

Be Indiscrete: The Benefits of Learning Continuous Spine Degeneration Severity Scores

Maria Monzon^{1,2}, Andrew Zisserman³, Robin Y. Park³, Catherine R. Jutzeler^{1,2}, and Amir Jamaludin³

¹ Biomedical Data Science Lab, Dept. D-HEST, ETH Zurich, Zurich, Switzerland

² Swiss Institute of Bioinformatics (SIB), Lausanne, 1015, Switzerland

³ Visual Geometry Group, Dept. of Engineering Science, University of Oxford, UK
mmonzon@ethz.ch

Abstract. Lumbar spine degeneration is a major contributor to chronic low back pain and is routinely assessed on MRI using ordinal grading systems, e.g. normal, mild, moderate, severe. Consequently, most approaches to train models to grade these MRIs formulate grading as a multi-class classification problem, treating ordinal grades as categorical, ignoring differences in misclassification severity, and imposing hard decision boundaries on a continuous disease process. This work explores modeling spinal degeneration as a continuous severity ranking problem. We introduce SpineRankNet, a framework that learns scalar severity scores from lumbar spinal MRI, and compare it against multi-class classification and ordinal regression. Using multiple degeneration measures from the Genodisc dataset, we show that a model trained using a ranking loss to produce a continuous score enables fine-grained ordering of MRI scans. Furthermore, the ordinal grading classes can be recovered from the score with comparable accuracy to those from a model trained directly for classification. The score learned by ranking even improves discrimination between more distant classes. Source code is available at <https://github.com/spinetools/spineranknet>.

Keywords: Pairwise Ranking · Ordinal Regression · Spine Degeneration · MRI

1 Introduction

Lumbar spine degeneration is a major cause of chronic low back pain, expected to affect over 800 million people by 2050 [4]. MRI enables non-invasive assessment of intervertebral disc (IVD) degeneration [20], central canal and foraminal stenosis [9], and vertebral endplate defects, among other conditions routinely graded on ordinal scales reflecting a continuous, progressive disease process.

Clinically, spinal degeneration is assessed using ordinal gradings: for example, Pfirrmann grading [20] assigns intervertebral discs (IVDs) to one of five categories reflecting progressive signal loss and disc height reduction on T2-weighted sagittal MRI, while central canal stenosis is graded on a four-point

scale (“normal”, “mild”, “moderate”, “severe”) [9]. Borderline cases often fall between adjacent categories, contributing to the low-to-moderate inter-rater agreement reported in spinal MRI assessment [5]. Recent deep learning methods have achieved expert-level performance in automating radiological grading of spinal MRI [7, 25, 26, 3]. However, most of these approaches are trained using categorical cross-entropy, treating ordinal grades as independent classes (which fundamentally mismatches the underlying biology) and penalising all misclassifications equally, despite the continuous nature of degeneration (e.g. “severe” misclassified as “moderate” vs “normal”). Moreover, IVD degeneration is a continuous biological process, and collapsing it into discrete labels discards nuance, potentially grouping biologically distinct disease states under a single label. Many spinal MRI pathologies also exhibit extreme class imbalance, where severe cases are rare, leading standard classifiers to bias toward the majority class and underperform on clinically critical minority cases. Simplifying the problem to binary classification (normal vs. abnormal) as in [7] also sacrifices clinically meaningful granularity.

This work models spinal degeneration as a continuous value, which overcomes the problems identified above with categorical scoring. First, we reformulate training as a ranking problem, learning a continuous scoring function that preserves the relative ordering of disease severity, instead of predicting absolute labels. Penalisation for error in ranking during training preserves the severity gap, making large errors (e.g. “normal” vs “severe”) explicitly more costly than adjacent misclassifications (e.g. “mild” vs “moderate”). The resulting score, learned by ranking, also captures the within-grade variability and naturally mitigates class imbalance, since severe cases appear in many training pairs regardless of their absolute frequency. Despite these advantages, ranking has been largely unexplored in structured ordinal grading of spinal pathology.

A continuous severity score is also more practical than discrete bins. It can serve as a more sensitive measure for detecting subtle longitudinal changes, supporting earlier identification of disease progression or treatment response. It would also facilitate relative comparisons during annotation, allowing clinicians to characterise the nuance of disease, rather than forcing an absolute categorical decision. By preserving ordinal structure on a continuous scale, the framework allows for more clinically-consistent transitions across grades that better reflect the gradual nature of spinal degeneration.

