International Journal of Research and Review
[Indexed, Peer-Reviewed Journal]
Volume 13; Issue: 6; June 2026
Website: www.ijrrjournal.com
Research Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

Validity and Reliability of the Pfirrmann and Schizas Criteria Degrees in Lumbar Degenerative Disease Patients Using the Deep Learning Method

Ivan Alexander Liando[1], I Wayan Suryanto Dusak[1], I Gusti Lanang Ngurah Agung Artha Wiguna[1], I Ketut Suyasa[1], Elysanti Dwi Martadiani[2], Made Bramantya Karna[1], I Gusti Ngurah Wien Aryana[1], Cokorda Gde Oka Dharmayuda[1], I Gede Eka Wiratnaya[1], Anak Agung Gde Yuda Asmara[1], I Wayan Subawa[1]

[1]Department of Orthopaedic and Traumatology, Faculty of Medicine, Udayana University/Ngoerah Hospital, Denpasar, Indonesia
[2]Department of Radiology, Faculty of Medicine, Udayana University/Ngoerah Hospital, Denpasar, Indonesia

Corresponding Author: Ivan Alexander Liando

DOI: https://doi.org/10.52403/ijrr.20260659

ABSTRACT

Traditional diagnosis of lumbar degenerative disease relies on clinical evaluation and MRI imaging. Machine learning (ML) and deep learning (DL) have potential in automating the assessment of spinal conditions. This study aims to evaluate the validity and reliability of deep learning models in determining the Pfirrmann and Schizas grade for lumbar degenerative disease using MRI. A retrospective study was conducted using MRI scans of lumbar spine patients. A deep learning model was trained to classify degenerative changes based on the Pfirrmann and Schizas scoring systems. Diagnostic accuracy was assessed using a Receiver Operating Characteristic (ROC) curve, and reliability was measured by interobserver agreement. A total of 170 patients were included, with a mean age of 55.20 ± 13.34 years (range 20–>60 years) and a near-equal sex distribution (48.8% male, 51.2% female). The deep learning model demonstrated good-to-excellent diagnostic validity for both Pfirrmann and Schizas classification across all five lumbar levels (L1–L2 to L5–S1), with sensitivity ranging from 80.85% to 96.30% and specificity from 80.17% to 95.24% for Pfirrmann, and sensitivity 82.76%–94.74% and specificity 90.15%–96.79% for Schizas. AUC-ROC values indicated good accuracy for Pfirrmann (0.815–0.890) and good-to-excellent accuracy for Schizas (0.880–0.929). Reliability was acceptable for both classifications (Cronbach's Alpha: Pfirrmann 0.792, Schizas 0.684). PPV was relatively lower across levels, likely reflecting class imbalance toward mild-to-moderate grades in the study cohort. Deep learning models have the potential to improve the diagnosis of LDD, enhance early intervention, and improve patient outcomes.

Keywords: Artificial intelligence, deep learning, lumbar degenerative disease, Pfirrmann classification, schizophrenia classification.

INTRODUCTION

Degenerative diseases of the lumbar spine are a leading cause of chronic low back pain, activity limitations, and reduced quality of

life, with a significant impact on disability globally. Their clinical impact varies from axial pain to functional impairment, necessitating comprehensive assessment through magnetic resonance imaging (MRI), which excels in evaluating lumbar spine structures. However, MRI assessment is often subject to subjectivity, which can result in variability in interpretation between radiologists, potentially impacting treatment decisions. Therefore, artificial intelligence (AI) with deep learning models offers a solution to improve consistency, effectiveness, and accuracy in MRI image analysis, reduce inter-examiner variation, and enhance standardization and accuracy in the evaluation of degenerative diseases of the lumbar spine.[1]

Degenerative spinal diseases and their structures can be effectively evaluated using magnetic resonance imaging, particularly T2-weighted sequences, which are instrumental in depicting discs, dural sacs, cerebrospinal fluid, and neural structures more clearly than other modalities.[2, 3] This anatomical clarity and signal information drive the need for a standardized classification system to consistently communicate the severity of degenerative changes between clinicians and between facilities. Standardization is necessary because treatment decisions, both conservative and operative, are often influenced by the severity of degeneration and stenosis reported on imaging. This need for standardization underlies the use of the Pfirrmann classification for disc degeneration and the Schizas classification for lumbar canal stenosis.[4, 5]

The use of the Pfirrmann and Schizas classifications is expected to help standardize the perception of disease severity in daily clinical practice. The Pfirrmann classification assesses disc degeneration based on signal intensity, structure, and disc height on magnetic resonance imaging, and several studies have reported a significant correlation between the degree of disc degeneration and clinical outcomes such as the Oswestry Disability Index (ODI) and the Visual Analog Scale (VAS) for pain.[4, 6] The Schizas classification categorizes lumbar canal stenosis based on the morphological appearance of the dural sac on axial T2-weighted sequences, taking into account the distribution of cerebrospinal fluid and nerve fibers. This makes it practical because it does not require special measurement tools.[5] Although helpful, the application of these two classifications still requires visual interpretation by the examiner, so assessment results can vary and are potentially influenced by subjectivity.[7]

