Received: 9 April 2024 | Revised: 20 June 2024 | Accepted: 25 June 2024

DOI: 10.1002/jsp2.1353

RESEARCH ARTICLE

JOR Spine OPEN ACCESS

# Automated magnetic resonance imaging-based grading of the lumbar intervertebral disc and facet joints

Maryam Nikpasand[1] | Jill M. Middendorf[2] | Vincent A. Ella[3] | Kristen E. Jones[4] | Bryan Ladd[4] | Takashi Takahashi[5] | Victor H. Barocas[1,3] | Arin M. Ellingson[6,7]

[1]Department of Mechanical Engineering, University of Minnesota, Minneapolis, Minnesota, USA

[2]Department of Mechanical Engineering, Johns Hopkins University, Baltimore, Maryland, USA

[3]Department of Biomedical Engineering, University of Minnesota, Minneapolis, Minnesota, USA

[4]Department of Neurosurgery, University of Minnesota, Minneapolis, Minnesota, USA

[5]Department of Radiology, University of Minnesota, Minneapolis, Minnesota, USA

[6]Department of Orthopedic Surgery, University of Minnesota, Minneapolis, Minnesota, USA

[7]Division of Physical Therapy and Rehabilitation Science, Department of Family Medicine and Community Health, University of Minnesota, Minneapolis, Minnesota, USA

**Correspondence**
Arin M. Ellingson, 420 Delaware St SE, MMC 388, Minneapolis, MN 55455, USA.
Email: ellin224@umn.edu

**Funding information**
National Center for Complementary and Integrative Health, Grant/Award Numbers: U01 AT010326, U24 AT011978

## Abstract

**Background:** Degeneration of both intervertebral discs (IVDs) and facet joints in the lumbar spine has been associated with low back pain, but whether and how IVD/joint degeneration contributes to pain remains an open question. Joint degeneration can be identified by pairing T1 and T2 magnetic resonance imaging (MRI) with analysis techniques such as Pfirrmann grades (IVD degeneration) and Fujiwara scores (facet degeneration). However, these grades are subjective, prompting the need to develop an automated technique to enhance inter-rater reliability. This study introduces an automated convolutional neural network (CNN) technique trained on clinical MRI images of IVD and facet joints obtained from public-access Lumbar Spine MRI Dataset. The primary goal of the automated system is to classify health of lumbar discs and facet joints according to Pfirrmann and Fujiwara grading systems and to enhance inter-rater reliability associated with these grading systems.

**Methods:** Performance of the CNN on both the Pfirrmann and Fujiwara scales was measured by comparing the percent agreement, Pearson's correlation and Fleiss kappa value for results from the classifier to the grades assigned by an expert grader.

**Results:** The CNN demonstrates comparable performance to human graders for both Pfirrmann and Fujiwara grading systems, but with larger errors in Fujiwara grading. The CNN improves the reliability of the Pfirrmann system, aligning with previous findings for IVD assessment.

**Conclusion:** The study highlights the potential of using deep learning in classifying the IVD and facet joint health, and due to the high variability in the Fujiwara scoring system, highlights the need for improved imaging and scoring techniques to evaluate facet joint health. All codes required to use the automatic grading routines described herein are available in the Data Repository for University of Minnesota (DRUM).

KEYWORDS
automated grading, deep learning, facet joint, Fujiwara, intervertebral disc, machine learning, Pfirrmann, spine

## 1 | INTRODUCTION

Degeneration of both intervertebral discs (IVDs) and facet joints in the lumbar spine has been associated with low back pain,[1–6] but whether and how IVD or facet joint degeneration contributes to pain remains an open question.[7] Exploration of IVD and facet joint degeneration requires reliable, objective, quantitative, noninvasive measures of the degeneration state. Nondestructive lumbar spine imaging is achievable by numerous techniques, including computed tomography (CT), x-ray, and magnetic resonance imaging (MRI). Each of these techniques offers unique advantages and disadvantages for visualizing features of tissue morphology and health better than the others. Here, we focus on MRI analysis due to its ability to capture soft tissue detail, which has led to its popularity for grading spinal health. Specifically, T1- and T2-weighted sequence MRI images[8–10] that are routinely available in clinical setting allow quantification of degeneration using the Fujiwara[8,11] grading system for the facet joints and the Pfirrmann[10] grading system for the IVD. Unfortunately, current MRI-based analysis techniques are subjective in nature and therefore prone to high variability and poor inter-rater and intra-rater reliability.[12,13]

The Fujiwara and Pfirrmann grading systems both use an integer scale to describe tissue degeneration based on expert rater analysis of visible features in the MRI image. The Fujiwara scale assigns facet joint health scores from 1 to 4, where 1 is a healthy joint while 4 is a severely degenerated joint.[8,11] This assessment is made by qualitatively analyzing the thickness of the articular cartilage, the hydration of the joint, and the presence of bone spurs—all of which are known signs of osteoarthritis.[14,15] Since the facet joint is often only a few pixels wide in the image, and since the light–dark spectrum of the grayscale image can be skewed by other pixel intensities, assigning a Fujiwara score is often difficult. In previous research studies, the Fujiwara scale has shown low (~30%–40%) inter-rater agreement,[13] although most scores differ between graders by only 1 point on the 4-point scale.[16] Similar to the Fujiwara system, the Pfirrmann grading system is used to assess IVD health on a scale from 1 to 5, where 1 is healthy.[10] The Pfirrmann grade is based on the IVD height, clear delineation the nucleus pulposus, and nucleus pulposus hydration. In general, Pfirrmann grading has been found to be more robust than Fujiwara grading, partially due to the IVD's larger size. Thus, the IVD MRI image contains more pixels, a greater range of tissue hydration values, and more details than the facet joint images. These details lead to a higher inter-rater reliability (range from 55 to 83%).[10,17,18] Since both the Fujiwara and the Pfirrmann grading scales ask humans to assign an integer value based on qualitative examination of images, they are both subjected to human errors and inconsistencies. For these reasons, the systems are not frequently used in the clinical setting. Therefore, such scoring systems could be enhanced with objective automated systems.

