Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quantitative Grading of Disc Health and Degeneration from Routine MRI: AI-Based Prediction of T2 Relaxometry Values using Radiomics and a proposal for New Grading

Pushpa Bhari Thippeswamy
Ganga Hospital
Mogan Kaviprawin
Ganga Hospital
Lekha Sree
Independent Researcher, Medical Imaging and Artificial Intelligence
Gnanaprakash Gurusamy
Ganga Hospital
Sri Vijay Anand K S
Ganga Hospital
Ajoy Prasad Shetty
Ganga Hospital
Rishi Mugesh Kanna
Ganga Hospital
Shanmuganathan Rajasekaran
rajasekaran.orth@gmail.com

Ganga Hospital

Research Article

Keywords: Lumbar Disc Degeneration, T2 Relaxometry, Radiomics, Artificial Intelligence, Machine Learning

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11393272/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Quantitative Grading of Disc Health and Degeneration from Routine MRI: AI-Based Prediction of T2 Relaxometry Values using Radiomics and a proposal for New Grading.

Abstract

Purpose: Identifying Disc degeneration (DD) before it is obvious on routine Magnetic Resonance Imaging (MRI) is the focus of spine research. Quantitative T2 relaxometry of the nucleus pulposus (T2r) offers objective biochemical evaluation of DD but has disadvantages of high costs, long scan times, and limited availability. We evaluated whether radiomic features derived from standard sagittal T2-weighted(T2W) MRI could accurately predict degeneration status and estimate continuous T2r values, thereby making T2r values available from normal T2W images.

Methods: We analyzed 621 lumbar intervertebral discs from 160 individuals. A data-derived T2r degeneration threshold was established using a reference cohort of 334 volunteer discs. Discs were partitioned at the patient level into training (353), validation (154), and external validation test set (114 discs). 111 radiomic features were extracted using 3D Slicer. A quantile-transformed logistic regression model was trained to predict binary degeneration (T2r $\leq$ 82.2 ms), while a Support Vector Regression (SVR) model estimated continuous T2r values.

Results: The binary radiomics model demonstrated good prediction with an AUC of 0.928 (95% CI: 0.87–0.98) and accuracy of 86.0% on the held-out test set. Importantly, the model correctly identified 70.8% of biochemically altered discs within Pfirmann I and II discs. For continuous modelling, SVR achieved a strong Spearman correlation ($\rho = 0.79$) with actual T2r values, yielding a low median absolute deviation of 12.2 ms.

Conclusion: T2r radiomics provides a promising surrogate for advanced quantitative mapping and provides universal access to biochemical disc profiling without need for specialised software, increased time, and costs.

Keywords: Lumbar Disc Degeneration, T2 Relaxometry, Radiomics, Artificial Intelligence, Machine Learning

Introduction

The cascading pathophysiology of Lumbar Disc Degeneration (LDD) initiates long before macro-structural alterations become visible on standard imaging. It begins at the molecular level, characterized by a progressive depletion of proteoglycans, the loss of type-II collagen integrity, and a concomitant reduction in the water-binding capacity of the nucleus pulposus (NP)[1]. In the early stages of biochemical alteration, discs classified as Pfirrmann grades I and II may retain a normal appearance, despite underlying changes in disc composition, leading to diagnostic delays. Conversely, advanced age or benign mechanical adaptations can cause a disc to appear morphologically degraded (Pfirrmann Grades III-V) while maintaining a biochemically stable, highly hydrated nucleus matrix, leading to inappropriate surgical interventions targeting asymptomatic, structurally altered but biochemically functional discs. This diagnostic blind spot introduces significant clinical ambiguity.

To bridge this diagnostic gap, advanced quantitative MRI sequences, such as T2 relaxometry of the nucleus pulposus (T2r), have been developed. By measuring absolute T2 relaxation times in milliseconds, T2 relaxometry directly quantifies hydrogen ion mobility, serving as a highly sensitive, objective metric for collagen architecture and fluid volume[2, 3]. Unfortunately, despite its documented usefulness, T2 relaxometry has failed to achieve widespread clinical adoption due to increased scan acquisition times, financial burden, specialized software requirements, and non-availability outside of academic research institutions[4] . This makes objective, quantitative biochemical evaluation and identification of early degeneration out of scope for most surgeons worldwide.

Radiomics offers a unique computational solution for converting standard medical images into high-dimensional, mineable quantitative data; radiomics extracts hundreds of microstructural, spatial, and textural features that escape human visual perception[5, 6]. When paired with machine learning architectures, these radiomic signatures can decode the underlying histopathological and biochemical status of tissues using routine scans[6, 7].

This study sought to develop a radiomics-based artificial intelligence (AI) framework capable of converting universally available routine sagittal T2W MRI sequences into objective biochemical metrics and predict T2r value. We were able to arrive at a continuous predictive model that allowed us to evaluate the disc health status into entire spectrum and propose a quantitative grading system. This, we hope, will make universal availability of quantitative disc

profiling accessible without the demands of extra scan time, specialized software, and added patient cost.

Methods

Study Population and Reference Threshold Establishment

We included 621 lumbar intervertebral discs (T12/L1 to L5/S1 levels) from 160 individuals, comprising patients with LDD and volunteers. All these included discs had conventional MRI-T2W sequence based Pfirrmann grade and T2 relaxometry values. Among 621 discs, we used 334 discs from volunteers with available Pfirrmann grading to estimate the T2r cut-off for discriminating normal (Pfirrmann grades I and II) from disc degeneration (Pfirrmann grades III, IV, and V). We used area under the receiver operating characteristic (ROC) curve analysis and the Youden index in JAMOVI (version 2.7.17) to determine the optimal T2r threshold that maximises sensitivity and specificity to identify the disc degeneration.

Image Segmentation and High-Dimensional Radiomic Extraction

Standard, routinely acquired sagittal T2W spine MRI sequences were used. We performed manual region-of-interest (ROI) segmentation encompassing the full boundaries of the intervertebral disc in the open-source 3D Slicer environment. After segmentation, we extracted high-dimensional radiomic features using PyRadiomics. A comprehensive panel of 111 unique radiomic features was extracted per disc. These features served as micro-structural digital biomarkers, capturing spatial pixel distributions, gray-level heterogeneity, and subtle geometric alterations. Each disc radiomic dataset consisted of 111 features categorised into eight domains ***(Table 1, Figure 1/2)***

Partitioning and Algorithmic Balancing

We developed two AI models based on the type of outcome variable. 1) Binary outcome classification (T2r cut-off-based normal vs abnormal disc); 2) Continuous T2r values.

We partitioned the 621 discs into three independent datasets: a training set of 353 discs (56.8%), a model validation set of 154 discs (24.7%), and a locked external validation/ held-out test set of 114 discs (18.5%). To prevent data leakage, we partitioned data strictly at the patient level rather than the individual disc level. We kept the locked external test set completely isolated

during all stages of model selection and accessed it only once at the study's conclusion to assess diagnostic performance.

Initial feature engineering revealed that 62 of the 111 extracted radiomic features exhibited significant data skewness. To stabilise algorithmic weights, we quantile-transformed each feature, ranked it, and mapped it onto a standard normal distribution using 200 quantiles learned exclusively from the training dataset. To address the inherent clinical imbalance between healthy and diseased discs within the patient population, an automated class-weighting matrix (approximately 1.74 Diseased: 1 Normal).

