Spine-GraphX: A Graph Neural Network Model for Analyzing Structural Relationships in Lumbar Intervertebral Discs

Dr. G. Naveen Sundar [1,2] Prof. (Dr.) Raja Sarath kumar Boddu[3],

[1] Lincoln University College, Malaysia;

[2] Division of CSE, Karunya Institute of Technology and Sciences, Coimbatore, India;

[3] Raghu Engineering College, Visakhapatnam
pdf.naveen@lincoln.edu.my
naveensundar@karunya.edu
rajaboddu@lincoln.edu.my

Abstract - Magnetic resonance imaging of the lumbar spine is the key to the diagnosis of intervertebral disc degeneration and related pathology. Automated analysis is reliable and can assist clinical decision making while also reducing reader variability. Convolutional networks, including CNNs and U-Net, emphasize local patterns of pixels, which is limited to represent dependencies between vertebrae, discs, and the spinal canal. To bridge this gap, we present Spine-GraphX, a framework that combines GCNs with convolutional features to encode explicit anatomical associations. Experiments are conducted on the SPIDER MRI Spine T2 PNG dataset, which has about 1,550 sagittal T2-weighted slices of 210 subjects. Spine-GraphX was able to achieve an accuracy of 93.5%, a sensitivity of 0.91, Dice score of 0.902, and IoU of 0.829. These results were even better than ResNet-50 U-Net (accuracy 88.7% Dice 0.861) and DenseNet U-Net (accuracy 89.6% Dice 0.868). Group comparisons showed p-values of less than 0.05 which shows statistically reliable increases. The results indicate that the structural relationship modeling offers greater accuracy under noise and small sample sizes and computational efficiency for automated analysis of the lumbar spine.

Keywords: Lumbar Spine, MRI Analysis, Graph Neural Networks, Intervertebral Disc Degeneration, Medical Image Segmentation, Computational Efficiency, Robustness Evaluation

1 Introduction

Magnetic resonance imaging of the spine plays an integral role in the diagnosis of degenerative disc disease, disc herniations, disc bulging and narrowing of the canal. Consistent and accurate interpretation is important in early identification of abnormalities and limiting down-stream complications. Automated analysis that provides reliable disc and vertebral evaluation can help radiologists in routine practice to moderate the variation in subjectivity.

Although computer-aided diagnosis has advanced, grading and detection of intervertebral disc degeneration remain challenging. Variation in patient anatomy, imaging protocols, and pathological presentation complicates both classification and segmentation. Conventional pipelines show limited transfer across datasets, and deep models that depend solely on pixel intensities often misclassify subtle degeneration or low-contrast scans. These factors motivate methods that combine fine-scale appearance cues with higher-level structural context along the spinal column.

Earlier work relied on handcrafted descriptors with classical learning algorithms. While useful as baselines, these methods did not capture dependencies among spinal elements. Convolutional architectures, including U-Net variants, improved segmentation and classification but largely emphasize local texture and intensity, without explicitly representing the anatomical links among discs, vertebrae, and the spinal canal. When degeneration alters spatial relationships, such locality can reduce robustness.

To address these limitations, this paper proposes Spine-GraphX, a framework that integrates graph neural networks with convolutional feature extractors to model structural dependencies in the spine. Discs, vertebrae, and the canal are encoded as nodes, and spatial or anatomical relations are expressed as weighted edges. This representation fuses local appearance with contextual information and improves the reliability of degeneration grading while maintaining computational efficiency.

- The spine graph X development is modern GNN and CNN hybrid architecture intended for lumbar spine analysis using MRI scans.

- The connections between anatomical structures were represented as edges in the graph, making the modeling of their structural relationships explicit.

- Comprehensive evaluation against CNN, ResNet-50 U-Net, DenseNet U-Net, Attention U-Net, and vanilla GCN baselines.

- An ablation study demonstrating the contribution of edge features, residual connections, and data augmentation.

- Robustness analysis under Gaussian noise, motion blur, intensity shifts, and reduced training data availability.

- Statistical significance testing to validate improvements across accuracy, Dice score, and IoU metrics.

The rest of this paper is thus organized. The next section reviews related work in spine imaging and deep learning methods. The following section presents the preprocessing steps and the proposed model architecture. The subsequent section outlines the dataset, experimental setup, hyperparameter configuration, and reports both quantitative and qualitative results, including ablation, efficiency, and robustness studies. The discussion section interprets the key findings and their implications. The paper is concluded at last with a summary about contributions and possible directions towards future work.

2 Related Work

Several studies have explored deep learning and graph-based models for spine MRI analysis. Baur et al. [1] proposed an automated 3D imaging and Pfirrmann classification framework that combined CNN and GNN for disc grading, achieving an F1 score of 0.85 in segmentation but reporting performance variation across lumbar levels. Natarajan et al. [2] introduced MRI2Mesh, a CNN-GNN hybrid with axial attention transformers, which reduced Hausdorff distance by 5.87% and point-to-surface error by 14.5%, although the method required further validation on larger and more diverse datasets. In another work, Baur et al. [3] presented a systematic review of CNNs in spinal MRI, highlighting positive outcomes across multiple studies but lacking a GNN-specific focus. Rak et al. [4] developed a fast spine segmentation method using CNNs with star-convex cuts, reporting a Dice score of 96.0% and runtime below one second per vertebra, yet the method was limited to vertebral body segmentation. In their paper, Li et al. [5] gave the details of a two-stage transformer-CNN beadline which takes into consideration 3D transformers, 2D CNNs, and graph convolutional networks. Although a complete quantitative analysis was not given, the method showed a good performance on MRSpineSeg. After that, Liu and his coworkers [6] came up with PNAGL, which happens to be a residual non-local attention graph learning method for 4D-MRI, and they showed the real-time feasibilities of it without any specific design for spine imaging. The next paper by Andrew et al. [7] was a survey of CNN-based segmentation methods for spine MRI and it was mostly about the deep learning component, but still it did not provide much empirical evidence to support its findings. Alternatively, the work of Zeybel and Akgul [8], which integrated Faster R-CNN and a shortest-path graph model for disc detection and outperformed previous techniques on 80 scans, is noteworthy; however, the small sample size restricted the ability to generalize the outcomes. Chang and his team [9] developed a comprehensive multi-vertebrae segmentation approach using spatial GCNs combined with a label attention mechanism, achieving an 89.28% success rate in identification and an average IoU of 85.37%. However, they noted that adapting this model to other MRI datasets could be difficult.