Automated scoring techniques have been developed to reduce inter-rater variability and improve reliability of IVD diagnoses and subjective scoring systems. For example, multiple automatic scoring techniques have been implemented in which the degree of IVD degeneration, including the degree of disc herniation is detected using convolutional neural networks (CNN) and semantic segmentation networks.[18–22] These techniques can implement algorithms to isolate the tissue of interest[19,21,23,24] and then identify whether the tissue of interest is damaged, bulging, or degenerated.[18–22] Similarly, CNNs have been implemented to understand and interpret the severity of spinal stenosis[24] and modic changes in the spine.[23] Finally, Pfirrmann grading of the IVD has been automated using deep learning techniques including a CNN.[18,23] The use of a CNN to score the health of the IVD using the Pfirrmann system has been shown to improve the inter-rater and intra-rater reliability.[18,23] Despite the promise of automated scoring techniques and automated identification systems for the IVD, automation of the scoring system for the facet joints (Fujiwara)[8,11] has yet to be completed. Improving the reliability of the facet joint (Fujiwara) scoring system is possibly more important than doing so for the Pfirrmann grading system because of the higher inter-rater[13] and intra-rater variability[25,26] in Fujiwara grading.

Additionally, grading a series of discs and joints for research purposes is time-consuming and requires special expertise. Due to the sheer volume of work needed for a large-scale study and busy schedules of specialists, it can be difficult for researchers to find a specialist to grade disc and joint images within a targeted time frame. An automated grading system trained based on specialist assigned grades could help facilitate this process.

Therefore, the goal of this study was to develop, verify, and apply an automated CNN technique to MRI images of the IVD and facet joint. We hypothesized that the CNN could improve the reliability of the Fujiwara scoring system over the standard expert grading in terms of inter-rater and intra-rater reliability of MRI facet joint scoring.

## 2 | METHODS

### 2.1 | Overview

A deep learning technique (deep CNN) was used to analyze MRI images and to classify health of lumbar discs and facet joints according to Pfirrmann and Fujiwara grading systems, respectively. Performance of the CNN on both the Pfirrmann and Fujiwara scales was assessed by comparing the predicted results from the classifier to the grades assigned by an experienced radiologist. To quantify inter-rater reliability of the system, percent agreement (PA), Pearson's correlation and Fleiss kappa values for the predicted results of the test data were compared with the multiple human grader scores on the same data.

### 2.2 | Dataset

Clinical MRI images of lumbar spines were obtained from the public-access Lumbar Spine MRI Dataset.[27] This dataset consists of anonymized clinical MRI scans of 515 symptomatic patients suffering from back pain. Each image stack contains slices of axial and sagittal views of the lower 3 or 5 lumbar vertebrae and IVDs. All image collection was performed using T2- and T1-weighted imaging during standard

care; thus, the scan procedures may have varied across subjects. Images were excluded if they exhibited signs of severe stenosis or showed the presence of instrumentation in the lumbar spine. For the purpose of this study, T2-weighted sagittal and T1-weighted axial views were used as the input data for scoring on the Pfirrmann and Fujiwara grading scales, respectively. The number of slices for each patient ranged from 12 to 20 for both sagittal and axial views.

First, the mid-sagittal and mid-axial slices of each motion segment were selected (total of 2633 disc images of motion segments from L1-L2 to L5-S1 and 2377 joint images from both sides of the L1-L2 through L5-S1 motion segments) from T2-weighted sagittal and T1-weighted axial views. Four graders (two PhD trained spine researchers, one spine surgeon, and one musculoskeletal radiologist) graded the IVDs using the Pfirrmann grading system while 5 graders (two PhD trained spine researchers, one spine surgeon, one neurosurgeon, and one musculoskeletal radiologist) graded the facet joints on both sides of the spine using the Fujiwara scoring system. A total of 2366 graded IVD images (90% of total input images) were used to train the automated algorithm, while 267 graded IVD images (10% of total input images) were kept aside and used later to test the accuracy of the automated network solution. For the facet joint, 2135 graded images (90% of total input images) were used to train the CNN, while 242 images (10% of total input images) were used to test the accuracy of the automated network solution.

## 2.3 | Preprocessing

A custom tool written in MATLAB was used to crop the selected T2/T1-weighted images and define a region of interest enclosing the disc or one of the facet joints from each level in the center of the image and approximately 30% of their surroundings (Figure 1). Images were then grouped based on their assigned graders for Pfirrmann (graders A–D) and Fujiwara (graders A–E) grading. For each joint, the grader who had graded the most images was assigned as the lead grader, and the performance of the rest of graders was compared with that grader to measure inter-rater reliability. B.L., a neurosurgery trained spine surgeon, served as the lead grader for disc images, and T.T. a radiologist specializing in the musculoskeletal imaging, served as the lead grader for facet joint images. The images graded by the lead grader were also used as ground truth labels in training our classifier algorithms.

To reduce the effect of signal inhomogeneity across the MRI scans, we normalized the disc images using the mean and standard deviation pixel intensity of each image. Due to the broader range of pixel intensity in facet joint images, we first limited the range to one standard deviation around the mean of pixel intensity values ($\mu \pm 0.5\sigma$) and then normalized the modified image to the mean and standard deviation of the preprocessed image. A series of preprocessing functions were implemented to resize each disc and facet joint input image to $64 \times 64$ and $32 \times 32$ pixel image, respectively, as the discs are larger than the fact joints, providing a consistent image resolution for the CNN. These relatively coarse pixel counts were required to maximize the number of images from the Lumbar Spine MRI Dataset. Higher resolution images could have been used for some cases, but doing so would have reduced the number of usable images or introduced inconsistency in resolution across samples. The resized images were shuffled and divided into training (80% of image set), developing (10%), and test datasets (10%). Next, more training images were generated by flipping the original images horizontally and rotating them $\pm 36^{\circ}$ (see[28–30] for discussion of flipping, rotating, and other methods for augmenting image datasets). This rotation and flip is believed to reduce the sensitivity of our model to rotations and mirrored images that may naturally occur while acquiring MRI images.

## 2.4 | Inter-rater reliability

A subset of disc and facet joint images were evaluated independently by four graders for the Pfirrmann scores and by five graders for the Fujiwara scores. Images which included all motion segments and levels of health were selected randomly for each pair of grader $X$ (where $X = $ B, C, or D for Fujiwara-based grades and B, C, D, or E for Pfirrmann-based grades) and the lead grader (grader A). Table 1 shows the number of images graded by the lead grader and each of the other

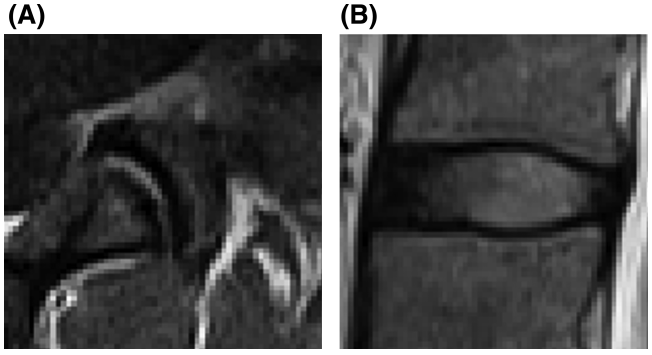


**FIGURE 1** Examples of cropped images for the (A) facet joint and (B) intervertebral disc.

**TABLE 1** Number of images graded by each grader pair of each type.

| | Lead grader (A) | |
|---|---|---|
| Second grader–IVD | B | 2620 |
| | C | 500 |
| | D | 788 |
| | Classification Model | 267 |
| Second grader–facet | B | 68 |
| | C | 301 |
| | D | 134 |
| | E | 246 |
| | Classification Model | 242 |

Abbreviations: IVD, intervertebral disc.