Model Development, Training, and Architecture Selection

In the binary classification AI model, we evaluated ten machine learning algorithms using repeated, patient-grouped cross-validation on the training and validation sets. Among these, we chose logistic regression for the final predictive model based on AUC ***(Table 2, Supplementary Table 1)***. The detailed study flow was depicted in ***Figure 3.***

For the continuous numerical outcome T2r model, we compared ten regression models on the training set using patient-level five-fold cross-validation, ranking them by mean absolute error. Support vector regression (SVR) was selected as the final model ***(Supplementary Table 2).***

We conducted all machine learning development in Python (version 3.12.10) using scikit-learn (version 1.5.2) and XGBoost (version 2.1.1). Model interpretability was evaluated via SHAP (SHapley Additive exPlanations, version 0.46.0) using the LinearExplainer module. To assess human diagnostic limitations, an inter-observer agreement analysis was performed on a highly complex subset of 51 discs that were biochemically normal (> 82.2 ms) but classified as degenerated by initial morphological Pfirrmann grading. We estimated the agreement for assessing the Pfirrmann grades among two radiologists using Krippendorff's alpha.

Ethics considerations: We de-identified the patient-related details for the analysis. We obtained an Institutional Review Board waiver of consent to conduct the study (approval number: IRB-2026-31).

Results:

Validation of the Baseline T2r Cut-off Threshold

We used 334 discs to estimate the T2r cut-off. Of these, 293 discs were classified as normal (Pfirrmann grades I–II) and 41 as morphologically degenerated (Pfirrmann grades III–V) based on conventional T2W MRI sequences. We derived a T2r threshold of 82.2; values below this threshold were classified as T2r-defined degeneration. The sensitivity and specificity were 95.12% and 89.42%, respectively ***(Table 3).*** Thus, for subsequent analyses, we categorised T2r values ≤82.2 as T2r-defined abnormal, whereas values >82.2 as T2r-defined normal discs.

Baseline Characteristics

Among the 621 discs analysed, 228 (36.7%) were classified as T2r-defined abnormal and 393 (63.3%) as normal. Participants' mean age was 43.6 ± 14.6 years, and 60.0% were male. The overall distribution of participant characteristics and disc-level features was broadly similar across the training, validation, and external validation sets **(Table 4).** Within the morphologically normal discs (424, Pfirrmann Grades I–II), 82 (19.3%) had abnormal T2r values. However, within the morphologically degenerated discs (197, Pfirrmann Grades III–V), 51 (25.8%) had a normal T2r value ***(Table 5).*** As the Pfirrmann grade increases, the proportion of discs with abnormal T2r value increases ***(Figure 4).***

Binary Radiomics AI Model Performance and Feature Contributions

The quantile-transformed logistic regression model demonstrated an ROC Curve (AUC) of 0.928 in the training set and 0.889 in the model validation set. In the locked external validation test set representing entirely unseen discs, the model maintained an AUC of 0.928 ***(Table 6).*** Among the 82 abnormal T2r discs that were graded as Pfirrmann 1–2, 24 were included in the external validation set. Of these 24 discs, the AI model correctly identified 70.8% as abnormal. Among the 111 features, ten features based on SHAP analysis contributed to the T2r-based model outcome ***(Figure 5)***

In the external validation set analysis, misclassification was more frequent among discs with T2r values close to the predefined cutoff of 82.2. Discs with values within 5 units of the threshold had the highest misclassification proportion (4/14, 28.6%), followed by those 5–10 units from the threshold (2/12, 16.7%). Discs with T2r values at a distance of over 20 units from 82.2 had lower misclassification ***(Figure 6).***

Continuous Radiomics AI Model Performance (T2r Value Prediction)

Across all evaluation datasets, approximately 80% of the true, independently measured T2r values fell within the exact predictive confidence ranges generated by the AI model ***(Table 7, Supplementary Table 2).***

The scatter plot demonstrates that predicted T2r value increased with measured actual T2r value (Spearman $\rho = 0.80$). The 80% prediction range contained the measured value for 91 of 114 discs (79.8%), consistent with its intended coverage ***(Figure 7)***.

Inter-Radiologist Morphological Disagreement Analysis: A total of 51 discs were identified as non-degenerated according to the T2r threshold but classified as degenerated by conventional Pfirrmann grading (grades III–V). The Krippendorff's alpha was 0.313, indicating limited agreement between the two independent radiologists.

Discussion

Low Back Pain affected an estimated 619 million people in 2020 and is projected to affect 843 million people by 2050[8] . Disc degeneration is the most common cause of low back pain, and hence spine research globally has focused on the cure and prevention of disc degeneration. Understanding the pathology of LDD at the molecular level and identifying degeneration even before it is obvious on routine MRI has been the main focus[9, 10] . This is important as a wide range of regeneration and reversal strategies, including cell therapy, growth factors, gene therapy and biomaterials, have been advocated to halt or even reverse degeneration[11] .

Beyond Pfirrmann Grade – The need for Quantitative assessment

The most widely used Pfirrmann grade[12] is an ordinal morphological grading system on a five-point visual scale and has only moderate inter-observer reliability[13] . Being image-based and visually assessed, it overlooks the considerable biological heterogeneity that may exist within individual Pfirrmann grades. It cannot fully characterize the biochemical alterations that precede structural degeneration obvious on MRI. T2r is useful as it provides quantitative information sensitive to NP hydration, proteoglycan-associated water binding, and collagen organization[14–16] . T2r is sensitive to the organisation of collagen and the mobility of water within the collagen-proteoglycan network. In patients with low back pain, T2 mapping at 3T has been compared with morphological grading and proposed as a non-invasive tool to assess

water content and collagen fiber integrity and early degeneration in different disc pathologies[15–17] .

Despite ample evidence of superiority to Pfirrmann grading, T2r is still not applied universally as it is limited by the need for a dedicated multi-echo acquisition, prolonged scan time, the need for specialised processing software available mainly in academic centres, and additional cost. It would be a game changer for the spine community if these barriers are overcome so that T2r information can be made available universally from normal MRI.

Beyond Pfirrmann and T2r – Radiomics

Radiomics extracts hundreds of micro-structural, spatial, and textural features that escape human visual perception[5]. When paired with machine learning architectures, these radiomic signatures can decode the underlying histopathological and biochemical status of tissues using routine scans[5, 6].

The strength of radiomics has been tested in the field of imaging LDD. Previous studies used texture analysis to quantify lumbar disc degeneration from conventional T2-weighted images[18, 19], built semi-automated classifiers of degenerative lumbar disease, derived histogram-based continuous indices[20], and trained deep-learning and radiomics models for automated grading[21, 22]. In the largest population-based example, 737 radiomic features from 1,397 participants classified Pfirrmann grade with a balanced accuracy of 76.7% and a kappa of 0.70[22]. Other texture studies have successfully utilised radiomics to evaluate low back pain status or painful annular fissures[23–25].

There is, however, a major limitation that is common to all previous literature. They mainly aimed to match radiologist classification and used Pfirrmann grade as the reference. A model trained to reproduce Pfirrmann grade inherits the subjectivity and inter-observer variability of the scale. They cannot, by design, detect early biochemical change that is not represented in Pfirrmann grading or identify subsets within the same Pfirrmann grade. Applying radiomics in this manner has therefore not overcome the disadvantages of grading, and has added little scientific advance beyond automation of what a radiologist already does. It was suggested that future work should look for subvisual texture changes that reflect the biochemical alterations which precede visible changes over the entire spectrum of LDD[22] . This was the precise aim of the present study.

Our study adopted a unique approach. Instead of predicting the Pfirrmann grade as the aim, we used radiomics from routine T2W sagittal images to predict T2r values directly as a numerical value. To our knowledge, this is the first study to predict T2r on a continuous scale from routine T2-weighted MRI by radiomics. The initial binary model identified T2r-defined abnormal discs in a locked test set with an AUC of 0.928, and the subsequent continuous model successfully ranked discs in line with measured T2r (Spearman ρ 0.8). Our model was also able to identify 70.8% of discs that were originally classified as Pfirrmann I & II but with abnormal T2r, documenting its ability to identify early biochemical changes.