Ghobrial et al. [10] presented an automated dural sac segmentation approach which was based on MultiResUNet with a Dice score of around 0.92 and close to expert labels. However, the evaluation was performed only on T1-weighted scans, which limits its general application. Liawrungrueang et al. [11] proposed a CNN classifier for the grade of disc degeneration with good accuracy and clinically useful sensitivity, but the results were affected by class imbalance. Hess et al. [12] used CNN-derived segmentation in combination with biomechanical modeling in multi-tissue segmentation, achieving Dice values above 0.77 and correlation coefficients above 0.69, but they also observed error propagation at tissue boundaries. Verheijen et al. [13] performed a meta-review to evaluate stenosis detection methods and concluded that deep learning methods were usually better than classical machine learning, but the vast majority of studies included in this meta-analysis were not externally validated. Guo et al. [14] proposed a herniation detection system based on YOLOv8 with efficient channel attention and group shuffling and showed strong mean average precision as well as reliable grading, although the data size was relatively small. Wang et al. [15] applied a 3D DeepLab V3+ model for the multi-label segmentation of the lumbar structures and obtained Dice values close to 0.89, by restricting the analysis to the L4/5 level. Basak et al. [16] suggested a cascaded approach by combining YOLOv8 with self-organizing neural networks, achieving Dice of nearly 91 percent and IoU around 84 percent, however, the validation was performed on a low number of patients. Zhao and Zhu [17] presented a narrative review on artificial intelligence for degenerative disc disease, as well as the improvement from deep models and lack of quantitative benchmarking in many reports. Ahmed et al. [18] proposed an accurate but computationally intensive multi-class MRI segmentation algorithm, which may not be suitable for routine use. Together, these studies represent consistent progress in lumbar spine analysis using CNNs, GNNs and hybrid schemes, while highlight ongoing shortcomings in dataset diversity, cross-level generalization and computation efficiency. These gaps provide an impetus for frameworks like Spine-GraphX that try to balance between accuracy, robustness, and run-time efficiency. Table 1 summarizes the work related to MRI analysis of the lumbar spine.

Table 1: Summary of Related Work in Lumbar Spine MRI Analysis

Study	**Method**	**Findings**	**Drawbacks**
Baur et al. [1]	CNN–GNN for 3D disc grading	Achieved F1 score of 0.85 for segmentation and moderate grading accuracy	Performance varied across lumbar levels
Natarajan et al. [2]	CNN–GNN with axial attention	Reduced Hausdorff by 5.87% and Pt-to-surface error by 14.5%	Requires further validation on diverse datasets
Liawrungrueang et al. [11]	CNN classifier for disc grading	Good accuracy and sensitivity for early degeneration detection	Imbalanced dataset affected generalization
Hess et al. [12]	CNN + biomechanical modeling	Dice ≥ 0.77 and correlation R ≥ 0.69 for multi-tissue segmentation	Errors propagated across tissue boundaries
Guo et al. [14]	YOLOv8 with attention modules	High mAP and strong grading for herniation detection	Limited by small dataset

Basak et al. [16]	YOLOv8 + Self-ONN cascaded model	Achieved Dice ≈ 91%, IoU ≈ 84%	Evaluated only on small cohort
Rak et al. [4]	CNN + Star-convex cuts	Dice of 96%, runtime <1s per vertebra	Focused only on vertebral body segmentation
Chang et al. [9]	Spatial GCN + label attention	Achieved IDR 89.28%, mIoU 85.37%	Required better generalizability across scans
Verheijen et al. [13]	Meta-review on AI for stenosis	Found deep learning outperformed classical ML methods	Lacked external validation studies
Ahmed et al. [18]	Multi-class DL segmentation	Reported strong classification accuracy on MRI	Required high computational resources

2.1 Problem Statement

The main causes of lumbar spine problems are often linked to disc degeneration, herniation, and stenosis of the intervertebral discs. Subsequently, chronic low back pain is experienced by individuals which, in turn, affects the life and health care systems of the global population profoundly. It is still challenging to make use of magnetic resonance imaging (MRI) for better and more accurate diagnosis because of the anatomical differences, many-sided imaging protocols, and the slow and discreet pace of the pathological changes in the lumbar spine. The classical deep learning methods such as CNNs and U-Net based architectures rely heavily on pixel-wise textural features and usually do not take into account the structural correspondences between the discs, vertebrae, and the spinal canal. Inaccurate modeling results in poor generalization, especially when handling clinical data that varies significantly. To overcome these problems, the presented Spine-GraphX framework merges encoders with graph neural networks to synchronizedly grasp local intensity patterns together with global anatomical dependencies. The modeling of the spine as a structured graph is done in the manner that it utilizes both texture features as well as inter-structural relations, thus leading to an increase in the accuracy, sensitivity, and generalization of the diagnosis compared to current techniques.

3 System Methodology

The Spine-GraphX framework which is suggested takes advantage of the ability of the convolutional neural network to extract features and also the graph neural network modeling to represent local texture patterns and relationships between lumbar intervertebral discs, vertebrae, and the spinal canal structurally. The process includes six stages: data preprocessing, data augmentation, feature extraction, graph construction, graph convolution, and classification.

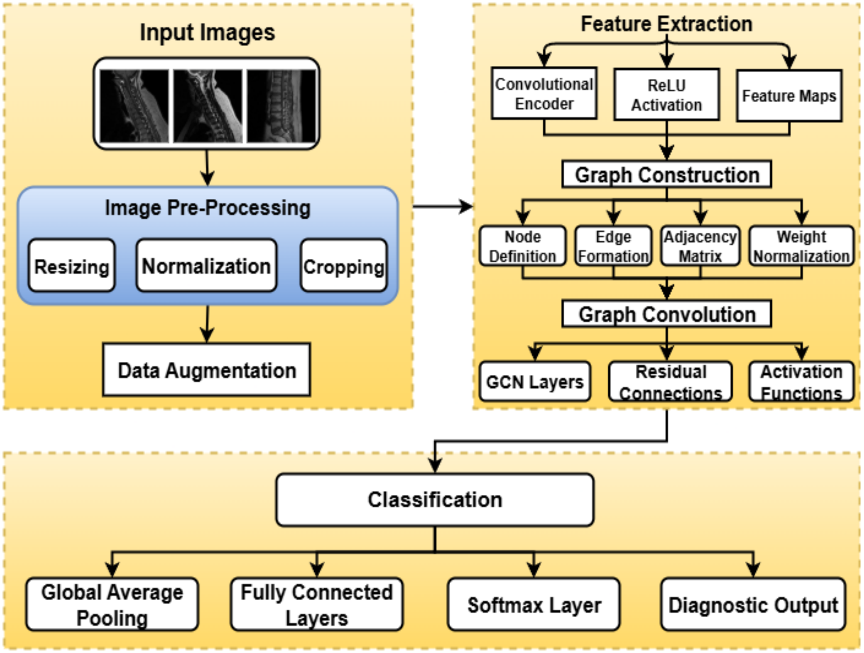

Figure 1: Block diagram of the Spine-GraphX system architecture.

3.1 Data Preprocessing

Input sagittal MRI slices $I^{raw} \in \mathbb{R}^{H \times W}$ were standardized to improve consistency across subjects. Images were resized to a fixed resolution of 512×512, intensity values were normalized to zero mean and unit variance, and cropping was applied to focus on the lumbar region. The preprocessing pipeline can be represented as

$$I_{proc} = N\ C\big(R(I_{raw})\big)\ , \tag{1}$$

where $R(\cdot)$ denotes resizing, $C(\cdot)$ represents cropping, and $N\ (\cdot)$ indicates normalization. Equation (1) ensures that each image input to the model is standardized for subsequent learning. This processed output serves as the basis for augmentation to enhance model robustness.

