DiffuSpine: Diffusion-Based Data Augmentation Framework for Grading Lumbar Degenerative Disc Disease

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 (MRI)-based lumbar degenerative disc disease grading can aid triage and more consistent spine reporting. The goal is robust three-grade severity classification for routine use. Severe cases are scarce and grade boundaries are visually subtle. Multi-center protocols also shift contrast and can miscalibrate confidence. Geometric and intensity augmentation are often used for training. MixUp, CutMix, and Generative Adversarial Network (GAN) synthesis introduce variation but can also decrease realism. DiffuSpine is a diffusion-based data augmentation framework with level conditioning. Quality filtering and balanced sampling aim to achieve fidelity and rarity. The experiments are performed on the dataset RSNA Lumbar Degenerative Imaging Spine Classification (LumbarDISC) available on Kaggle, which contains 2697 patients and 8593 MRI series. The cohort includes 300,000+ slices from L1-L2 to L5-S1 and Normal/Mild, Moderate and Severe labels. With DiffuSpine, Shifted Window Transformer Tiny (Swin Transformer-Tiny) achieves 94.6% accuracy, 93.8% Macro F1-score and an Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.974. Expected Calibration Error (ECE) drops from 0.034 (standard augmentation) to 0.020, while Severe F1-score improves from 83.5% to 92.3%. These gains facilitate prioritisation, second reading and audit-friendly confidence. The biggest constraint is external validation between sequences and between sites.

Keywords**:** Diffusion model, data augmentation, lumbar degenerative disc disease, Magnetic Resonance Imaging (MRI), severity grading, Swin Transformer-Tiny, calibration

1 INTRODUCTION

Lumbar degenerative disc disease is a very common cause of chronic low back pain and decreased mobility. Magnetic Resonance Imaging (MRI) is commonly used to evaluate spinal canal stenosis, foraminal narrowing, subarticular recess stenosis, and disc-level degeneration. Normal/Mild, Moderate, or Severe grades are commonly used by radiologists from L1-L2 to L5-S1. When imaging volume is high, automated grading can help with triage, second reading, longitudinal review and consistent reporting. Moderate and Severe cases are difficult to grade, because they can only be differentiated by subtle changes in morphology. The image contrast also varies according to the scanner manufacturer, acquisition protocol, coil configuration and patient anatomy. Another problem is class imbalance, as Severe cases are less frequent than Normal/Mild cases. A classifier can have a high overall accuracy and fail to recognise clinically important minority patterns. There is also a need for confidence calibration, as high predicted probabilities should indicate a high probability of being correct. Conventional systems use a combination of convolutional or transformer backbones with flips, rotations, intensity perturbation, or noise injection. MixUp and CutMix mix images or patches as a form of regularisation. Synthetic samples can be generated using Generative Adversarial Networks. These approaches have the potential to increase diversity; however, the images they create can suffer from a loss of anatomical consistency or a lack of fine-grained features. Bad synthetic samples can also boost confidence without enhancing correctness. DiffuSpine overcomes these restrictions by using conditional diffusion-based augmentation. Severity and disc-level information guide synthesis, and confidence/entropy filtering removes uncertain samples. Class-balanced sampling leads to a higher proportion of Moderate and Severe cases. Note that the augmentation process is only applied during training and the Swin Transformer-Tiny inference pathway is not changed. Discrimination and calibration are assessed on the LumbarDISC dataset using Accuracy, Macro F1-score, Weighted F1-score, Area Under the Receiver Operating Characteristic Curve and Expected

Calibration Error. The primary novel contributions are: a level-conditioned diffusion augmentation strategy, a filtering rule for low-quality synthetic images, targeted sampling for minority grades, and a multi-level evaluation with comparison and ablation analysis.