Interpretation of the main findings

Binary prediction: In our 621 discs, the radiomics model was able to place the disc on either side of the reference point of 82.2 with a sensitivity of 91.5% in training, 86% in validation and 83.3% in held-out validation. Misclassification was clustered around the threshold: discs within 10 ms of 82.2 ms had an error rate of 23.1% (6 of 26), compared with 11.4% (10 of 88) for the rest. This implies that a binary classification was more accurate at the extremes of the range and less accurate for assessing discs around the reference point. This highlights the disadvantage of a binary classification utilised in previous studies.

Continuous prediction model:

Our model estimated T2r value as a continuous quantitative variable with an accuracy of 80% and an excellent correlation of 0.8 between the calculated T2r and predicted T2r by radiomics, with a median absolute error of 12-15ms. Although not a replacement for a measured value in an individual disc, the estimates prove that they are good for ranking discs and a useful result for screening and for stratifying disc health.

Our results opened up the possibility of two major improvements over the previous radiomics: the possibility of diagnosing disc degeneration at a preclinical stage and evaluating discs over a continuous spectrum of health and degeneration.

Identification of preclinical nucleus pulposus changes

The discordance between Pfirrmann grading and quantitative values noted in our results may help to identify early degeneration at the preclinical stage. Among 424 discs classified as Pfirrmann I–II, 82 (19.3%) had measured nucleus pulposus T2r values at or below the study

threshold of 82.2 ms. This suggests that approximately 1/5th of the discs evaluated as Pfirrmann Grade I/II and normal may already suffer from biochemical changes not appreciated by normal MRI. Conversely, among 197 discs classified as Pfirrmann III–V, 51 (25.9%) retained T2r values above the threshold indicating that these discs may not be completely abnormal. This was true for the held-out test sets also. 24 discs were Pfirrmann I–II despite low measured T2r values, and it was interesting that the radiomics model correctly identified 70.8% of these as abnormal. This adds weightage to our proposal that a continuous prediction model may help to flag discs that appear relatively preserved but may already exhibit reduced hydration or altered matrix organization.

The spectrum of disc health – Quantitative Grading System

Disc matrix deterioration does not occur in abrupt biological steps at arbitrary rigid boundaries. Degeneration involves changes in hydration, proteoglycan – associated water retention, collagen organisation and structural integrity, which may all change at different rates and may also be partially dissociated. A numerical score has the added advantage of categorising the entire alterations. To facilitate clinical interpretation and easy application on the basis of continuous T2r estimate, we propose a five-tier descriptive grading system.

Grade	T2R value (ms)	Proposed category
1	>100	High T2r normal
2	91-100	Low T2r Normal
3	71-90	Transient Zone
4	61-70	Low T2r range
5	≤60	Markedly low T2r range

The basis of the classification is as follows: The area under the curve constructed for 334 discs from volunteers with a spread of all Pfirrmann categories showed a T2r threshold of 82.2 ms between Pfirrmann I/II and Pfirrmann III/IV/V. Hence, the nearest round off value of 80 ms was taken as the reference point and a transitional zone from 71-90 was proposed to include discs with T2r values of 10ms on either side. The justification was that our results showed that 38% of the misclassification happened within 10ms on either side of the reference point and so this transitional zone will accommodate the biological and measurement uncertainty inherent

in dichotomisation. We then built two categories on either side to include discs with two levels of higher and lower T2r values.

Although the proposed boundaries are provisional, our results showed proof of its justification. The incidence of Pfirrmann Grade 1 was 77% in Grade I compared to 7% in Grade II, 15% in Grade III, 2% in Grade IV, and 0% in Grade V ($p<0.01$). Similarly, the incidence of Pfirrmann Grade IV and V was only 8% in Grade I compared to 2% in Grade II, 10% in Grade III, 12% in Grade IV, and 68% in Grade V ($p<0.01$) **(Table 8).** As there was a significant spillover between each proposed T2r grade in the incidence of Pfirrmann I to V, there is proof that the proposed T2r classification is valid and probably more accurate.

Pfirrmann Vs Quantitative T2r Grade

Although our T2r grading has five grades similar to Pfirrmann, the fact that 19.3% of Pfirrmann Grade I and II discs demonstrated T2r values below the normal showed that quantitative classification may be more sensitive and reflect truly the biochemical changes and conventional MRI. This was also true for Pfirrmann IV and V discs, where 2% of the discs showed values above 90 ms and 8% of the discs showed >100 ms. These observations suggest that morphological Pfirrmann grading does not always reflect the underlying quantitative characteristics and highlights the potential of T2r assessment to provide complementary information on disc health.

Clinical implications

The study has important clinical contributions. Disc degeneration being the primary cause of low back pain, the current attempts for accurate evaluation and early identification are thwarted by the inability of normal T2W sequences to go beyond objective image classifications[26]. Our study has shown that radiomics has an accuracy of over 80% in identifying degeneration. We have also developed a five-tier T2r classification based on the continuous prediction that will help in clinical communication. The continuous score has potential for longitudinal measurement and will help to define study population and provide objective outcome measures for biological and regenerative therapy trials.

Another salient contribution of the study is proving that radiomics is an effective surrogate for quantitative MRI and making T2r evaluation universally available. This will overcome the hurdles of the need for specialised software, increased acquisition times, and increased costs.

Limitations

Several limitations require consideration. First, T2 relaxation time is influenced not only by nucleus pulposus parameters, but also by magnetic strength, coil configuration, and post-processing methods[27]. Hence, it remains to be seen whether the findings are directly transferable across scanners and protocols. Further, although the predictor and measured T2r values demonstrated a strong correlation, their numerical agreement and clinical interchangeability remain unestablished. Absolute prediction errors, calibration, and agreement analysis need to be pursued, particularly near the 82.2 ms threshold.

Our proposed classification and transition zone are exploratory and require further and independent validation. Their boundaries must be evaluated against age, spinal level, morphological degeneration and clinical outcomes. Prospective multicentre longitudinal studies are necessary to determine whether the proposed quantitative characteristics predict degeneration progression and are clinically meaningful.

Conclusion:

Our study was successful in developing a radiomics-based model to provide a quantitative grading of disc health and degeneration using routine T2W MRI sequences. The proposed T2r-based quantitative grading could be a superior alternative to the morphological Pfirrmann grading in assessing biochemical status of disc health and identification of early degeneration. With multicentre validation, the proposed radiomics model can overcome the current limitations of universal usage of T2r and help clinical scientists and surgeons working in resource-limited settings.

References

Mwale F, Iatridis JC, Antoniou J (2008) Quantitative MRI as a diagnostic tool of intervertebral disc matrix composition and integrity. European Spine Journal 2008 17:4 17:432–440. https://doi.org/10.1007/S00586-008-0744-4

2.1 Marinelli NL, Haughton VM, Anderson PA (2010) T2 Relaxation Times Correlated with
2 Stage of Lumbar Intervertebral Disk Degeneration and Patient Age. American Journal
3 of Neuroradiology 31:1278–1282. https://doi.org/10.3174/AJNR.A2080

3.4 Marinelli NL, Haughton VM, Muñoz A, Anderson PA (2009) T 2 relaxation times of
5 intervertebral disc tissue correlated with water content and proteoglycan content. Spine
6 (Phila Pa 1976) 34:520–524. https://doi.org/10.1097/BRS.0B013E318195DD44

4.7 Jeeva M, Sundararajan S, Hari Baskar S, et al (2026) Role of T2 mapping MRI in early
8 lumbar intervertebral disc degeneration: a narrative review. Egyptian Journal of
9 Radiology and Nuclear Medicine 2026 57:1 57:191-. https://doi.org/10.1186/S43055-026-
10 01855-4