**TABLE 2** Summary of interrater reliability between the graders *X* (=A, B, C, D, and E) and the lead grader.

| | A versus B | | | A versus C | | | A versus D | | | A versus E | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | % Agr | *r* | *κ* | % Agr | *r* | *κ* | % Agr | *r* | *κ* | % Agr | *r* | *κ* | % Agr | *r* | *κ* |
| Fujiwara scores (facet joint) | 34 | 0.4 | 0.04 | 43 | 0.51 | 0.12 | 40 | 0.44 | 0.16 | 44 | 0.52 | 0.19 | 40 | 0.47 | 0.13 |
| Pfirrmann scores (IVD) | 65 | 0.7 | 0.47 | 50 | 0.8 | 0.33 | 53 | 0.7 | 0.34 | - | - | - | 56 | 0.73 | 0.38 |

*Note*: % Agr is the percent agreement, *r* is Pearson's correlation coefficient, and *κ* is Fleiss kappa value.
Abbreviations: IVD, intervertebral disc.

graders. PA, Pearson correlation, and Fleiss kappa values were calculated between each pair of graders and averaged across all the graders to assess inter-rater reliability among the graders (Table 2).

## 2.5 | Network architectures

The CNN-based classifier systems take preprocessed 64 × 64/32 × 32 pixel grayscale images of lumbar discs/facet joints as input and maps to one of the 5 Pfirrmann or 4 Fujiwara grades. The machine learning pipeline was created using the Tensorflow[31] (v.2.9.2) library through the Google Colab Pro platform. Each hidden layer consists of a two-dimensional (2D) convolution layer followed by batch normalization, activation, and pooling layers. These hidden layers were followed by a series of fully connected (FC) layers. The Bayesian Optimization class of the Keras Tuner library was used on the development datasets to pick the optimal set of hyperparameters for each model. Specifically, the number of hidden layers (1 to 6 layers), number of units in each hidden layer (32 to 512 with steps of 32), activation function choices (relu, tanh, and elu), and optimizer learning rate ($10^{-4}$ to $10^{-2}$ with a "log" type sampling method) were automatically adjusted to find the best set of hyperparameters for each system in order to maximize the accuracy of the models. The maximum number of trials for the Bayesian optimizer was set to 100 while the remaining parameters were kept at their default values. A SoftMax output layer was used to calculate the probability vector over the 4 Pfirrmann or 5 Fujiwara labels. The L1-L2 (Lasso-Ridge) regularization method and dropout layers were added to the systems to overcome overfitting to training data. The optimal neural networks for both systems were trained using Adam optimizer, and checkpointing was performed by saving the best model at the end of each epoch if the accuracy was improved compared with the previous saved model.

## 2.6 | Validation/performance evaluation

The performance of the system was evaluated by PA, Pearson's correlation coefficient, and Fleiss kappa value between the predicted scores and ground truth scores (i.e., lead grader's scores). A two-proportion *z*-test was adopted to compare the performance of the CNN-based algorithms with the human results. The null hypothesis for this comparison was that the probability of grader *X* (where *X* = B, C, or D for Fujiwara-based grades and B, C, D, or E for Pfirrmann-based grades) agreeing with the lead grader (grader A) is the same as that of the CNN-based algorithm agreeing with the lead grader (grader A). $p < 0.05$ was considered statistically significant.

We categorized the classification results into six groups:

1. **Correct classification, high confidence**: The algorithm prediction was the same as lead grader's score, and the probability value for the predicted grade was greater than 75%.
2. **Correct classification, low confidence**: The algorithm prediction was the same as lead grader's score, and the probability value for the predicted grade was lower than 75%
3. **Small error, high confidence**: The algorithm prediction was off by one grading point, and the probability value for the predicted grade was greater than 75%.
4. **Small error, low confidence**: The algorithm prediction was off by one class and the probability value for the predicted class was lower than 75%.
5. **Off by 2 or more grades, high confidence (>1 grade off)**: The algorithm prediction was off by more than one grading point, and the probability value for the predicted grade was greater than 75%.
6. **Off by 2 or more grades, low confidence (>1 grade off)**: The algorithm prediction was off by more than one grading point and the probability value for the predicted class was lower than 75%.

## 2.7 | Availability of software

All grading software described in this article, as well as the detailed outputs, are available in the Data Repository for University of Minnesota (DRUM:https://hdl.handle.net/11299/264061).

# 3 | RESULTS

## 3.1 | Human graders

The distribution of Fujiwara scores and Pfirrmann grades assigned to the MRI images of the entire dataset by the human graders (Figure 2) show a strongly skewed distribution of values for the Fujiwara scoring system (skewness = 0.63) and a more symmetric distribution for the