The subjectivity of this assessment is a critical issue because low back pain is multifactorial, and imaging findings do not always align with the severity of symptoms. Several studies have shown that even individuals without symptoms can have degenerative findings on magnetic resonance imaging, so the radiological and clinical correlations are not always linear.[8, 9] This situation has given rise to debate regarding the use of magnetic resonance imaging as a sole predictor, particularly in lumbar canal stenosis, due to the risk of bias in both image interpretation and clinical conclusions.[7] These limitations, such as inter-examiner variation and potential radiological and clinical discrepancies, emphasize the need for a more objective, consistent, and replicable approach to assessing the degree of disc degeneration and lumbar canal stenosis.

The need for a more objective and consistent approach aligns with advances in artificial intelligence (AI), particularly deep learning, which shows great potential in medical image analysis to recognize complex patterns that may be missed by human observers.[1] The use of T2-weighted magnetic resonance imaging sequences in deep learning models has the potential to improve standardization in the assessment of degeneration and stenosis severity, while supporting early detection before severe clinical manifestations appear, given that current diagnoses often rely on advanced symptomatic presentations that can hinder proactive intervention.[1, 10]

Therefore, scientific evidence is needed to demonstrate that the developed models are truly accurate and consistent with expert assessment references. Therefore, this study aims to assess the validity and reliability of AI-based deep learning models in determining Pfirrmann and Schizas classification grades in patients with lumbar degenerative disease (LDD).

MATERIALS & METHODS

Research Design

This is a diagnostic study with validity and reliability analysis. It aims to evaluate the performance of a deep learning-based artificial intelligence model in classifying the degree of intervertebral disc degeneration (Pfirman criteria) and the degree of spinal canal stenosis (Schizas criteria) in patients with Lumbar Degenerative Disease (LDD), using T2-weighted MRI images as the primary modality and specialist radiologist expertise as the gold standard. The study was conducted at Ngoerah Hospital, Denpasar, using secondary data from medical records from January 2018 to November 2025.

Study Population

The study population consisted of all MRI images of patients undergoing radiology examinations at Ngoerah Hospital, with a purposive sampling of patients who met eligibility criteria. The sample size was calculated using a prevalence-based diagnostic formula,[11] with an expected sensitivity of 90%, precision of 10%, type I error rate of 5% ($Z\alpha = 1.96$), and a prevalence of LDD of 37%,[12] resulting in a minimum sample size of 93 cases.

Eligibility Criteria

Inclusion criteria included: a confirmed diagnosis of LDD based on clinical and MRI findings, age ≥ 20 years, willingness to participate in the study, and complete medical records and MRI results. Subjects were excluded if they had a history of spinal trauma or fracture, had received previous spinal treatment or surgery, had comorbidities that could affect MRI findings (autoimmune, malignant, or infectious), had de novo scoliosis, or had metastatic bone disease of the spine.

Research Variables

The gold standard variable was the expertise of a radiologist, while the variables studied were the classification results of a deep learning model. There were two dependent variables: the Pfirman classification based on sagittal T2W MRI (dichotomously categorized as mild-moderate (grades 1–3), and severe (grades 4–5), and the Schizas classification based on axial T2W MRI (categorized as minor-moderate (Schizas A–B), and severe-extreme (Schizas C–D). Control variables included age, gender, systemic disease, and MRI image quality with a minimum resolution of 320 × 320 pixels (Nagaraj et al., 2024).

Data Collection Procedure

Lumbar spine T2W MRI images were acquired using a Siemens Magnetom SKYRA 3 Tesla MRI machine in the radiology department of Ngoerah Hospital. Each image underwent preprocessing (normalization and noise reduction), followed by YOLO-based automatic segmentation to generate regions of interest per disc level (sagittal sections) and spinal cord area (axial sections). The cropped ROIs were then classified by the Pfirmann classifier and Schizas classifier models trained on expert-labeled data. In parallel, all images were independently read by a radiologist as the gold standard for comparison purposes.

Data Analysis

Data were analyzed in three stages: (1) descriptive analysis to describe sample characteristics; (2) diagnostic validity testing using the Receiver Operating Characteristic (ROC) method with Area Under the Curve (AUC) as a measure of overall accuracy, along with calculations of sensitivity, specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV); and (3) internal consistency

reliability testing using the Cronbach's Alpha coefficient. All analyses were performed using MedCalc software version 14.

