Artificial Intelligence Classification for Detecting and Grading Lumbar Intervertebral Disc Degeneration

Wongthawat Liawrungrueang[1)], Watcharaporn Cholamjiak[2)], Peem Sarasombath[3)], Khanathip Jitpakdee[4)] and Vit Kotheeranurak[5)6)]

1) Department of Orthopaedics, School of Medicine, University of Phayao, Phayao, Thailand
2) Department of Mathematics, School of Science, University of Phayao, Phayao, Thailand
3) Department of Orthopaedics, Phramongkutklao Hospital and College of Medicine, Bangkok, Thailand
4) Department of Orthopedics, Queen Savang Vadhana Memorial Hospital, Chonburi, Thailand
5) Department of Orthopaedics, Faculty of Medicine, Chulalongkorn University, and King Chulalongkorn Memorial Hospital, Bangkok, Thailand
6) Center of Excellence in Biomechanics and Innovative Spine Surgery, Chulalongkorn University, Bangkok, Thailand

Abstract:

Introduction: Intervertebral disc degeneration (IDD) is a primary cause of chronic back pain and disability, highlighting the need for precise detection and grading for effective treatment. This study focuses on developing and validating a convolutional neural network (CNN) with a You Only Look Once (YOLO) architecture model using the Pfirrmann grading system to classify and grade lumbar intervertebral disc degeneration based on magnetic resonance imaging (MRI) scans.

Methods: We developed a deep learning model trained on a dataset of anonymized MRI studies of patients with symptomatic back pain. MRI images were segmented and annotated by radiologists according to the Pfirrmann grading for the datasets. The segmentation MRI-disc image dataset was prepared for three groups: a training set (1,000), a testing set (500), and an external validation set (500) to assess model generalizability without overlapping images. The model's performance was evaluated using accuracy, sensitivity, specificity, F1 score, prediction error, and ROC-AUC.

Results: The AI model showed high performance across all metrics. For Grade I IDD, the model achieved an accuracy of 97%, 95%, and 92% in the training, testing, and external validation sets, respectively. For Grade II, the sensitivity was 100% in both training and testing sets and 98% in the validation set. For Grade III, the specificity was 95.4% in the training set and 94% in both testing and validation sets. For Grade IV, the F1 score was 97.77% in the training set and 95% in both testing and validation sets. For Grade V, the prediction error was 2.3%, 2%, and 2.5% in the training, testing, and validation sets, respectively. The overall ROC-AUC was 97%, 92%, and 95% in the training, testing, and validation sets, respectively.

Conclusions: The AI-based classification model exhibits high accuracy, sensitivity, and specificity in detecting and grading lumbar IDD using the Pfirrmann grading. AI has significantly enhanced diagnostic precision and reliability, providing a powerful tool for clinicians in managing IDD. The potential impact is substantial, although further clinical validation is necessary before integrating this model into routine practice.

Keywords:

Artificial Intelligence, Machine Learning, Intervertebral Disc Degeneration, Lumbar Spine, MRI, Pfirrmann Grading

Spine Surg Relat Res 2024; 8(6): 552-559
dx.doi.org/10.22603/ssrr.2024-0154

Introduction

Degenerative disc disease (DDD) is a significant burden on healthcare systems globally, accounting for a substantial portion of patient visits to medical practitioners[1)]. With an estimated prevalence impacting up to 80% of individuals, DDD stands out as a pervasive condition intricately linked to the degenerative process of intervertebral discs[2,3)]. The

Corresponding author: Wongthawat Liawrungrueang, mint11871@hotmail.com
Received: May 27, 2024, Accepted: June 15, 2024, Advance Publication: August 6, 2024

progressive degradation of these discs underscores the pressing need for accurate diagnosis, which is a cornerstone of effective disease management[4]. Symptomatic intervertebral disc diseases, often manifested through chronic back pain and functional limitations, underscore the critical importance of precise diagnostic methods[2,5]. Timely and accurate identification of DDD facilitates tailored treatment strategies, thus optimizing patient outcomes and minimizing the disease's socioeconomic impact.

The Pfirrmann grading system[3] is a widely used method for classifying and grading lumbar intervertebral disc degeneration based on magnetic resonance imaging (MRI) findings[3]. It categorizes discs into five grades (I-V) based on the appearance of the nucleus pulposus and annulus fibrosus and the distinction between them. Grade I represents a healthy disc with a distinct nucleus pulposus and annulus fibrosus, while Grade V indicates severe degeneration with collapse of the disc space and loss of distinction between the nucleus and annulus. Grades II to IV denote progressively increasing levels of degeneration, characterized by changes in signal intensity, loss of disc height, and alterations in disc morphology[3]. The Pfirrmann grading system provides a standardized and reproducible means of assessing disc degeneration, facilitating clinical decision-making and treatment planning for patients with lumbar spine disorders[3]. Considering the high prevalence and clinical significance of DDD, there is a compelling demand for advanced diagnostic approaches that can reliably detect and characterize the extent of disc degeneration. This underscores the importance of exploring innovative technologies, such as artificial intelligence (AI), which hold promise for augmenting diagnostic accuracy and enhancing patient care[6].