2. Related Work

Deep learning has been recently used in lumbar and cervical spine imaging. Shakir et al. [6] fused deep learning and machine learning in lumbar bone marrow classification and found that limited data and image variation are the main constraints. Wang et al. [8] created a multitask model for facet joint osteoarthritis (FJOA) on computed tomography (CT), and McSweeney et al. [7] applied segmentation and radiomic features to intervertebral disc degeneration. Kowlagi et al. [9] studied multimodal low back pain stratification, and Ram et al. [10] proposed a multi-branch residual network for intervertebral disc classification. These studies provide helpful information; however, there are issues with cohort dependence, scanner variation, labelling consistency, and external validation. Other studies of the cervical spine have also demonstrated the benefits and drawbacks of automated reporting. Lee et al. [11] found stenosis and cord signal changes on Magnetic Resonance Imaging. A subsequent transformer-based system enhanced reporting efficiency and reader agreement [12]. A two-stage framework for cervical stenosis on computed tomography [14] and detection-based grading with AFFM-YOLOv8 for lumbar disc herniation [13] have also been investigated. Most works focus on classifier architectures or localisation. There is less focus on grade-aware synthetic augmentation and confidence calibration. DiffuSpine addresses these two problems without changing the final classifier.

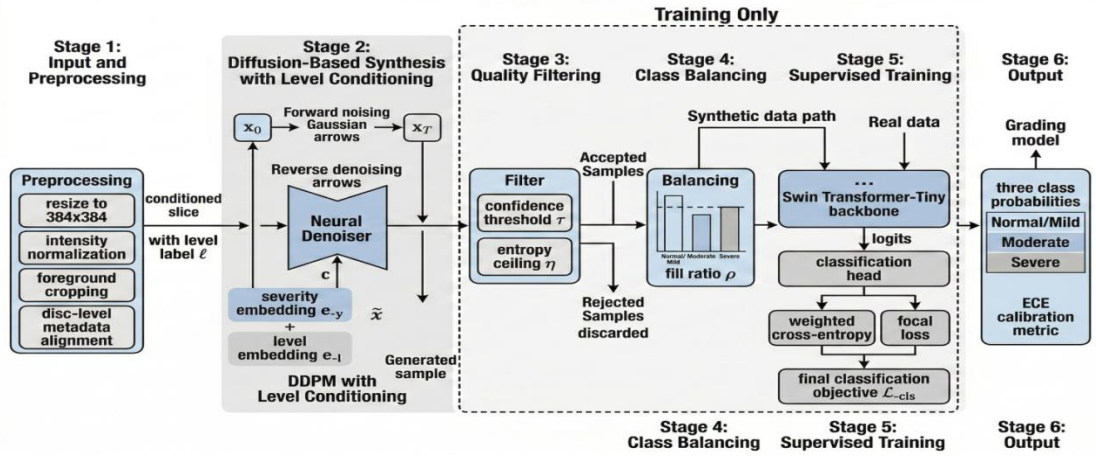


Figure 1: Overview of Proposed framework

3. SYSTEM METHODOLOGY

3.1 Preprocessing and Grading Network

Each sagittal or axial T2 slice is resized to 384×384 pixels. Study-based intensity normalization is applied, and foreground cropping can be used to focus on the spinal region. Severity labels are $y \in \{0, 1, 2\}$ for Normal/Mild, Moderate, and Severe. Disc-level labels are $\ell \in \{1, \ldots, 5\}$ for L1-L2 through L5-S1.

The grading model uses Swin Transformer-Tiny followed by a three-class classification head, as shown in figure 1. The backbone captures local and long-range image relations that are useful when grade boundaries are subtle. DiffuSpine does not add any component to the test-time network.

3.2 Level-Conditioned Diffusion Synthesis

A conditional denoising diffusion model creates additional training images. The conditioning vector combines severity and disc-level embeddings, $c = [e_y; e_l]$. During training, Gaussian noise is added to a clean image x_0, and the denoiser learns to estimate the added noise. The diffusion objective is

$$L_{diff}(\theta) = \mathbb{E}_{t,x_0,\epsilon}[\| \epsilon - \epsilon_\theta(x_t, t, c) \|_2^2]. \quad (1)$$

This conditioning encourages generated images to retain both grade-related appearance and level-specific morphology. Synthetic samples are produced only for the training split.

