RADIOMICS AND MACHINE LEARNING FOR PFIRRMANN GRADE CLASSIFICATION OF INTERVERTEBRAL DISCS IN LUMBAR MRI

*

Sofía González-Martínez[1], Jesús Alejandro Alzate-Grisales[1], Joaquim Montell-Serrano[1], Francisco García-García[2], Julio Domenech-Fernandez[3], Carlos Mayor de Juan[3], and María de la Iglesia-Vayá[4]

[1]*Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana, Unidad Mixta de Imagen Biomédica e Inteligencia Artificial FISABIO-CIPF, 46020, Valencia, Spain*
[2]*Computational Biomedicine Laboratory, Principe Felipe Reseach Center (CIPF), 46012, Valencia, Spain*
[3]*Servicio Cirugía Ortopédica Hospital Arnau de Vilanova, 46015, Valencia, Spain*
[4]*Departamento de Sistemas Informáticos y Computación (DSIC). Universidad Politécnica de València (UPV). 46022, Valencia, Spain*

ABSTRACT

Intervertebral disc degeneration (IDD) is a leading cause of chronic low back pain, yet its clinical grading with the Pfirrmann scale is subjective and prone to variability. This study evaluates a radiomics-based machine learning framework for automatic classification of Pfirrmann grades from lumbar T2-weighted MRI. Radiomic features of disc shape, intensity, and texture were extracted under IBSI guidelines and classified using six machine learning models with patient-level cross-validation, Bayesian hyperparameter optimization, and probability calibration. Gradient Boosting achieved the best overall performance, with a mean AUC of 0.87 in multiclass classification and 0.94 in binary classification, the latter improving sensitivity to advanced degeneration and alleviating class imbalance. SHapley Additive exPlanations (SHAP) identified texture descriptors and shape sphericity as key predictors, with feature patterns aligning with physiologic degeneration stages. These results demonstrate that radiomics combined with machine learning provides accurate and interpretable grading of disc degeneration, offering a reproducible and clinically meaningful alternative to subjective visual assessment.

***K*eywords** Intervertebral disc degeneration · Pfirrmann grading · Lumbar spine MRI · Radiomics · Machine Learning

1 Introduction

Intervertebral disc degeneration (IDD) is a progressive condition that affects the structural integrity and biomechanical function of the spinal column, representing one of the leading causes of chronic low back pain worldwide [1]. This process involves a gradual loss of water and essential extracellular components in the nucleus pulposus, alterations in the annulus fibrosus, and reduced disc height, ultimately impairing the ability of the disc to absorb and distribute mechanical loads [2]. To standardize the assessment of disc degeneration on MRI, the Pfirrmann grading system is widely used, classifying discs from Grade I (healthy) to Grade V (severe degeneration) based on signal intensity, disc structure, and disc height [3]. Accurate grading is crucial for guiding treatment decisions, from conservative management to surgical interventions, and for monitoring disease progression. However, the subjective nature of visual grading and variability among radiologists highlight the need for more objective, reproducible methods to evaluate disc health. Pfirrmann grading shows only moderate interobserver agreement, with frequent discrepancies between adjacent

Citation: **Gonzalez-Martinez, S., Alzate-Grisales, J. A., Montell-Serrano, J., García-García, F., Domenech-Fernandez, J., Mayor de Juan, C., & de la Iglesia-Vayá, M. (2025). Radiomics and Machine Learning for Pfirrmann Grade CLassification of Intervertebral Discs in Lumbar MRI**
Corresponding Author: Sofía González-Martínez
Contact: sofia.gonzalez@fisabio.es

categories. This variability, observed even among experienced readers and across sites, introduces measurement noise that weakens statistical power and hinders multi-center comparability [4].

Recent years have seen growing interest in radiomics – the extraction of quantitative features from medical images as a tool for objective and reproducible assessment of disc degeneration. Radiomic features—including shape, intensity, and texture descriptors—can capture sub-visual patterns associated with early degeneration that are not discernible through conventional visual grading [5]. This aligns with the radiomic hypothesis that imaging phenotypes can reflect underlying biological processes, offering potential for more accurate phenotyping of IDD.

Several studies have demonstrated the feasibility of radiomics in intervertebral disc analysis. McSweeney et al. [6] developed a robust radiomics signature for lumbar IDD using deep learning-based segmentation and T2-weighted MRI from the Northern Finland Birth Cohort (n = 1,397). By extracting 737 radiomic features and applying sparse partial least squares discriminant analysis, the model achieved a balanced accuracy of 76.7% (Cohen's Kappa = 0.70), outperforming conventional indices such as disc height index and peak signal intensity difference (balanced accuracy = 66.0%).

Similarly, Xie et al. [5] proposed an MRI radiomics-based decision support tool for cervical disc degeneration in a two-center study including 2,610 cervical discs from 435 patients. Using a fine-tuned MedSAM model for automated segmentation and extracting 924 features from T1- and T2-weighted images, the authors trained and compared multiple machine learning models. The final combined radiomics model achieved an AUC of 0.95 and accuracy of 89.5%.

These algorithms have recently been explored in literature. Vector Support Machines (SVMs) are popular for radiomics due to their effectiveness in high-dimensional spaces. For example, Fan et al. (2024) [7] reported that SVMs achieved the highest accuracy when integrating radiomic and deep features. Tree-based ensembles (Random Forest, Gradient Boosting) can handle many features and reduce overfitting via ensembling. Although recent research specifically focused on Pfirrmann grading is limited, related studies on low-back pain prediction have reported strong performance, with random forests using disc features achieving AUCs of 0.83–0.88 [8]. A prospective study by Climent et al. [9] found that MRI texture analysis, combined with machine learning, showed moderate ability to predict which patients with chronic low back pain would benefit less from rehabilitation. Approaches to modeling Pfirrmann grades vary: logistic regression typically treats them as separate categories, whereas ordinal regression accounts for their natural ordering by emphasizing the severity of misclassifications. Interestingly, Niemeyer et al. (2021) found that advanced ordinal loss functions did not consistently outperform a well-tuned multiclass baseline, suggesting that conventional classifiers can be highly competitive even when labels are ordered [10].

Despite these advances, important challenges remain, including standardization of feature extraction, validation in multi-center cohorts, and integration with clinical workflows. Furthermore, most studies have focused on binary or simplified classifications, underscoring the need for more models that capture the full spectrum of degeneration severity.

Building on these advances, the present study focuses on the development and evaluation of a radiomics-based machine learning model for automated Pfirrmann grading of lumbar intervertebral discs using routine MRI. Although previous work has applied radiomics and deep learning for disc degeneration classification, most approaches analyze all lumbar levels collectively, potentially overlooking anatomical and biomechanical differences that influence degeneration patterns. This study addresses this limitation by performing a level-specific analysis across L1–L5 discs, providing insight into how radiomic signatures vary along the lumbar spine.

The objective of this study was to investigate whether radiomic features extracted from lumbar intervertebral discs on T2-weighted MRI, combined with machine learning algorithms, can accurately classify disc degeneration according to the Pfirrmann grading system. A dataset of lumbar spine MRI examinations, manually graded by an expert radiologist to establish a reliable reference standard, was analyzed. From each intervertebral disc, a comprehensive set of quantitative radiomic features was extracted, including first-order intensity descriptors, second-order texture metrics, and three-dimensional shape measures. Several machine learning algorithms were trained and evaluated, including Support Vector Machines, Logistic Regression, Random Forest, Naïve Bayes, k-Nearest Neighbors, and Gradient Boosting. By leveraging a large and curated dataset and comparing multiclass and binary prediction strategies, we aimed to evaluate both the feasibility and reliability of AI-based automated grading as a standardized alternative to subjective visual assessment.

2 Material and Methods

2.1 Dataset

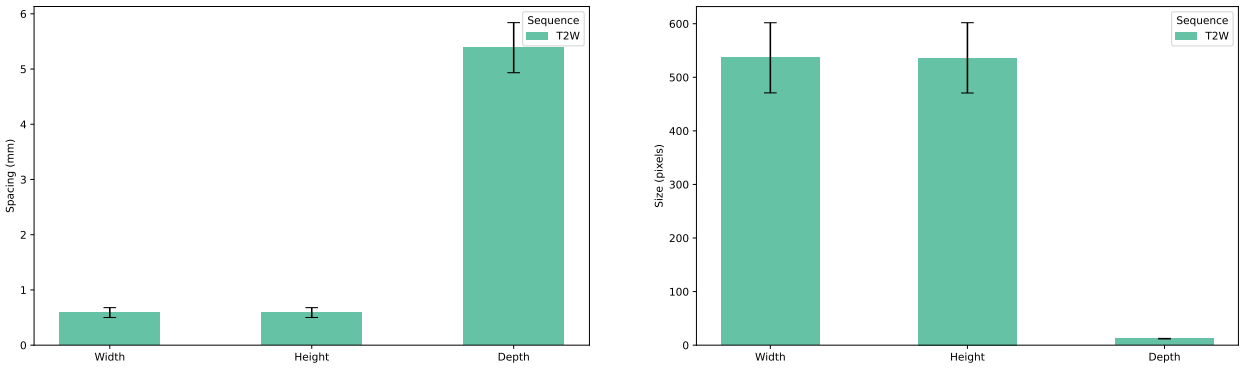
The MIDAS (Massive Image Data Anatomy of the Spine) dataset is a comprehensive collection of lumbar spine MRI scans, developed through a collaborative initiative of the Valencian Region Health Public System, Spain, which

aggregates all spine MRI examinations from the 25 hospitals across the region. This dataset encompasses over 23,600 studies and approximately 124,800 MRI images from patients aged 20 to 80 years. Each imaging session is accompanied by a corresponding radiology report. Data acquisition was approved by the local ethics committee, and all information was anonymized by the BIMCV [11]. For standardization, the subset employed in this study was converted to NIfTI format for images and JSON for reports and metadata, following the MIDS protocol [12], ensuring that scans from the same patient were organized together.