In summary, we make the following contributions: *(i)* we introduce *SpineRankNet*, a multi-pathology framework that produces continuous severity scores for lumbar spine MRI grading; *(ii)* we conduct a large-scale comparison of classification, ordinal regression, and ranking across 11 concurrent spinal grading tasks on the multicentre Genodisc cohort; *(iii)* we show that ranking consistently reduces large-grade errors; and *(iv)* we demonstrate that a more granular, continuous severity scale allows for better separation of borderline cases.

1.1 Related Work

Radiological assessment of lumbar degeneration has traditionally relied on geometric biomarkers, i.e. disc height ratios, canal diameters, and foraminal widths derived from manual or semi-automated measurements [20, 14]. Recent models use either segmentation-derived metrics [16, 15, 27] or end-to-end deep learning and on Pfirrmann grading [8, 13], stenosis [12], herniation [10], and Modic changes [3], but most existing models have focused only on individual pathologies. Among these, SpineNet [7, 25] and the Context-Aware Spinal Transformer [24, 26] are notable exceptions, performing concurrent multi-pathology grading within a unified architecture. However, they still treat ordinal severity labels as independent nominal classes and ignore the ordering and magnitude of grading errors.

Recent ordinal regression methods [18, 22] have partially addressed this issue by enforcing rank consistency, demonstrating improvements over nominal cross-entropy in medical grading tasks [21]. However, they still discretise the severity continuum using fixed thresholds, leading to unstable decisions at grade boundaries and collapsing within-grade variability into a single label. Pairwise ranking offers an alternative by learning a continuous scoring function from relative severity comparisons rather than categorical labels [6]. Such relative-attribute frameworks are well established in computer vision [17, 23, 2, 1], where continuous scores from relative comparisons are highly discriminative for ordered attributes. In medical image analysis, ranking has been applied to preserve clinical severity ordering from imaging data [19], to learn continuous scores for single conditions such as knee osteoarthritis [11]. However, these applications remain limited to individual pathologies and have not been systematically compared against classification and ordinal regression baselines. To our knowledge, pairwise ranking has never been applied to structured multi-pathology severity grading of the spine.

2 Spine Degeneration Continuous Score

This section describes our method to predict a continuous score for each grading task from a single IVD input. Unlike classification, which treats grades as categorical, or ordinal regression, which discretises at fixed thresholds, our ranking formulation preserves the continuous disease spectrum while learning from relative severity comparisons. At inference, our model outputs both a continuous score and a corresponding radiological grade (see Fig. 1). We train this model using pairwise ranking procedure, which is described next.

2.1 Ranking to be indiscrete

For each grading task (e.g. Pfirrmann or narrowing) our aim is to learn a scalar scoring function for the IVD x :

$$s_t^{(\theta)} : x \rightarrow [0, S],$$

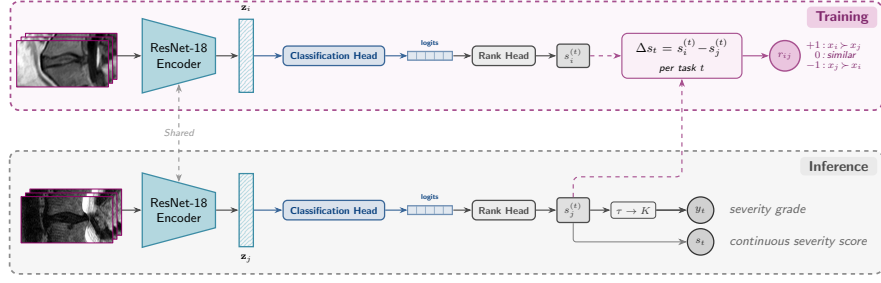


Fig. 1: **Multi-task architecture for joint classification and ranking.** A shared 3D ResNet-18 encoder produces an embedding $\mathbf{z}$, followed by task-specific classification and ranking heads. Each ranking head outputs a continuous severity score $s^{(t)} \in [0, 10]$; for clarity, the figure shows one representative task head. Pairwise supervision is derived from ground-truth grade orderings $r_{ij}^{(t)} = \text{sign}(y_i^{(t)} - y_j^{(t)})$ and optimized with a severity-aware hinge loss.

such that the score is aligned with the K ordinal clinical grades for that task $y_t \in \{0, \dots, K_{t-1}\}$. We set $S=10$ across all tasks, as this range aligns with common clinical degeneration scales.