3.3 Quality Filtering and Class Balancing

Generated images are evaluated by the current classifier. A candidate $\tilde{x}$ is retained when the predicted probability of its intended class is at least tau and its entropy is below eta:

$$keep(\tilde{x}) = \mathbb{I}[\hat{p}_y \geq \tau]\mathbb{I}[-\sum_{k=1}^{3} \hat{p}_k \log(\hat{p}_k) \leq \eta]. \quad (2)$$

The rule removes uncertain or inconsistent samples before they enter classifier training. The number of retained synthetic samples for class y is controlled by

$$m_y = \text{floor}[\rho(n_{max} - n_y)], \quad (3)$$

where n_y is the number of real samples in class y, n_{max} is the largest class count, and rho controls the balancing strength. This process mainly increases exposure to Moderate and Severe cases without allowing synthetic data to dominate the training set.

3.4 Classification Objective

The classifier uses weighted cross-entropy and focal loss. Class weights reduce the effect of imbalance, while the focal term gives more attention to difficult samples. The final objective is

$$L_{cls}(\phi) = \lambda L_{wce}(\phi) + (1 - \lambda) L_{focal}(\phi). \quad (4)$$

Mini-batches contain real images and filtered synthetic images. The diffusion model is used only to prepare training samples.

4. EXPERIMENTAL RESULTS

4.1 Dataset and Experimental Setup

The experiments use the publicly available RSNA Lumbar Degenerative Imaging Spine Classification dataset. It includes 2,697 patients, 8,593 MRI series and over 300,000 sagittal and axial T2 slices from eight international centers. Labels provided by the radiologist are Normal/Mild (L1-L2), Moderate (L3-L4) and Severe (L5-S1). The split is made at patient level with 70% for training, 15% for validation and 15% for testing.

Table 1: Dataset and training configuration

Item	Setting
Dataset	LumbarDISC
Patients	2,697
MRI series	8,593
Total slices	300,000+
Institutions	8 international centers
Imaging	Sagittal T2 and Axial T2 MRI
Classes	Normal/Mild, Moderate, Severe
Disc levels	L1-L2 to L5-S1

Train/Validation/Test	1,887 / 405 / 405 patients
Backbone	Swin Transformer-Tiny
Input size	384×384
Optimizer	AdamW
Learning rate	3×10^{-4}
Batch size	316
Epochs	40
Scheduler	Cosine decay
Loss	Weighted Cross-Entropy + Focal Loss
Metrics	Accuracy, Macro-F1, Weighted-F1, AUROC, ECE

4.2 Overall Comparison

Figure 2 shows the same backbone under different augmentation strategies. Standard augmentation improves accuracy from 89.2% to 91.0%. MixUp, CutMix, and Generative Adversarial Network augmentation provide smaller additional gains. DiffuSpine reaches 94.6% accuracy, 93.8% Macro F1-score, 94.4% Weighted F1-score, and an AUROC Curve of 0.974. It also gives the lowest Expected Calibration Error of 0.020. The run-to-run deviation decreases to +/-0.3, compared with +/-0.7 without augmentation.