RESULT

Sample Characteristics

A total of 170 subjects were included in this study. The mean age of participants was 55.20 ± 13.34 years. The age distribution was as follows: 20–30 years, 9 subjects (5.3%); 31–40 years, 16 subjects (9.4%); 41–50 years, 27 subjects (15.9%); 51–60 years, 55 subjects (32.4%); and above 60 years, 63 subjects (37.1%). Regarding sex distribution, 83 subjects (48.8%) were male, and 87 subjects (51.2%) were female. Full characteristics are presented in Table 1.

Table 1. Sample Characteristics

Characteristic	n = 170 (%)
Age (years)	55.20 ± 13.34
20–30 years	9 (5.3%)
31–40 years	16 (9.4%)
41–50 years	27 (15.9%)
51–60 years	55 (32.4%)
> 60 years	63 (37.1%)
Sex	
Male	83 (48.8%)
Female	87 (51.2%)

Diagnostic Validity of the Deep Learning Model for Pfirrmann Classification

The diagnostic validity of the deep learning model was assessed against the radiologist's interpretation as the gold standard across all five lumbar disc levels (L1–L2 through L5–S1), using sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio LR (+), and negative likelihood ratio LR (−). At all levels, the model demonstrated good validity, with sensitivity ranging from 80.85% to 96.30% and specificity from 80.17% to 95.24%.

At the L1–L2 level, 54 subjects were classified as Pfirrmann grade 4–5 (severe disc degeneration) and 116 as grade 1–3 (mild-moderate) by the radiologist. The deep learning model correctly identified 48 true positives and 101 true negatives, yielding a sensitivity of 88.89%, a specificity of 87.07%, a PPV of 76.19%, an NPV of 94.39%, an LR (+) of 6.87, and an LR (−) of 0.13. At the L2–L3 level, 42 subjects were classified as grade 4–5 and 128 as grade 1–3. The model identified 40 true positives and 106 true negatives, yielding a sensitivity of 82.81%, specificity of 95.24%, PPV of 64.52%, NPV of 98.15%, an LR+ of 5.54, and LR (−) of 0.06. At the L3–L4 level, 47 subjects were classified as grade 4–5 and 123 as grade 1–3. The model identified 38 true positives and 101 true negatives, yielding a sensitivity of 80.85%, specificity of 82.11%, PPV of 63.33%, NPV of 91.82%, LR (+) of 4.52, and LR (−) of 0.23. At the L4–L5 level, 54 subjects were classified as grade 4–5 and 116 as grade 1–3. The model identified 52 true positives and 93 true negatives, yielding a sensitivity of 96.30%, a specificity of 80.17%, a PPV of 69.33%, an NPV of 97.89%, an LR (+) of 4.86, and an LR (−) of 0.05. At the L5–S1 level, 59 subjects were classified as grade 4–5 and 111 as grade 1–3. The model identified 52 true positives and 94 true negatives, yielding a sensitivity of 88.14%, a specificity of 84.68%, a PPV of 75.36%, an NPV of 93.07%, an LR (+) of 5.75, and an LR (−) of 0.14 (Table 2).

Table 2. Diagnostic Validity of the Deep Learning Model for Pfirrmann Classification

Variable		Radiology		Se	Sp	PPV	NPV	LR(+)	LR(−)
		Grade 4–5	Grade 1–3						
Diagnostic Validity of the Deep Learning Model for Pfirrmann Classification at L1–L2									
AI	Grade 4–5	48 (76.2%)	15 (23.8%)	88.89%	87.07%	76.19%	94.39%	6.87	0.13
	Grade 1–3	6 (5.6%)	101 (94.4%)						
Diagnostic Validity of the Deep Learning Model for Pfirrmann Classification at L2–L3									
AI	Grade 4–5	40 (64.5%)	22 (35.5%)	82.81%	95.24%	64.52%	98.15%	5.54	0.06
	Grade 1–3	2 (1.9%)	106 (98.1%)						
Diagnostic Validity of the Deep Learning Model for Pfirrmann Classification at L3–L4									

AI	Grade 4–5	38 (63.3%)	22 (36.7%)	80.85%	82.11%	63.33%	91.82%	4.52	0.23
	Grade 1–3	9 (8.2%)	101 (91.8%)						
Diagnostic Validity of the Deep Learning Model for Pfirrmann Classification at L4–L5									
AI	Grade 4–5	52 (69.3%)	23 (30.7%)	96.30%	80.17%	69.33%	97.89%	4.86	0.05
	Grade 1–3	2 (2.1%)	93 (97.9%)						
Diagnostic Validity of the Deep Learning Model for Pfirrmann Classification at L5–S1									
AI	Grade 4–5	52 (75.4%)	17 (24.6%)	88.14%	84.68%	75.36%	93.07%	5.75	0.14
	Grade 1–3	7 (6.9%)	94 (93.1%)						

*Se = sensitivity; Sp = specificity; PPV = positive predictive value; NPV = negative predictive value; LR (+) = positive likelihood ratio; LR (−) = negative likelihood ratio

Diagnostic Validity of the Deep Learning Model for Schizas Classification

The diagnostic validity of the deep learning model for Schizas classification was similarly evaluated across all five lumbar levels (L1–L2 through L5–S1). The model distinguished severe-extreme stenosis (Schizas C and D) from minor-moderate stenosis (Schizas A and B). Across all levels, sensitivity ranged from 82.76% to 94.74% and specificity from 90.15% to 96.79%, indicating good to excellent diagnostic performance.