By leveraging cutting-edge AI techniques and comprehensive datasets, we aim to develop and validate a novel framework for precisely identifying and grading degenerative disc disease. Through rigorous evaluation against established diagnostic standards, our research endeavors to contribute to the advancement of diagnostic capabilities in the management of DDD, ultimately improving patient outcomes and quality of life. This study aims to address this pressing need by investigating the potential of AI-based approaches to diagnose DDD accurately.

Materials and Methods

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee and Institutional Review Board (Institutional Review Board (IRB) approval, IRB number: HREC-UP-HSST 1.1/033/67). Informed consent was not required because the dataset did not show the identity of the patient. We developed a computer-assisted diagnosis with a deep learning model to classify lumbar intervertebral disc degeneration using MRI scans. Deep learning model development uses a CNN with the YOLO (You Only Look Once) model used for object detection tasks[7]. YOLO was chosen over other architectures primarily due to its speed, single-stage detection approach, and end-to-end training capability. By directly predicting bounding boxes and class probabilities in a single pass of the network, YOLO achieves real-time performance, making it suitable for applications where speed is crucial. Its unified approach simplifies the detection pipeline while maintaining high localization accuracy and effectiveness even for small objects. In addition, YOLO's flexibility and adaptability have made it a popular choice for various computer vision tasks beyond standard object detection[8,9]. Unlike traditional object detection methods that require multiple region proposals and subsequent classification, YOLO divides the input image into a grid and simultaneously predicts bounding boxes and class probabilities for each grid cell[7]. We developed our model using Python programming (Python 3.6 and TensorFlow). The model architecture was optimized for the classification and grading of IDD based on T2-weighted MRI images. The CNN architecture underwent training on the training dataset using stochastic gradient descent optimization with backpropagation. Fig. 1 shows the detection flowchart for the evaluation of the model training dataset, testing set, and external validation of the deep learning model using a CNN with a YOLO architecture.

Data acquisition and preprocessing by sagittal T2-weighted MRI images of lumbar spines were obtained from an open-access dataset by Sudirman et al. The public dataset contains anonymized clinical MRI studies of patients with symptomatic back pain that do not show any patient's identity[10]. The dataset used for model training consisted of 515 lumbar spine MRI scan images meticulously annotated by expert radiologists according to the Pfirrmann grading system, a widely accepted classification scheme for disc degeneration. To ensure the integrity of the dataset, inclusion criteria were strictly adhered to, encompassing MRI scans from adult patients with symptomatic low back pain. Conversely, exclusion criteria excluded scans showing tumors, infections, inflammatory disorders, congenital diseases, or lumbar spine fractures. Regarding the quality of the MRI images and acquisition settings. The dataset includes high-resolution images primarily at 320×320 pixels with 12-bit per pixel precision, captured using MRI machines with 1.5 and 3 Tesla (T) field strengths, which are standard in clinical practice. Most of the scans were performed with patients in the head-first supine position, ensuring consistent image quality. Imaging parameters include a slice thickness of 4 mm, slice spacing of 4.4 mm, and pixel spacing of 0.6875 mm uniformly across all axial-view slices. Each study lasts between 15 and 45 min and includes at least the last seven vertebrae and the first two sacral links in sagittal view, ensuring comprehensive coverage. Representative images demonstrating the high clarity and detail essential for accurate Pfirrmann grading are included in the manuscript, showcasing the dataset's suitability for developing deep learning models for lumbar intervertebral disc degeneration detection and grading[10].

In our study, we first augmented the image to a number of 1,500. We randomly divided 1,500 images into two sets:

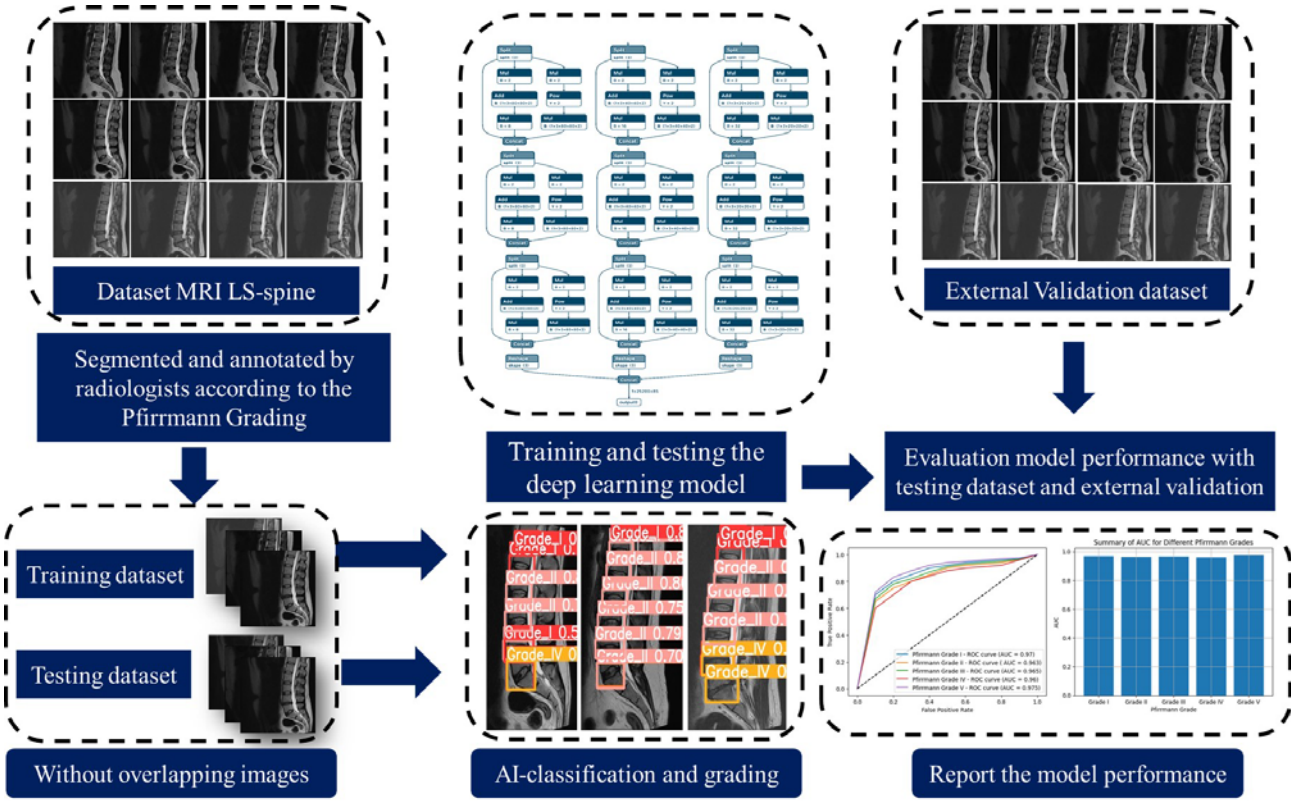

Figure 1. Schematic illustration and detection flowchart evaluation model training dataset, testing set, and external validation of the deep learning model.

a training set (1,000 scans) and a testing set (500 scans) using computer randomization without overlapping images. The external validation data (500 scans) was sourced from another available dataset[11]. The test dataset assessed the model's performance on unseen data post-training, while the external validation dataset confirmed the model's generalizability and robustness in a separate subset, ensuring its effectiveness in real-world clinical settings. Data augmentation was employed to prevent overfitting, using the Augmenters Python package. Augmentation techniques were horizontal flip, crop (zoom 0%-20%), rotation (−15° to 15°), and shear (±15° horizontal and ±15° vertical).

During training, the deep learning model adjusted its parameters iteratively through stochastic gradient descent optimization with backpropagation to minimize the disparity between predicted and ground-truth labels. Subsequently, the model's performance was rigorously assessed on the testing set using a range of standard metrics such as accuracy, sensitivity, specificity, F1 score, prediction error, and receiver operating characteristic area under the curve (ROC-AUC). Furthermore, the model's generalizability was evaluated on the external validation set. External validation by an additional external validation set (500 MRI images) was used to assess the model's generalizability beyond the training and testing datasets. Model performance was evaluated using standard metrics, such as accuracy, sensitivity, specificity, F1 score, prediction error, and receiver operating characteristic area under the curve (ROC-AUC). The study employed Python programming language and deep learning frameworks for model development and training, using computational resources such as personal computers equipped with appropriate hardware specifications to facilitate efficient model training and evaluation.

The study categorizes the number of degenerative disc levels into five Pfirrmann grades, which indicate varying degrees of disc degeneration from Grades I (the least severe) to V (most severe). The data is in three distinct groups: model, testing, and external validation datasets (Table 1). Each group contains a balanced number of instances per grade, with the model datasets having 200 instances per grade, totaling 1,000. The testing and external validation datasets each have 100 instances per grade, totaling 500 for each group. This consistent and equal distribution across all grades and datasets ensures comprehensive coverage and reliability when training, testing, and externally validating models. Such a balanced approach is essential to accurately assess model performance and generalizability, ensuring that each severity level of disc degeneration is adequately represented and evaluated.

Results

Table 2 shows the performance model of deep learning compared with the testing set and external validation.