The specific subset analyzed in this work consisted of 717 subjects with T2-weighted lumbar MRI. Semantic segmentation of the lumbar images was automatically generated using the model described by Sáenz-Gamboa et al. (2023) [13], retaining only the masks corresponding to intervertebral discs for radiomic analysis. Subsequently, three expert (two orthopedic surgeons and one neuroradiologist) manually reviewed the selected cases and assigned Pfirrmann grades to each disc.

2.2 Preprocessing and Exploratory Analysis

Initial technical characterization of the T2-weighted lumbar MRI scans was performed to assess spatial resolution and consistency prior to feature extraction. The average in-plane resolution was approximately 0.6 mm × 0.6 mm, with an average slice thickness of 5.4 mm (see Figure 1a). The image dimensions in pixels were highly uniform, with an average width and height of 540 pixels; however, the number of slices per volume was comparatively low (see Figure 1b). This configuration produces anisotropic volumes with high sampling density in the in-plane direction, but coarse resolution in the through-plane direction due to limited slice counts and large inter-slice spacing. Such anisotropy has important implications for radiomic analysis: consistent voxel spacing is required for fair feature comparison across subjects, but poor z-resolution undermines the reliability of 3D texture descriptors. Consequently, radiomic features were extracted in 2D from sagittal slices and aggregated at the disc level, and 3D descriptors were interpreted with caution.



(a) Spacing distribution for T2-weighted sequences.

(b) Pixel size distribution for T2-weighted sequences.

Figure 1: Initial technical characterization of T2-weighted lumbar MRI scans. (a) Average in-plane resolution and slice thickness. (b) Distribution of image dimensions in pixels.

Figure 2 shows an example of the anatomical segmentation. The left panel displays a mid-sagittal T2-weighted image of the lumbar spine, and the right panel overlays the corresponding disc masks from L1–L2 through L5–S1. This visualization highlights both the anatomical localization of intervertebral discs and the precise regions of interest used for radiomic feature extraction.

Class distribution was examined to assess prevalence of Pfirrmann grades at each lumbar level (Figure 3). Across all levels, Grades 3 and 4 were dominant, with Grade 3 most frequent in L1–L4 and Grade 4 more prevalent in L4–L5 and L5–S1. Grades 1 and 5 were consistently rare, each representing fewer than 5% of discs at every level. This imbalance has two main consequences. First, minority classes (Grades 1 and 5) provide limited data for training, making them more challenging to classify accurately and potentially biasing models toward the majority classes. Second, the level-specific prevalence of degeneration, such as the higher frequency of Grade 4 at L4–L5, indicates that anatomical context influences grade distribution and should be considered in subsequent analyses.

Before quantitative feature extraction, all images underwent standardized preprocessing to enhance consistency and suppress acquisition artifacts. Low-frequency intensity inhomogeneities were corrected using the N4ITK bias field correction algorithm, reducing scanner-related variations and improving the reliability of first-order and texture feature [14]. Random noise was suppressed with Perona–Malik anisotropic diffusion filtering, which preserves relevant

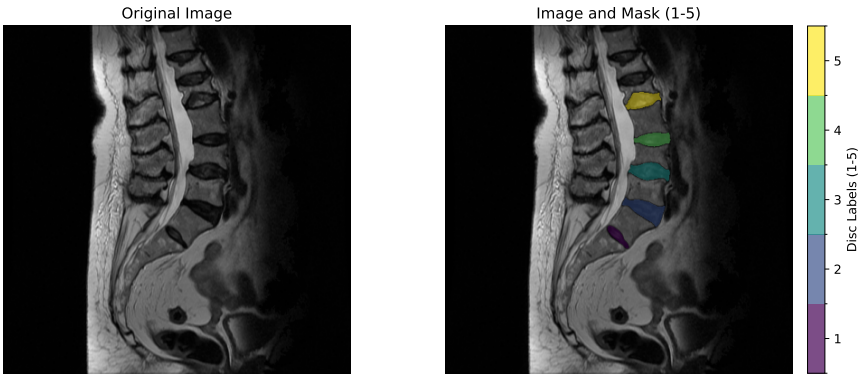


Figure 2: Original Sagittal T2 Image and Labeled Disc Masks (Grades 1–5)

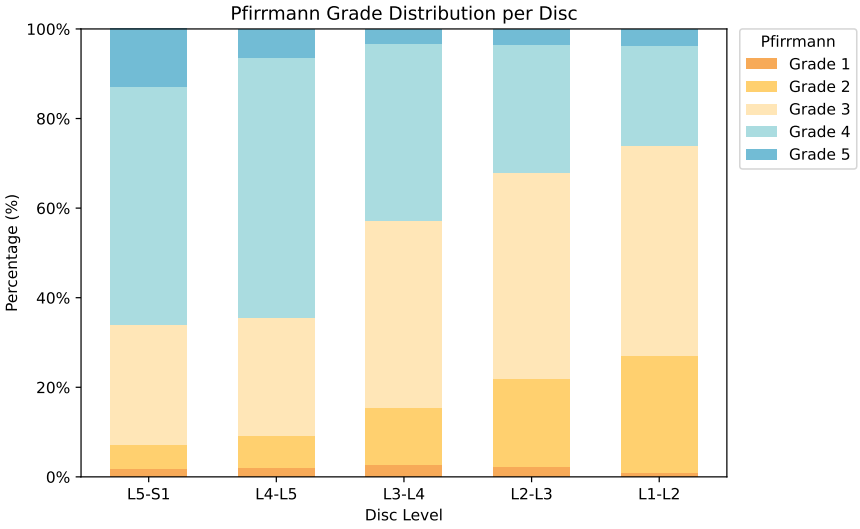


Figure 3: Distribution of Pfirrmann grades across lumbar disc levels (L1–L2 to L5–S1). Grades 3 and 4 dominate the dataset, with Grade 3 most prevalent at upper levels (L1–L4) and Grade 4 more common at lower levels (L4–L5, L5–S1). Grades 1 and 5 are consistently underrepresented (<5% at each level), creating a pronounced class imbalance with implications for model training and evaluation.

structural boundaries while smoothing background variation. To prevent numerical artefacts during subsequent processing steps, voxel intensities were converted to 32-bit floating-point precision. Figure 4 illustrates the effect of bias field correction on a representative scan.

2.3 Radiomic Feature Extraction

Radiomics was employed to transform the segmented intervertebral disc regions into high-dimensional quantitative descriptors of shape, intensity, and texture. Feature extraction was performed with PyRadiomics (v3.1.0) in Python (v3.11.11), following the recommendations of the Image Biomarker Standardization Initiative (IBSI) [15].

A critical step in radiomic texture analysis is the discretization of voxel intensities, since binning strongly influences both reproducibility and interpretability of texture features. For this dataset, an empirical, cohort-specific strategy was adopted. The intensity range of each disc ROI was computed, and the mean range across all discs was used to define a uniform bin width. Several bin counts (16, 32, 64, and 128) were evaluated to balance robustness against noise with preservation of fine structural detail. The final bin width selected from this analysis was applied consistently across all scans, ensuring comparable quantization for both first-order and texture-based features. The complete parameter configuration used in PyRadiomics is summarized in Table 1.

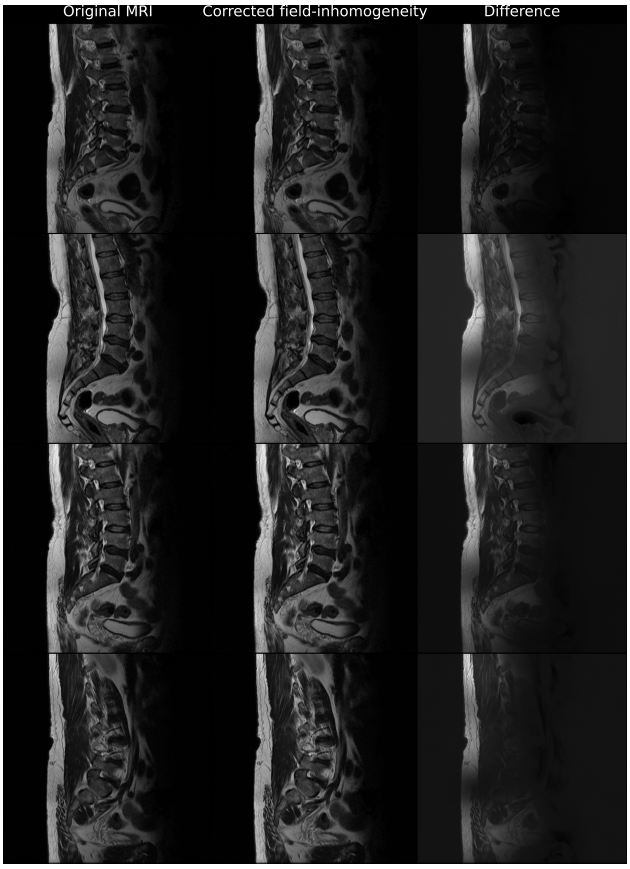


Figure 4: Effect of N4ITK bias field correction on T2-weighted lumbar MRI scans. The first column shows the original images with visible low-frequency intensity inhomogeneities, the second column shows the corrected images after bias field correction, and the third column highlights the difference between original and corrected scans. Correction reduces scanner-related intensity variations while preserving anatomical structures, thereby improving the reliability of subsequent radiomic feature extraction.