We use pairwise ranking supervision to learn the scoring function, and extend the ranking SVM loss [17, 1] with severity-aware components to achieve this.

Ranking loss. To train the model on pairwise comparisons, we construct pairs (x_i, x_j) with relative labels $r_{ij}^t \in \{-1, 0, +1\}$ indicating whether IVD x_i is less severe than, equally severe to, or more severe than disc x_j for task t .

A severity-weighted hinge loss is applied to ordinal pairs $\mathcal{P}^t = (x_i, x_j)$ with different grades ($y_i^t \neq y_j^t$):

$$\mathcal{L}_{\text{hinge},m}^t = \frac{1}{|\mathcal{P}^t|} \sum_{(x_i, x_j) \in \mathcal{P}^t} w_{ij}^t [\max(0, m_{ij}^t - r_{ij}^t \Delta s_{ij}^t)]^2 \quad \text{if } d_{ij}^t \neq 0$$

where $r_{ij}^t = \text{sign}(y_i^t - y_j^t) \in \{-1, +1\}$ is the label direction, and $\Delta s_{ij}^t = s_i^t - s_j^t$ is the score difference. This term penalizes both ranking errors (wrong direction) and violations of the margin requirement. The severity-aware margin, m_{ij}^t , and weight, w_{ij}^t are defined below.

To prevent score spread within a single grade from fragmenting the learned severity axis, same-grade pairs are penalised by a squared score difference [17]:

$$\mathcal{L}_{\text{sim}}^t = \frac{1}{|\mathcal{S}^t|} \sum_{(x_i, x_j) \in \mathcal{S}^t} (\Delta s_{ij}^t)^2 \quad \text{if } d_{ij}^t = 0 \quad (1)$$

The total training loss sums over all T tasks combining the two terms for each task t :

$$\mathcal{L} = \sum_{t=1}^T \mathcal{L}_{\text{rank}}^t = \sum_{t=1}^T \left(\lambda_{\text{dis}} \mathcal{L}_{\text{hinge}_m}^t + \lambda_{\text{sim}} \mathcal{L}_{\text{sim}}^t \right) \quad (2)$$

where $\lambda_{dis} = 1$, $\lambda_{sim} = 0.25$ represents the trade-off coefficients between enlarging the margin and satisfying the pairwise relative constraints [1].

Difference-specific weight. To reflect the clinical significance of grade differences, each ordinal pair is weighted by its normalized grade distance:

$$w_{ij}^t = |y_i^t - y_j^t| \cdot \frac{S}{K_t - 1}, \quad (3)$$

where y_i^t, y_j^t are ground-truth gradings, and the denominator $K_t - 1$ normalizes weights across tasks with different grade granularities.

Severity-aware margin. Unlike standard ranking SVM [17, 1] that applies a fixed margin m to all pairs, we adapt the margin to both the severity gap and the task-specific grade structure

$$m_{ij}^t = m_{\text{base}} + m_{\text{scale}} \cdot \frac{w_{ij}^t}{S}$$

with $m_{\text{base}} = 0.5$ and $m_{\text{scale}} = 1.5$, adjacent grades (small w_{ij}) require minimal margin, while extreme pairs (large w_{ij}) or severe misrankings require proportionally larger score separations (to overcome the poor scaling critiqued in [23]).