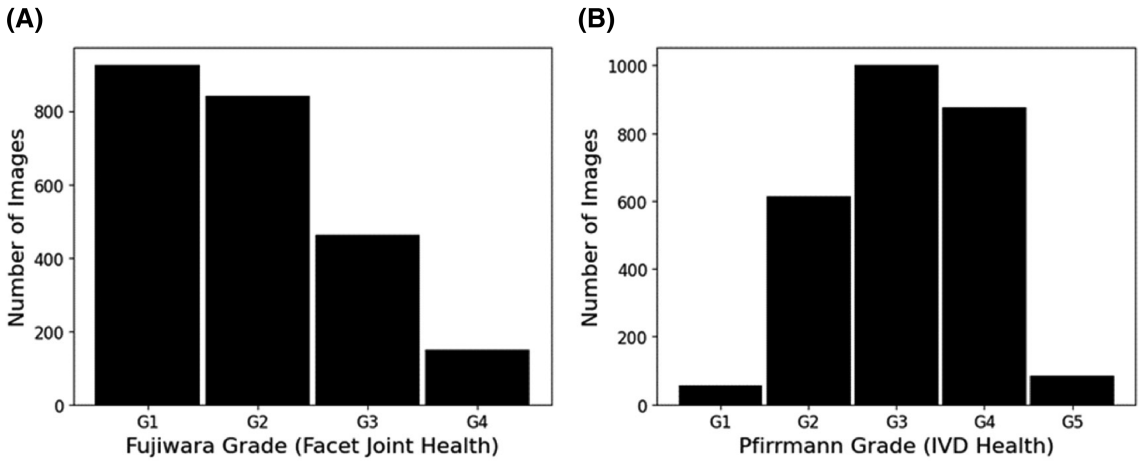


**FIGURE 2** Histogram of dataset for (A) Fujiwara scores and (B) Pfirrmann scores. IVD, intervertebral disc.

Pfirrmann grading system (skewness = −0.15). Fujiwara scores of grade 4 were identified in only 6.2% of all data (Figure 2A). All other Fujiwara scores came from between 19.4% and 38.9% of the training dataset images. For the Pfirrmann grading system, very few images were graded with a score of 1 or 5, the highest and lowest scores (2.1% and 3.2% of images for scores of 1 and 5, respectively, Figure 2B). The other Pfirrmann grades were each assigned to 23.3% to 38.1% of the images (Figure 2B). This variability in the number of images assigned to each group may have negatively affected the accuracy of our deep learning algorithm.

Comparisons between the lead grader (grader A) and each of the other graders (Table 2) showed a large amount of inter-rater variability for the Fujiwara grading system (facet joint) but better performance for the Pfirrmann grading system (IVD). The agreement rate between graders ranged from 34 to 44% (average of 0.4) for Fujiwara grading and 50 to 65% (average of 0.56) for Pfirrmann grading, while the Pearson's correlations coefficient ranged from 0.4 to 0.52 (average of 0.47) for Fujiwara grading and 0.69 to 0.79 (average of 0.73) for Pfirrmann grading. The Fleiss kappa value ranged from 0.04 to 0.19 (average of 0.13) for Fujiwara and 0.33 to 0.47 (average of 0.38) for Pfirrmann grading (Table 2). Under the broad categorizations of Landis and Koch,[32] the Fleiss kappa results correspond to *slight* agreement for the Fujiwara grading and *fair to moderate* agreement for the Pfirrmann grading.

## 3.2 | Classification results—Fujiwara grading of the facet joint

The CNN classification algorithm performed comparably to the human graders for both tasks. For the facet joint images, the CNN-generated Fujiwara scores agreed with the lead grader on 49% of the images (Table 3), which was slightly higher than the average 40% match rate for the human graders. Similarly, the Fleiss kappa statistic for the CNN versus the lead grader was 0.18, which was larger than the human graders' average $\kappa$ value of 0.13. In contrast, the correlation between the CNN result and the lead grader's (grader A) was lower than for the human graders (0.3 vs. 0.47 for CNN and human graders average respectively; Table 2 and Table 3). These results mean that on average the Fujiwara-based model was as good as a human grader at picking the same score as grader A but produced larger or less consistent errors when it was not in agreement with grader A.

**TABLE 3** Summary of deep learning algorithm reliability.

| | Classification model | | |
|---|---|---|---|
| | % Agr | *r* | $\kappa$ |
| Fujiwara scores (facet joint) | 49 | 0.3 | 0.18 |
| Pfirrmann scores (IVD) | 78 | 0.82 | 0.68 |

*Note*: % Agr is the percent agreement, *r* is Pearson's correlation coefficient, and $\kappa$ is Fleiss kappa value.
Abbreviations: IVD, intervertebral disc.

Figure 3 contains intensity plots showing the degree of accuracy both between graders and for CNN in assigning Fujiwara scores. These plots show that most differences between graders are 1 level or less. In some cases, a disagreement by one Fujiwara scale point was more likely than agreement between two graders. For example, between graders A and B, the most common combination of scores was a 3 for grader A and a 2 for grader B (22.1%, Figure 3A). Similarly, between graders A and C, the most common combination of scores was a 1 for grader A and a 2 for grader B (28%, Figure 3B). The most popular combination between graders A and D and between graders A and E were both selecting a score of 2 (16.4%, Figure 3C and 21%, Figure 3D). The model was also able to accurately predict Fujiwara scores of 1 and 2 the most (25% for score 1 and 23.6% for score 2, Figure 3E). In addition, differences in scores that were greater than 2 occurred for all comparisons provided. Fujiwara grade 4 was underrepresented in the initial data (only 6.2% of all the facet joint images were labeled as grade 4). This imbalanced dataset caused the CNN

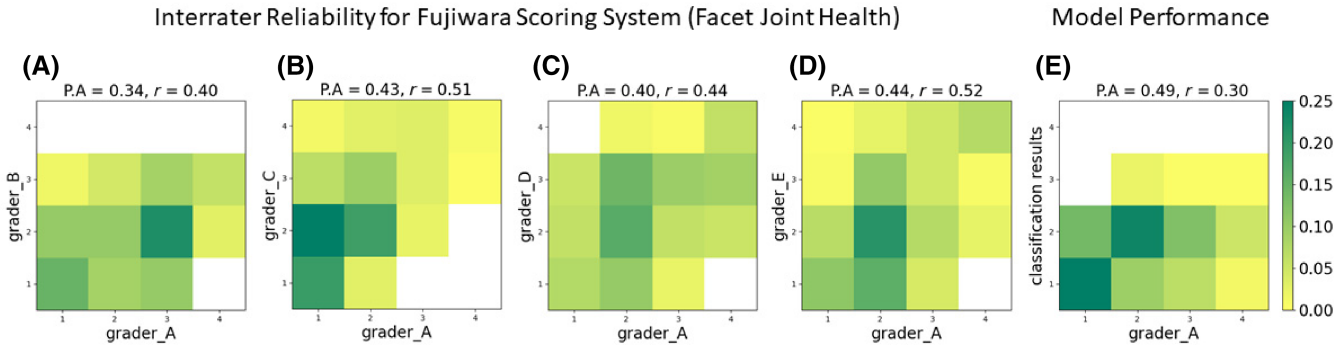


**FIGURE 3** Intensity plots showing the fraction of images that were given a unique combination of Fujiwara scores between 2 graders or between the lead grader and the neural network. (A) Grader A versus grader B. (B) Grader A versus grader C. (C) Grader A versus grader D. (D) Grader A versus grader E. (E) grader A versus the model.