The extracted features were grouped into standard categories. First-order statistics characterized the intensity distribution within each disc, capturing metrics such as mean, variance, skewness, and kurtosis. Shape-based features quantified geometric properties, including disc volume, surface area, compactness, and elongation. Texture descriptors were computed from multiple matrix formulations of spatial intensity relationships, including the grey-level co-occurrence matrix (GLCM), grey-level run-length matrix (GLRLM), grey-level size-zone matrix (GLSZM), grey-level dependence matrix (GLDM), and neighborhood grey-tone difference matrix (NGTDM), each designed to capture heterogeneity at different scales. Finally, filtered image types such as wavelet decompositions, Laplacian of Gaussian (LoG) transforms, and non-linear intensity mappings were incorporated to highlight multi-scale and non-linear textural patterns (Table 2).

2.4 Predictive Modeling

The predictive task was defined in two complementary formulations. The primary formulation was a multiclass classification problem in which the model directly assigned discs to Pfirrmann grades 1 through 5. This setup preserved the full resolution of the grading scale but was challenged by the scarcity of extreme grades (1 and 5). Class imbalance of this type is a well-documented issue in machine learning, where underrepresented classes tend to be poorly modeled while majority classes dominate predictive performance [16]. To address both class imbalance and the observed separation in feature distributions, a secondary binary classification task was implemented. In this formulation, Grades 1–3 were grouped as low-to-moderate degeneration, while Grades 4–5 were grouped as advanced degeneration. This binary partition reduced sparsity, simplified the learning problem, and aligned with a clinically meaningful threshold distinguishing early from advanced disc degeneration.

Both tasks were approached with the same panel of machine learning classifiers selected for their complementary strengths in handling high-dimensional and heterogeneous radiomic features. These included Support Vector Machines

Parameter	Value	Description
normalize	true	Normalizes intensities to mean 0 and SD 1 before scaling
normalizeScale	100	Standard deviation scaling factor applied after normalization
preCrop	true	Crops images to the mask bounding box for efficiency
interpolator	sitkBSpline	Interpolator used for image resampling
resampledPixelSpacing	[0.5901, 0.5901, 0]	Target voxel spacing (0 in slice dimension due to 2D analysis)
force2D	true	Extracts textures in 2D per axial slice
force2Ddimension	2	Dimension (axial) used for 2D textures
geometryTolerance	1.0e+03	Tolerance for geometric mask validation
correctMask	true	Corrects minor mismatches between mask and image
binWidth	12.55	Fixed bin width for intensity discretization
voxelArrayShift	300	Intensity shift to avoid negative values after normalization
label	1	Primary ROI label for disc segmentation

Table 1: PyRadiomics parameter configuration used for feature extraction. Normalization, interpolation, voxel spacing, and discretization were standardized across all scans to ensure reproducibility and comparability of radiomic descriptors.

Type	Parameters	Purpose
Original	Default	Unfiltered image for baseline feature extraction
LoG	sigma: [0.6, 2, 3]	Enhances edges and multi-scale textural structures
Wavelet	level: 2	Decomposes image into frequency sub-bands
Square	Default	Amplifies higher intensity values
SquareRoot	Default	Reduces high dynamic range by compressing intensities
Logarithm	Default	Expands low intensities and compresses high values
Exponential	Default	Amplifies medium-to-high intensity regions

Table 2: Image types used for radiomic feature extraction. Each filtered representation emphasizes different intensity characteristics or spatial scales, providing complementary texture information to the original T2-weighted images.

(SVM), Logistic Regression, Random Forest, Naïve Bayes, k-Nearest Neighbors (k-NN), and Gradient Boosting. SVM and Logistic Regression provide strong baselines for linear and kernel-based separation in high-dimensional spaces. Random Forest and Gradient Boosting capture non-linear interactions and offer robustness to feature correlations. Naïve Bayes serves as a simple probabilistic baseline, while k-NN provides instance-based classification sensitive to local feature similarity. Using this diverse set of algorithms allowed comparison across linear, probabilistic, ensemble, and distance-based approaches.

All classifiers were embedded in a cross-validation framework designed to preserve patient independence and class proportions. Hyperparameters were kept fixed across folds within each experimental setup, ensuring that observed differences in performance could be attributed to the formulation of the prediction task and the feature selection strategy rather than to model tuning.

2.5 Feature Selection and Experimental Design

Feature selection was conducted independently within the training folds to prevent information leakage, with strategies adapted to the structure of the prediction task. In the multiclass formulation, where five Pfirrmann grades were compared, each feature was first tested for normality across the full set of classes using the Shapiro–Wilk test. Features with normally distributed values were evaluated with one-way ANOVA, while those deviating from normality were assessed with the non-parametric Kruskal–Wallis test. These methods are specifically designed for more than two groups, producing a single p-value that reflects whether a feature shows significant variation across the entire grade spectrum. Because the outcome had more than two categories, the area under the ROC curve (AUC) was not used at this stage, as it is formally defined for binary outcomes and would require one-vs-rest or pairwise reformulations. Such reformulations would complicate interpretation and inflate the number of statistical tests, increasing the risk of false discoveries. Instead, features were ranked by corrected p-values using the Benjamini–Hochberg procedure to control the false discovery rate. To limit overfitting in the high-dimensional feature space, the number of selected features was capped at one per fifteen samples.

In the binary formulation, where discs were grouped into low-to-moderate (Grades 1–3) versus advanced (Grades 4–5) degeneration, pairwise comparisons were possible. After testing for normality with Shapiro–Wilk, features with normally distributed values in both groups were evaluated using the two-sample Student's t-test, while the Mann–Whitney U-test was applied to features not meeting normality assumptions. In addition to p-values, discriminative power was quantified by computing the AUC for each feature, which directly measures its ability to separate the two groups. For features with AUC below 0.5, class orientation was inverted to maintain interpretability. Sensitivity, specificity, and optimal thresholds were derived using Youden's J index, providing clinically relevant indicators alongside statistical significance. As in the multiclass setup, the number of features retained was constrained to one per fifteen samples, with selection based on lowest corrected p-values.

Model training and evaluation were performed within a repeated StratifiedGroupKFold cross-validation framework. Stratification preserved the relative distribution of classes in each fold, while grouping by patient identifier ensured that all discs from the same subject were assigned to the same split, thereby preventing information leakage across training and validation data. Repeated resampling provided a more stable estimate of generalization performance by averaging across different fold compositions. This design ensured that comparisons between multiclass and binary formulations reflected differences in problem definition and feature selection rather than biases in data partitioning.

2.6 Model Optimization

After cross-validation, the classifier with the best overall performance was selected for further refinement. This model was retrained on the training partition and optimized using a Bayesian hyperparameter search. The search was carried out under the same group- and stratification-aware splitting strategy to prevent patient leakage, and optimization was guided by a multi-metric objective that prioritized the area under the ROC curve (AUC) while simultaneously tracking F1-score and balanced accuracy. The configuration that maximized AUC was retained for final evaluation.

A hold-out test set, defined using a patient-wise split, was reserved for independent evaluation of the optimized model. Performance on this set was quantified with a comprehensive panel of discrimination, calibration, and agreement metrics, including accuracy, sensitivity, specificity, predictive values, Cohen's κ, and the Matthews correlation coefficient. A confusion matrix was also generated to provide a transparent overview of classification errors across classes.

2.7 Interpretability Analysis

Model interpretability was assessed using SHAP (SHapley Additive exPlanations), which assigns each feature a contribution value based on its influence on the model's output. Global interpretability was explored through summary plots, beeswarm diagrams, and feature–value scatterplots, which ranked features by importance and showed how their values shifted predictions toward lower or higher Pfirrmann grades. Class-specific heatmaps further illustrated how the impact of features varied across different levels of degeneration.

To provide statistical support, SHAP value distributions were compared between classes using the Mann–Whitney U-test with Holm correction, identifying features with significantly different contributions across groups. This combined approach offered both a visual and quantitative understanding of how the optimized classifier reached its decisions. The full methodological workflow, from preprocessing to interpretability, is summarized in Figure 5.

3 Results

The results are presented for two formulations: multiclass Pfirrmann grading (Grades 1–5) and binary classification (Grades 1–3 vs. Grades 4–5). Both were evaluated under patient-level cross-validation, with AUC as the primary metric and per-class F1 scores to capture balance between precision and recall.

3.1 Multiclass Classification

In the multiclass setting, discs were classified into the five Pfirrmann grades. Six classifiers were compared using patient-level cross-validation, with the area under the curve (AUC) as the primary outcome and per-class F1 scores as secondary metrics to capture class-specific balance between precision and recall. This dual evaluation was necessary due to the significant class imbalance, particularly the scarcity of Grades 1 and 5, which can lead to inflated global metrics while masking poor sensitivity to rare classes.