Score Parameterization For a given input IVD volume $x^{(i)}$, we first extract a shared feature representation using a convolutional encoder.

$$\mathbf{z}^{(i)} = f_{\theta}(x^{(i)}) \in \mathbb{R}^d, \quad (4)$$

shared across all T grading tasks. Per-task linear heads first map $\mathbf{z}^{(i)}$ to grade logits $\hat{y}_t^{(i)} \in \mathbb{R}^{K_t}$ (K_t ordinal grades), pretrained with inverse-frequency-weighted cross-entropy; the ranking heads are then trained on top to produce continuous severity scores (Fig. 1). For each task t , we learn a lightweight projection head

$$\varphi_t : \mathbb{R}^d \rightarrow \mathbb{R},$$

implemented as a small two-layer MLP, that maps the shared embedding to a scalar latent score

$$h_t^{(i)} = \varphi_t(\mathbf{z}^{(i)}). \quad (5)$$

The latent score is then transformed into a bounded continuous severity score

$$s_t^{(i)} = \sigma_S(h_t^{(i)}) = S \cdot \frac{\text{sp}(\beta h_t^{(i)})}{\text{sp}(\beta h_t^{(i)}) + \text{sp}(\beta (S - h_t^{(i)}))}, \quad (6)$$

where $\text{sp}(z) = \ln(1 + e^z)$ is the softplus function and $\beta = 5$ controls the boundary sharpness. This soft-clamp guarantees $s_t^{(i)} \in [0, S]$.

To obtain a discrete severity grade from the continuous severity score, we introduce $K-1$ ordered thresholds $\boldsymbol{\tau} = \{\tau_1, \dots, \tau_{K-1}\}$ that partition each task-specific score into K ordinal categories.

3 Dataset and Implementation Details

Dataset. The Genodisc dataset [7] includes T1w and T2w sagittal lumbar MRIs from $\sim 2,000$ subjects across five European centres with heterogeneous acquisition parameters (0.2–3.0 T, 0.25–0.9 mm in-plane, 2.6–6.0 mm slice spacing). Each subject has up to six lumbar IVD levels (T12-L1 to L5-S1). Expert radiological annotations cover 11 grading tasks per IVD level, spanning disc degeneration (Pfirrmann grading, disc narrowing, herniation, bulging), spinal canal and nerve compression (central canal stenosis, foraminal stenosis), structural abnormalities and vertebral degeneration (spondylolisthesis, endplate changes). The dataset exhibits severe class imbalance: 60–96% normal prevalence, with severe cases comprising only 2–10% (e.g., severe stenosis: 2%).

IVD volumes ($128 \times 256 \times 12$ voxels) are extracted via SpineNetv2 [25] vertebra localization, disc-centered rotation, and resampling. Data is split at the patient level into train (80%), validation (10%), and test (10%), stratified by acquisition centre and pathological grade distribution to ensure representative class prevalence across all partitions.

Evaluation. We report balanced accuracy (BA), quadratic weighted Cohen’s kappa (QWK), receiver operating characteristic area under the curve (ROC-AUC), Matthews correlation coefficient (MCC) and mean absolute error (MAE) on a held-out test set. Deterministic test-time augmentation (TTA) evaluates each scan across 54 views ($2 \text{ flips} \times 3 \text{ horizontal translations} \times 3 \text{ vertical translations} \times 3 \text{ slice shifts}$). Final predictions average the softmax probabilities or predicted severity scores across views.

Architecture and training. All implementations share a 3D ResNet-18 encoder with asymmetric convolutions (strides $(1, s, s)$) that preserve the slice dimension, reflecting the anisotropic resolution of spinal MRI. For each task t , a linear head maps the pooled features $\mathbf{z} = \text{GAP}(\Phi(\mathbf{x})) \in \mathbb{R}^{512}$ to logits $\hat{\mathbf{y}}_t \in \mathbb{R}^{K_t}$, trained with inverse-frequency-weighted cross-entropy ($w_c = N_t / (K_t \cdot N_{t,c})$). T1w/T2w sequences are randomly sampled during training, with Pfirrmann/disc narrowing labels masked for T1w. Encoder and classification heads are trained from scratch (AdamW, $\text{lr} = 10^{-3}$, $\text{wd} = 10^{-4}$) for up to 300 epochs (early stopping: patience 30, min 50 epochs; batch size 32). Ranking heads are then optimized for 200 epochs while fine-tuning the encoder ($\text{lr} = 10^{-4}$) to prevent catastrophic forgetting. Thresholds τ are grid-searched on validation data.

Training augmentations include sagittal slice shifts (± 2 slices, with boundary slices zeroed), left-right slice-order flips ($p=0.5$), intensity offset (± 0.1), random translations (up to 32 px horizontally, 24 px vertically), isotropic scaling ($\pm 10\%$), and in-plane rotation ($\pm 15^\circ$).