This study (Table 2) provides a detailed comparison of a deep learning model's performance across three datasets: training, testing, and external validation sets. Accuracy, which represents the proportion of correct predictions out of

Table 1. The Number of Degenerative Disc Levels Is Based on Pfirrmann Grading Deviation for Three Groups: Model, Testing, and External Validation Datasets.

Pfirrmann grade	Model dataset	Testing dataset	External validation dataset
Grade I	200	100	100
Grade II	200	100	100
Grade III	200	100	100
Grade IV	200	100	100
Grade V	200	100	100
Total	1000	500	500

Table 2. The Performance Model of Deep Learning Compared with Testing Set and External Validation.

Metric	Training Set	Testing Set	External Validation
Accuracy	97%	95%	92%
Sensitivity	100%	100%	98%
Specificity	95.4%	94%	94%
F1 Score	97.77%	95%	95%
Prediction Error	2.3%	2%	2.5%
ROC curve (receiver operating characteristic curve)	97%	92%	95%

all predictions made by the model, is highest in the training set at 97% and slightly decreases to 95% in the testing set and further to 92% in the external validation set. This decline indicates a typical reduction in performance when the model encounters new data. Sensitivity, measuring the model's ability to identify positive cases correctly, remains perfect at 100% for training and testing sets but slightly decreases to 98% for the external validation set. This suggests that the model is highly effective at detecting positive instances across all datasets, with only a minor decrease in the external validation set. Specificity, which assesses the model's ability to identify negative cases correctly, is 95.4% for the training set and slightly lower but consistent at 94% for testing and external validation sets. This consistency suggests that the model reliably identifies negative cases across different datasets. The F1 score, which is the harmonic mean of precision and recall, is 97.77% for the training set and decreases to 95% for both testing and external validation sets. This metric indicates that the model maintains a balanced performance in terms of precision and recall, even when evaluated on new data. Prediction error, which indicates the proportion of incorrect predictions, is low across all datasets, increasing marginally from 2.3% in the training set to 2.5% in the external validation set. This reflects a minor increase in incorrect predictions when the model is applied to unseen data. The receiver operating characteristic (ROC) curve, measuring the model's ability to distinguish between classes, is 97% for the training set, decreases to 92% for the testing set, and improves to 95% for the external validation set. Fig. 2 shows the response ROC curve analysis of the deep learning model in the internal training dataset, testing dataset, and external validation.

The comprehensive performance of a deep learning model across various grades of intervertebral disc degeneration is categorized by the Pfirrmann grading system (Table 3). Each grade, from Grades I to V, is assessed based on several key metrics. Accuracy, representing the proportion of correct predictions, is consistently high across all grades, ranging from 95.7% to 97.2%. Sensitivity, or the model's ability to correctly identify true positives, slightly varies between grades but generally remains above 95%. Specificity, which measures the model's capacity to identify true negatives correctly, exhibits similarly high percentages, ranging from 95.4% to 97%. The F1 score, a harmonic mean of precision and recall, demonstrates robust performance across grades, hovering around 96%-97%. Prediction error, reflecting the proportion of incorrect predictions, remains relatively low for each grade, ranging from 1.5% to 2.3%. Lastly, the ROC curve, indicative of the model's overall discrimination ability, consistently shows strong performance, with areas under the curve ranging from 96% to 97.5%. These metrics collectively illustrate the deep learning model's effectiveness in accurately classifying intervertebral disc degeneration across different grades according to the Pfirrmann system that the author reports in the heatmap chart in Fig. 3, 4. Fig. 5 shows the ROC curve and barchart analysis of the deep learning model for detecting disc degenerative changes from Grades I to V.

Discussion

Artificial intelligence (AI) propels new discoveries in spinal therapies, facilitates large-scale data processing, and provides sophisticated simulation tools for surgeon training. Moreover, it enhances research and education[12-14].

This study demonstrates AI's potential for accurately detecting and grading lumbar IDD using the Pfirrmann grading system. The deep learning model, developed using a CNN