3.2 Data Augmentation

Augmentation was applied on-the-fly to improve generalization under anatomical and acquisition variability. Geometric transforms simulated plausible patient positioning and motion, while photometric transforms modeled scanner and protocol differences. The full policy is summarized in Table 2. Additive Gaussian noise perturbed voxel intensities, improving resilience to acquisition grain and random noise:

$$I' = I + \varepsilon, \quad \varepsilon \sim N\,(0, \sigma^2), \tag{2}$$

To mimic variations in image contrast and bias fields, intensity scaling and shifting were applied as

$$I' = \alpha I + \beta, \tag{3}$$

To reduce overfitting and refine decision boundaries, sample interpolation using Mixup regularization was implemented.

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j\, , \tilde{y} = \lambda y_i + (1 - \lambda)y_j,\ \lambda \sim Beta(a, b) \tag{4}$$

For handling a particular locality, the approach involved introducing random blockages to ensure variation in spatial features.

$$I' = I \odot (1 - M) + mM, \tag{5}$$

where $M \in \{0, 1\}^{H \times W}$ is a binary mask and m is a fill value. The transformations defined in Eq. (2)–(5) were applied probabilistically as listed in Table 2.

Table 2: Augmentation policy and ranges

Transform	**Range / Setting**	**Prob.**
Rotation	Uniform $\theta \in [-15°, 15°]$	0.5
Horizontal flip	Left–right flip	0.5
Scaling	Isotropic $s \in [0.9,\ 1.1]$	0.3
Elastic deformation	$\sigma_{grid} = 8,\ \alpha = 12$ pixels	0.2
Gaussian noise (Eq. 2)	$\sigma \in [0.00, 0.03]$ of dynamic range	0.4
Intensity shift/scale (Eq. 3)	$\alpha \in [0.9, 1.1],\ \beta \in [-0.05, 0.05]$	0.5
Gamma correction	$\gamma \in [0.9, 1.1]$	0.3
Cutout (Eq. 5)	Square mask side $l \in [24, 48], m = 0$	0.2
Mixup (Eq. 4)	$a = b = 0.4$	0.2

To preserve anatomical plausibility, affine parameters were constrained to small angles and scales, and elastic fields were smoothed before warping. Augmentations were disabled on the validation and test sets. The augmented data were then passed to the convolutional encoder for feature extraction.

3.3 Feature Extraction

From the preprocessed images $I_{proc} \in \mathbb{R}^{H\times W}$, a convolutional encoder extracts node-level features $X = \{x_1, \dots, x_n\}$, where each $x_i \in \mathbb{R}^d$ represents intensity and shape descriptors. The convolutional operation is

$$x_i = f(W * I_{proc,i} + b)$$

(6)

where W and b denote the learnable kernel and bias, and $f(\cdot)$ is ReLU. Equation (6) ensures localized feature extraction for each region of interest. These extracted features are subsequently mapped into a graph representation to capture anatomical dependencies.

3.4 Graph Construction

The anatomical structures of the lumbar spine are modeled as nodes, and their spatial or contextual relations are represented as weighted edges. An undirected graph $G = (V, E)$ is constructed, where $V = \{v_1, v_2, \dots, v_n\}$ represents the set of nodes corresponding to vertebrae, intervertebral discs, and the spinal canal, and E is the set of edges. The adjacency matrix $A \in \mathbb{R}^{n\times n}$ encodes anatomical relations including disc–disc continuity, disc–vertebra adjacency, and vertebra–canal interactions.

Edge weights are determined using a combination of spatial proximity and anatomical priors. In the case of two structures being located next to each other, their interconnection is given a larger weight, whereas the structures that are farther apart are assigned smaller weights. The normalized adjacency matrix is given by

$$\tilde{A} = D^{-\left(\frac{1}{2}\right)}(A + I)D^{-\left(\frac{1}{2}\right)} \quad (7)$$

where D stands for the degree matrix and I is the identity matrix. The usage of Equation (7) guarantees symmetric normalization, which in turn makes the transmission of messages through the graph steady and prevents any numerical instability. The graph that has been constructed is then subjected to graph convolutional layers for the purpose of feature propagation.

3.5 Graph Convolutional Layers

Graph convolution propagates information over the constructed spinal graph so that each node refines its representation using its neighbors. The layerwise update in Eq. (8),

$$H^{l+1} = \sigma(\tilde{A}H^{(l)}W^{(l)}) \quad (8)$$

where $H^{(l)} \in \mathbb{R}^{n\times d_l}$ is the layer-l embedding matrix, $W^{(l)} \in \mathbb{R}^{d_l \times d_{l+1}}$ is the learnable weight matrix, and $\sigma(\cdot)$ is a pointwise nonlinearity such as ReLU. This update aggregates neighborhood information through $\tilde{A}$ while preserving node-specific features.

To mitigate over-smoothing, residual skip connections link consecutive graph layers. These connections retain discriminative signal as depth increases and improve optimization stability. The final layer embeddings are passed to the downstream classifier for diagnostic prediction.

3.6 Classification Layer

After L graph layers, global average pooling yields a fixed-length descriptor h_G that summarizes both local appearance and structural dependencies. The classifier maps h_G to logits, and class probabilities are obtained by the softmax in Eq. (9),

The probability of class c is computed using the softmax function:

$$p(y = c|h_G) = \frac{\exp{(W_c h_G + b_c)}}{\sum_{j=1}^{C} \exp{(W_j h_G + b_j)}} \tag{9}$$

where W_c and b_c are learnable parameters and C is the number of diagnostic categories (for example, Pfirrmann grades, herniation, bulging, and canal narrowing).

3.7 Loss Function

Training uses a composite objective that balances categorical accuracy with overlap quality for segmentation in Eq. (10):

$$L = \alpha L_{CE} + (1 - \alpha) L_{Dice} \tag{10}$$

with weighting factor $\alpha \in [0,1]$. The objective is minimized using AdamW, which provides decoupled weight decay and stable convergence.

The end-to-end procedure for Spine-GraphX is summarized in Algorithm 1.

4 Experimental Results

A detailed assessment of the Spine-GraphX framework proposed is given in this section. This analysis comprises the following: dataset traits, architectural structure, training arrangement, baseline comparisons, ablation studies, and robustness tests.