Across 50 evaluations, Gradient Boosting was found to be the most effective method, achieving the highest mean AUC and consistently strong per-class F1 scores (Table 3). A Friedman test, followed by Wilcoxon signed-rank tests with Holm correction, confirmed significant overall differences among classifiers ($p < 1 \times 10^{-36}$). The ROC curves

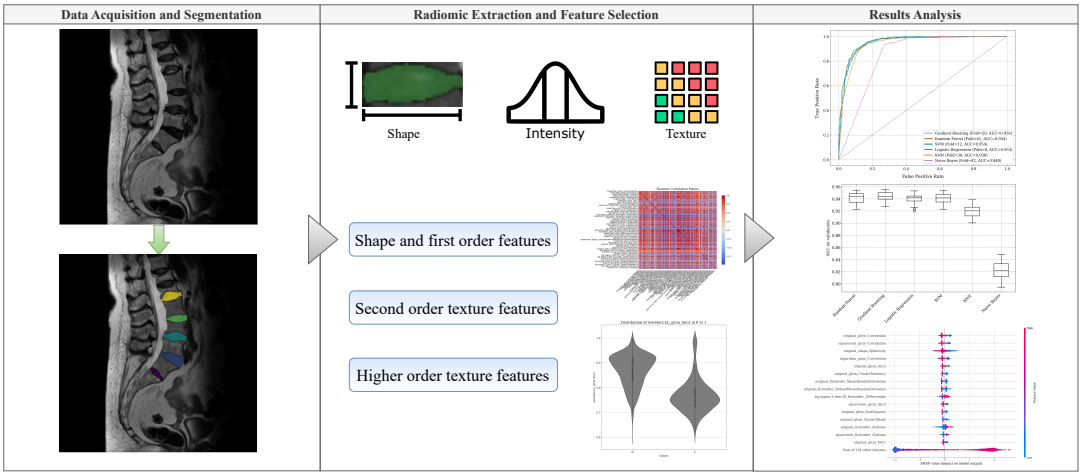


Figure 5: Overall methodological workflow. The pipeline begins with **data acquisition and segmentation**, where intervertebral discs are identified on sagittal T2-weighted lumbar MRI scans. In the next step, **radiomic features** are extracted and selected, covering shape descriptors, first-order intensity statistics, second-order texture metrics, and higher-order filtered features. Finally, **results analysis and interpretability** includes model training, cross-validation, performance evaluation (ROC curves, boxplots of metrics), and SHAP-based feature attribution. This workflow illustrates the end-to-end process from raw image data to interpretable predictive models.

(Figure 6) illustrate model performance for the median-performing fold and the optimal fold, showing that Gradient Boosting and Random Forest maintain higher true-positive rates across the range of false-positive rates than linear models such as Logistic Regression and SVM, which are less flexible with non-linear feature interactions.

Classifier	AUC val (mean ± std)	F1 class 0	F1 class 1	F1 class 2	F1 class 3	F1 class 4
Gradient Boosting	**0.873 ± 0.014**	0.066 ± 0.096	0.400 ± 0.051	**0.674 ± 0.024**	**0.811 ± 0.016**	**0.595 ± 0.064**
SVM	**0.877 ± 0.013**	0.117 ± 0.060	**0.485 ± 0.039**	0.562 ± 0.032	0.784 ± 0.020	0.584 ± 0.058
Random Forest	0.865 ± 0.017	0.010 ± 0.036	0.361 ± 0.060	0.676 ± 0.023	0.807 ± 0.018	0.418 ± 0.070
Logistic Regression	0.838 ± 0.027	**0.128 ± 0.075**	0.484 ± 0.043	0.569 ± 0.026	0.776 ± 0.019	0.567 ± 0.049
KNN	0.778 ± 0.020	0.069 ± 0.081	0.384 ± 0.051	0.624 ± 0.027	0.786 ± 0.018	0.476 ± 0.066
Naive Bayes	0.806 ± 0.022	0.143 ± 0.141	0.447 ± 0.051	0.437 ± 0.035	0.750 ± 0.021	0.254 ± 0.050

Table 3: Performance comparison of six classifiers for multiclass Pfirrmann grading. Mean validation AUC (± standard deviation) and per-class F1 scores are reported. Best results per column are highlighted in bold. Gradient Boosting achieved the highest overall mean AUC and strong F1 performance across the majority of classes.

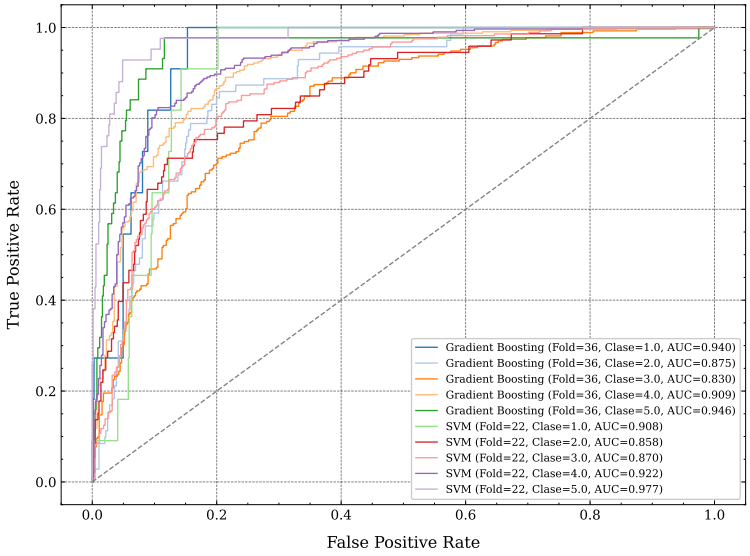


Figure 6: ROC curves for multiclass Pfirrmann grading. Optimal folds highlight superior discrimination by Gradient Boosting and Random Forest, compared to linear classifiers.

Despite these strong global results, analysis by class revealed inconsistent outcomes across the Pfirrmann spectrum. Grades 3 and 4 dominated the dataset and were predicted with relatively high F1 scores, reflecting the availability of sufficient training examples. In contrast, grades 1 and 5 were associated with low F1 scores, indicating poor sensitivity for rare cases. This imbalance highlights a limitation of multiclass Pfirrmann grading: while mid-range classes can be reliably identified, extreme grades remain difficult due to a lack of examples.

To refine the best-performing model, the Gradient Boosting algorithm was retrained using the training partition, with the hyperparameters optimised via Bayesian search. The final configuration is summarized in Table 4.

Hyperparameter	**Selected Value**
Learning rate	0.0094
Maximum depth	2
Maximum features	`sqrt`
Number of estimators	707
Subsample	0.5

Table 4: Best hyperparameters for Gradient Boosting in the multiclass classification task, obtained through Bayesian optimization.

Model interpretability analyses provided additional insights. SHAP beeswarm plots (Figure 7) ranked radiomic features according to their overall impact on different classes. This revealed that texture- and intensity-derived descriptors were most effective in distinguishing degeneration grades. Features with high values tended to push predictions towards advanced grades, while lower values were associated with early degeneration.

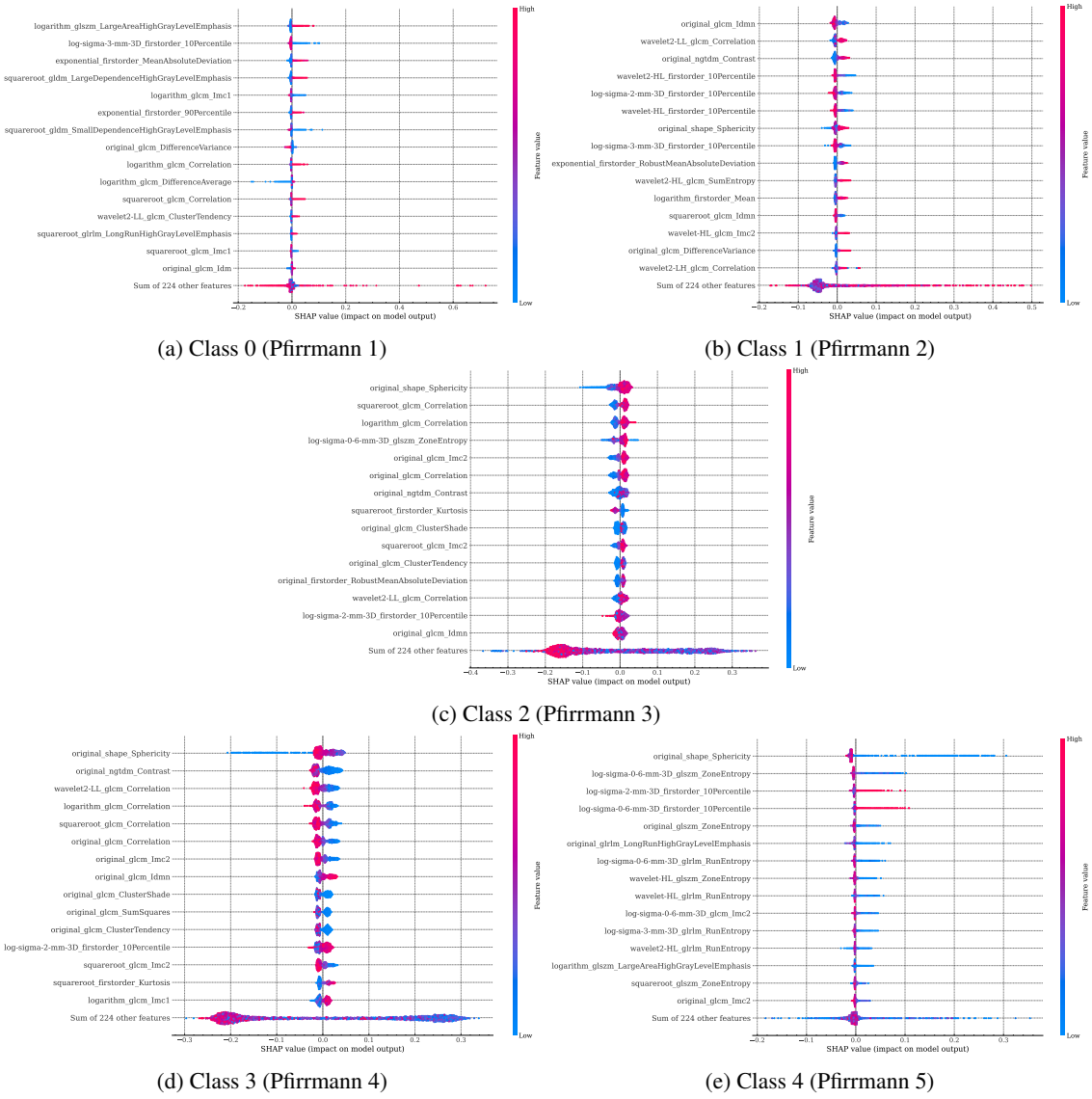
The heatmaps in Figure 8 further decomposed these patterns at the patient level. Each row corresponds to a feature and each column to a patient, enabling the visualisation of how individual features influence class assignment. The consistent presence of high SHAP values for specific features in mid-range grades demonstrates their stable discriminative power. Meanwhile, the more diffuse patterns for Grades 1 and 5 confirm the challenges posed by limited training data.