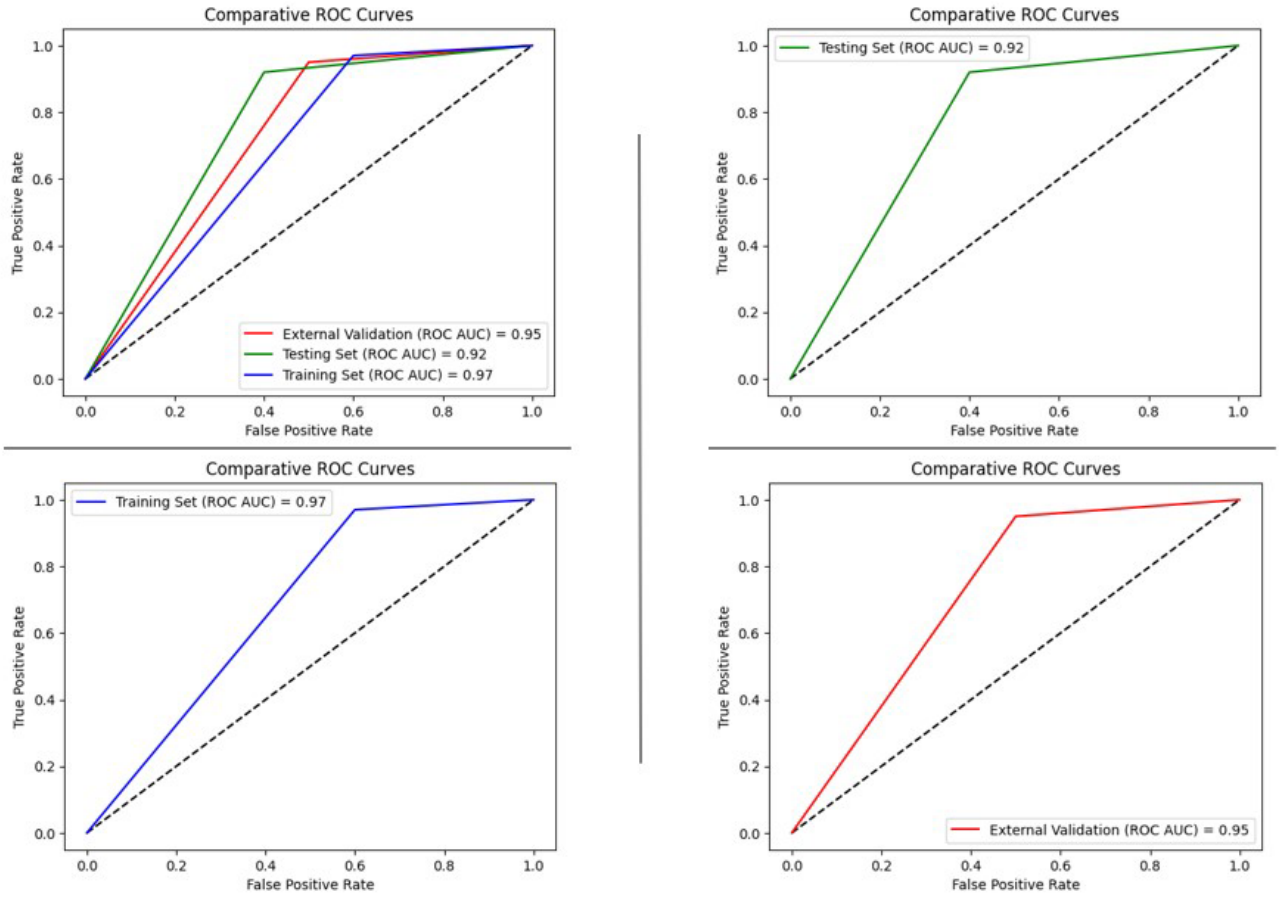

Figure 2. Response receiver operating characteristic (ROC) curve analysis of the deep learning model in the internal training dataset, testing dataset, and external validation.

Table 3. The Overall Performance of This Model for the Deep Learning Model.

Pfirrmann Grade	Accuracy	Sensitivity	Specificity	F1 Score	Prediction Error	ROC
Grade I	97%	100%	95.4%	97.77%	2.3%	97%
Grade II	96.5%	98%	96.2%	96.8%	1.8%	96.3%
Grade III	96.8%	97.5%	96.7%	96.9%	1.6%	96.5%
Grade IV	95.7%	96.2%	95.8%	95.9%	2.1%	96%
Grade V	97.2%	95.8%	97%	97.1%	1.5%	97.5%

with the YOLO architecture, exhibited high performance across various metrics, such as accuracy, sensitivity, specificity, F1 score, prediction error, and receiver operating characteristic area under the curve (ROC-AUC). These results underscore the efficacy of AI in augmenting diagnostic precision for IDD and providing a reliable tool for clinicians. The model's performance across all grades of IDD was robust, with overall accuracy rates exceeding 90% in testing and external validation sets. This high level of accuracy indicates that the AI model can reliably distinguish between different grades of disc degeneration, which is critical for effective clinical decision-making. Sensitivity and specificity rates were similarly high, suggesting that the model can accurately identify both positive (presence of degeneration) and negative cases (absence of degeneration). These attributes are crucial for minimizing false positives and false negatives, thus enhancing the reliability of diagnostic outcomes. However, the observed decline in performance metrics from the training to the external validation set across 50 epochs suggests potential overfitting of the model. While the model demonstrates high accuracy, sensitivity, and specificity on the training data, its effectiveness slightly diminishes when applied to new, unseen data in the external validation set. This discrepancy indicates that the model might have learned to memorize patterns specific to the training data, rather than generalizing well to unseen cases. To address this issue, techniques such as regularization or dropout could be employed during training to prevent overfitting and improve the model's generalizability to real-world clinical scenarios.