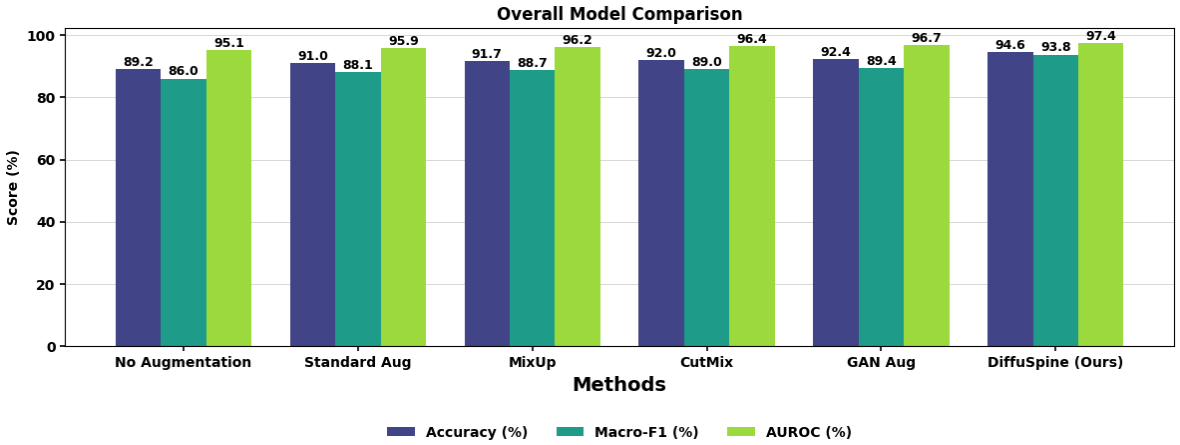


Figure 2: Grading performance under different augmentation strategies.

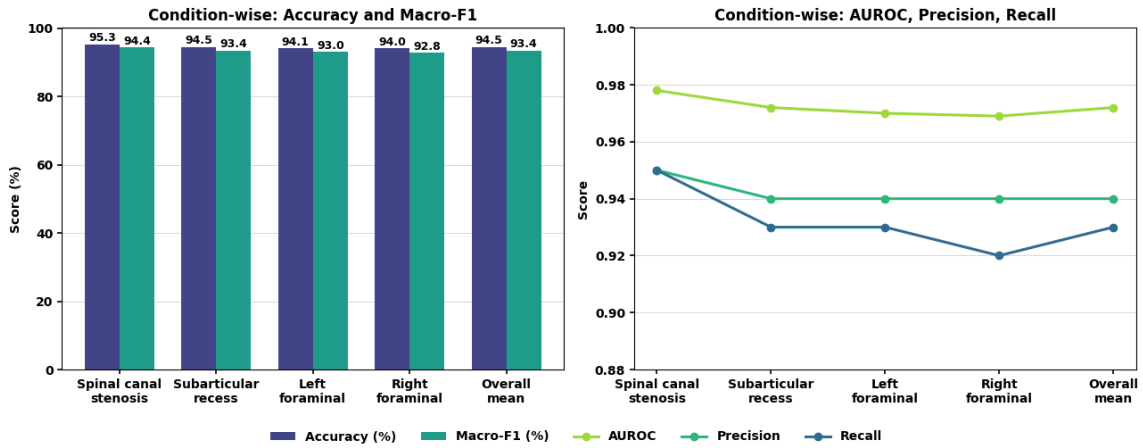


Figure 3: Condition-wise evaluation of DiffuSpine.

4.3 Condition-Wise and Level-Wise Results

DiffuSpine performs consistently across the evaluated lumbar conditions. Spinal canal stenosis reaches 95.3%

accuracy, 94.4% Macro F1-score, and 0.978 AUROC Curve. Subarticular recess stenosis reaches 94.5% accuracy, 93.4% Macro F1-score, and 0.972 Area Under the Receiver Operating Characteristic Curve, as shown in Figure 3. Left and right foraminal narrowing achieve 94.1% and 94.0% accuracy, with Macro F1-scores of 93.0% and 92.8%. The mean across conditions is 94.5% accuracy, 93.4% Macro F1-score, 0.972 AUROC Curve, 0.94 precision, and 0.93 recall. Level-wise analysis shows gains at all five lumbar levels. The largest improvement occurs at L5-S1, where Macro F1-score rises from 87.0% to 93.7%. L4-L5 records the highest accuracy of 95.4%. The mean Macro F1-score increases from 88.0% to 93.8%, giving a 5.8-point improvement, as shown in Figure 4.