Pair construction for training. To construct pairs of training examples, we sample candidate pairs from B samples in a mini-batch. For a given pair (i, j) , we calculate the signed difference between the grades, $d_{ij}^t = y_t^{(i)} - y_t^{(j)}$, for each grading task t . From the $\binom{B}{2}$ candidates in a batch of $B=32$, 70% of pairs per task are selected by descending violation magnitude (hard-negative mining) and 30% sampled uniformly to preserve training stability.

Table 1: Comparison on the Genodisc test set (ResNet-18), aggregated over 11 grading tasks. Mean $\pm$ std over the individual task performances (single run)

Loss	BA (%) $\uparrow$	ROC-AUC $\uparrow$	QWK $\uparrow$	MAE $\downarrow$	MCC $\uparrow$
CE	64.1 $\pm$ 7.9	0.94 $\pm$ 0.02	0.69 $\pm$ 0.12	0.18 $\pm$ 0.09	0.54 $\pm$ 0.08
CORN	56.9 $\pm$ 9.8	—*	0.72 $\pm$ 0.12	0.15 $\pm$ 0.10	0.55 $\pm$ 0.10
MAE $_{\ell}$	58.8 $\pm$ 15.7	0.84 $\pm$ 0.13	0.54 $\pm$ 0.24	0.39 $\pm$ 0.30	0.38 $\pm$ 0.20
RankNet	58.7 $\pm$ 8.3	0.90 $\pm$ 0.03	0.73 $\pm$ 0.11	0.16 $\pm$ 0.09	0.57 $\pm$ 0.07
DeepRankSVM	61.0 $\pm$ 8.4	0.88 $\pm$ 0.05	0.74 $\pm$ 0.11	0.16 $\pm$ 0.09	0.57 $\pm$ 0.07
SpineRankNet	63.3 $\pm$ 9.5	0.94 $\pm$ 0.04	0.76 $\pm$ 0.10	0.14 $\pm$ 0.09	0.59 $\pm$ 0.08

*ROC for CORN not available

Classification, regression, and ranking. We benchmark against representative methods from three objective families under identical 3D ResNet-18 encoders and evaluation protocols: (i) Classification (inverse-frequency-weighted CE); (ii) Ordinal regression (CORN [22] with cumulative binary decomposition, direct MAE $_{\ell}$ regression); (iii) Pairwise ranking (RankNet [2], DeepRankSVM [1]). All ranking methods use two-phase training: classification heads pretrained with weighted CE, then ranking heads are finetuned.

4 Experimental Results

Our experiments aim to establish that ranking-trained continuous scores match direct classification performance when discretized, and to highlight their added clinical value over discrete ordinal grades.

Classification and ordinal baselines vs pairwise ranking. Table 1 shows that classification (CE) and ordinal regression (CORN) impose a trade-off between accuracy and ordinal concordance, whereas pairwise ranking methods yield the best ordinal metrics. SpineRankNet achieves superior ordinal agreement (QWK 0.76, MAE 0.14) while maintaining classification-level discrimination (ROC-AUC 0.94). Notably, its lowest MAE demonstrates systematic reduction of large-grade prediction errors—critical for clinical safety—while highest MCC confirms robustness under severe class imbalance, all without compromising overall accuracy. It further outperforms prior ranking models (RankNet, DeepRankSVM), validating the severity-adaptive margin and same-grade similarity loss.

Ranking Layer Selection and Thresholding. For threshold selection, coordinate-descent grid search (coarse 200-point sweep, two refinements, Nelder–Mead polish) achieves BA 63.1%/QWK 0.762, surpassing isotonic regression fitting monotonic score $\rightarrow$ grade mappings (55.7%/0.744), Youden’s index optimizing per-boundary ROC thresholds (58.8%/0.685), and GMM clustering scores into K Gaussians (54.3%/0.448). For ranking head architecture, a two-layer MLP achieves BA 63.1%/QWK 0.762, outperforming linear projection (58.3%/0.733), Kolmogorov-Arnold Networks (61.0%/0.698), and transformer heads (41.8%/0.303).