Finally, the SHAP summary bar plot (Figure 9) aggregates the absolute feature contributions across Pfirrmann grades. This representation highlights a core group of radiomic descriptors, particularly sphericity and several GLCM-derived correlations, that consistently influenced classification across all classes. At the same time, distinct subsets of features were more relevant for transitions between specific grades, such as between Grades 2 and 3 or between Grades 3 and 4. This overlap in feature relevance across adjacent grades provides a mechanistic explanation for the misclassification patterns observed in the confusion matrices Figure 10, where most errors occurred between neighboring categories rather than between the extremes of degeneration.

3.2 Binary Classification

Given the strong class imbalance observed in the multiclass setting, particularly the scarcity of Grades 1 and 5, the problem was reformulated into a binary task. This grouping alleviates data imbalance and reflects a clinically meaningful threshold. Pfirrmann grades 1–3 correspond to none-to-moderate degeneration (class 0), while grades 4–5 represent advanced degeneration (class 1). The multiclass SHAP heatmaps further supported the decision to collapse categories (Figure 8), as radiomic features consistently separated lower-to-intermediate grades from advanced degeneration, suggesting a natural division at this threshold.

Using the same patient-level cross-validation protocol as in the multiclass task, binary classification produced markedly better results. Table 5 summarizes the comparative performance of the six classifiers, with Gradient Boosting achieving the highest mean AUC (0.944 ± 0.007) and balanced F1 scores across both classes. SVM, Random Forest, and Logistic Regression performed similarly well, all exceeding an AUC of 0.94, whereas k-Nearest Neighbors (KNN) and Naïve



(a) Class 0 (Pfirrmann 1)

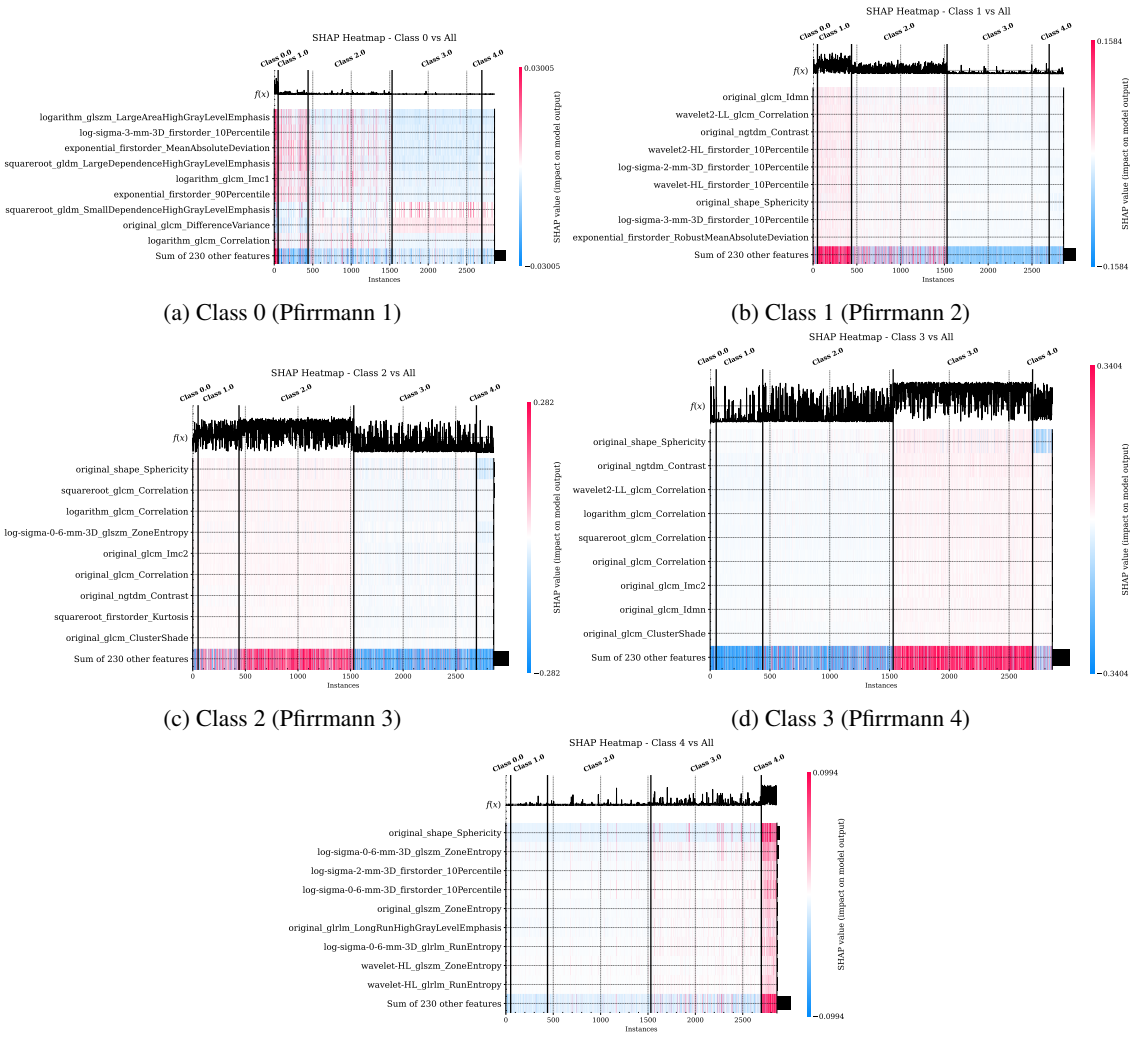
(b) Class 1 (Pfirrmann 2)

(c) Class 2 (Pfirrmann 3)

(d) Class 3 (Pfirrmann 4)

(e) Class 4 (Pfirrmann 5)

Figure 7: SHAP beeswarm plots showing the most influential radiomic features across Pfirrmann grades. Each subfigure corresponds to one class, with features ranked by mean absolute SHAP value. Point position indicates the direction and magnitude of feature contributions, while color encodes feature values (red = high, blue = low). Texture- and intensity-derived descriptors dominate the separation between classes, whereas Grades 1 and 5 display more diffuse patterns due to class imbalance.



(a) Class 0 (Pfirrmann 1)

(b) Class 1 (Pfirrmann 2)

(c) Class 2 (Pfirrmann 3)

(d) Class 3 (Pfirrmann 4)

(e) Class 4 (Pfirrmann 5)

Figure 8: SHAP heatmaps illustrating patient-level feature contributions across Pfirrmann grades. Each column corresponds to a patient, while rows represent radiomic features ranked by importance. Color intensity reflects the magnitude and direction of SHAP values, highlighting which features drive predictions for each grade. Consistent patterns emerge for mid-range grades (2–3), whereas extreme classes (1 and 5) show more heterogeneous feature attributions due to limited sample size.

Bayes lagged behind. Notably, the per-class F1 scores revealed a significant improvement in sensitivity to advanced degeneration (Grades 4–5, Class 1) compared to the multiclass formulation, while maintaining strong performance in the none-to-moderate group (Grades 1–3, Class 0).

Figure 11 shows ROC curves for both the median-performing fold (Figure 11a) and the optimal fold (Figure 11b). The clear separation between true-positive and false-positive rates demonstrates the stability of Gradient Boosting, which consistently dominated across folds. Together with the class-specific F1 results in Table 5, these findings confirm the advantage of the binary grouping for both discrimination and robustness.

To refine the best-performing model, the Gradient Boosting algorithm was retrained using the training partition, with the hyperparameters optimised via Bayesian search. The final configuration is summarized in Table **??**.

These results were reinforced by interpretability analyses. The SHAP beeswarm plot (Figure 12a) ranked features according to their predictive value for advanced degeneration, identifying texture- and intensity-based descriptors as the dominant classification drivers. Features such as `GLCM Correlation`, `ClusterTendency`, and `Imc2` consistently appeared as top contributors. Higher values of correlation- and homogeneity-related features were associated