Compared to previous studies[15], our model shows significant advancements in the automated detection and grading of IDD. Earlier works have often struggled with maintaining high performance across diverse datasets and varying levels

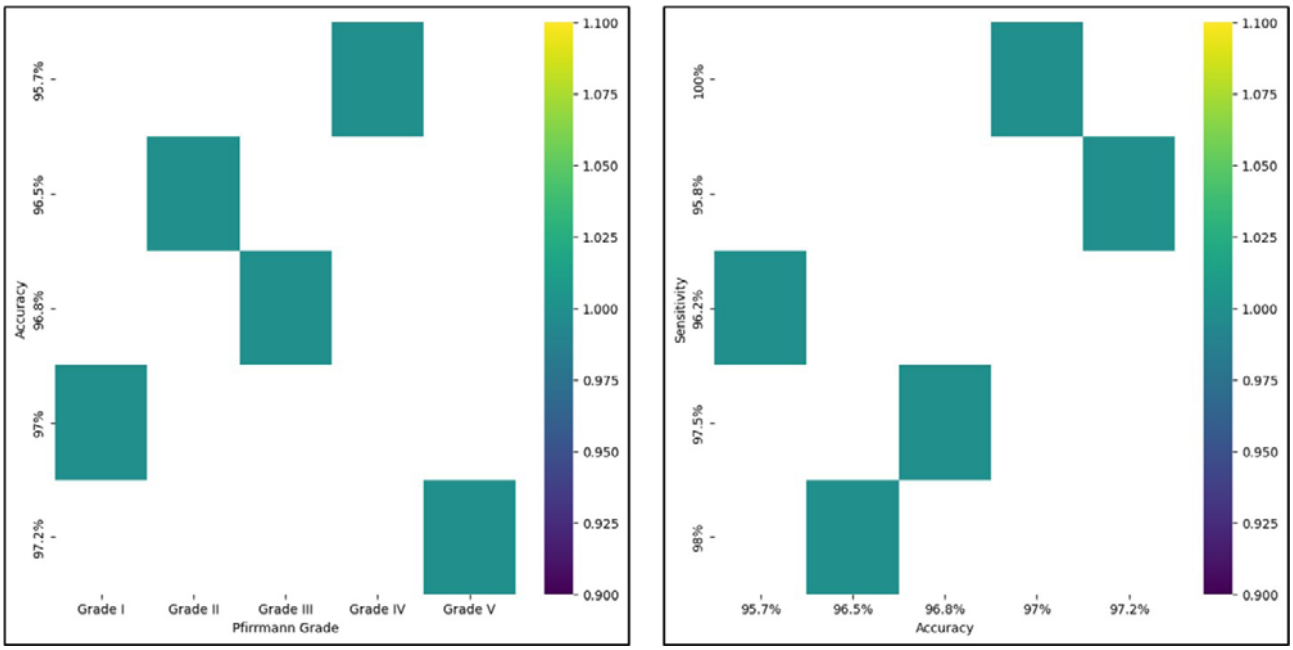

Figure 3. Heatmap comparison: accuracy-Pfirrmann grade (left) and sensitivity-accuracy (right).