At L1–L2, 11 subjects were classified as Schizas C or D, and 159 as A or B. The model identified 10 true positives and 151 true negatives, yielding a sensitivity of 90.91%, specificity of 94.97%, PPV of 55.56%, NPV of 99.34%, LR (+) of 18.07, and LR (−) of 0.10. At L2–L3, 15 subjects were classified as Schizas C or D, and 156 as A or B. The model identified 13 true positives and 151 true negatives, yielding a sensitivity of 86.67%, specificity of 96.79%, PPV of 72.22%, NPV of 98.69%, LR (+) of 27.04, and an LR (−) of 0.14. At L3–L4, 20 subjects were classified as Schizas C or D, and 150 as A or B. The model identified 18 true positives and 143 true negatives, yielding a sensitivity of 90.00%, a specificity of 95.33%, a PPV of 72.00%, an NPV of 98.62%, LR (+) of 19.29, and LR (−) of 0.10. At L4–L5, 38 subjects were classified as Schizas C or D, and 132 as A or B. The model identified 36 true positives and 119 true negatives, yielding a sensitivity of 94.74%, specificity of 90.15%, PPV of 73.47%, NPV of 98.35%, LR (+) of 9.61, and LR (−) of 0.06. At L5–S1, 29 subjects were classified as Schizas C or D, and 141 as A or B. The model identified 24 true positives and 132 true negatives, yielding a sensitivity of 82.76%, a specificity of 93.69%, a PPV of 72.73%, an NPV of 96.35%, an LR (+) of 13.12, and an LR (−) of 0.18 (Table 3).

Table 3. Diagnostic Validity of the Deep Learning Model for Schizas Classification

Variable		**Radiology**		**Se**	**Sp**	**PPV**	**NPV**	**LR(+)**	**LR(−)**
		C & D	**A & B**						
Diagnostic Validity of the Deep Learning Model for Schizas Classification at L1–L2									
AI	C & D	10 (55.6%)	8 (44.4%)	90.91%	94.97%	55.56%	99.34%	18.07	0.10
	A & B	1 (0.7%)	151 (99.3%)						
Diagnostic Validity of the Deep Learning Model for Schizas Classification at L2–L3									
AI	C & D	13 (76.5%)	5 (23.5%)	86.67%	96.79%	72.22%	98.69%	27.04	0.14
	A & B	2 (1.3%)	151 (98.7%)						
Diagnostic Validity of the Deep Learning Model for Schizas Classification at L3–L4									
AI	C & D	18 (72.0%)	7 (28.0%)	90.00%	95.33%	72.00%	98.62%	19.29	0.10
	A & B	2 (1.4%)	143 (98.6%)						
Diagnostic Validity of the Deep Learning Model for Schizas Classification at L4–L5									
AI	C & D	36 (73.5%)	13 (26.5%)	94.74%	90.15%	73.47%	98.35%	9.61	0.06
	A & B	2 (1.7%)	119 (98.3%)						
Diagnostic Validity of the Deep Learning Model for Schizas Classification at L5–S1									
AI	C & D	24 (72.7%)	9 (27.3%)	82.76%	93.69%	72.73%	96.35%	13.12	0.18
	A & B	5 (3.6%)	132 (96.4%)						

*Se = sensitivity; Sp = specificity; PPV = positive predictive value; NPV = negative predictive value; LR(+) = positive likelihood ratio; LR(−) = negative likelihood ratio

Reliability and Accuracy of the Deep Learning Model

Internal consistency reliability of the deep learning model for Pfirrmann classification was assessed using Cronbach's Alpha, yielding a value of 0.792, indicating acceptable reliability (Table 4). Diagnostic accuracy was further evaluated using the area under the receiver operating characteristic curve (AUC-ROC) at each lumbar level. The model demonstrated good accuracy across all levels, with AUC values ranging from 0.815 to 0.890 (all $p < 0.001$), as summarized in Table 5. Internal consistency reliability of the deep learning model for Schizas classification yielded a Cronbach's Alpha of 0.684, indicating acceptable reliability (Table 6). AUC-ROC analysis demonstrated excellent accuracy at most levels, with AUC values ranging from 0.880 to 0.929 (all $p < 0.001$). Notably, the model achieved excellent accuracy (AUC ≥ 0.900) at levels L1–L2, L2–L3, L3–L4, and L5–S1, while good accuracy (AUC = 0.880) was observed at L4–L5 (Table 5).