5.1 Lambin P, Rios-Velazquez E, Leijenaar R, et al (2012) Radiomics: Extracting more
12 information from medical images using advanced feature analysis. Eur J Cancer
13 48:441–446. https://doi.org/10.1016/J.EJCA.2011.11.036

6.4 Chen Y, Wang F, Yu L, et al (2026) Four MRI-Based Radiomics Models for Diagnosis of
15 Lumbar Intervertebral Disc Degeneration. Journal of Imaging Informatics in Medicine
16 2026 1–13. https://doi.org/10.1007/S10278-026-01966-4

7.7 Zheng HD, Sun YL, Kong DW, et al (2022) Deep learning-based high-accuracy
18 quantitation for lumbar intervertebral disc degeneration from MRI. Nature
19 Communications 2022 13:1 13:841-. https://doi.org/10.1038/s41467-022-28387-5

8.0 Ferreira ML, De Luca K, Haile LM, et al (2023) Global, regional, and national burden
21 of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: a
22 systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol
23 5:e316–e329. https://doi.org/10.1016/S2665-9913(23)00098-X

9.4 Rider SM, Mizuno S, Kang JD (2018) Molecular Mechanisms of Intervertebral Disc
25 Degeneration. Spine Surg Relat Res 3:1. https://doi.org/10.22603/SSRR.2017-0095

10.6 Gradisnik L, Kocivnik N, Maver U, Velnar T (2025) Degenerative Disease of
27 Intervertebral Disc: A Narrative Review of Pathogenesis, Clinical Implications and
28 Therapies. Bioengineering 2026, Vol 13, Page 40 13:40.
29 https://doi.org/10.3390/BIOENGINEERING13010040

11.1 Sono T, Shima K, Shimizu T, et al (2024) Regenerative therapies for lumbar
31 degenerative disc diseases: a literature review. Front Bioeng Biotechnol 12:1417600.
32 https://doi.org/10.3389/FBIOE.2024.1417600/TEXT

12.3 Pfirrmann CWA, Metzdorf A, Zanetti M, et al (2001) Magnetic resonance classification
34 of lumbar intervertebral disc degeneration. Spine (Phila Pa 1976) 26:1873–1878.
35 https://doi.org/10.1097/00007632-200109010-00011

13. Rustenburg C, Emanuel K, Holewijn R, et al (2019) LUMBAR INTERVERTEBRAL DISC DEGENERATION: CAN WE RELY ON CURRENT CLASSIFICATION SYSTEMS? Orthopaedic Proceedings 101-B:3–3. https://doi.org/10.1302/1358-992X.2019.10.003

14. Bhari Thippeswamy P, Rajasekaran S, Ramachandran K, et al (2025) Role of Magnetic Resonance Spectroscopy and T2 Relaxometry as Imaging Biomarker of Early Lumbar Intervertebral Disc Degeneration. Global Spine J 15:2736–2743. https://doi.org/10.1177/21925682241311515

15. Trattnig S, Stelzeneder D, Goed S, et al (2010) Lumbar intervertebral disc abnormalities: comparison of quantitative T2 mapping with conventional MR at 3.0 T. European Radiology 2010 20:11 20:2715–2722. https://doi.org/10.1007/S00330-010-1843-2

16. Welsch GH, Trattnig S, Paternostro-Sluga T, et al. Parametric T2 and T2* mapping techniques to visualize intervertebral disc degeneration in patients with low back pain: initial results on the clinical use of 3.0 Tesla MRI. Skeletal Radiol. 2011 May;40(5):543-51. doi: 10.1007/s00256-010-1036-8

17. Welsch GH, Trattnig S, Paternostro-Sluga T, et al (2010) Parametric T2 and T2* mapping techniques to visualize intervertebral disc degeneration in patients with low back pain: initial results on the clinical use of 3.0 Tesla MRI. Skeletal Radiology 2010 40:5 40:543–551. https://doi.org/10.1007/S00256-010-1036-8

18. Michopoulou S, Costaridou L, Vlychou M, et al (2011) Texture-based quantification of lumbar intervertebral disc degeneration from conventional T2-weighted MRI. Acta radiol 52:91–98. https://doi.org/10.1258/AR.2010.100166

19. Mayerhoefer ME, Stelzeneder D, Bachbauer W, et al (2012) Quantitative analysis of lumbar intervertebral disc abnormalities at 3.0 Tesla: Value of T 2 texture features and geometric parameters. NMR Biomed 25:866–872. https://doi.org/10.1002/NBM.1803

20. Waldenberg C, Hebelka H, Brisby H, Lagerstrand KM (2017) MRI histogram analysis enables objective and continuous classification of intervertebral disc degeneration. European Spine Journal 2017 27:5 27:1042–1048. https://doi.org/10.1007/S00586-017-5264-7

21. Zheng HD, Sun YL, Kong DW, et al (2022) Deep learning-based high-accuracy quantitation for lumbar intervertebral disc degeneration from MRI. Nature Communications 2022 13:1 13:841-. https://doi.org/10.1038/s41467-022-28387-5

22. McSweeney T, Tiulpin A, Kowlagi N, et al (2025) Robust Radiomic Signatures of Intervertebral Disc Degeneration From MRI. Spine (Phila Pa 1976) 50:1737. https://doi.org/10.1097/BRS.0000000000005435

23. Abdollah V, Parent EC, Dolatabadi S, et al (2021) Texture analysis in the classification of T2-weighted magnetic resonance images in persons with and without low back pain. Journal of Orthopaedic Research 39:2187–2196. https://doi.org/10.1002/JOR.24930

24. Eriksson S, Waldenberg C, Torén L, et al (2022) Texture Analysis of Magnetic Resonance Images Enables Phenotyping of Potentially Painful Annular Fissures. Spine (Phila Pa 1976) 47:430–437. https://doi.org/10.1097/BRS.0000000000004160

25. Ketola JHJ, Inkinen SI, Karppinen J, et al (2021) T2-weighted magnetic resonance imaging texture as predictor of low back pain: A texture analysis-based classification pipeline to symptomatic and asymptomatic cases. Journal of Orthopaedic Research 39:2428–2438. https://doi.org/10.1002/JOR.24973

26. Russo F, Ambrosio L, Giannarelli E, et al (2023) Innovative quantitative magnetic resonance tools to detect early intervertebral disc degeneration changes: a systematic review. The Spine Journal 23:1435–1450. https://doi.org/10.1016/J.SPINEE.2023.05.011

27. Conner MN, Newman HR, Peloquin JM, et al (2026) MRI T2 Relaxation Time of the Intervertebral Disc: Repeatability, the Effect of Time of Day, and Unloaded Rest. JOR Spine 9:e70193. https://doi.org/10.1002/JSP2.70193

Figure Legends:

Figure 1. Segment editor module of the 3D Slicer application: The disc volume is selected by color-coding the segmented disc and excluding the adjacent structures. The 3D volume representation is shown in the volume rendering box in the top left corner of the image.

Figure 2. Radiomics feature extraction: Final step of radiomics feature extraction with input being the segmented disc. Manual customization is used to include all the features. The first, second, and higher-order features are then downloaded for further analysis (shown as numbered parameters under the segmentation image).