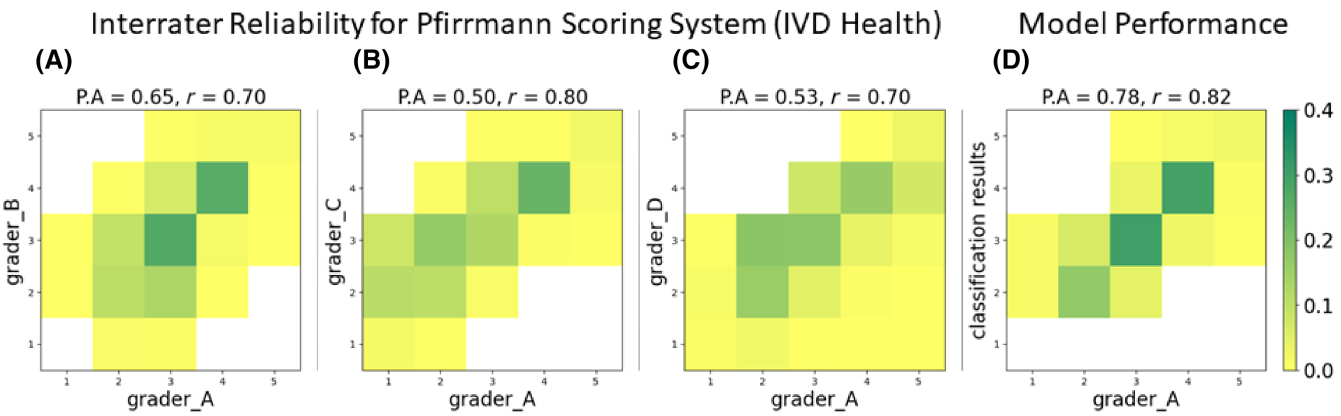


**FIGURE 4** Intensity plots showing the fraction of images that were given a unique combination of Pfirrmann scores between 2 graders or between our 'gold standard' grader and the deep learning model. (A) Grader A versus grader B. (B) Grader A versus grader C. (C) Grader A versus grader D. (D) Grader A versus the model.

algorithm to become biased toward grades 1, 2, and 3 and disregard grade 4, with the final result being that the CNN algorithm did not assign any test images a Fujiwara grade of 4 (Figure 3E).

### 3.3 | Classification results—Pfirrmann grading of the IVD

The classification CNN method was able to improve the reliability of the Pfirrmann scoring system. For the IVD images, the CNN-generated Pfirrmann scores agreed with the lead grader (grader A) on 78% of the images (Table 3), which was significantly better than the human graders ($p < 10^{-6}$). The Fleiss kappa statistic for the CNN was 0.68, which was, again, much higher than the human graders and indicative of substantial agreement.[32] The correlation between the CNN results and the lead grader's (grader A) was also higher than for the human graders (0.82 vs. an average of 0.73; Tables 2 and 3).

Inter-rater agreement plots show the most common combinations of Pfirrmann scores between graders of MRI images were for scores in agreement with each other (Figure 4). Grader A versus B showed the most images in a group with 3 s (Figure 4A), grader A versus C showed the most images in a group with 4 s (Figure 4B), grader A versus D showed the most in a group with 3 s (Figure 4C), while our model often predicted 3 and 4 s in agreement with grader A (Figure 4D). Similar to the Fujiwara scores but to a lesser extent, all inter-rater comparisons showed some scoring image differences greater than 2 values different.

### 3.4 | Error analysis

Differences were seen in the type of errors that occurred with the Fujiwara scoring system versus the Pfirrmann grading. For the Fujiwara grading of facet joint images (Figure 5A), 49.2% of the data were labeled correctly. A substantial fraction of the correct and incorrect grades were cases of low classifier confidence (< 75% assigned probability). 51% of all high-confidence predictions were correct and 49% of all low-confidence predictions were correct. The values were comparable to the inter-rater match rate of 40%.

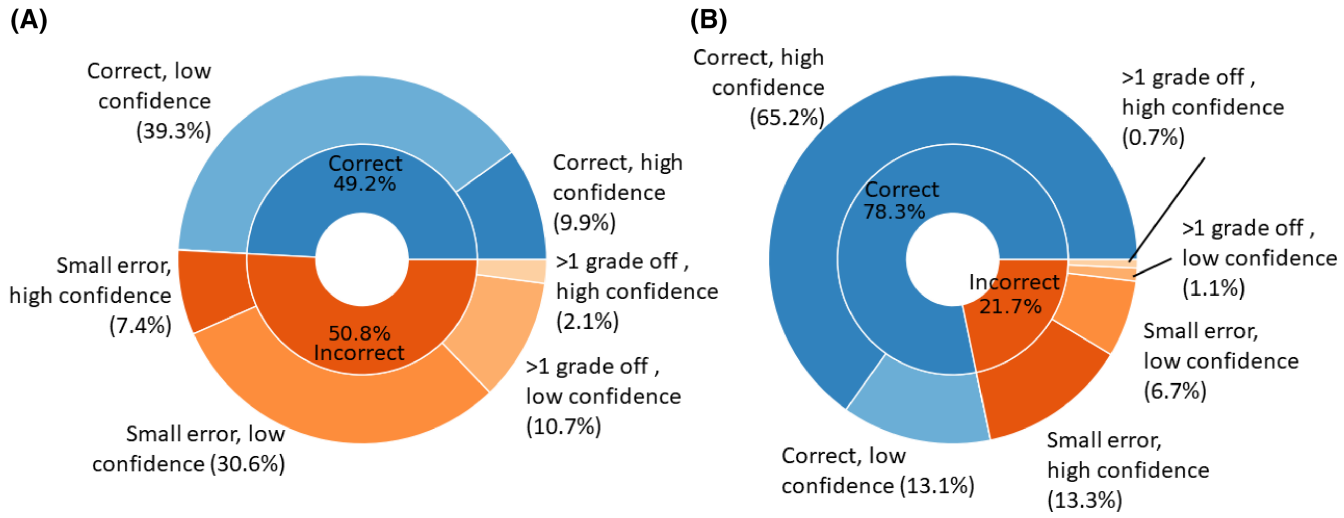


**FIGURE 5** Analysis of the type of errors generated by the CNN classification system algorithm for the (A) Fujiwara grading of the facet joint and (B) Pfirrmann grading of the intervertebral disc. The inner rings represent the overall accuracy of the models on predicting the grades same as lead grader. The outer rings show subdivisions of the inner rings that represent the itemized error based on explanation in Section 2.6.

