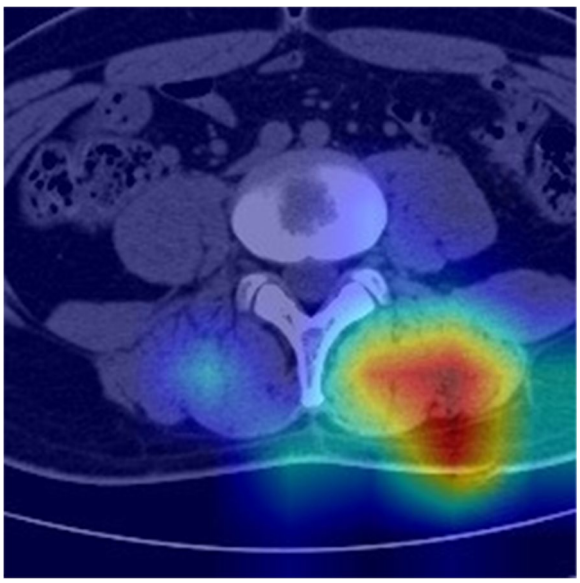

Fig. 11 An interesting prediction

Although Grad-CAM provided valuable insights into the focus areas of the proposed model, it occasionally highlighted regions that clinicians typically do not prioritize, such as the erector spinae muscle. This raises concerns about whether the focus of the model aligns with clinical expectations. Future improvements should refine the attention mechanism of the model to better focus on clinically relevant regions, particularly those that are directly associated with the diagnosis of LDH. However, the unexpected focus on the erector spinae could also suggest the potential of the model to identify subtle, previously unexplored patterns that may contribute to the diagnosis of LDH. This highlights the possibility of identifying new diagnostic markers that have not been traditionally considered in clinical practice.

Several areas for improvement can be identified. First, expanding the size of the dataset will help improve the ability of the developed model to generalize and enhance its robustness across diverse patient populations. Second, incorporating multicentre datasets will aid in evaluating the performance of the model across different health care settings. Third, exploring alternative model architectures,

such as hybrid CNN-ViT models, could further enhance the performance of the model by leveraging the strengths of both convolutional networks and transformers. These hybrid models can capture both local features (via CNNs) and long-range dependencies (via ViTs), improving their diagnostic accuracy in complex medical imaging tasks.

Conclusion

The current study presents a pioneering approach for diagnosing LDH by employing ViTs in conjunction with CT imaging and demonstrates the potential of ViTs to improve the LDH diagnosis process using CT imaging. The use of Grad-CAM for model interpretability purposes further enhances the clinical applicability of the proposed approach. However, challenges related to data availability, clinical integration, and model interpretability must be addressed to ensure the widespread adoption of this technology in clinical practice. Future work should focus on expanding the existing datasets, exploring alternative model architectures, and addressing regulatory and clinical integration challenges to maximize the impact of AI in health care.

Abbreviations

LDH	Lumbar Disc Herniation
ViTs	Vision Transformers
CNNs	Convolutional Neural Networks
MRI	Magnetic Resonance Imaging
CT	Computed Tomography
Grad-CAM	Gradient-Weighted Class Activation Mapping
AI	Artificial Intelligence
DL	Deep Learning
MLP	Multilayer Perceptron
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
TPR	True-Positive Rate
FPR	False-Positive Rate
TP	True Positives
TN	True Negatives
FP	False Positives
FN	(False Negatives)

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12891-025-08602-2.

Supplementary Material 1. [33, 34]

Acknowledgements

We sincerely thank the reviewers for their valuable feedback and constructive suggestions, which have helped improve the quality of this manuscript.

Clinical trial number

Not applicable.

Authors' contributions

QS.C conceptualized the study, curated the data, developed the methodology, wrote the original draft, visualized the results, and coordinated the project. XY.W, H.L, and Y.Z contributed to the conceptualization, methodology, data curation, original draft writing, and visualization. T.J supervised the study, revised the manuscript, and acquired funding. All authors read and approved the final manuscript.

Funding

This work was supported by the Anhui Provincial Natural Science Foundation (2308085MH294) and the Natural Science Research Project for Anhui Universities (2022 AH050510).

Data availability

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

Declarations

Ethics approval and consent to participate

Our study adhered to the Declaration of Helsinki. This study received approval from the Ethics Committee of the First Affiliated Hospital of Anhui University of Traditional Chinese Medicine (no. 2024MCZQ28), and the ethics committee waived the need to consent to participate because of the minimal risk involved.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 January 2025 Accepted: 31 March 2025
Published online: 29 April 2025