Figure 3: Study Workflow for Radiomics-Based Model Development

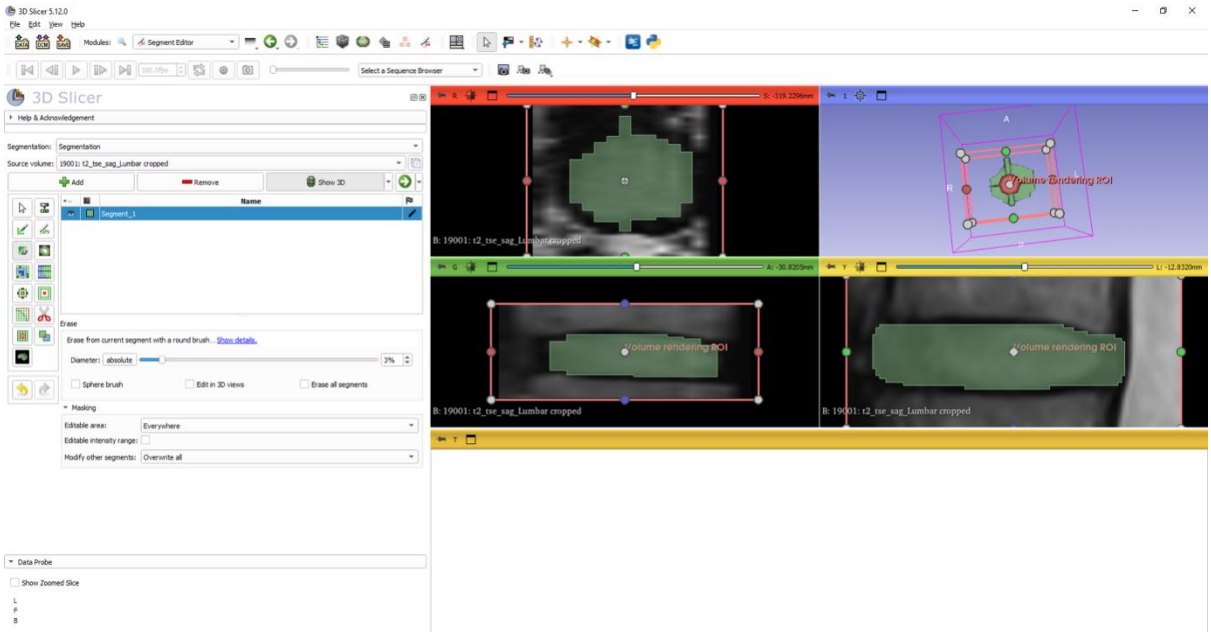
Figure 4: Distribution of normal (T2r >82.2) and abnormal (T2r ≤82.2) values across Pfirrmann grades

Figure 5: Top 10 radiomic features by SHAP contribution

Figure 6: Misclassification of 114 external validation discs according to the absolute distance between T2r values and the derived cutoff of 82.2

Figure 7: Predicted probability of degeneration against T2r

1 Figure 1:



2

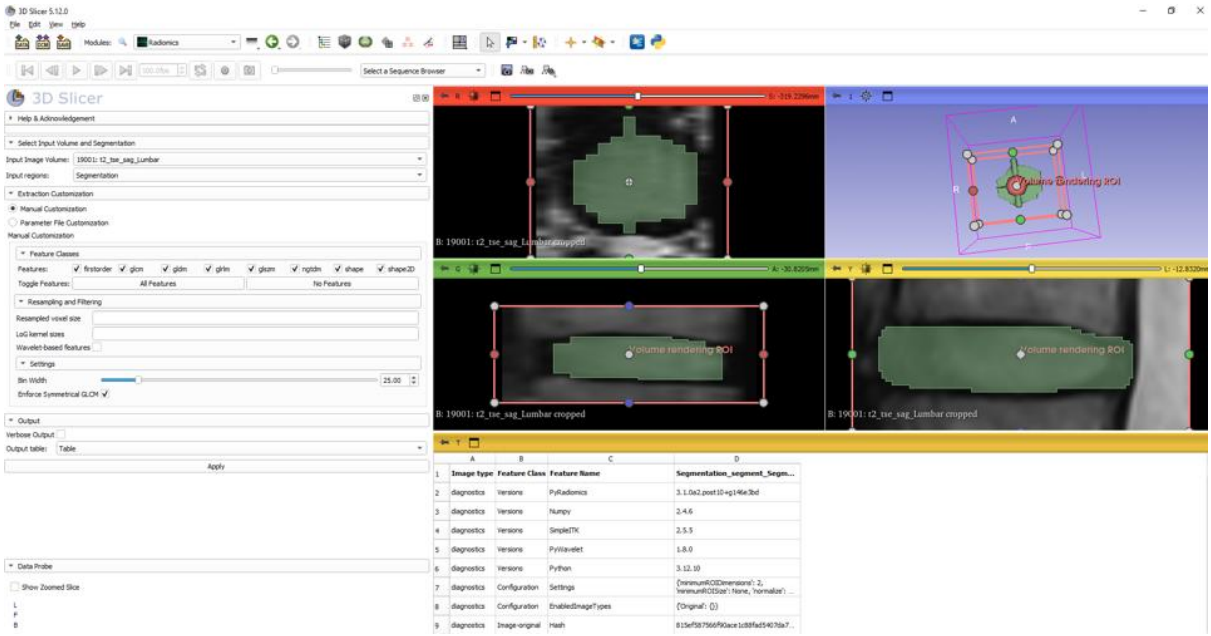
3

4

5

6

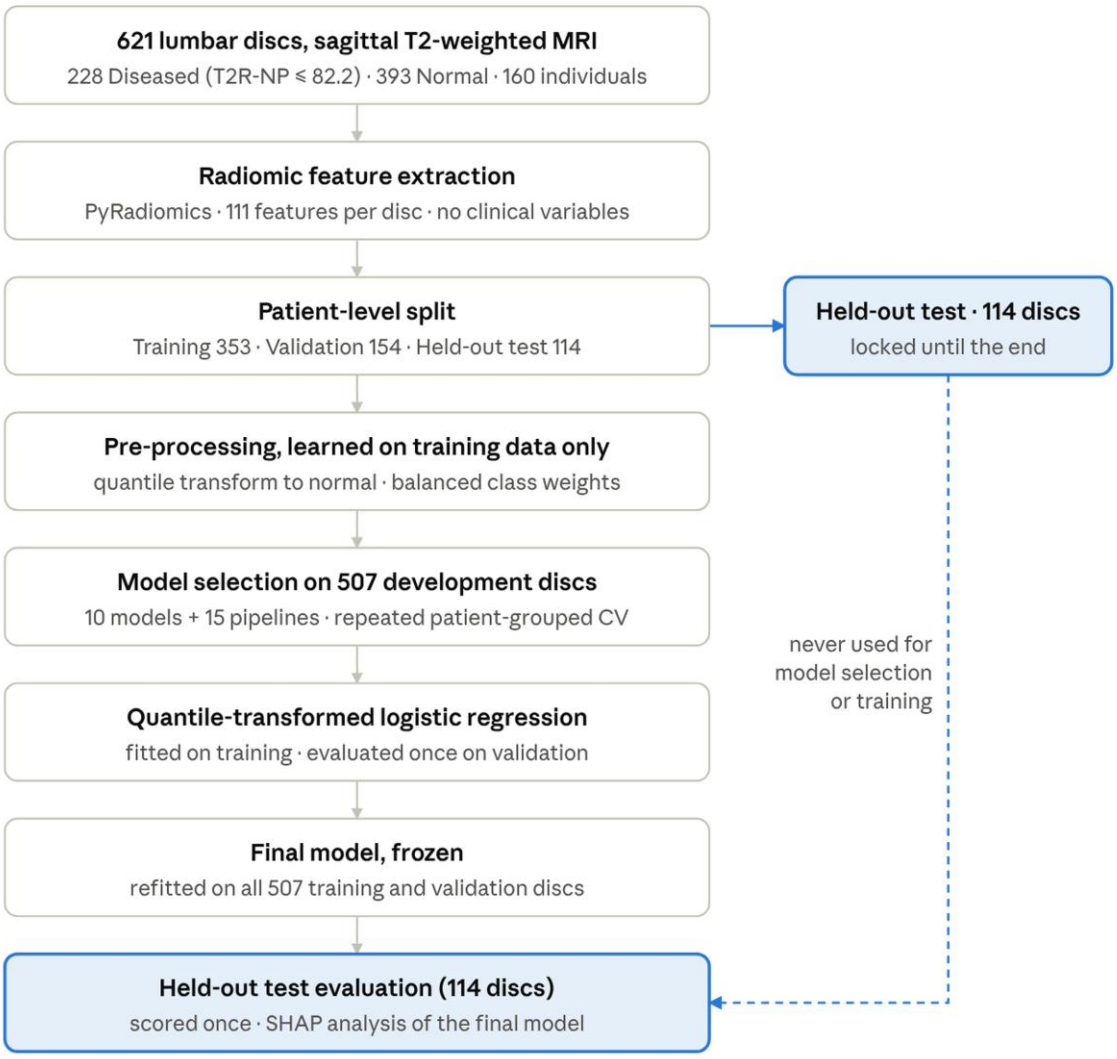
1 Figure 2:


3D Slicer
Volume rendering ROI
B: 19001: t2_tse_sag_Lumbar cropped

2

3

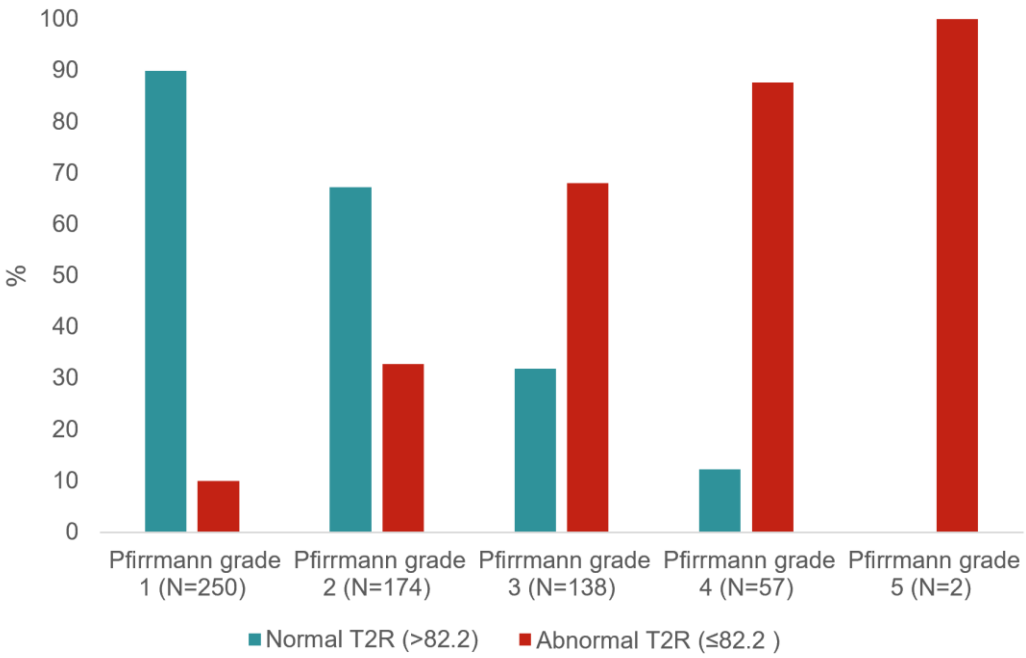
1 Figure 3:


621 lumbar discs, sagittal T2-weighted MRI
228 Diseased (T2R-NP ≤ 82.2) · 393 Normal · 160 individuals
Radiomic feature extraction
PyRadiomics · 111 features per disc · no clinical variables
Patient-level split
Training 353 · Validation 154 · Held-out test 114
Held-out test · 114 discs
locked until the end
Pre-processing, learned on training data only
quantile transform to normal · balanced class weights
Model selection on 507 development discs
10 models + 15 pipelines · repeated patient-grouped CV
never used for model selection or training
Quantile-transformed logistic regression
fitted on training · evaluated once on validation
Final model, frozen
refitted on all 507 training and validation discs
Held-out test evaluation (114 discs)
scored once · SHAP analysis of the final model

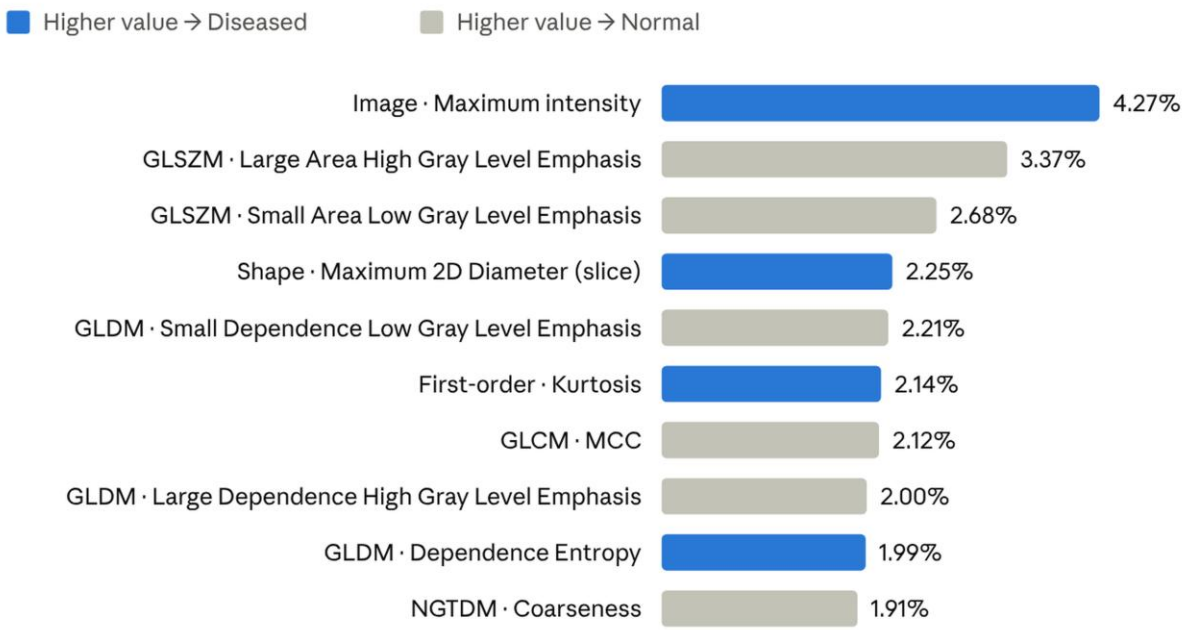
2

3

Figure 4:


100
90
80
70
60
50
40
30
20
10
0
%
Pfirrmann grade 1 (N=250)
Pfirrmann grade 2 (N=174)
Pfirrmann grade 3 (N=138)
Pfirrmann grade 4 (N=57)
Pfirrmann grade 5 (N=2)
Normal T2R (>82.2)
Abnormal T2R (≤82.2)