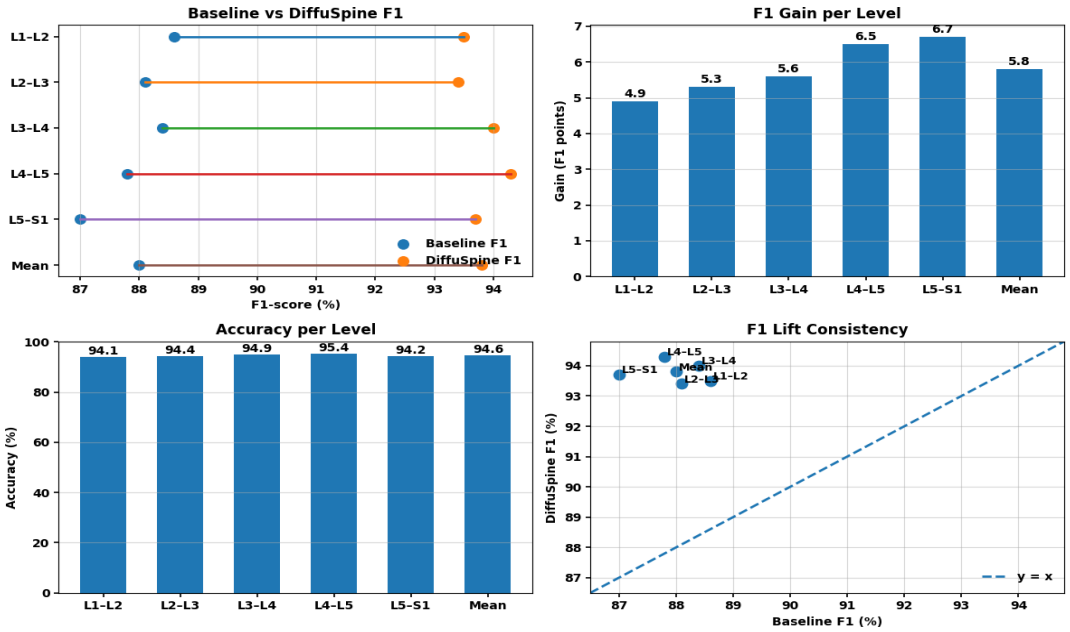


Figure 4: Level-wise grading performance across L1-L2 to L5-S1.

4.4 Ablation Study

The ablation results show a steady contribution from each DiffuSpine component. Diffusion synthesis raises Macro F1-score from 88.1% to 90.6%. Quality filtering improves it to 92.0% and reduces Expected Calibration Error to 0.024. Class balancing raises accuracy to 94.0%. Level conditioning gives the final result of 93.8% Macro F1-score, 0.974 AUROC Curve, 0.020 Expected Calibration Error, and 94.6% accuracy. The pattern suggests that diffusion synthesis is the primary gain, with filtering, balancing and level information enhancing reliability and minority-grade learning.

Table 4: Ablation study of DiffuSpine components

Variant	Macro-F1	AUROC	ECE	Accuracy (%)
Standard Augmentation only	88.1	0.959	0.034	91.0
+ Diffusion samples	90.6	0.966	0.028	92.6
+ Quality filtering	92.0	0.970	0.024	93.5
+ Class balancing	92.8	0.972	0.022	94.0
+ Level conditioning	93.8	0.974	0.020	94.6

5. CONCLUSION

DiffuSpine performs level-conditioned diffusion augmentation, quality filtering, and class-balanced sampling to improve the grading of lumbar degenerative disc disease. The method achieves 94.6% accuracy, 93.8% Macro F1-score and 0.974 AUROC Curve on the LumbarDISC dataset. Expected Calibration Error goes down from 0.034 with standard augmentation to 0.020. Severe F1-score also rises from 83.5% to 92.3%. These results show an improvement in minority-grade recognition and more consistent prediction confidence.

REFERENCES

[1] Y. Xu, S. Zheng, Q. Tian, Z. Kou, W. Li, X. Xie, and X. Wu, "Deep learning model for grading and localization of lumbar disc herniation on magnetic resonance imaging," J. Magn. Reson. Imaging, vol. 61, no. 1, pp. 364-375, 2025, doi: 10.1002/jmri.29403.

[2] Z. Soydan, E. Bayramoglu, R. Karasu, I. Sayin, S. Salturk, and H. Uvet, "An automatized deep segmentation and classification model for lumbar disk degeneration and clarification of its impact on clinical decisions," Global Spine J., vol. 15, no. 2, pp. 554-563, Mar. 2025, doi: 10.1177/21925682231200783.