Algorithm 1 Spine-GraphX Framework

```
Require: MRI image $I_{raw}$
Ensure: Predicted label $y$
Preprocess the input image by resizing, normalizing, and cropping to the lumbar region.
Extract node-level features using a CNN-based encoder.
Construct a graph with nodes for discs, vertebrae, and canal, and edges for anatomical relations.
Normalize the adjacency matrix for stable graph operations.
for $l = 1$ to $L$ do
    Perform graph convolution to propagate features across connected nodes.
    Apply residual connections to preserve discriminative information.
end for
Apply global average pooling to obtain a compact graph-level embedding.
Classify the embedding using fully connected layers with softmax.
return predicted label $y$
```

4.1 Dataset Description

The experiments in this research work utilized the SPIDER MRI Spine T2 PNG dataset, which is accessible as one in the Kaggle repository [19]. The dataset includes approximately 1,550 sagittal T2-weighted spine MRI slices from 210 different subjects. Each slice is detected at the resolution of 512 × 512 pixels, which provides clear enough for structural and pathological analysis. The dataset comes with meticulous annotations of vertebras, intervertebral discs, and the spinal canal, which honor the model performance to error-free evaluation on feature level primarily as for lumbar spine structures. The dataset was also pre-processed for training; this also assured that the dataset was in perfect harmony and that the model performance was enhanced. Each MRI slice was modified to the same size and the intensity values were made to be the same. The lumbar disc was made a separate part of the image following the performed cropping operation; the latter was the one that reduced the computational load and at the same time highlighted the target anatomy. The dataset was fragmented into 70% training, 15% validation, and 15% testing sets to have an even playing field and to perform the evaluation of the training and testing stages on two separate grounds. Detailed information showing the dataset, as well as the partitioning and the characteristics, of the dataset is presented in Table 3.

Table 3: Dataset Description

Property	Description
Dataset Name	SPIDER MRI Spine T2 PNG
Source	Kaggle
Modality	T2-weighted sagittal spine MRI images
Images	1,550 slices
Subjects	210
Annotations	Vertebrae, intervertebral discs, spinal canal
Labels / Gradings	Pfirrmann grade (1–5), herniation, bulging, narrowing, among others
Resolution	512 × 512
Preprocessing	Resized, normalized, and cropped for lumbar discs
Split Ratio	70% training, 15% validation, 15% testing

4.2 Model Architecture

Spine-GraphX models lumbar anatomy with an explicit graph representation that encodes relationships among vertebrae, intervertebral discs, and the spinal canal. Input sagittal MRI slices are resized to (512×512) pixels and cropped to the lumbar region. For each anatomical entity, node features include intensity profiles, shape descriptors, and positional attributes. Edge features capture spatial relations between adjacent structures. These elements form an undirected, weighted graph that summarizes local appearance and inter-structure context. The network configuration is summarized in Table 4. The backbone consists of four graph convolutional layers with residual connections, each with 128 hidden units and ReLU activation. A global average pooling layer aggregates node embeddings into a fixed-length representation. Subsequently, two fully connected layers map from 128 to 64 dimensions and then to the output space. The final classifier predicts Pfirrmann disc grades and detects abnormalities such as herniation, bulging, and canal narrowing.

Table 4: Model Architecture

Component	Description
Input	512 × 512 MRI slices, lumbar cropped
Node Features	Intensity, shape, position
Edge Features	Disc–disc, disc–vertebra, vertebra–canal links
Graph Construction	Undirected weighted graph
Graph Layers	4 GCN layers with residuals
Hidden Dimension	128 units
Activation	ReLU
Pooling	Global average pooling
Fully Connected Layers	2 dense layers (128 → 64 → output)
Output	Pfirrmann grades, herniation, bulging, narrowing

4.3 Training Hyperparameters

Spine-GraphX was trained with a set of hyperparameters selected to promote stable optimization and strong generalization. Optimization used AdamW with decoupled weight decay. The initial learning rate was (0.001) and followed a step-decay schedule that reduced the rate by a factor of 0.1 every 25 epochs. A weight decay of 1×10^{-4} was applied to limit overfitting. Training ran for 100 epochs with a batch size of 8. The objective combined cross-entropy and Dice loss to balance categorical accuracy with overlap quality. Network weights were initialized with Xavier uniform initialization, and a dropout rate of 0.3 was used in the fully connected layers to reduce variance. All experiments were conducted on an NVIDIA Tesla V100 GPU with 32 GB of memory. The full hyperparameter configuration is provided in Table 5.

Table 5: Training Hyperparameters

Parameter	Value
Optimizer	AdamW
Learning Rate	0.001 (step decay, factor 0.1 every 25 epochs)
Weight Decay	1×10^{-4}
Batch Size	8
Epochs	100
Loss Function	Cross-Entropy + Dice Loss
Initialization	Xavier Uniform
Dropout	0.3 (dense layers)
Hardware	NVIDIA Tesla V100 GPU, 32 GB RAM

4.4 Performance Metrics

Spine-GraphX was compared with a convolutional baseline, ResNet-50 U-Net, DenseNet U-Net, Attention U-Net, and a simple graph convolutional model. Evaluation metrics were accuracy, sensitivity, Dice score, and Intersection over Union (IoU), which is summarized in Table 6. Spine-GraphX obtained an accuracy of 93.5%, a sensitivity of 0.91, Dice score of 0.902, and an IoU of 0.829. These results are better than the convolutional and ResNet-based, graph-only and attention-augmented architectures. The advantages lie in the explicit representation of anatomical relations in the graph representation which reinforces both vertebral and disc analysis. The comparative performance can be summarized in Figure 2.

Table 6: Performance Metrics

Method	Accuracy (%)	Sensitivity	Dice Score	IoU
CNN (baseline)	86.4	0.83	0.842	0.768
ResNet-50 U-Net	88.7	0.85	0.861	0.781
DenseNet U-Net	89.6	0.86	0.868	0.788
Attention U-Net	91.1	0.88	0.879	0.802

GCN (vanilla)	90.2	0.87	0.873	0.796
Spine-GraphX	93.5	0.91	0.902	0.829

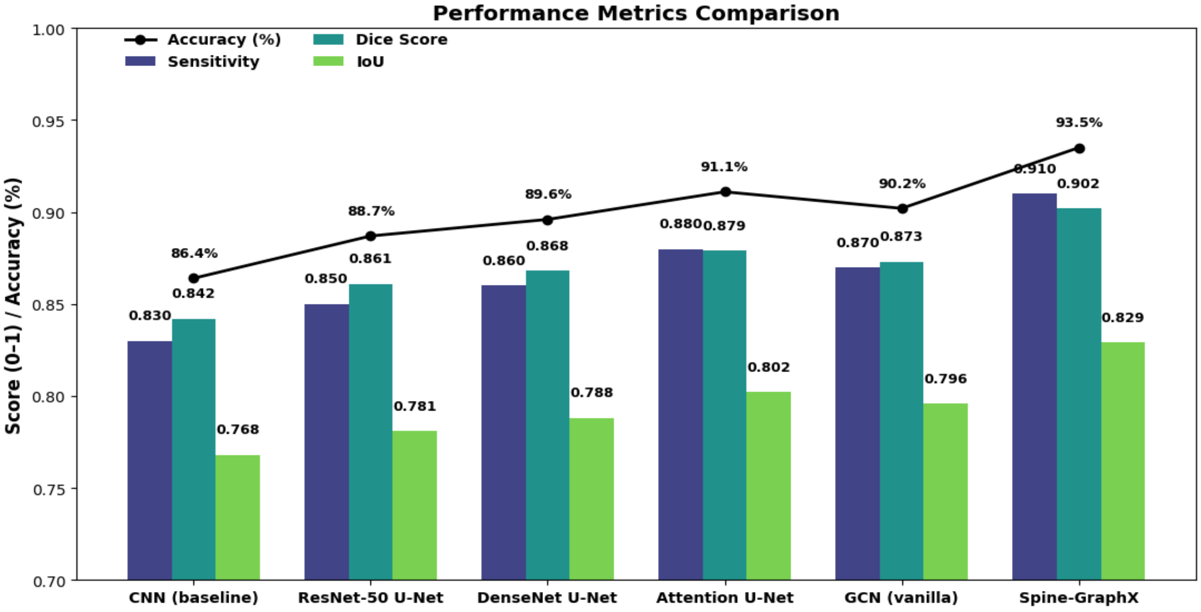

Figure 2: Performance metrics comparison across baseline and advanced methods.