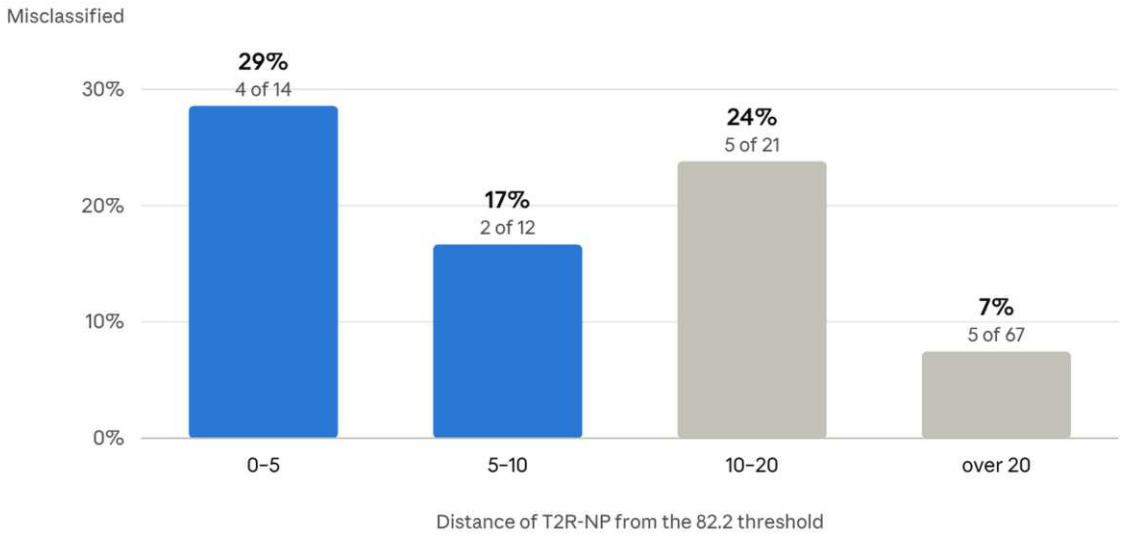
1 Figure 5:


Higher value → Diseased
Higher value → Normal
Image · Maximum intensity
4.27%
GLSZM · Large Area High Gray Level Emphasis
3.37%
GLSZM · Small Area Low Gray Level Emphasis
2.68%
Shape · Maximum 2D Diameter (slice)
2.25%
GLDM · Small Dependence Low Gray Level Emphasis
2.21%
First-order · Kurtosis
2.14%
GLCM · MCC
2.12%
GLDM · Large Dependence High Gray Level Emphasis
2.00%
GLDM · Dependence Entropy
1.99%
NGTDM · Coarseness
1.91%

2

3

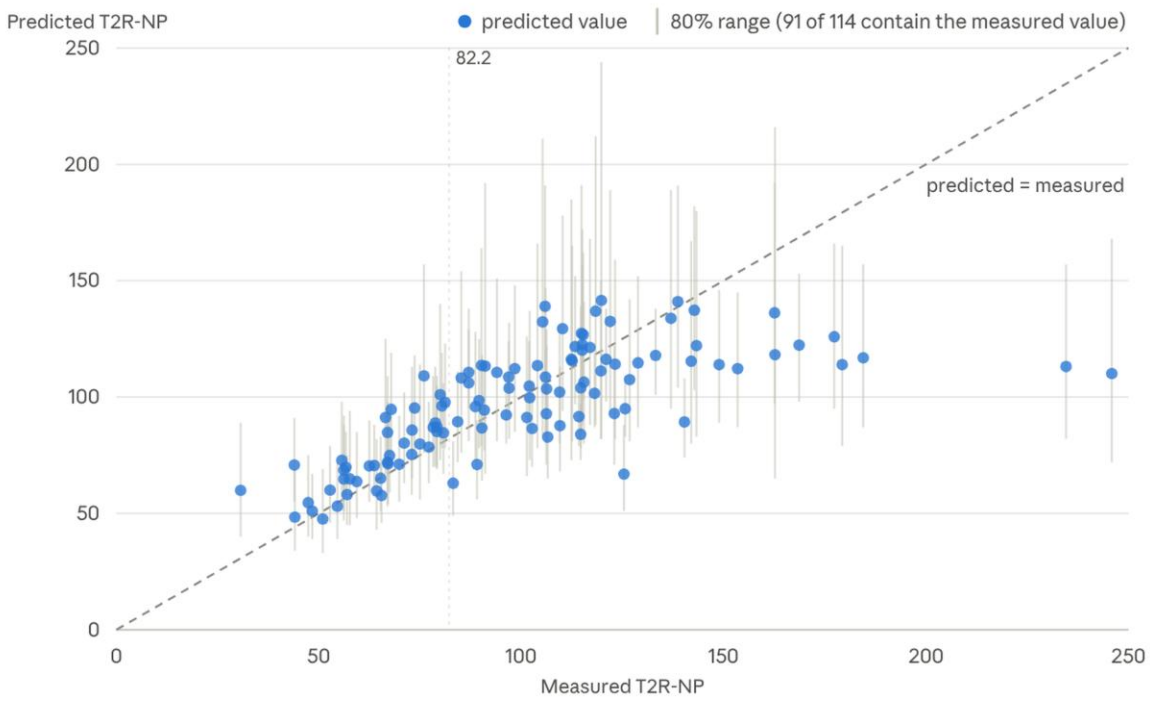
1 Figure 6:


Misclassified
30%
20%
10%
0%
29%
4 of 14
17%
2 of 12
24%
5 of 21
7%
5 of 67
0–5
5–10
10–20
over 20
Distance of T2R-NP from the 82.2 threshold

2

3

1 Figure 7:


Predicted T2R-NP
predicted value
80% range (91 of 114 contain the measured value)
250
200
150
100
50
0
82.2
predicted = measured
0
50
100
150
200
250
Measured T2R-NP

2

3

Table 1: Distribution of 111 radiomic features as per categories

Feature group	**Features**	**%**
Gray Level Co-occurrence Matrix (GLCM)	24	21.6%
Gray Level Size Zone Matrix (GLSZM)	16	14.4%
First-order intensity	18	16.2%
Gray Level Dependence Matrix (GLDM)	14	12.6%
Shape (3D)	14	12.6%
Gray Level Run Length Matrix (GLRLM)	16	14.4%
Neighbouring Gray Tone Difference Matrix (NGTDM)	5	4.5%
Image diagnostics (whole-image intensity, voxel count)	4	3.6%
Total	**111**	**100%**

Table 2: Parameters of the chosen logistic regression model

Parameter	Value	Explanation
Feature transform	Quantile → normal, 200 quantiles	Maps each feature onto a bell curve, removing skew and extreme values; learned from training data only
Model	Logistic regression (L2 penalty)	Combines the 111 features in a weighted sum
Regularisation (C)	0.1	Shrinks the weights so the model cannot overfit correlated features
Class weight	Balanced (≈ Diseased 1.74: Normal 1)	Offsets the smaller number of Diseased discs
Decision threshold	0.339	A disc is called Diseased at or above this probability; set to retain at least 92% sensitivity in training (median over 20 patient-level splits)
Random seed	42	Fixes the randomness, so the model is reproducible

Table 3: Diagnostic performance of T2R-NP at the derived cut-off

Diagnostic measure	Estimate (%)	95% CI
Sensitivity	95.12	83.47–99.40
Specificity	89.42	85.32–92.70
Prevalence	12.28	8.95–16.28
Positive predictive value (PPV)	55.71	47.24–63.87
Negative predictive value (NPV)	99.24	97.13–99.80
Accuracy	90.12	86.40–93.10

Table 4: Distribution of patients and discs across the training, validation and external validation sets