Table 4. Internal Consistency Reliability

Classification	Cronbach's Alpha	Number of Items
Pfirrmann Classification	0.792	10
Schizas Classification	0.684	10

Table 5. AUC-ROC Values of the Deep Learning Model

Level	AUC	Std. Error	p-value	95% CI
Pfirrmann Classification				
L1–L2	0.880	0.031	< 0.001	0.820 – 0.940
L2–L3	0.890	0.028	< 0.001	0.836 – 0.945
L3–L4	0.815	0.039	< 0.001	0.739 – 0.891
L4–L5	0.882	0.027	< 0.001	0.829 – 0.936
L5–S1	0.864	0.031	< 0.001	0.802 – 0.926
Schizas Classification				
L1–L2	0.929	0.051	< 0.001	0.829 – 1.030
L2–L3	0.920	0.052	< 0.001	0.819 – 1.022
L3–L4	0.927	0.040	< 0.001	0.848 – 1.006
L4–L5	0.880	0.040	< 0.001	0.800 – 0.959
L5–S1	0.914	0.059	< 0.001	0.797 – 1.030

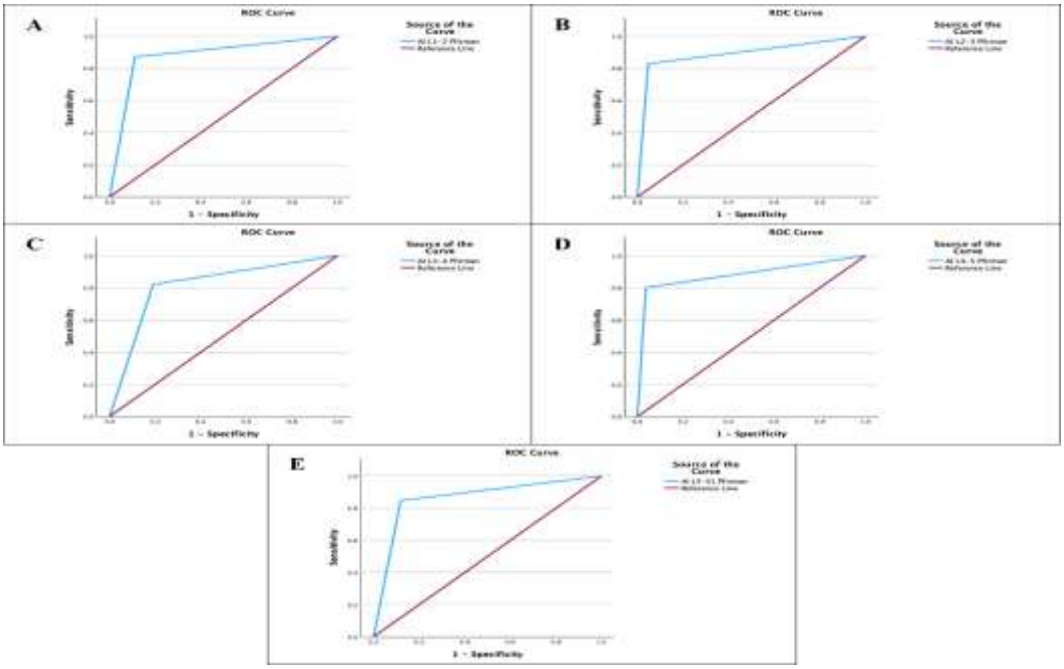

Figure 1. ROC Curve of Deep Learning Model as Predictor of Pfirmann Classification. (A) Lumbar Level L1-L2, (B) Lumbar Level L2-L3, (C) Lumbar Level L3-L4, (D) Lumbar Level L4-L5, (E) Lumbar Level L5-S1