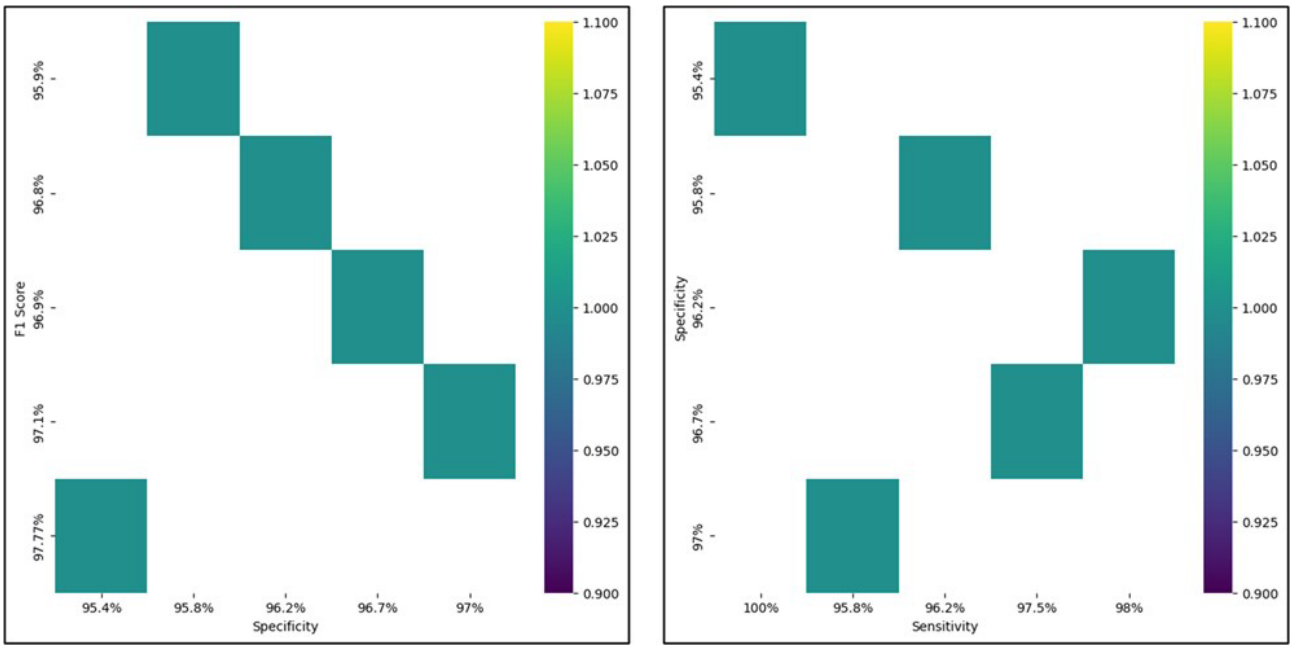

Figure 4. Heatmap comparison: F1 score-specificity (left) and specificity-sensitivity (right).

of disc degeneration. Our study addresses these challenges by employing a comprehensive dataset and a rigorous validation process, which includes an external validation set to assess generalizability. Our model demonstrates robust performance across various metrics, such as accuracy, sensitivity, specificity, and F1 score, with thorough evaluation on both testing and external validation datasets. This study addresses the limitations of previous methods by employing a comprehensive dataset, rigorous validation process, and external validation set to assess generalizability.

This study applied clinical relevance and potential integration to the Web-based application tool for assigned diagnoses. The integration of AI into routine clinical practice could revolutionize the management of IDD. Accurate and rapid grading of disc degeneration allows for timely and appropriate treatment planning, improving patient outcomes and reducing the burden of chronic back pain. The AI model can serve as a decision-support tool for radiologists, enhancing their diagnostic capabilities and potentially reducing diagnostic variability. Furthermore, the ability to automate the grading process can save time and resources, allowing clinicians to focus on more complex aspects of patient care.

The limitations and future directions of this study have some limitations that warrant discussion. Although compre-

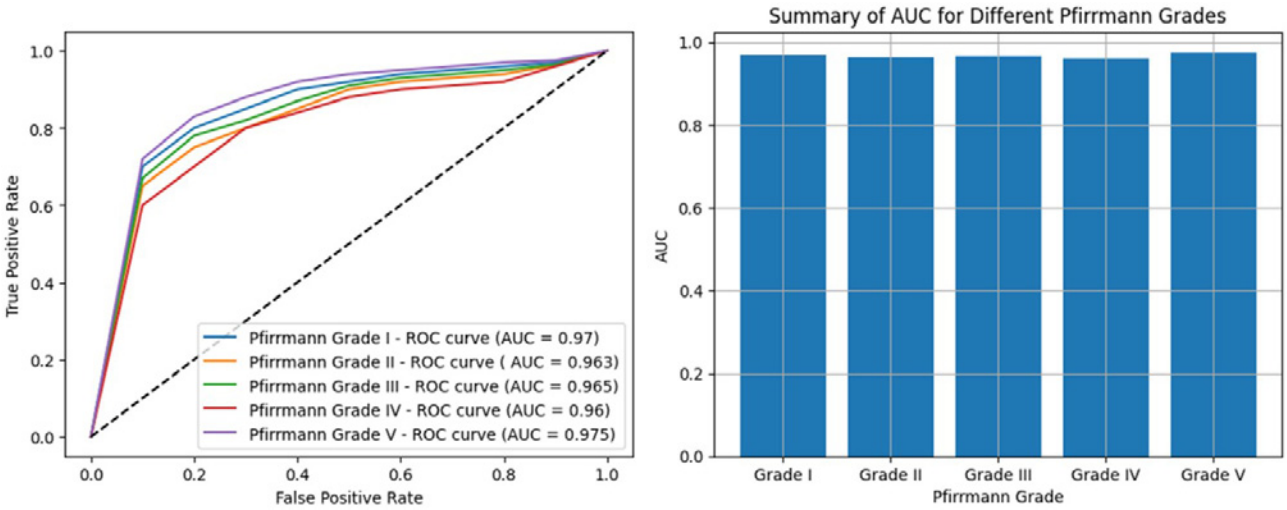

Figure 5. The response receiver operating characteristic (ROC) curve (left) and barchart (right) analysis of the deep learning model for detecting disc degenerative changes from Grades I to V.