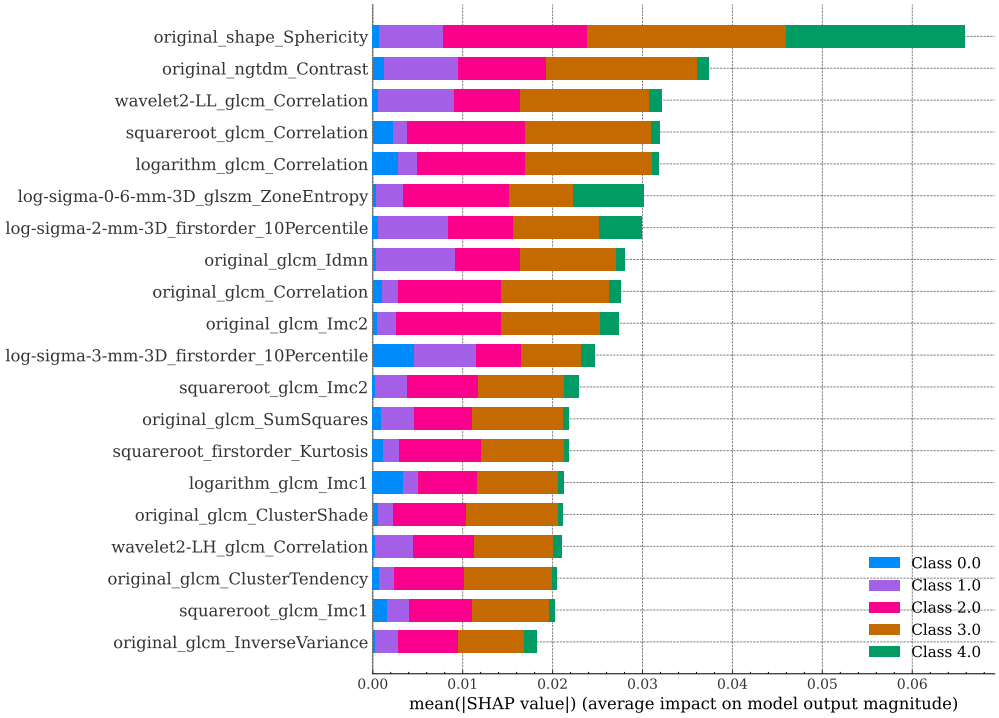


Figure 9: SHAP summary bar plot showing the mean absolute contribution of the top radiomic features across Pfirrmann grades. Each bar is color-coded by grade, indicating the relative importance of features for each class. While some features (e.g., sphericity, GLCM correlation) are broadly influential, others are class-specific and help discriminate between adjacent grades.

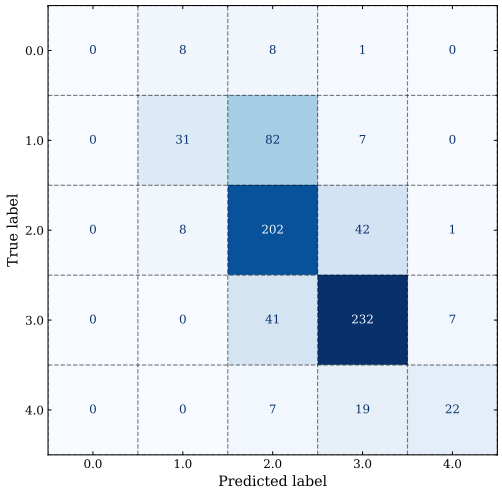


Figure 10: Confusion Matrix for Multiclass Classification. The model achieved good performance for the mid-range classes (1-3), which dominates the dataset, while extreme classes showed lower sensitivity due to their scarcity. Most misclassifications occurred between adjacent grades (e.g., 2 vs. 3, 3 vs. 4), reflecting the gradual nature of disc degeneration and the difficulty of sharply separating neighboring categories.

with non-degenerated or moderately degenerated discs (Class 0), while lower values pushed predictions towards advanced degeneration (Class 1). Conversely, dispersion-based features such as `MeanAbsoluteDeviation` and `RobustMeanAbsoluteDeviation` increased the probability of advanced degeneration when elevated, reflecting the greater heterogeneity present in late-stage disc degeneration.

Classifier	AUC val (mean ± std)	F1 class 0	F1 class 1
Gradient Boosting	**0.944 ± 0.007**	**0.878 ± 0.014**	**0.864 ± 0.016**
SVM	0.941 ± 0.009	0.877 ± 0.013	0.870 ± 0.014
Random Forest	0.941 ± 0.008	0.877 ± 0.014	0.863 ± 0.016
Logistic Regression	0.940 ± 0.007	0.877 ± 0.014	0.866 ± 0.014
KNN	0.919 ± 0.010	0.868 ± 0.015	0.850 ± 0.017
Naïve Bayes	0.823 ± 0.013	0.777 ± 0.017	0.810 ± 0.017

Table 5: Binary classification performance across classifiers. Gradient Boosting achieved the best results overall, with the highest mean AUC and balanced F1 scores for both non-advanced (Class 0) and advanced (Class 1) degeneration.

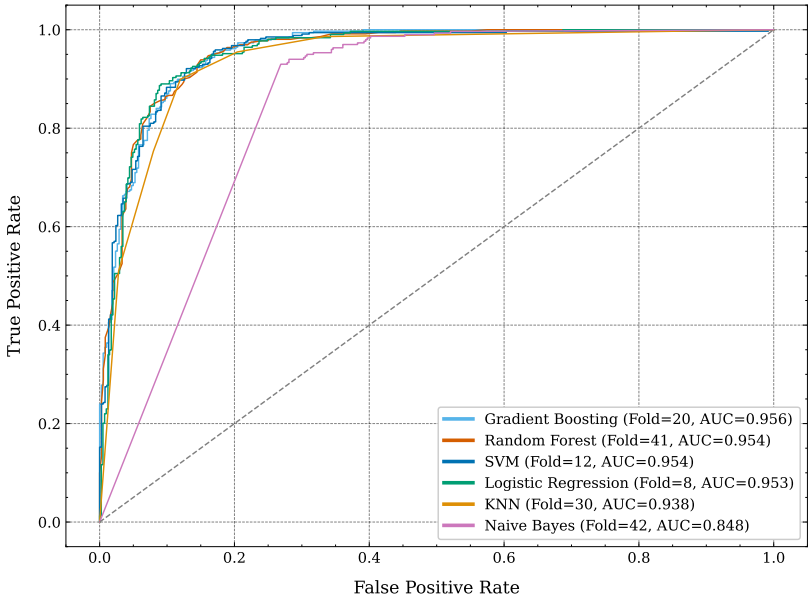


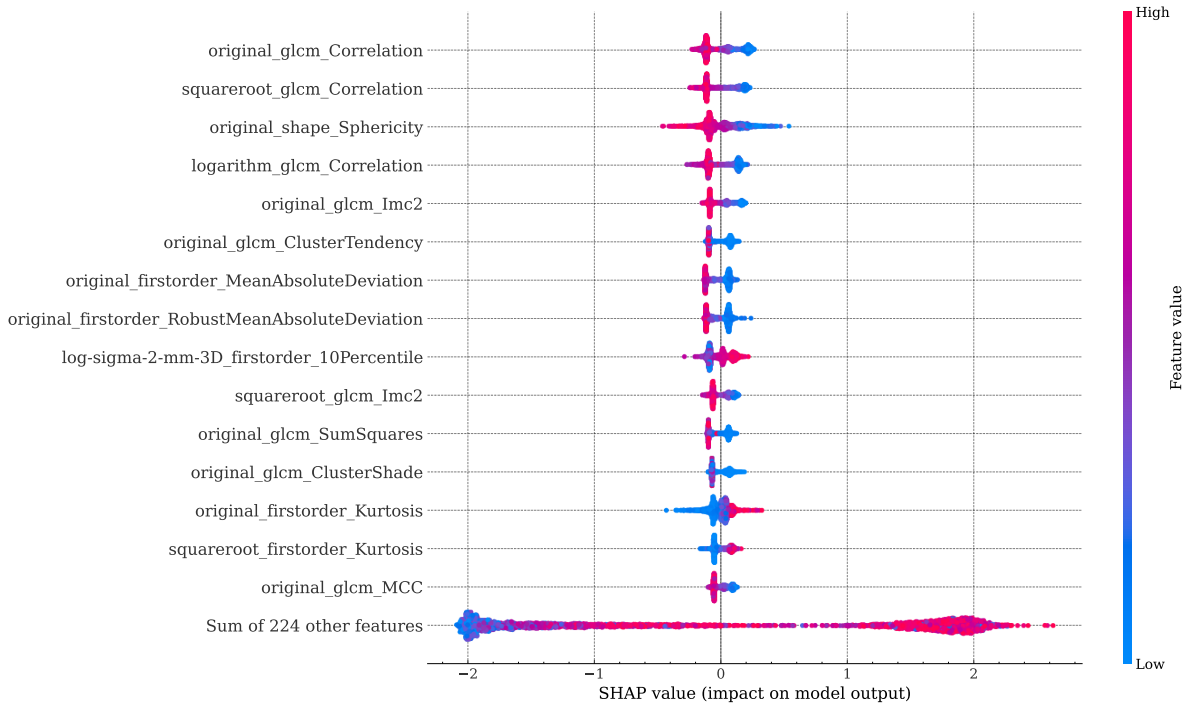
Figure 11: ROC curves for binary classification. Gradient Boosting consistently achieved the highest discrimination, outperforming other classifiers in optimal folds.

Hyperparameter	Selected value
Learning rate	0.0250
Max depth	5
Max features	log2
Number of estimators	747
Subsample fraction	0.952

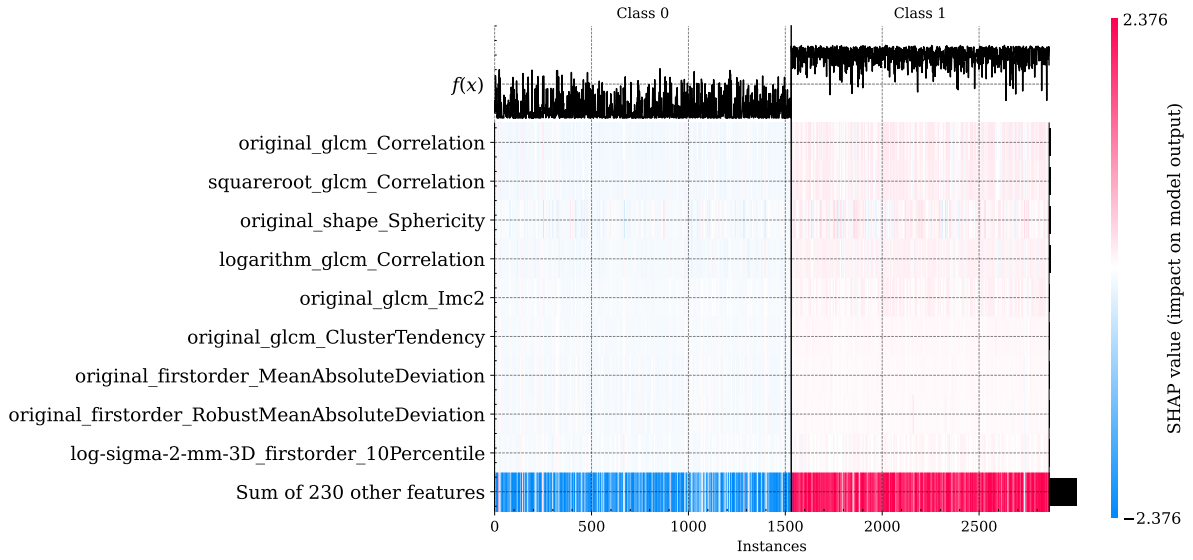
Table 6: Best hyperparameters of the Gradient Boosting model after Bayesian optimization.