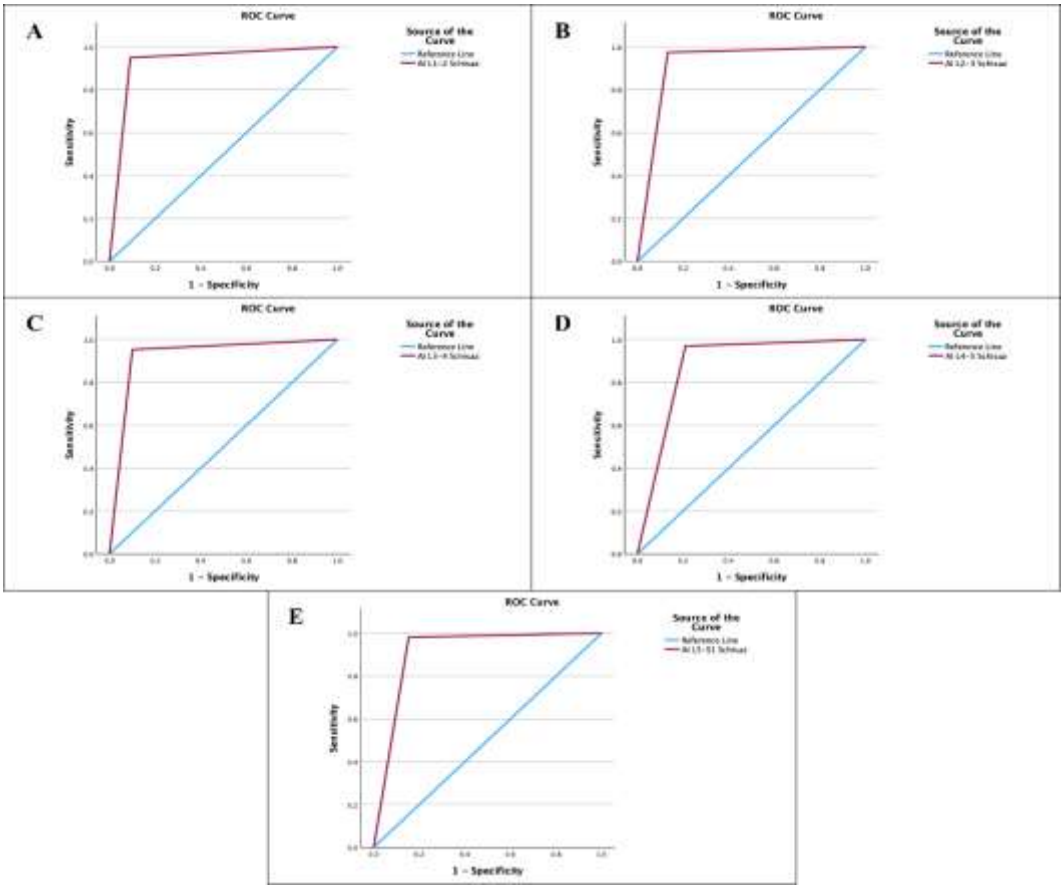

Figure 2. ROC Curve of Deep Learning Model as Predictor of Schizosis Classification. (A) Lumbar Level L1-L2, (B) Lumbar Level L2-L3, (C) Lumbar Level L3-L4, (D) Lumbar Level L4-L5, (E) Lumbar Level L5-S1

DISCUSSION

This study demonstrated that a deep learning model combining a convolutional neural network (CNN) architecture and You Only Look Once (YOLO)-based object detection is a valid and reliable instrument for automatically classifying the degree of intervertebral disc degeneration according to the Pfirrmann grading system and the degree of lumbar spinal canal stenosis according to the Schizas classification in patients with Lumbar Degenerative Disease (LDD). Evaluated against the radiologist's interpretation as the gold standard, the model exhibited consistently good to excellent diagnostic performance across all five lumbar levels (L1-L2 through L5-S1) for both classification tasks. Reliability assessment confirmed acceptable internal consistency for both grading systems, and AUC-ROC analysis further demonstrated good to excellent overall diagnostic accuracy at every level, supporting the potential clinical utility of the model as an adjunct tool in radiological assessment of LDD.

The deep learning model demonstrated strong diagnostic validity for Pfirrmann classification, with the highest sensitivity observed at the L4-L5 level. This finding aligns with clinical expectations, as L4-L5 is a biomechanically loaded segment disproportionately susceptible to advanced disc degeneration, making it an area where high model sensitivity is particularly valuable. These results are consistent with Liawrungrueang et al., who reported a sensitivity of 100% for detecting Grade II intervertebral disc degeneration using a CNN-based model,[13] and with Wang et al., who reported a specificity of approximately 87% using a YOLO-based deep learning framework.[14] The highest specificity was recorded at the L2-L3 level, reflecting the model's capacity to accurately exclude mild-to-moderate degeneration

where imaging characteristics are relatively more distinct. The highest NPV at L2-L3 indicated near-certain exclusion of severe disc degeneration when the model yielded a negative result, directly mirroring values reported by Liawrungrueang et al. LR(+) and LR(-) values further reinforced the model's clinical discriminatory power: the highest LR(+) at L1-L2 and the lowest LR(-) at L4-L5 are particularly noteworthy, as a very low LR(-) at the level most prone to degeneration effectively rules out severe disease.[13] This finding is consistent with Bharadwaj et al., who reported a similarly low LR(-) in AI-based stenosis detection.[15] Minor discrepancies in performance metrics across studies are attributable to differences in imaging protocols, dataset composition, model architecture, and training sample size.