[3] L. Ding, R. Gu, X. Liu, Q. Li, H. Cui, S. Cheng, Y. Shi, T. Zhu, Z. Hu, and Y. Zhang, "Exploring the relationship between severity of intervertebral disc degeneration and dyslipidemia: A deep learning approach," Research Square, May 13, 2025, doi: 10.21203/rs.3.rs-6250034/v1.

[4] P. Giaccone, F. D'Antoni, F. Russo, L. Ambrosio, G. F. Papalia, O. D'Angelis, et al., "Prevention and management of degenerative lumbar spine disorders through artificial intelligence-based decision support systems: A systematic review," BMC Musculoskelet. Disord., vol. 26, no. 1, Art. no. 126, 2025, doi: 10.1186/s12891-025-08356-x.

[5] A. Trento, S. Rapisarda, N. Bresolin, A. Valenti, and E. Giordan, "Artificial intelligence and its impact on the management of lumbar degenerative pathology: A narrative review," Medicina, vol. 61, no. 8, Art. no. 1400, 2025, doi: 10.3390/medicina61081400.

[6] Y. H. Shakir, T. S. Kiong, C. P. Chen, and S. S. A. Kumar, "Hybrid DL and ML approach for MRI-based classification of bone marrow changes in lumbar vertebrae," Bull. Electr. Eng. Inform., 2025, doi: 10.11591/eei.v14i5.10617.

[7] T. McSweeney, A. Tiulpin, N. Kowlagi, J. Maatta, J. Karppinen, and S. Saarakkala, "Robust radiomic signatures of intervertebral disc degeneration from MRI," Spine, vol. 50, no. 24, pp. 1737-1746, 2025, doi: 10.1097/BRS.0000000000005435.

[8] Y. Wang, Z. Chen, J. Huang, Q. He, D. Leng, L. Yang, et al., "Comprehensive evaluation of facet joints osteoarthritis radiological features on lumbar CT: A multitask deep learning approach," JOR Spine, vol. 8, no. 3, Art. no. e70115, 2025, doi: 10.1002/jsp2.70115.

[9] N. Kowlagi, E. Heikkala, S. Saarakkala, J. Karppinen, and A. Tiulpin, "An unsupervised learning approach for multimodal low back pain stratification," Spine, early access, Dec. 2025, doi: 10.1097/BRS.0000000000005593.

[10] I. Ram, S. Kumar, and A. K. Keshri, "Classification of intervertebral disc using novel multi-branch convolutional residual network model," Biomed. Signal Process. Control, vol. 102, Art. no. 107332, Apr. 2025, doi: 10.1016/j.bspc.2024.107332.

[11] A. Lee, J. Wu, C. Liu, A. Makmur, Y. H. Ting, et al., "Deep learning model for automated diagnosis of degenerative cervical spondylosis and altered spinal cord signal on MRI," Spine J., vol. 25, no. 2, pp. 255-264, Feb. 2025, doi: 10.1016/j.spinee.2024.09.015.

[12] A. Lee, J. Wu, C. Liu, A. Makmur, Y. H. Ting, et al., "Using deep learning to enhance reporting efficiency and accuracy in degenerative cervical spine MRI," Spine J., vol. 25, no. 9, pp. 1942-1950, Sep. 2025, doi: 10.1016/j.spinee.2025.03.009.

[13] Y. Wang, Z. Yang, S. Cai, W. Wu, and W. Wu, "MRI grading of lumbar disc herniation based on AFFM-YOLOv8 system," Sci. Rep., vol. 15, no. 1, Art. no. 32880, Sep. 2025, doi: 10.1038/s41598-025-18417-9.

[14] Y.-L. Zhang, J.-W. Huang, K.-Y. Li, H.-L. Li, X.-X. Lin, et al., "Automated classification of cervical spinal stenosis using deep learning on CT scans," Spine, early access, Jun. 2025, doi: 10.1097/BRS.0000000000005414.