The SHAP heatmap (Figure 12b) provided a patient-level perspective, showing how feature contributions aggregated across individuals. A clear separation was visible between the two classes: discs in Class 0 were characterized by high homogeneity and brightness-related features, whereas discs in Class 1 exhibited elevated heterogeneity and shape irregularities. This consistency across patients confirms that the selected subset of radiomic features generalizes well to unseen data and captures physiologically meaningful patterns of degeneration.

Moreover, the edge cases merge with their physiologically closer neighbours in the binary confusion matrix (Figure 13), improving balance and clinical utility. Errors are more evenly distributed, with far fewer occurring relative to class size.



(a) SHAP beeswarm plot ranking features by contribution to binary classification.



(b) SHAP heatmap showing patient-level feature contributions for binary classification.

Figure 12: SHAP-based interpretability analysis for binary classification. Texture- and intensity-derived features dominate predictions, with consistent separation between non-advanced and advanced degeneration across patients.

4 Discussion

4.1 Multiclass Classification

The comparative from Table 3 and the ROC curves (Figure 6) shows a clear split between models. The perfomance of Gradient Boosting and SVM achieve the highest overall AUC values, indicating superior ability to distinguish between Pfirrmann grades, while maintaining competitive F1 scores in most classes. Instead, the remaining models—particularly

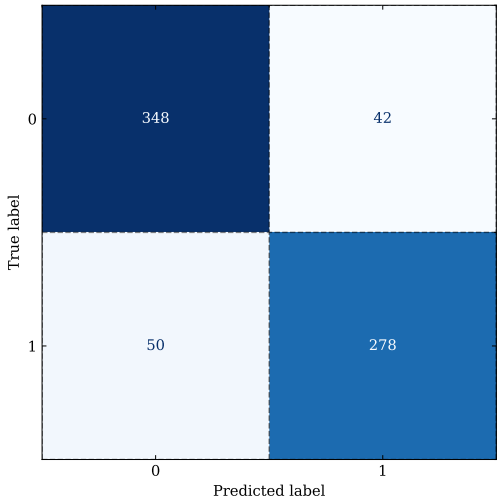


Figure 13: Confusion Matrix for Binary Classification. Where Pfirrmann Grades 1–3 were grouped as non-advanced (Class 0) and Grades 4–5 as advanced degeneration (Class 1). The matrix shows that the model achieved strong sensitivity and specificity, with relatively balanced classification between the two groups.

KNN and Naïve Bayes—exhibit lower AUC values and greater variability in F1 scores, indicating less consistent and less reliable classification performance, especially for the extreme grades.

Gradient Boosting was selected for further analysis due to its ability to achieve the highest mean AUC 0.87 and strong per-class F1 for Grades 2-4 where most cases reside. Across all models, F1 scores are lowest for Class 0 (Pfirrmann 1) and highest for mid-range grades (Classes 2–3), this pattern reflects the strong class imbalance.

The SHAP beeswarm plots (Figure 7) for each Pfirrmann grade show consistent patterns in the contribution of specific radiomic features to classification across the degeneration range. For lower grades (Pfirrmann 1–2), shape-related features such as sphericity and certain GLCM correlations generally show higher values, pushing predictions towards these classes, suggesting that more regular disc shapes and higher texture uniformity are associated with healthy or mildly degenerated discs. For intermediate grades (Pfirrmann 3), texture descriptors such as GLCM, NGTDM and entropy-based metrics have more importance, with a balanced distribution of SHAP contributions from high and low values. This reflects a transitional stage where both increases in heterogeneity and losses in uniformity influence the classification. For higher grades (Pfirrmann 4–5), shape features like sphericity also show an inversion in impact, low values now move predictions toward these grades, consistent with morphological irregularities in advanced disc degeneration.

In the heatmaps (Figure 8) a clear natural split can be seen between lower grades (Pfirrmann 1–3) and higher grades (Pfirrmann 4–5). For instance, the SHAP heatmaps of Classes 3 vs. All and 4 vs. All show that the features that move predictions towards advanced grades (e.g. reduced sphericity, higher entropy and lower percentiles) differ markedly from those favouring earlier grades (e.g. higher correlation, higher sphericity and brighter intensities). This transition point closely aligns with the later adopted binary grouping, which reinforces the idea that the shift in radiomic feature patterns around Grades 3–4 is both physiologically meaningful and statistically separable.

Furthermore, in the SHAP summary bar plot can be seen the relative importance of features across classes. Texture-based descriptors dominate the ranking, reflecting the central role of image heterogeneity and homogeneity in grading degeneration. Interestingly, shape sphericity is among the top contributors, despite the limited number of samples in Grades 1 and 5. This indicates that morphological changes can strongly influence classification, even in underrepresented classes.

4.2 Binary Classification

The binary reformulation addressed the imbalance in the multiclass setup, particularly the lack of Grades 1 and 5, and matched a clinically meaningful division.

As shown in Table 5 and Figure 11, it is possible to improve the performance of the model when transitioning from multiclass to binary classification. Gradient Boosting, SVM, Random Forest, and Logistic Regression all achieved near-identical AUC values above 0.94, indicating excellent discrimination between non-advanced and advanced degeneration.

The F1 scores for both classes were also well balanced, suggesting that the models maintained sensitivity to advanced cases (Class 1) without compromising the accurate identification of non-advanced discs (Class 0).

Table 2 summarizes the comparative results, showing that Gradient Boosting, SVM, Random Forest, and Logistic Regression all achieved very high AUC values (> 0.94) and balanced F1 scores for both classes, while Naive Bayes performed less well. Gradient Boosting again emerged as the selected model due to its combination of highest AUC, stable per-class F1 scores, and consistent ROC curve performance across folds.

Compared to the multiclass setup, the binary grouping reduced the challenge of class imbalance and increased class separability, as reflected in the sharper ROC curves and smaller performance variability. This improvement is consistent with the SHAP heatmap observations from the multiclass analysis, where features already showed a natural division between Class 0 and Class 1, indicating that this threshold captures the most robust radiomic separation in the dataset.

For the SHAP values analysis, in the beeswarm (Figure 12a) feature contributions align with the expected imaging patterns of disc degeneration. High values of homogeneity and brightness-related features (e.g., GLCM Correlation, IMC2, ClusterTendency) predominantly appear on the left side in blue for Class 1, indicating that when these features are high, the model is more likely to predict Class 0. Conversely, higher measures of heterogeneity and dispersion (MeanAbsoluteDeviation, RobustMAD, ClusterShade, SumSquares) appear as red points to the right for Class 1, showing their association with advanced degeneration. Sphericity also plays a role, with lower values pushing predictions toward Class 1, reflecting morphological changes in degenerated discs.

Also, the heatmap, Figure 12b, clearly shows the separation between the two classes, with patients naturally grouping according to their degeneration status. In Class 0, the dominant pattern is high values for homogeneity and brightness-related features, which are shown as blue zones that push predictions towards the 'non-degenerated' label. In contrast, Class 1 shows higher levels of dispersion and heterogeneity features, such as MeanAbsoluteDeviation, RobustMAD and texture variance, which appear as red regions that drive predictions towards the advanced degeneration label.

4.3 Comparison between classifications

The multiclass formulation (Grades 1–5) provided a fine-grained view of disc degeneration but suffered from severe class imbalance, particularly for Grades 1 and 5. As shown in the ROC curves, Gradient Boosting emerged as the most reliable classifier, yet its performance was uneven across classes: mid-range grades (2–4) were identified with relatively high sensitivity, while extreme grades were poorly captured.

Across the five classes, a coherent progression of radiomic patterns is observed. Early grades (1–2) are characterised by higher homogeneity/brightness (e.g., GLCM correlation, IMC2/ClusterTendency, higher low-percentile intensities) and more regular shape (higher sphericity), which collectively push predictions toward these classes. Grade 3 shows a transitional texture, with rising heterogeneity (e.g., MAD/RobustMAD, entropy, NGTDM contrast) alongside partial preservation of homogeneity features—consistent with mixed dehydration and early fissuring. Advanced grades (4–5) display marked heterogeneity (greater dispersion and entropy, stronger run/zone complexity) and loss of sphericity, together with lower intensity percentiles, which drive predictions toward severe degeneration. These trends appear consistently in the class-specific beeswarms and are echoed by the class-wise SHAP distributions.