For Schizas classification, the model similarly achieved good-to-excellent diagnostic validity across all lumbar levels, with the highest sensitivity at L4-L5. This is anatomically consistent, as L4-L5 is among the most frequently stenosed segments, and the greater prevalence of severe stenosis at this level likely contributes to the model's superior sensitivity there. These results compare favorably with Chen et al., who reported a sensitivity of 88% for Schizas-based deep learning classification of severe stenosis, suggesting that the current model offers a marginal improvement, possibly attributable to dataset diversity and broader representation of stenotic patterns.[2] Specificity was highest at L2-L3, consistent with the more distinctive dural sac morphology at this level. NPV was highest at L1-L2, where the low prevalence of severe stenosis allows the model to confidently exclude significant canal compromise. The LR(+) was highest at L2-L3, indicating that a positive model output at this level is a particularly strong confirmatory signal for severe stenosis, in line with values reported by Chen et al., while the lowest LR(-) at L4-L5,[2] corroborates Bharadwaj et al. and underscores the model's utility in safely ruling out severe stenosis at the most clinically critical lumbar segment.[15] The relatively lower PPV values across Schizas levels reflect the inherent class imbalance. Severe stenosis (Schizas C/D) is less prevalent than minor-to-moderate stenosis in this cohort, a limitation expected to improve with larger, more balanced training datasets.

The reliability of the deep learning model was assessed through internal consistency analysis using Cronbach's Alpha. The value for Pfirrmann classification (alpha = 0.792) exceeded the commonly accepted threshold of 0.70, while that for Schizas classification (alpha = 0.684) approached this threshold, together indicating that the model produces consistent outputs across repeated measurements. These findings are consistent with the broader literature demonstrating that well-trained deep learning models can achieve radiologist-level consistency in structured grading tasks.[16, 17] AUC-ROC analysis confirmed good diagnostic accuracy for Pfirrmann grading across all five levels, and excellent accuracy for Schizas grading at four of five levels, which is comparable to the AUC range of 0.80-0.90 reported by Ghauri et al.,[18] and values documented by Chen et al. for T2-weighted MRI-based deep learning classification.[2] The comparatively lower AUC at L3-L4 for Pfirrmann grading may reflect greater heterogeneity of degenerative changes at this transitional spinal segment, a challenge also acknowledged in prior studies. Nonetheless, the consistently high AUC values across both classification systems affirm that the model's diagnostic accuracy is robust and clinically meaningful.

From an architectural standpoint, the YOLO-based segmentation and classification pipeline employed in this study offers a practical advantage in terms of real-time processing speed and ease of clinical deployment, consistent with findings reported by Liawrungrueang et al.[13] This contrasts with more complex multi-task architectures such as that

proposed by Bharadwaj et al., which integrates V-Net segmentation and Big Transfer (BiT) for comprehensive detection of central canal stenosis, foraminal stenosis, and facet arthropathy (AUROC 0.92-0.93), but at the cost of greater computational demand and data requirements.[15] The two approaches thus represent complementary points on the speed-versus-comprehensiveness spectrum, with the model used in this study being more suited to rapid, level-specific grading in routine clinical workflow. Regarding interpretability, the current model, like most deep learning systems, functions as a black box, which limits the transparency of its decision-making process, a recognized challenge in medical AI adoption. The future integration of gradient-weighted class activation mapping (Grad-CAM) or saliency map visualization would allow clinicians to identify which regions of the MRI drive the model's classification decisions, thereby increasing clinical trust and enabling more meaningful human-AI collaboration.[16, 17]

The sample characteristics of this study, predominantly patients older than 51 years with a near-equal sex distribution, are epidemiologically consistent with the known age-related trajectory of LDD and the absence of a strong sex predilection, as reported in the literature, further supporting the representativeness of the study population.[13, 17]

This study has several strengths. It is among the first to simultaneously validate a deep learning model for both Pfirrmann and Schizas grading across all five lumbar disc levels in a single cohort, providing a comprehensive performance profile directly applicable to clinical practice. The use of a 3-Tesla MRI system with standardized image quality criteria and a two-stage deep learning pipeline, YOLO-based segmentation followed by level-specific classification, contributes to methodological rigor, and the sample size exceeded the minimum calculated threshold while encompassing a broad age range and balanced sex distribution. Nevertheless, several limitations must be acknowledged. First, this was a single-center study, which may limit generalizability to institutions with different MRI systems, imaging protocols, or patient demographics. Second, the class imbalance between severe and mild-to-moderate grades, particularly for Schias' classification, likely constrained PPV, and future studies should employ larger, class-balanced training datasets to address this. Third, the Cronbach's Alpha for Schizas classification, while approaching acceptability, was below 0.70, indicating that further model refinement and larger validation cohorts are warranted. Finally, the absence of Grad-CAM or equivalent visualization in the current implementation limits interpretability; incorporating such tools in future work will be essential for safe and transparent clinical integration. Prospective multicenter validation studies with diverse patient populations are recommended to confirm and extend these findings.

CONCLUSION

The deep learning model proved valid and reliable in determining Pfirrmann and Schizas grade in patients with Lumbar Degenerative Disease, thus potentially assisting with objective and consistent radiological evaluation. Further research is recommended to expand the dataset variation, address data imbalance, refine model training by involving medical teams, and integrate additional clinical and imaging data to enable the model to support more optimal diagnosis and treatment planning.