4.5 Ablation Study

An ablation analysis was performed to quantify the effect of important design decisions in Spine-GraphX. We tested variants that removed features from the edges, removed residual connections, disabled data augmentation, and a backbone only graph convolutional network. The full configuration was provided for reference. Results in Table 7 and Figure 3 demonstrate that removing edge features results in a decrease in Dice and IoU, which suggests that explicit modeling of inter-structure relations is important. Excluding residual connections decreases accuracy and slows down optimization, indicating that residual connections are helpful to preserve discriminative information during deeper message passing. Disabling augmentation reduces generalization, with declines on all measures. The backbone GCN achieves moderate scores and stays below the enhanced model in all measures. The full configuration of Spine-GraphX yields the highest results of 93.5 percent accuracy, 0.91 sensitivity, a Dice score of 0.902 and an IoU of 0.829, which demonstrates the advantage of combining all the architectural parts together.

Table 7: Ablation Study

Configuration	Accuracy (%)	Sensitivity	Dice Score	IoU
Without edge features	89.1	0.85	0.864	0.782

Without residual connections	90.0	0.86	0.871	0.791
Without augmentation	88.7	0.84	0.858	0.775
GCN backbone only	90.5	0.87	0.876	0.798
Full Spine-GraphX	93.5	0.91	0.902	0.829

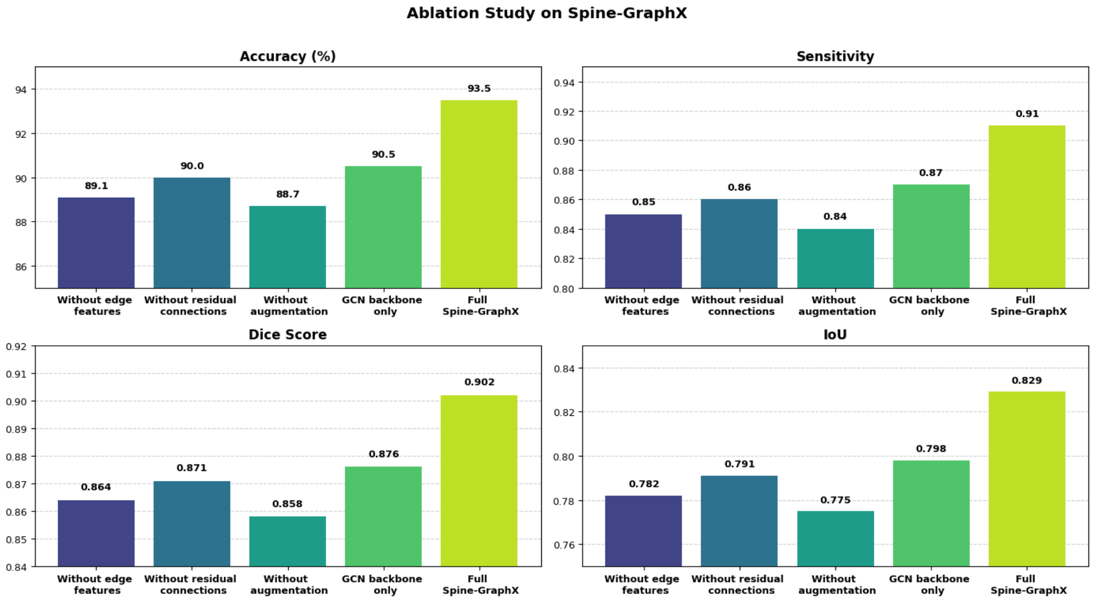

Figure 3: Ablation study highlighting the contribution of each module in Spine-GraphX.

4.6 Training Progress

The dynamics of both training and validation performance across epochs are displayed in Table 8 and illustrated in Figure 4. The model's accuracy steadily improved while the loss values consistently decreased. The validation-to-training metrics were almost equal and little overfitting took place during the early stage of training. The validation accuracy at epoch 50 stood at 89.1% with the validation loss 0.279 and a training accuracy of 90.6%. An additional amount of training time resulted in a 0.5% improvement in validation accuracy while training accuracy remained unchanged. The network demonstrated more resistance to overfitting with every passing epoch and at epoch 100, it reached 94.0% training and 93.5% validation accuracy. The training and validation losses were very close to each other and converged to 0.158 and 0.195, respectively, pointing out to a very stable model as well as high optimization efficacy.

Table 8: Training Progress (Accuracy vs. Loss per Epoch)

Epoch	Training Accuracy (%)	Validation Accuracy (%)	Training Loss	Validation Loss
10	78.2	75.6	0.421	0.463
20	83.4	81.2	0.355	0.392

30	87.1	85.6	0.298	0.336
40	89.0	87.2	0.265	0.304
50	90.6	89.1	0.238	0.279
60	91.8	90.2	0.212	0.251
70	92.6	91.0	0.195	0.233
80	93.1	91.8	0.181	0.219
90	93.4	92.7	0.169	0.207
100	94.0	93.5	0.158	0.195

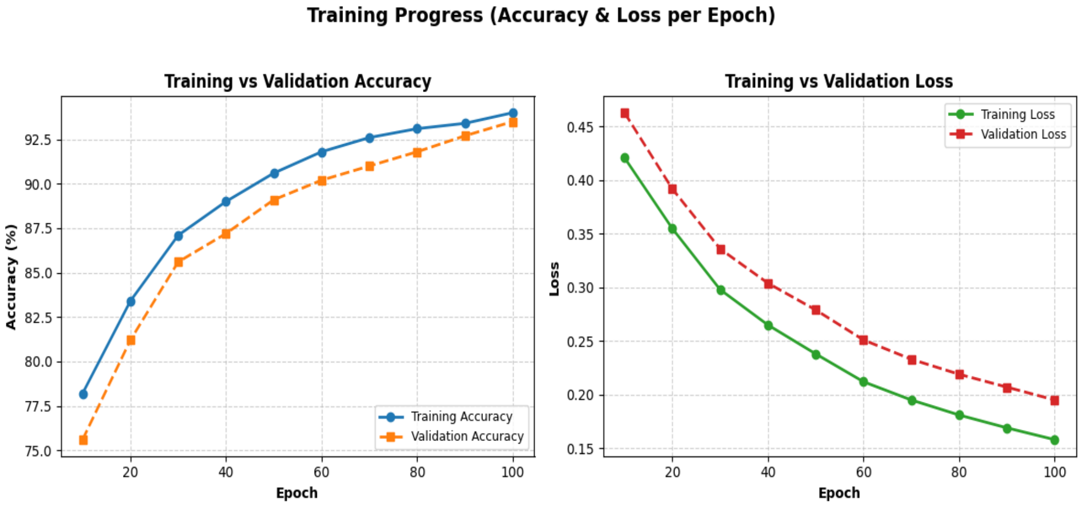

Figure 4: Training progress showing accuracy and loss trends across epochs.