References

1. Knezevic NN, Candido KD, Vlaeyen JWS, Van Zundert J, Cohen SP. Low back pain. Lancet. 2021;398(10294):78–92.
2. Lee JH, Choi KH, Kang S, Kim DH, Kim DH, Kim BR, et al. Non-surgical treatments for patients with radicular pain from lumbosacral disc herniation. Spine J. 2019;19(9):1478–89.
3. Schmid AB, Dove L, Ridgway L, Price C. Early surgery for sciatica. BMJ. 2023;381:791.
4. Kim J, van Rijn RM, van Tulder MW, Koes BW, de Boer MR, Ginai AZ, et al. Diagnostic accuracy of diagnostic imaging for lumbar disc herniation in adults with low back pain or sciatica is unknown; a systematic review. Chiropr Man Ther. 2018;26:1–14.
5. Wassenaar M, van Rijn RM, van Tulder MW, Verhagen AP, van der Windt DAWM, Koes BW, et al. Magnetic resonance imaging for diagnosing lumbar spinal pathology in adult patients with low back pain or sciatica: a diagnostic systematic review. Eur Spine J. 2012;21(2):220–7.
6. Hall FM. Back pain and the radiologist. Radiology. 1980;137(3):861–3.
7. Lurie JD. What diagnostic tests are useful for low back pain? Best Pract Res Clin Rheumatol. 2005;19(4):557–75.
8. van Rijn RM, Wassenaar M, Verhagen AP, Ostelo RWJG, Ginai AZ, de Boer MR, et al. Computed tomography for the diagnosis of lumbar spinal pathology in adult patients with low back pain or sciatica: a diagnostic systematic review. Eur Spine J. 2012;21(2):228–39.
9. Zubair M, Rais HM, Al-Tashi Q, Ullah F, Faheeem M, Khan AA. Enabling predication of the deep learning algorithms for low-dose CT scan image denoising models: a systematic literature review. IEEE Access. 2024;12:79025–50.
10. Zubair M, Rais HM, Alazemi T. A novel attention-guided enhanced u-net with hybrid edge-preserving structural loss for low-dose ct image denoising. IEEE Access. 2025;13:6909–23.
11. Hu Y, Liu Y, Zhang S, Zhang T, Dai B, Peng B, et al. A cross-space CNN with customized characteristics for motor imagery EEG classification. IEEE T Neural Sys Reh Eng. 2023;31:1554–65.
12. Patil SS, Ramteke M, Verma M, Seth S, Bhargava R, Mittal S, et al. A domain-shift invariant CNN framework for cardiac MRI segmentation across unseen domains. J Digit Imaging. 2023;36(5):2148–63.
13. Tang C, Zhang W, Li H, Li L, Li Z, Cai A, et al. CNN-based qualitative detection of bone mineral density via diagnostic CT slices for osteoporosis screening. Osteoporos Int. 2021;32(5):971–9.
14. Qian J, Su G, Shu X, Shen K, Chen B, Wang X. Lumbar disc herniation diagnosis using deep learning on MRI. J Radiat Res Appl Sci. 2024;17(3):100988.

15. Duan X, Xiong H, Liu R, Duan X, Yu H. Enhanced deep learning model for detection and grading of lumbar disc herniation from MRI. Med Biol Eng Comput. 2024;62:3709–19.
16. Azad R, Kazerouni A, Heidari M, Aghdam EK, Molaei A, Jia Y, et al. Advances in medical image analysis with vision transformers: a comprehensive review. Med Image Anal. 2024;91:103000.
17. Takahashi S, Sakaguchi Y, Kouno N, Takasawa K, Ishizu K, Akagi Y, et al. Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. J Med Syst. 2024;48(1):84.
18. Liu Z, Lv Q, Yang Z, Li Y, Lee CH, Shen L. Recent progress in transformer-based medical image analysis. Comput Biol Med. 2023;164:107268.
19. Pu Q, Xi Z, Yin S, Zhao Z, Zhao L. Advantages of transformer and its application for medical image segmentation: a survey. Biomed Eng Online. 2024;23(1):14.
20. Kassis I, Lederman D, Ben-Arie G, Giladi RM, Shelef I, Zigel Y. Detection of breast cancer in digital breast tomosynthesis with vision transformers. Sci Rep. 2024;14(1):22149.
21. Vinayahalingam S, van Nistelrooij N, Rothweiler R, Tel A, Verhoeven T, Troltzsch D, et al. Advancements in diagnosing oral potentially malignant disorders: leveraging Vision transformers for multi-class detection. Clin Oral Investig. 2024;28(7):364.
22. Huang F, Qiu A. Ensemble vision transformer for dementia diagnosis. IEEE J Biomed Health Inform. 2024;28(9):5551–61.
23. Khaniki MA, Golkarieh A, Manthouri M. Brain tumor classification using vision transformer with selective cross-attention mechanism and feature calibration. Arxiv. 2024. https://doi.org/10.48550/arXiv.2406.17670.
24. Wu Y, Qi S, Sun Y, Xia S, Yao Y, Qian W. A vision transformer for emphysema classification using CT images. Phys Med Biol. 2021;66(24):245016.
25. Taye GD, Sisay ZH, Gebeyhu GW, Kidus FH. Thoracic computed tomography (CT) image-based identification and severity classification of COVID-19 cases using vision transformer (ViT). Discover Appl Sci. 2024;6(8):384.
26. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128(2):336–59.
27. Teng Q, Liu Z, Song Y, Han K, Lu Y. A survey on the interpretability of deep learning in medical diagnosis. Multimed Syst. 2022;28(6):2335–55.
28. Cui Y, Zhu J, Duan Z, Liao Z, Wang S, Liu W. Artificial intelligence in spinal imaging: current status and future directions. Int J Environ Res Public Health. 2022;19(18):11708.
29. Ren G, Yu K, Xie Z, Wang P, Zhang W, Huang Y, Wang Y, Wu X. Current applications of machine learning in spine: from clinical view. Global Spine J. 2022;12(8):1827–40.
30. Zhang R. A state-of-the-art survey of deep learning for lumbar spine image analysis: X-ray, CT, and MRI. AI Medicine. 2024;1(1):3.
31. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H. Transformers in medical imaging: a survey. Med Image Anal. 2023;88:102802.
32. Yazici A, Yerlikaya T. The relationship between the degeneration and asymmetry of the lumbar multifidus and erector spinae muscles in patients with lumbar disc herniation with and without root compression. J Orthop Surg Res. 2022;17(1):541.
33. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
34. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.

Publisher's Note