hensive, the dataset used was derived from a specific patient population with symptomatic low back pain, which may not fully represent the broader spectrum of IDD seen in the general population. Moreover, the imaging devices, patient demographics, or geographic locations were anonymous. Future studies should include more diverse datasets to validate the model further and ensure its applicability across different demographic groups. This limits the applicability of the AI model across diverse clinical settings and patient demographics, reducing the generalizability and practical utility of the findings and potentially biasing the dataset. Another limitation is the reliance on the Pfirrmann grading system, which, while widely accepted, is subject to some degree of subjectivity. The primary differential between the five-graded Pfirrmann disc degeneration grade is the brightness and differentiation of the nucleus pulposus from the disc membrane; this classification is easily modified by other observers or investigators. Incorporating additional imaging biomarkers and combining AI predictions with clinical data could further enhance the model's diagnostic accuracy and clinical relevance. In addition, prospective clinical trials are needed to evaluate the real-world impact of integrating AI into the diagnostic workflow and to understand how it influences clinical decision-making and patient outcomes.

Conclusion

This study highlights the potential of AI in enhancing the diagnostic process for lumbar intervertebral disc degeneration. The developed deep learning model demonstrates high accuracy, sensitivity, and specificity in detecting and grading IDD according to the Pfirrmann grading system. However, further research is needed to improve the model in more diverse populations and explore its integration into clinical practice. The continued advancement of AI technologies promises to transform the landscape of diagnostic radiology, ultimately improving patient care and outcomes.

Conflicts of Interest: The authors declare that there are no relevant conflicts of interest.

Sources of Funding: None of this research received any external funding.

Author Contributions: W.L., W.C., P.S., K.J., and V.K. designed the study. W.L., W.C., P.S., K.J., and V.K. performed the experiments and analyzed the data. W.L., W.C., and P.S. provided critical reagents. W.L. supervised the experiments. W.L., W.C. and P.S. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Ethical Approval: This study was conducted in accordance with the Declaration of Helsinki and with approval from the Ethics Committee and Institutional Review Board of University of Phayao (approval code: HREC-UP-HSST 1.1/033/67).

Informed Consent: Informed consent for publication was obtained from all participants in this study.

References

1. Dowdell J, Erwin M, Choma T, et al. Intervertebral disk degeneration and repair. Neurosurgery. 2017;80(3S):S46-54.
2. Farshad-Amacker NA, Farshad M, Winklehner A, et al. MR imaging of degenerative disc disease. Eur J Radiol. 2015;84(9):1768-76.
3. Pfirrmann CW, Metzdorf A, Zanetti M, et al. Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine (Phila Pa 1976). 2001;26(17):1873-8.
4. Kos N, Gradisnik L, Velnar T. A Brief review of the degenerative intervertebral disc disease. Med Arch. 2019;73(6):421-4.
5. Kim HS, Wu PH, Jang IT. Lumbar degenerative disease part 1: anatomy and pathophysiology of intervertebral discogenic pain and radiofrequency ablation of basivertebral and sinuvertebral nerve treatment for chronic discogenic back pain: a prospective

case series and review of literature. Int J Mol Sci. 2020;21(4): 1483.
6. Bajwa MH, Samejo AA, Zubairi AJ. Clinical applications of AI-prediction tools in spine surgery: a narrative review. J Pak Med Assoc. 2024;74(Suppl_4):S97-9.
7. Gündüz MŞ, Işık G. A new YOLO-based method for real-time crowd detection from video and performance analysis of YOLO models. J Real Time Image Process. 2023;20(1):5.
8. Duman ŞB, Çelik Özen D, Bayrakdar IŞ, et al. Second mesiobuccal canal segmentation with YOLOv5 architecture using cone beam computed tomography images. Odontology. 2024;112(2): 552-61.
9. Majumder M, Wilmot C. Automated vehicle counting from pre-recorded video using you only look once (YOLO) object detection model. J Imaging. 2023;9(7):131.
10. Sudirman S, Al Kafri A, Natalia F, et al. Lumbar spine MRI dataset. Mendeley Data, V2. 2019;2.
11. van der Graaf JW. SPIDER - Lumbar spine segmentation in MR images: a dataset and a public benchmark. Boston: Zendo; 2023:
12. Liawrungrueang W, Cho ST, Sarasombath P, et al. Current trends in artificial intelligence-assisted spine surgery: a systematic review. Asian Spine J. 2024;18(1):146-57.
13. Siemionow KB, Katchko KM, Lewicki P, et al. Augmented reality and artificial intelligence-assisted surgical navigation: technique and cadaveric feasibility study. J Craniovertebr Junction Spine. 2020;11(2):81-5.
14. Morrow E, Zidaru T, Ross F, et al. Artificial intelligence technologies and compassion in healthcare: A systematic scoping review. Front Psychol. 2023;13.
15. D'Antoni F, Russo F, Ambrosio L, et al. Artificial intelligence and computer aided diagnosis in chronic low back pain: a systematic review. Int J Environ Res Public Health. 2022;19(10):5971.