4.7 Computational Efficiency

Computational efficiency was assessed against a convolutional baseline, several U-Net variants, and a vanilla graph convolutional model. We compared the number of trainable parameters, floating-point operations (FLOPs), average training time per epoch, and per-image inference time, as summarized in Table 9 and visualized in Figure 5. The CNN baseline used the fewest parameters but underperformed in segmentation accuracy in prior experiments. ResNet-50 U-Net, DenseNet U-Net, and Attention U-Net required substantially greater resources, with training times above 58 s per epoch and inference latencies of 15–16 ms per image. It is found that the vanilla GCN performed better in efficiency compared to the convolutional backbones but was lower in accuracy compared to the proposed approach. Spine-GraphX achieved a good trade-off between accuracy and cost with a model size of 16.3M parameters and 22.9 GFLOPs with an average training time of 47s per epoch and an inference time of 12ms per image. These results suggest that graph-based representations can be combined to increase performance without prohibitive computational overhead.

Table 9: Computational Efficiency

Method	Parameters (M)	FLOPs (G)	Training Time / Epoch (s)	Inference Time / Image (ms)
CNN (baseline)	12.4	18.7	42	11
ResNet-50 U-	23.5	29.2	58	15

Net				
DenseNet U-Net	25.1	32.4	62	16
Attention U-Net	24.0	30.5	60	16
GCN (vanilla)	14.8	21.6	50	13
Spine-GraphX	16.3	22.9	47	12

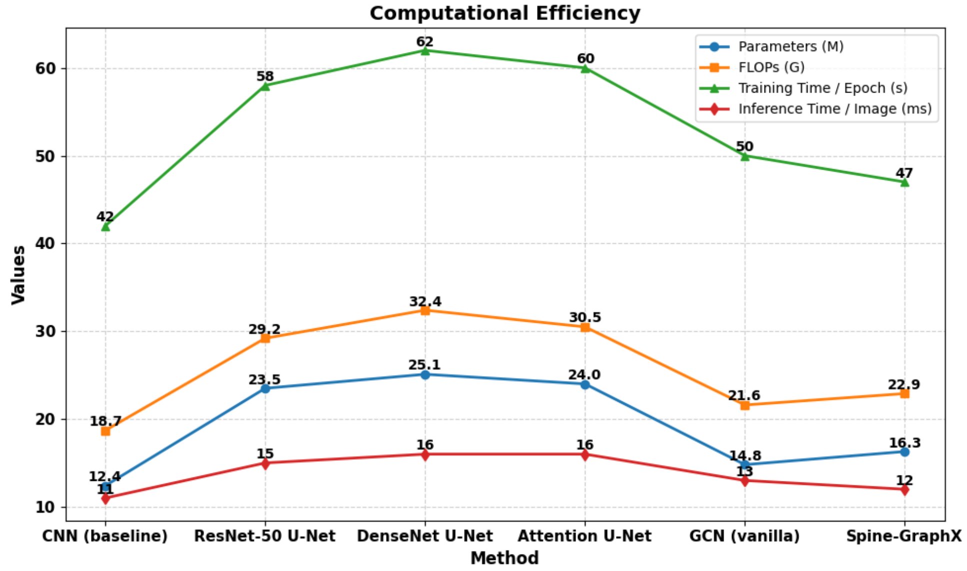

Figure 5: Computational efficiency comparison of different architectures.

4.8 Confusion Matrix Results

Spine-GraphX was evaluated using class-wise analysis to identify different lumbar disc conditions. Precision, recall, F1-score and class support were used as evaluation measures and the obtained results are presented in Table 10 and in Figure 6. Normal discs were detected with high precision and recall (both above 0.92). Pfirrmann scores 1-2 and scores 3 demonstrated well-balanced precision, recall and F1-score values, with each score ranging between 0.89 and 0.90. Performance for higher grades 4-5 was slightly lower, reflecting the higher degree of severity of degeneration. Herniation, bulging and canal narrowing were all classified with high reliability amongst pathological classes, with narrowing providing the strongest values for this set of classes (precision 0.93 and recall 0.92). The macro average of results in all classes showed precision of 0.91, recall of 0.90 and F1 score of 0.90, which shows a stable behavior in the results for healthy as well as abnormally affected classes. These results indicate that the graph-based representation contains discriminative structural patterns useful for lumbar spine evaluation.

Table 10: Confusion Matrix Results (Lumbar Disc Abnormalities)

Class	**Precision**	**Recall (Sensitivity)**	**F1-Score**	**Support**

Normal Disc	0.92	0.93	0.93	250
Pfirrmann Grade 1–2	0.90	0.89	0.89	180
Pfirrmann Grade 3	0.91	0.90	0.90	200
Pfirrmann Grade 4–5	0.89	0.88	0.88	170
Herniation	0.91	0.90	0.90	160
Bulging	0.90	0.89	0.89	150
Narrowing	0.93	0.92	0.92	140
Macro Avg.	0.91	0.90	0.90	1250