Declaration by Authors

Ethical Approval: Approved

Acknowledgement: We express our deepest gratitude to all parties for their support of this study.

Source of Funding: All funding was provided by the authors without external sources.

Conflict of Interest: No conflicts of interest declared.

REFERENCES

1. Yi W, Zhao J, Tang W, et al. Deep learning-based high-accuracy detection for lumbar and cervical degenerative disease on T2-weighted MR images. Eur spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc 2023; 32: 3807–3814.
2. Chen Z, Lei F, Ye F, et al. MRI-based vertebral bone quality score for the assessment of osteoporosis in patients undergoing surgery for lumbar degenerative diseases. J Orthop Surg Res 2023; 18: 257.
3. El-karamany M, Maghawry A, Bakr A. Management of Degenerative Lumbar Spine Disease by Posterior Lumbar Interbody Fusion and Percutaneous Pedicular Fixation. Benha Med J. Epub ahead of print 29 May 2023. DOI: 10.21608/bmfj.2023.198090.1774.
4. Pfirrmann CW, Metzdorf A, Zanetti M, et al. Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine (Phila Pa 1976) 2001; 26: 1873–1878.
5. Schizas C, Theumann N, Burn A, et al. Qualitative grading of severity of lumbar spinal stenosis based on the morphology of the dural sac on magnetic resonance images. Spine (Phila Pa 1976) 2010; 35: 1919–1924.
6. El-Hady A, Molla S, Elwan S, et al. Evaluation of health related quality of life with the use of Oswestry disability index in degenerative discogenic low back pain. Egypt Rheumatol Rehabil; 50. Epub ahead of print 17 January 2023. DOI: 10.1186/s43166-022-00166-6.
7. Joern L, Kongsted A, Thomassen L, et al. Pain cognitions and impact of low back pain after participation in a self-management program: a qualitative study. Chiropr Man Therap 2022; 30: 8.
8. Jensen MC, Brant-Zawadzki MN, Obuchowski N, et al. Magnetic resonance imaging of the lumbar spine in people without back pain. N Engl J Med 1994; 331: 69–73.
9. Brinjikji W, Luetmer PH, Comstock B, et al. Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. AJNR Am J Neuroradiol 2015; 36: 811–816.
10. Hilton B, Gardner EL, Jiang Z, et al. Establishing Diagnostic Criteria for Degenerative Cervical Myelopathy [AO Spine RECODE-DCM Research Priority Number 3]. Glob spine J 2022; 12: 55S-63S.
11. Dahlan MS. Besar Sampel Dalam Penelitian Kedokteran Dan Kesehatan.No Title. 4th edn. Jakarta: Epidemiologi Indonesia, 2016.
12. Ravindra VM, Senglaub SS, Rattani A, et al. Degenerative Lumbar Spine Disease: Estimating Global Incidence and Worldwide Volume. Glob spine J 2018; 8: 784–794.
13. Liawrungrueang W, Cholamjiak W, Sarasombath P, et al. Artificial Intelligence Classification for Detecting and Grading Lumbar Intervertebral Disc Degeneration. Spine Surg Relat Res 2024; 8: 552–559.
14. Wang A, Wang T, Liu X, et al. Automated diagnosis and grading of lumbar intervertebral disc degeneration based on a modified YOLO framework. Front Bioeng Biotechnol 2025; 13: 1526478.
15. Bharadwaj UU, Christine M, Li S, et al. Deep learning for automated, interpretable classification of lumbar spinal stenosis and facet arthropathy from axial MRI. Eur Radiol 2023; 33: 3435–3443.
16. Won D, Lee H-J, Lee S-J, et al. Spinal Stenosis Grading in Magnetic Resonance Imaging Using Deep Convolutional Neural Networks. Spine (Phila Pa 1976) 2020; 45: 804–812.
17. Lehnen NC, Haase R, Faber J, et al. Detection of Degenerative Changes on MR Images of the Lumbar Spine with a Convolutional Neural Network: A Feasibility Study. Diagnostics (Basel, Switzerland); 11. Epub ahead of print May 2021. DOI: 10.3390/diagnostics11050902.
18. Ghauri MS, Reddy AJ, Tak N, et al. Utilizing Deep Learning for X-ray Imaging: Detecting and Classifying Degenerative Spinal Conditions. Cureus 2023; 15: e41582.

How to cite this article: Ivan Alexander Liando, I Wayan Suryanto Dusak, I Gusti Lanang Ngurah Agung Artha Wiguna, I Ketut Suyasa, Elysanti Dwi Martadiani, Made Bramantya Karna et al. Validity and reliability of the Pfirrmann and Schizas criteria degrees in lumbar degenerative disease patients using the deep learning method. *International Journal of Research and Review*. 2026; 13(6): 611-620. DOI: *https://doi.org/10.52403/ijrr.20260659*

******