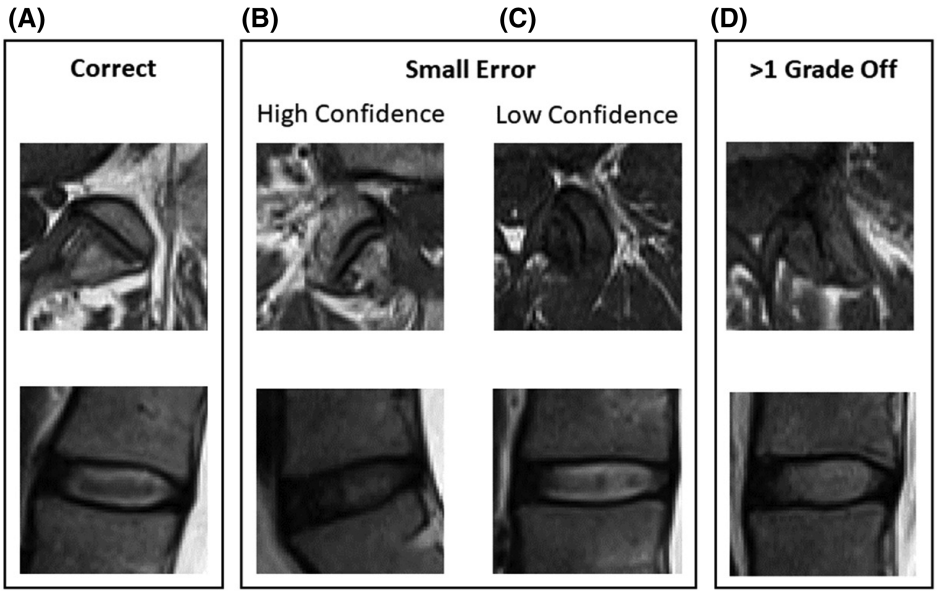


**FIGURE 6** Representative magnetic resonance imaging (MRI) images of (A) images that were accurately labeled by the algorithm (assigned Fujiwara and Pfirrmann grade: 2 and 3, predicted Fujiwara and Pfirrmann grade: 2 and 3) (B) images that were off by 1 grade with high confidence (assigned Fujiwara and Pfirrmann grade: 3 and 3, predicted Fujiwara and Pfirrmann grade: 2 and 4) (C) off by 1 grade with low confidence (assigned Fujiwara and Pfirrmann grade: 2 and 2, predicted Fujiwara and Pfirrmann grade: 1 and 3) (D) and off by more than 1 grade (assigned Fujiwara and Pfirrmann grade: 4 and 1, predicted Fujiwara and Pfirrmann grade: 2 and 3).

In contrast, 78.3% of IVD images were labeled correctly (Figure 5B), and the Pfirrmann grade error was most commonly off by 1 grade point (small error). The 1-point errors occurred with both high confidence (13.1% of all cases, 60.3% of all errors) and low confidence (6.7% of all cases, 30.8% of all errors). The CNN was off by more than 1 Pfirrmann grade point on only 1.9% of cases (8.7% of all errors). The CNN for Pfirrmann grading also showed less severe errors than the Fujiwara classifier. In addition, 82.5% of high-confidence CNN predictions agreed with the lead reviewer, whereas only 62.7% of the low-confidence predictions did; as for the Fujiwara grading, the low-confidence prediction results were comparable to the inter-rater performance (56.2% match). Representative MRI images associated with errors in the Fujiwara and Pfirrmann deep learning classification methods are shown in Figure 6.

## 4 | DISCUSSION

Our CNN-based deep learning algorithm showed similar reliability to the human graders for the Fujiwara grading system and was able to improve the reliability of the Pfirrmann grading system. The improvement in the reliability of the Pfirrmann grading system is consistent with previous results obtained for a deep learning algorithm used to assess the health of the IVD.[18,23] Using deep learning to understand the facet joint is a relatively new concept. Our deep learning model for Fujiwara scoring was able to capture and correctly identify a similar amount of MRI images to the average grader, and the high-confidence predictions from the model in particular matched the lead grader at a higher rate than the other human graders. Both the model reliability and the average inter-rater reliability scores (PA and

Pearson's correlation) were less than ideal for the Fujiwara scoring system. MRI-based assessment of facet joint health is difficult due to the low MRI resolution and smaller size of facet joints (compared with IVDs) in the lumbar spine. This low resolution may contribute to the high inter-rater variability and the high error content of our model. In the following paragraphs, we explore a series of potential reasons for the poor agreement in Fujiwara scoring: (1) lack of image detail, (2) multifaceted nature of joint degeneration and the grading criteria, (3) insufficient number of training images, and (4) nonuniform distribution of the training images. The first presents an imaging challenge, the second an assessment challenge, and the last two machine learning / training challenges.

We begin with the lack of image detail—the small size of the tissues in question and the limited resolution of the images resulted in a relatively small amount of information content, especially for the facet joint. Better resolution would obviously provide more information and could thus lead to more consistent grading (from both the humans and the CNN). Improvements to MRI imaging could also improve the reliability of a machine-learning tool such as ours. These improvements include quantitative weighted T1 and T2 imaging, as has been proposed for other diarthrodial joints and the IVD.[33–37] Due to the larger size of the IVDs in MRI images, the IVD images have more and larger features than the facet joint images. These larger and more detailed features of IVD images make it easier to detect degeneration more accurately in the IVD than in the facet joints.

A second challenge arises because degeneration is a multiscale, multitissue, and heterogeneous process that is difficult to capture in a single MRI image independent of the image resolution. Facet joint degeneration in particular is characterized by loss of proteoglycan content, thinning of the articular cartilage, increased joint calcification, and changes to the subchondral bone.[14,15,38] These features of arthritis can vary both across the surface of the tissue and through the depth of the articular cartilage and the bone.[39–41] If one grader puts more weight on calcification than cartilage thickness compared to another grader or examines one region of the image more carefully compared to another grader, they can easily assign different grades to the same image. It might be possible to address the challenge of inconsistent grading by limiting the CNN training to images on which the graders agreed, but doing so would necessarily omit the more difficult evaluations and thereby might reduce the robustness of the trained network. Another possible approach would be to train the CNN to estimate nominally objective quantities (e.g., IVD height or facet joint cartilage thickness), which could then be used to determine how different graders weight those quantities in their evaluations, but the intrinsic subjectivity of the grading systems would remain.