Characteristics		**Training**	**Test Validation**	**External validation test**	**Total**
Individual level (N)		**93**	**37**	**30**	**160**
Age (in years)	Mean ± SD (years)	45.1 ± 15.4	42.2 ± 12.5	40.8 ± 14.1	43.6 ± 14.6
	Median (IQR)	46 (34–56)	42 (34–49)	42 (28–50)	44 (32–54)
	Range	16–81	20–66	19–66	16–81
Sex	Male	58 (62.4%)	21 (56.8%)	17 (56.7%)	96 (60.0%)
	Female	35 (37.6%)	16 (43.2%)	13 (43.3%)	64 (40.0%)
Disc level (N)		**353**	**154**	**114**	**621**
Disc level	T12/L1	1 (0.3%)	0 (0.0%)	0 (0.0%)	1 (0.2%)
	L1/2	79 (22.4%)	31 (20.1%)	21 (18.4%)	131 (21.1%)
	L2/3	78 (22.1%)	30 (19.5%)	24 (21.1%)	132 (21.3%)
	L3/4	73 (20.7%)	31 (20.1%)	25 (21.9%)	129 (20.8%)
	L4/5	62 (17.6%)	32 (20.8%)	24 (21.1%)	118 (19.0%)
	L5/S1	60 (17.0%)	30 (19.5%)	20 (17.5%)	110 (17.7%)
Pfirrmann grade	Grade I	143 (40.5%)	57 (37.0%)	50 (43.9%)	250 (40.3%)
	Grade II	95 (26.9%)	44 (28.6%)	35 (30.7%)	174 (28.0%)
	Grade III	74 (21.0%)	41 (26.6%)	23 (20.2%)	138 (22.2%)
	Grade IV	40 (11.3%)	12 (7.8%)	5 (4.4%)	57 (9.2%)
	Grade V	1 (0.3%)	0 (0.0%)	1 (0.9%)	2 (0.3%)
T2r-based classification	Diseased (T2r ≤ 82.2)	129 (36.5%)	57 (37.0%)	42 (36.8%)	228 (36.7%)
	Normal (T2r > 82.2)	224 (63.5%)	97 (63.0%)	72 (63.2%)	393 (63.3%)

Table 5: Distribution of T2r-Defined Disc Degeneration and T2r Values Across Pfirrmann Grades

Pfirrmann grade	Normal T2r category (>82.2)			Abnormal T2r category (≤82.2)		
	Number of discs (%)	**T2R-NP, mean ± SD**	**Range**	**Number of discs (%)**	**T2R-NP, mean ± SD**	**Range**
Grade 1	225 (90.0)	141.22 ± 59.44	82.31–552.90	25 (10.0)	74.31 ± 7.29	53.20–82.20
Grade 2	117 (67.2)	111.64 ± 25.78	82.60–247.20	57 (32.8)	70.25 ± 9.80	43.97–82.16
Grade 3	44 (31.9)	102.46 ± 28.08	82.70–245.90	94 (68.1)	63.38 ± 10.20	33.70–82.00
Grade 4	7 (12.3)	112.24 ± 17.60	88.40–137.00	50 (87.7)	52.82 ± 12.68	15.10–78.60
Grade 5	0 (0.0)	—	—	2 (100)	46.75 ± 3.75	44.10–49.40
Total	**393 (63.3)**	**127.56 ± 50.64**	**82.31–552.90**	**228 (36.7)**	**63.84 ± 12.59**	**15.10–82.20**
Pfirrmann grade	**Normal T2r category (>82.2)**			**Abnormal T2r category (≤82.2)**		
	Number of discs (%)	**T2R-NP, mean ± SD**	**Range**	**Number of discs (%)**	**T2R-NP, mean ± SD**	**Range**
Grades 1–2	342 (80.7)	131.10 ± 52.39	82.31–552.90	82 (19.3)	71.49 ± 9.26	43.97–82.20
Grades 3–5	51 (25.8)	103.81 ± 26.96	82.70–245.90	146 (74.2)	59.54 ± 12.19	15.10–82.00

Table 6: Binary logistic regression model performance in the training, validation and external validation sets

Set	AUC (95% CI)	Sensitivity	Specificity	F1 score	Accuracy
Training, cross-validated	0.928 (0.90–0.96)	91.5% (118/129)	75.9% (170/224)	0.784	81.6% (288/353)
Model validation	0.889 (0.83–0.95)	86.0% (49/57)	81.4% (79/97)	0.790	83.1% (128/154)
External validation	0.928 (0.87–0.98)	83.3% (35/42)	87.5% (63/72)	0.814	86.0% (98/114)

Table 7: Prediction Error, Systematic Bias, and Classification Accuracy of the quantitative T2r Model

True T2R-NP	Number of discs in training set	Number of discs in test validation set	Number of discs in external validation	Bias (predicted − true)	True value in the estimated range	Accuracy (Diseased /Normal)
Below 60	44	26	15	+9.2	80%	100%
60–82.2	85	31	27	+9.5	89%	59%
82.2–100	63	30	17	+10.0	88%	88%
100–130	94	34	38	−1.4	84%	92%
Above 130	67	33	17	−38.8	65%	100%

Table 8: Distribution of recommended T2 relaxometry categories across Pfirrmann grades of lumbar intervertebral discs

Recommended T2R categories	Pfirrmann grade 1		Pfirrmann grade 2		Pfirrmann grade 3		Pfirrmann grade 4 & 5	
	n	%	n	%	n	%	n	%
<=60: Markedly Low T2r range	1	0%	10	6%	34	25%	40	68%
61-70: Low T2r range	5	2%	16	9%	36	26%	7	12%
71-90: Transient Zone	33	13%	51	29%	41	30%	6	10%
91-100: Low T2r normal	18	7%	27	16%	12	9%	1	2%
>100: High T2r normal	193	77%	70	40%	15	11%	5	8%

Supplementary Table 1: **Evaluation metrics of the ten regression models for binary T2r-based outcome**

Model	AUC	Average precision	Sensitivity	Specificity	Accuracy	F1 score
Logistic regression	**0.914**	**0.848**	**85.80%**	**85.20%**	**85.40%**	**0.812**
Linear SVM	0.912	0.851	85.50%	83.20%	84.00%	0.797
RBF SVM	0.908	0.826	85.50%	84.40%	84.80%	0.805
XGBoost	0.905	0.834	85.70%	80.30%	82.20%	0.78
Random forest	0.905	0.833	85.70%	81.60%	83.10%	0.788
k-nearest neighbours	0.903	0.826	85.70%	80.90%	82.60%	0.784
Linear discriminant analysis	0.902	0.844	85.80%	77.90%	80.80%	0.767
Extra Trees	0.9	0.834	85.80%	77.60%	80.60%	0.765
Gradient boosting	0.899	0.832	85.50%	77.60%	80.50%	0.763
Naive Bayes	0.845	0.714	85.50%	65.70%	73.00%	0.699

**Comparison of ten classification models on the training and validation sets (507 discs) using patient-level cross-validation, with each model's threshold set to a sensitivity of about 85% for a like-for-like comparison. Logistic regression was selected as the final model, as it achieved the highest AUC (0.914), specificity (85.2%), accuracy (85.4%) and F1 score (0.812).*

Supplementary Table 2: Evaluation metrics of the ten regression model for the quantitative T2r based predictive outcome

Model	MAE	Median_AE	RMSE	R2	Spearman	Accuracy
Support vector regression	**21.685**	**12.144**	**41.641**	**0.401**	**0.82**	**0.855**
Random forest	22.741	13.282	41.88	0.394	0.812	0.846
Extra Trees	22.866	13.441	42.024	0.39	0.811	0.851
XGBoost	22.99	13.059	42.732	0.369	0.811	0.852
Gradient boosting	23.37	13.621	42.466	0.376	0.805	0.848
Elastic net	23.673	15.193	42.454	0.376	0.8	0.824
k-nearest neighbours	23.776	15.351	42.621	0.372	0.792	0.823
Lasso regression	23.915	14.903	42.849	0.364	0.796	0.822
Ridge regression	24.008	14.626	43.823	0.335	0.806	0.828
Partial least squares	24.705	14.76	45.422	0.286	0.8	0.827

**AE - Absolute Error, MAE - Mean Absolute Error, RMSE - Root Mean Squared Error.*

**Support vector regression was selected because it predicted T2R-NP most accurately, with the lowest mean and median error, the highest R^2 and Spearman correlation, and the highest accuracy(85.5%) for separating degenerated from healthy discs. Errors are in T2R-NP units; accuracy is the proportion of discs whose predicted value fell on the correct side of 82.2.*