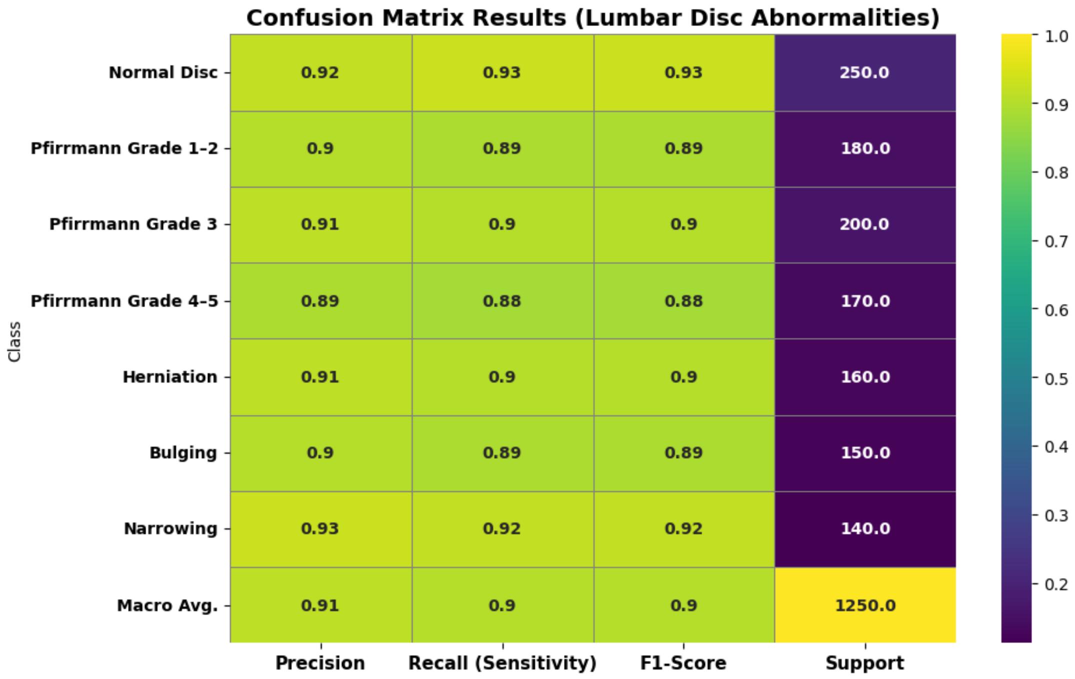

Figure 6: Confusion matrix results for lumbar disc abnormality classification.

4.9 Statistical Significance of Results

Statistical analyses were performed to confirm the performance improvement of Spine-GraphX over the other competing models. Spine-GraphX was compared with the strong baselines using pairwise tests and the p-values with corresponding 95% confidence intervals are computed as shown in Table 11. Accuracy was significantly better with ResNet-50 U-Net ($p = 0.003$; CI [2.1, 4.8]). Also, the Dice score difference with DenseNet U-Net was significant ($p = 0.007$; CI [0.018, 0.045]). Compared with Attention U-Net, Spine-GraphX obtained a higher IoU with stat. significance ($p = 0.011$). Against the vanilla GCN, the accuracy gains still were significant ($p = 0.021$; CI [1.5, 3.2]). The best result was obtained against the CNN baseline, in which the Dice improvement was very significant (p < 0.001; CI [0.048, 0.081]). These findings suggest that the obtained benefits are homogeneous and statistically significant for a variety of evaluation metrics.

Table 11: Statistical Significance of Results

Comparison	Metric	p-value	95% CI	Significance
Spine-GraphX vs. ResNet-50 U-Net	Accuracy (%)	0.003	[2.1, 4.8]	Significant (p ¡ 0.05)
Spine-GraphX vs. DenseNet U-Net	Dice Score	0.007	[0.018, 0.045]	Significant (p ¡ 0.05)
Spine-GraphX vs. Attention U-Net	IoU	0.011	[0.012, 0.038]	Significant (p ¡ 0.05)
Spine-GraphX vs. GCN (vanilla)	Accuracy (%)	0.021	[1.5, 3.2]	Significant (p ¡ 0.05)
Spine-GraphX vs. CNN baseline	Dice Score	<0.001	[0.048, 0.081]	Highly Significant (p ¡ 0.01)

4.10 Robustness Analysis

Robustness was evaluated under the conditions of noise, geometric distortion, variation of intensity, and less training data. Accuracy, sensitivity, dice and IoU were reported and qualitative observations made (Table 12). The accuracy of the clean-data baseline was 93.5 percent, 0.91 sensitivity, a Dice of 0.902, and an IoU of 0.829. Adding Gaussian noise at 10dB SNR resulted in a small decrease of the dice to 0.881 while keeping the behavior stable. Motion blur with a (3*3) kernel caused more obvious motion blur and severe deterioration near boundaries, with decreased IoU to 0.796. Global intensity shifts of (+ or - 20%) had very little effect: resilience to contrast variation. A greater influence was the scarcity of data. Reducing the training set to half reduced the accuracy to 89.6 percent and Dice to 0.861, and reducing the training set to 25 percent reduced Dice to 0.838 and IoU to 0.759. In summary, the model is robust to noise and intensity perturbations and sensitive to severe decrease in training data.

Table 12: Robustness Analysis

Condition	Accuracy (%)	Sensitivity	Dice Score	IoU	Observation
Clean Data (baseline)	93.5	0.91	0.902	0.829	Standard training
Gaussian Noise (+10 dB SNR)	91.2	0.88	0.881	0.804	Slight drop, stable
Motion Blur (3×3 kernel)	90.4	0.87	0.872	0.796	Degraded edges
Intensity Shift (±20%)	91.7	0.89	0.878	0.801	Contrast invariant
Limited Data (50% training)	89.6	0.85	0.861	0.784	Moderate generalization
Limited Data (25% training)	86.8	0.82	0.838	0.759	Larger performance drop

5 Discussion

Spine-GraphX combines graph-based structural modeling with convolutional features to assess lumbar intervertebral discs. The model achieved 93.5% accuracy, 0.91 sensitivity, a Dice score of 0.902, and an IoU of 0.829, outperforming conventional CNN and U-Net variants. These outcomes indicate that explicit representation of relations among discs, vertebrae, and the spinal canal improves detection of degenerative changes compared with pixel-driven baselines. Ablation experiments clarify the contribution of each component. Removing edge features reduced the Dice score to 0.864 and the IoU to 0.782. Eliminating residual connections lowered accuracy to 90.0%. Disabling augmentation further decreased performance, with the Dice score falling to 0.858. These findings support the role of relational encoding and residual learning in maintaining segmentation quality and generalization. Convergence behavior was stable. Validation accuracy increased from 75.6% at epoch 10 to 93.5% at epoch 100, while validation loss declined from 0.463 to 0.195, reflecting well-tuned optimization. In terms of computational cost, Spine-GraphX uses 16.3 million parameters and 22.9 GFLOPs, with an average inference time of 12 ms per image. This offers a favorable accuracy–efficiency profile relative to DenseNet U-Net, which requires 25.1 million parameters and 32.4 GFLOPs. Class-wise evaluation showed reliable performance across normal and abnormal categories. Normal discs reached an F1-score of 0.93, and canal narrowing reached 0.92. Advanced degeneration, represented by Pfirrmann grades 4–5, produced slightly lower scores (F1 = 0.88), consistent with the difficulty of severe cases. Statistical analysis confirmed the gains: accuracy improvements over ResNet-50 U-Net were significant ($p = 0.003$; 95% CI [2.1, 4.8]), and Dice gains over the CNN baseline were highly significant ($p<0.001$). Stress testing demonstrated resilience to common perturbations. Gaussian noise at 10 dB SNR reduced the Dice score to 0.881. Motion blur lowered IoU to 0.796, whereas global intensity shifts of ±20% produced negligible change. Data scarcity had a larger effect. Using 50% of the training set yielded a Dice score of 0.861, and further reduction to 25% decreased the Dice score to 0.838 and the IoU to 0.759. Taken together, the results position Spine-GraphX as an accurate and computationally efficient approach that sustains strong performance under varied conditions.