The spinal level (L1 through S1) also has the potential to affect the cartilage and joint mechanical and structural properties.[15] In our data, the grading system was generated based on a single MRI slice (mid-axial slice of the facet joints and mid-sagittal slice of the discs). A more thorough model that used the full volumetric data along the depth of the 3D MRI stack could result in a better performance. Furthermore, the inter-rater reliability for our data would go up to 87.2% if a difference of one grade were considered correct. These off-by-one-grade images could be a sign that the algorithm might be improved if we were to create more categories that clearly categorize more features of degeneration. However, the issue of low image resolution and inconsistent pixel intensities still exists with a larger range of Fujiwara values. Additionally, MRI images are not good at detecting bone health changes,[16] but CT imaging paired with MRI or x-ray imaging paired with MRI together may improve the accuracy of facet joint health assessments.[42] However, this approach may not be a practical solution, as exposing a patient to two imaging modalities may not be justified from a patient care perspective.

In light of the complexity of the problem, one must conclude that some amount of variability in the grading is inevitable.

Turning from issues inherent in grading spinal health to those specific to this project, the quantity of images in the training dataset and the distribution of the images in that training dataset will affect the ability of the model to predict the facet joint health and the IVD health. The distribution of Fujiwara scores assigned to our training dataset for the facet joints was skewed. The lowest scores (scores of 1 and 2), which represent the healthiest facet joints, were more common than the highly degenerated joint scores (scores of 3 and 4). This same pattern held true for our test dataset as well. Therefore, it is not surprising that our model accurately identified more facet joints with scores of 1 and 2 than degenerated facet joints with scores of 3 and 4. An even distribution over all possible values would help the model to learn features from all groups and would improve the accuracy of the deep learning model. In contrast, the distribution of Pfirrmann grades followed a less skewed distribution that did not contain many 'healthy' (grade of 1) or severely degenerated discs (grade of 5). This distribution of IVD grades for a clinical dataset has been seen previously.[18] These distributions of the training dataset, which were outside of our control, likely impacted the accuracy of the deep learning algorithm.

Additionally, the distributions of image scores seen in this study may be caused by the type of patient that visits the clinic and gets a standard MRI scan of their lumbar spine. Most of these patients are suffering from some form of spinal disease or pain. The high number of grade 1 Fujiwara images is surprising and may reflect the diffuse nature of low back pain, the challenges in assessing facet joint health, and/or a bias in our lead grader. Conversely, we expected to have a large amount of severely degenerated facet joints and IVDs from the patients in the database. Since the highest scores for both the Fujiwara and Pfirrmann grades were assigned to the least number of images, this may indicate that while most patients have some spinal degeneration, the degree of degeneration is not typically severe, at least as measured by MRI.[16]

From this study, we can conclude that, despite many challenges, deep learning algorithms have the potential to improve the reliability of MRI analysis of lumbar spine degeneration and can already perform at a level comparable to human graders in terms of inter-rater agreement. As clinical MRI imaging of the facet joint improves with steady technological advancement, the image resolution and the distinction of facet joint features will be enhanced. As these MRI improvements

are made, the deep learning techniques outlined in this paper (and new techniques that may be developed) for the facet joint can be retrained and improved concomitantly with the imaging technology. Moreover, it is important to note that using the models introduced in this paper, it is not clear what features of MRI images cause some facet joints and IVDs to be accurately labeled and other images to be not accurately labeled. However, there exist various interpretability and feature extraction methods, including but not limited to saliency maps, layer-wise relevance propagation, integrated gradients, and gradient-based approaches that could unveil critical features of the input data that significantly contribute to the neural network prediction process. Future work in this study could focus on implementation of these interpretability techniques to enhance the transparency and reliability of the deep learning methods.

## AUTHOR CONTRIBUTIONS

MN, VHB, and AME conceived the initial idea for this work. MN, JMM, and VAE developed the code and model development. KEJ, BL, and TT performed the manual grading assessments and clinical interpretation. MN, JMM, VAE, VHB, and AME contributed to conception of the work and to analysis and interpretation of the data. All authors prepared and edited the manuscript.

## ACKNOWLEDGMENTS

The authors acknowledge our funding sources NIH U01-AT010326 and U24-AT011978 (SPINEWORK). The authors thank Wesley Kochpatcharin for his help on preprocessing the images.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## ORCID