When grades are grouped into Binary Class 0 (1–3) and Binary Class 1 (4–5), the same signatures condense cleanly: Class 0 aggregates the homogeneous, bright, and spherical phenotype of early–moderate discs, whereas Class 1 concentrates the heterogeneous, darker, and less spherical phenotype of advanced degeneration. The binary heatmap shows a clear block separation of these patterns across patients, and the binary beeswarm preserves the same feature directions observed in multiclass—simply with greater stability. In short, the binary formulation captures the natural divide already visible in the multiclass features while mitigating the instability introduced by the under-represented extreme grades.

Overall, the multiclass approach enabled grade-specific interpretation. However, its clinical utility was limited by the scarcity of data in extreme grades. By contrast, the binary classification produced stronger, more reliable results that aligned with both radiological physiology and clinical practice. These patterns are consistent with prior radiomics studies: in a large lumbar cohort, McSweeney et al. [6] reported that 2D sphericity and intensity interquartile range are robust Pfirrmann correlates and that radiomics showed that quantitative radiomic signatures are stronger predictors of degeneration severity than the conventional "by-eye" or simple quantitative markers radiologists usually rely on. Also, in a two-centre cervical study, Xie et al. [5] showed that higher-order/texture features (GLCM, GLRLM, GLSZM, GLDM, NGTDM, LoG, wavelets) dominate the diagnostic signal. Together, these findings reinforce our observation that early grades are characterized by homogeneity/brightness and preserved shape, whereas advanced disease shows greater heterogeneity and loss of sphericity, supporting the 1–3 vs. 4–5 binary threshold.

This study has several strengths. The MRI data were collected across 25 hospitals of the Valencian Health System, providing heterogeneous but standardized material that enhances generalizability. The large dataset, centralized curation, and patient-level cross-validation strengthen the robustness of the analysis, while the use of interpretability methods (SHAP) increases transparency and supports clinical relevance. Nevertheless, some limitations should be acknowledged. The distribution of Pfirrmann grades was highly imbalanced, with very few cases of extreme degeneration, and the anisotropic resolution of the scans constrained the use of 3D texture features. In addition, the Pfirrmann system introduces interobserver variability, binary reformulation reduced granularity, and external validation was not performed. Finally, although all MRI exams were clinically indicated, the underlying reasons for imaging referral were not included in the analysis and may represent unmeasured confounders.

5 Conclusion

In conclusion, this study examined whether radiomic features extracted from T2-weighted lumbar spine MRIs could be used to classify intervertebral disc degeneration according to the Pfirrmann grading system. Gradient Boosting was found to be the most consistent model, offering both strong discrimination and physiologically coherent feature attributions. Although the multiclass approach identified grade-specific patterns, the limited number of Grades 1 and 5 resulted in poor performance. Reformulating the task into a binary classification (non-advanced vs. advanced degeneration) yielded stronger, more stable results that align more closely with clinical decision-making. Importantly, feature attribution analyses confirmed that homogeneity and shape metrics dominate early to moderate degeneration, whereas heterogeneity and irregularity features become increasingly relevant in advanced stages.

Despite limitations such as the use of data from a single centre and an imbalance in extreme grades, this work provides evidence that interpretable, radiomics-driven models can complement radiological assessment, reduce variability in grading and support a more consistent evaluation of lumbar disc degeneration. Future work will focus on addressing the scarcity of extreme grades to improve class balance and model robustness.

References

[1] Alize J Ferrari, Damian Francesco Santomauro, Amirali Aali, Yohannes Habtegiorgis Abate, Cristiana Abbafati, Hedayat Abbastabar, Samar Abd ElHafeez, Michael Abdelmasseh, Sherief Abd-Elsalam, Arash Abdollahi, Auwal Abdullahi, Kedir Hussein Abegaz, Roberto Ariel, Richard Gyan Aboagye, Hassan Abolhassani, Lucas Guimarães Abreu, Hasan Abualruz, Eman Abu-Gharbieh, Niveen ME Abu-Rmeileh, and Ilana N Ackerman. Global incidence, prevalence, years lived with disability (ylds), disability-adjusted life-years (dalys), and healthy life expectancy (hale) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet*, 403(10440):2133–2161, Apr 2024.

[2] P. P. Vergroesen, I. Kingma, K. S. Emanuel, R. J. Hoogendoorn, T. J. Welting, B. J. van Royen, J. H. van Dieën, and T. H. Smit. Mechanics and biology in intervertebral disc degeneration: a vicious circle. *Osteoarthritis and Cartilage*, 23(7):1057–1070, 2015.

[3] Christian W. A. Pfirrmann, Alexander Metzdorf, Marco Zanetti, Juerg Hodler, and Norbert Boos. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine*, 26(17):1873–1878, Sep 2001.

[4] L. P. Yu, W. W. Qian, G. Y. Yin, Y. X. Ren, and Z. Y. Hu. Mri assessment of lumbar intervertebral disc degeneration with lumbar degenerative disease using the pfirrmann grading systems. *PLoS One*, 7(12):e48074, 2012.

[5] Jun Xie, Yi Yang, Zekun Jiang, Kerui Zhang, Xiang Zhang, Yuheng Lin, Yiwei Shen, Xuehai Jia, Hao Liu, Shaofen Yang, Yang Jiang, and Litai Ma. Mri radiomics-based decision support tool for a personalized classification of cervical disc degeneration: a two-center study. *Frontiers in Physiology*, Volume 14 - 2023, 2024.

[6] Terence McSweeney, Aleksei Tiulpin, Narasimharao Kowlagi, Juhani Määttä, Jaro Karppinen, and Simo Saarakkala. Robust radiomic signatures of intervertebral disc degeneration from mri. *Spine*, Jun 2025.

[7] Zheng Fan, Tong Wu, Yang Wang, Zhuoru Jin, Tong Wang, and Da Liu. Deep-learning-based radiomics to predict surgical risk factors for lumbar disc herniation in young patients: A multicenter study. *Journal of Multidisciplinary Healthcare*, Volume 17:5831–5851, Dec 2024.

[8] Ali Muhaimil, Saikiran Pendem, Niranjana Sampathilla, Priya P S, Kaushik Nayak, Krishnaraj Chadaga, Anushree Goswami, Obhuli Chandran M, and Abhijit Shirlal. Role of artificial intelligence model in prediction of low back pain using t2 weighted mri of lumbar spine. *F1000Research*, 13:1035–1035, Oct 2024.

[9] V. J. Climent-Peris, L. Martí-Bonmatí, A. Rodríguez-Ortega, and J. Doménech-Fernández. Predictive value of texture analysis on lumbar MRI in patients with chronic low back pain. *European Spine Journal*, 32(12):4428–4436, Dec 2023.

[10] Frank Niemeyer, Fabio Galbusera, Youping Tao, Annette Kienle, Meinrad Beer, and Hans-Joachim Wilke. A deep learning model for the accurate and reliable classification of disc degeneration based on mri data. *Investigative Radiology*, 56(2):78–85, Jul 2020.

[11] None de, None Salinas, Gonzalo, Juan Carlos, None Llobet Rafael, Miguel Angel, None Martínez Jacobo, None Martí-Bonmatí Luis, None Blanquer Ignacio, None Regaña Manuel, and Miguel. Bimcv: Synergy between peta bytes of data in population medical imaging, computer aided diagnosis and avr. *Studies in health technology and informatics*, Jan 2015.

[12] J. M. Saborit-Torres, J. J. Saenz-Gamboa, J. À. Montell, J. M. Salinas, J. A. Gómez, I. Stefan, M. Caparrós, F. García-García, J. Domenech, J. V. Manjón, G. Rojas, A. Pertusa, A. Bustos, G. González, J. Galant, and M. de la Iglesia-Vayá. Medical imaging data structure extended to multiple modalities and anatomical regions, 2020.

[13] Jhon Jairo Sáenz-Gamboa, Julio Domenech, Antonio Alonso-Manjarrés, Jon A. Gómez, and Maria de la Iglesia-Vayá. Automatic semantic segmentation of the lumbar spine: Clinical applicability in a multi-parametric and multi-center study on magnetic resonance images. *Artificial Intelligence in Medicine*, 140:102559, 2023.

[14] Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010.

[15] Alex Zwanenburg, Martin Vallières, Mahmoud A. Abdalah, Hugo J. W. L. Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J. Beukinga, Ronald Boellaard, Marta Bogowicz, Luca Boldrini, Irène Buvat, Gary J. R. Cook, Christos Davatzikos, Adrien Depeursinge, Marie-Charlotte Desseroit, Nicola Dinapoli, Cuong Viet Dinh, Sebastian Echegaray, Issam El Naqa, Andriy Y. Fedorov, Roberto Gatta, Robert J. Gillies, Vicky Goh, Michael Götz, Matthias Guckenberger, Sung Min Ha, Mathieu Hatt, Fabian Isensee, Philippe Lambin, Stefan Leger, Ralph T.H. Leijenaar, Jacopo Lenkowicz, Fiona Lippert, Are Losnegård, Klaus H. Maier-Hein, Olivier Morin, Henning Müller, Sandy Napel, Christophe Nioche, Fanny Orlhac, Sarthak Pati, Elisabeth A.G. Pfaehler, Arman Rahmim, Arvind U.K. Rao, Jonas Scherer, Muhammad Musib Siddique, Nanna M. Sijtsema, Jairo Socarras Fernandez, Emiliano Spezi, Roel J.H.M. Steenbakkers, Stephanie Tanadini-Lang, Daniela Thorwarth, Esther G.C. Troost, Taman Upadhaya, Vincenzo Valentini, Lisanne V. van Dijk, Joost van Griethuysen, Floris H.P. van Velden, Philip Whybra, Christian Richter, and Steffen Löck. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020. PMID: 32154773.

[16] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane. Imbalanced data problem in machine learning: A review. *IEEE Access*, 13:13686–13699, January 2025.