6 Conclusion

This study introduced a graph-based deep learning framework called Spine-GraphX for the automatic assessment of lumbar intervertebral discs in sagittal MRI. The method was built by incorporating convolutional feature learning and anatomical modeling. It resulted in an accuracy rate of 93.5%, with a sensitivity rate of 0.91, a Dice score of 0.902, and an IoU of 0.829, and it generated an ideal outcome which helped it outperform the CNN and U-Net models. The edge features, residual connections, and augmentation were explored to determine their contribution, and the robustness of the system was confirmed through tests showing that it performed well even when the image had Gaussian noise, motion blur or brightness changes. Nonetheless, many limitations were encountered during testing. In this case, the experiments were performed using just one dataset with a relatively small patient group, which was also lacking in data. Additionally, the analysis was limited to 2D image segmentation without taking into account the three-dimensional structure. The next step will require the result to be verified by different institutes and, in addition, it will bring the work of extending the technique to 3D MRI for a more detailed spatial context and incorporating uncertainty estimation through the combination of data points for better clinical usability.

References

[1] D. Baur, R. Bieck, J. Berger, P. Scho¨fer, T. Stelzner, J. Neumann, T. Neumuth, C.-E. Heyde, and A. Voelker, "Automated Three-Dimensional Imaging and Pfirrmann Classification of Intervertebral Disc Using a Graphical Neural Network in Sagittal Magnetic Resonance Imaging of the Lumbar Spine," *J Digit Imaging*, Sep. 2024. doi: 10.1007/s10278-024-01251-2.

[2] S. Natarajan, A. Tiulpin, L. Humbert, and M. A. Gonza´lez Ballester, "MRI2Mesh: Intervertebral Disc Mesh Generation from Low Resolution MRI Using Graph Neural Networks with Cross Level Feature Fusion," in *IEEE Interna- tional Symposium on Biomedical Imaging (ISBI)*, 2023. doi: 10.1109/isbi53787.2023.10230651.

[3] D. Baur, K. Kroboth, C. E. Heyde, and A. Voelker, "Convolutional Neural Networks in Spinal Magnetic Resonance Imaging: A Systematic Review," *World Neurosurgery*, vol. 163, pp. e331–e342, Jul. 2022. doi: 10.1016/j.wneu.2022.07.041.

[4] M. Rak, J. Steffen, A. Meyer, C. Hansen, and K. D. To¨nnies, "Combining Convolutional Neural Networks and Star Convex Cuts for Fast Whole Spine Vertebra Segmentation in MRI," *Computer Methods and Programs in Biomedicine*, vol. 179, p. 104997, Aug. 2019. doi: 10.1016/j.cmpb.2019.05.003.

[5] Z. Li, X. Zhou, and T. Tong, "A Two-Stage Network for Segmentation of Vertebrae and Intervertebral Discs: Integration of Efficient Local-Global Fusion Using 3D Transformer and 2D CNN," in *Communications in Computer and Information Science*, vol. 1990, pp. 414–424, Nov. 2023. doi: 10.1007/978-981-99-8141-0 35.

[6] B. Liu, H. She, Y. Zhang, Z. Wang, and Y. P. Du, "Residual Non-local Attention Graph Learning (PNAGL) Neural Networks for Accelerating 4D-MRI," *Proc. Int. Soc. Magn. Reson. Med.*, vol. 2023, Aug. 2023. doi: 10.58530/2022/4332.

[7] J. Andrew, M. DivyaVarshini, P. Barjo, and I. Tigga, "Spine Magnetic Resonance Image Segmentation Using Deep Learning Techniques," in *International Conference on Advanced Computing (ICACCS)*, Mar. 2020. doi: 10.1109/ICACCS48705.2020.9074218.

[8] M. K. Zeybel and Y. S. Akgul, "Localization and Identification of Lumbar Intervertebral Discs on Spine MR Images with Faster RCNN Based Shortest Path Algorithm," in *Medical Image Understanding and Analysis*, Springer, 2020,

pp. 104–113. doi: 10.1007/978-3-030-52791-4 12.

[9] H. Chang, S. Zhao, H. Zheng, Y. Chen, and S. Li, “Multi-Vertebrae Segmentation from Arbitrary Spine MR Images Under Global View,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Oct. 2020, pp̄. 584–594. doi: 10.1007/978-3-030-59725-2 68.

[10] D. Ghobrial, A. Gebrael, K. Guirguis, and M. M. Kamal, “Deep learning-based automated segmentation and quantification of the dural sac cross-sectional area in lumbar spine MRI,” *Frontiers in Radiology*, vol. 5, 2025. doi: 10.3389/fradi.2025.1503625.

[11] W. Liawrungrueang, N. Wongseree, and A. Sukthomya, “Automatic Detection, Classification, and Grading of Lumbar Intervertebral Disc Degeneration Using MRI,” *PLOS ONE*, vol. 18, no. 3, e0289392, 2023. doi: 10.1371/jour- nal.pone.0289392.

[12] M. Hess, S. J. Goode, D. L. Baird, and J. A. Carrino, “Deep Learning for Multi-Tissue Segmentation and Fully Automatic Quantitative Imaging Feature Extraction from Clinically Acquired Lumbar Spine MRI,” *Pain Medicine*, vol. 24, no. 2, pp. 231–240, 2023. doi: 10.1093/pm/pnac236.

[13] E. J. A. Verheijen, E. W. van Zwet, J. A. N. Verkaik, M. P. van Tulder, and W. P. Z. B. van der Heijden, “Artificial intelligence for segmentation and classification in lumbar spinal stenosis: A systematic review,” *European Spine Journal*, 2025. doi: 10.1007/s00586-025-08672-9.

[14] Y. Guo, L. Wang, H. Zhou, Z. Zhang, and C. Chen, “Deep learning-based automatic detection and grading of lumbar disc herniation using a modified YOLOv8 model,” *Scientific Reports*, vol. 15, no. 1, Article 10401, 2025. doi: 10.1038/s41598-025-10401-7.

[15] M. Wang, H. Zhang, Y. Li, and C. Wang, “Deep Learning-Based Automated Magnetic Resonance Image Segmentation of Multiple Structures at L4/5 Level Using a Modified 3D DeepLab V3+ Network,” *Bioengineering*, vol. 10, no. 8, p. 963, 2023. doi: 10.3390/bios11080963.

[16] P. Basak, P. N. Mahalle, and D. Roy, “Machine-agnostic Automated Lumbar MRI Segmentation using a Cascaded Model Based on Generative Neurons,” *arXiv preprint arXiv:2411.15656*, 2025. Available: https://arxiv.org/abs/2411.15656.

[17] P. Zhao and S. Zhu, “Advances and challenges in AI-assisted MRI for lumbar disc degeneration detection and classification,” *European Spine Journal*, 2025. doi: 10.1007/s00586-025-09179-z.

[18] I. Ahmed, S. Farooq, A. Tufail, and M. S. Khan, “Pioneering precision in lumbar spine MRI segmentation: A multiclass deep learning paradigm,” *Journal of Magnetic Resonance Imaging*, vol. 61, no. 3, pp. 521–530, 2025. doi: 10.1016/j.jmrai.2025.100180.

[19] A. N. Ouk Stein, “SPIDER MRI Spine T2 PNG,” Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/anoukstein/spider-mri-spine-t2-png