*Arin M. Ellingson* https://orcid.org/0000-0001-6154-8035

## REFERENCES

1. O'Leary SA, Paschos NK, Link JM, Klineberg EO, Hu JC, Athanasiou KA. Facet joints of the spine: structure–function relationships, problems and treatments, and the potential for regeneration. *Annu Rev Biomed Eng*. 2018;20(1):145-170.
2. Manchikanti L, Singh V, Pampati V, Damron KS, Beyer CD, Barnhill RC. Is there correlation of facet joint pain in lumbar and cervical spine? An evaluation of prevalence in combined chronic low back and neck pain. *Pain Physician*. 2002;5(4):365-371.
3. Manchikanti L, Singh V, Rivera J, Pampati V. Prevalence of cervical facet joint pain in chronic neck pain. *Pain Physician*. 2002;5(3):243-249.
4. Peng B, Bogduk N. Cervical discs as a source of neck pain. An analysis of the evidence. *Pain Med*. 2019;20(3):446-455.
5. Pettersson K, Hildingsson C, Toolanen G, Fagerlund M, Björnebrink J. Disc pathology after whiplash injury. *Spine*. 1997;22:283-287.
6. Ellingson AM, Nuckley DJ. Altered helical Axis patterns of the lumbar spine indicate increased instability with disc degeneration. *J Biomech*. 2015;48(2):361-369.
7. Childs JD, Cleland JA, Elliott JM, et al. Neck pain: clinical practice guidelines linked to the international classification of functioning, disability, and health from the Orthopaedic section of the American Physical Therapy Association. *J Orthop Sports Phys Ther*. 2008;38(9):A1-A34.
8. Fujiwara A, Tamai K, Yamato M, et al. The relationship between facet joint osteoarthritis and disc degeneration of the lumbar spine: an MRI study. *Eur Spine J*. 1999;8(5):396-401.
9. Fujiwara A, Tamai K, An HS, et al. The relationship between disc degeneration, facet joint osteoarthritis, and stability of the degenerative lumbar spine. *J Spinal Disord*. 2000;13(5):444-450.
10. Pfirrmann CWA, Metzdorf A, Zanetti M, Hodler J, Boos N. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine*. 2001;26(17):1873-1878.
11. Fujiwara A, Lim T-H, An HS, et al. The effect of disc degeneration and facet joint osteoarthritis on the segmental flexibility of the lumbar spine. *Spine*. 2000;25(23):3036-3044.
12. Zhou X, Liu Y, Zhou S, et al. The correlation between radiographic and pathologic grading of lumbar facet joint degeneration. *BMC Med Imaging*. 2016;16(27):1-8.
13. Stieber J, Quirno M, Cunningham M, Errico TJ, Bendo JA. The reliability of computed tomography and magnetic resonance imaging grading of lumbar facet arthropathy in total disc replacement patients. *Spine*. 2009;34(23):833-840.
14. O'Leary SA, Link JM, Klineberg EO, Hu JC, Athanasiou KA. Characterization of facet joint cartilage properties in the human and interspecies comparisons. *Acta Biomater*. 2017;54:367-376.
15. Gupta S, Smith HE, Fainor M, Mauck RL, Gullbrand SE. Level dependent alterations in human facet cartilage mechanics and bone morphometry with spine degeneration. *J Orthop Res*. 2022;43:1-10.
16. Little JW, Grieve T, Cantu J, et al. Reliability of human lumbar facet joint degeneration severity assessed by magnetic resonance imaging. *J Manipulative Physiol Ther*. 2020;43(1):43-49.
17. Urrutia J, Besa P, Campos M, et al. The Pfirrmann classification of lumbar intervertebral disc degeneration: an independent inter- and intra-observer agreement assessment. *Eur Spine J*. 2016;25(9):2728-2733.
18. Niemeyer F, Galbusera F, Tao Y, Kienle A, Beer M, Wilke HJ. A deep learning model for the accurate and reliable classification of disc degeneration based on MRI data. *Invest Radiol*. 2021;56(2):78-85.
19. Mbarki W, Bouchouicha M, Frizzi S, Tshibasu F, Farhat LB, Sayadi M. Lumbar spine discs classification based on deep convolutional neural networks using axial view MRI. *Interdiscip Neurosurg*. 2020;22:100837.
20. Zheng H-D, Sun Y-L, Kong D-W, et al. Deep learning-based high-accuracy quantitation for lumbar intervertebral disc degeneration from MRI. *Nat Commun*. 2022;13(1):841.
21. Ketola JHJ, Inkinen SI, Karppinen J, Niinimäki J, Tervonen O, Nieminen MT. T2-weighted magnetic resonance imaging texture as predictor of low back pain: a texture analysis-based classification pipeline to symptomatic and asymptomatic cases. *J Orthop Res*. 2021;39(11):2428-2438.
22. Wang C, Yuan J, Huang Z, Shi Z. Deep learning-based correlation analysis between spine surgery lumbar facet joint and lumbar disc herniation using magnetic resonance images. *Sci Program*. 2021;2021:9623991.
23. Jamaludin A, Lootus M, Kadir T, et al. ISSLS PRIZE IN BIOENGINEERING SCIENCE 2017: automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. *Eur Spine J*. 2017;26(5):1374-1383.
24. Lu J-T, Pedemonte S, Bizzo B, et al. DeepSPINE: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. *Proc Mach Learn Res*. 2018;85:1-16.
25. Middendorf JM, Barocas VH. MRI-based degeneration grades for lumbar facet joints do not correlate with cartilage mechanics. *JOR Spine*. 2023;6:e1246.

26. Gupta S, Xiao R, Fainor M, Mauck RL, Smith HE, Gullbrand SE. Level dependent alterations in human facet cartilage mechanics and bone morphometry with spine degeneration. *J Orthop Res*. 2023;41(3):674-683.
27. Sudirman S, Al Kafri A, Natalia F, et al. Lumbar Spine MRI Dataset 2. 2019. doi:10.17632/k57fr854j2.2
28. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):60.
29. Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. *Array*. 2022;16:100258.
30. Goceri E. Medical image data augmentation: techniques, comparisons and interpretations. *Artif Intell Rev*. 2023;56(11):12561-12605.
31. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv*. 2016. doi:10.48550/arXiv.1603.04467
32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
33. Ellingson AM, Mehta H, Polly DW, Ellermann J, Nuckley DJ. Disc degeneration assessed by quantitative T2* (T2 star) correlated with functional lumbar mechanics. *Spine*. 2013;38(24):1533-1540.
34. Cutcliffe HC, Davis KM, Spritzer CE, DeFrate L. The characteristic recovery time as a novel, noninvasive metric for assessing in vivo cartilage mechanical function. *Ann Biomed Eng*. 2020;48(12):2901-2910.
35. Xia Y, Wang N, Lee J, Badar F. Strain-dependent T1 relaxation profiles in articular cartilage by MRI at microscopic resolutions. *Magn Reson Med*. 2011;65(6):1733-1737.
36. Takashima H, Takebayashi T, Yoshimoto M, et al. Investigation of intervertebral disc and facet joint in lumbar spondylolisthesis using T2 mapping. *Magn Reson Med Sci*. 2014;13(4):261-266.
37. Stelzeneder D, Messner A, Vlychou M, et al. Quantitative in vivo MRI evaluation of lumbar facet joints and intervertebral discs using axial T2 mapping. *Eur Radiol*. 2011;21(11):2388-2395.
38. Boszczyk BM, Boszczyk AA, Korge A, et al. Immunohistochemical analysis of the extracellular matrix in the posterior capsule of the zygapophysial joints in patients with degenerative L4–5 motion segment instability. *J Neurosurg Spine*. 2003;99(1):27-33.
39. Kim JS, Ali MH, Wydra F, et al. Characterization of degenerative human facet joints and facet joint capsular tissues. *Osteoarthr Cartil*. 2015;23(12):2242-2251.
40. Tischer T, Aktas T, Milz S, Putz RV. Detailed pathological changes of human lumbar facet joints L1-L5 in elderly individuals. *Eur Spine J*. 2006;15(3):308-315.
41. Tanno I, Murakami G, Oguma H, et al. Morphometry of the lumbar zygapophyseal facet capsule and cartilage with special reference to degenerative osteoarthritic changes: an anatomical study using fresh cadavers of elderly Japanese and Korean subjects. *J Orthop Sci*. 2004;9(5):468-477.
42. Berg L, Thoresen H, Neckelmann G, Furunes H, Hellum C, Espeland A. Facet Arthropathy evaluation: CT or MRI? *Eur Radiol*. 2019;29(9):4990-4998.


**How to cite this article:** Nikpasand M, Middendorf JM, Ella VA, et al. Automated magnetic resonance imaging-based grading of the lumbar intervertebral disc and facet joints. *JOR Spine*. 2024;7(3):e1353. doi:10.1002/jsp2.1353