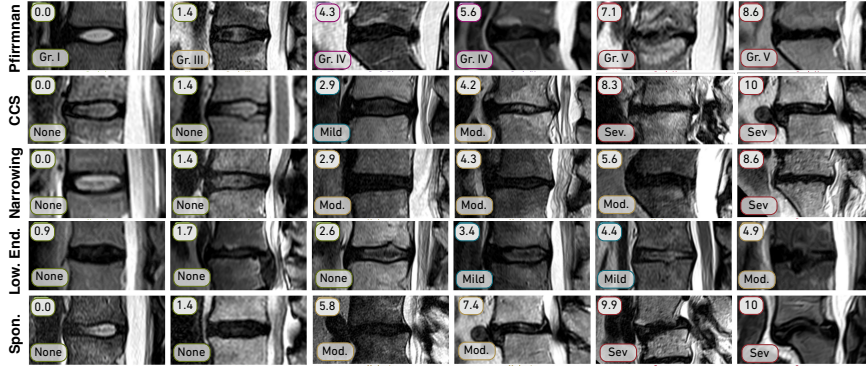


Fig. 2: Example visualization of severity ranking for the pathology scores. IVD volumes for Pfirrmann, central canal stenosis (CCS), Narrowing, and Lower Endplate defects (Low. End) and Spondylolisthesis, ordered by predicted score $s \in [0, 10]$ (top) with ground-truth ordinal grades (bottom). Note how the severity progressively increases with score. For example, for CCS, “None” cases (0.0 vs. 1.4) differentiate normal from subtle central canal compression, while “Severe” cases (8.3→10) capture progressive cord impingement

Continuous score degeneration. Fig. 2 further shows that the learned continuous severity scores capture intra-grade variability that categorical models cannot represent (e.g., mild-to-moderate discs span 2.7–4.1 despite identical grade labels; advanced degeneration cases cluster between 7.5–9.8), consistent with the quantitative ordinal improvements reported in Table 1.

5 Conclusion and Future Work

We provide a comparison of classification, ordinal regression, and pairwise ranking objectives for highly-imbalanced, multi-pathology spinal grading on lumbar MRI. Experiments on the multi-centre Genodisc cohort show that ranking-based models shift the error profile toward adjacent-grade confusions while reducing clinically distant misclassifications, reflected by consistent QWK improvements across the majority of 11 grading tasks and pairwise formulations. The continuous severity scores also identify borderline samples near inter-grade thresholds, offering a finer-grained view of disease progression than discrete labels alone.

While pairwise ranking shows promise for ordinal grading tasks where error magnitude matters more than per-class accuracy, all experiments use a single dataset. Future work should include external validation on truly independent cohorts, evaluation under longitudinal scenarios, and assessment of clinical utility by radiologists.

Acknowledgements. This project was supported by a grant (#380, Jutzeler, Manjaly) from the Strategic Focus Area “Personalized Health and Related Technologies (PHRT)” of the ETH Domain (Swiss Federal Institute of Technology), by the Swiss National Science Foundation (Ambizione Grant #PZ00P3_186101, CRJ). This project was also funded by the EPSRC programme grant Visual AI (EP/T025872/1) and EPSRC CDT in Health Data Science (EP/S02428X/1). We thank our clinical collaborators Prof. J. Fairbank, Dr. J. Urban, Dr. S. Ather, and Prof. I. McCall. AI-assisted tools (GitHub Copilot, Claude, Perplexity, Writefull) were used for coding and grammar refinement; the authors reviewed all output and take full responsibility for the content.

Disclosure of Interests. C. R. serves as a scientific consultant to AbbVie. This role is unrelated to the present work and had no influence on the design, analysis, interpretation, or reporting of the study. The remaining authors have no relevant competing interests to declare.

References

1. Ahmed, S.A.A., Yanikoglu, B.: Relative Attribute Classification with Deep-RankSVM. In: ICPR. p. 659–671 (2021)
2. Burges, C.J.C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.N.: Learning to rank using gradient descent. In: ICML. pp. 89–96 (2005)
3. Cina, A., Haschtmann, D., Damopoulos, D., Gerber, N., Loibl, M., Fekete, T., Mannion, A., Galbusera, F.: Modic change classification from sagittal MRI with 3D convolutional neural networks and comparative analysis of normalization techniques. *Brain and Spine* **3**, 102401 (2023)
4. Ferreira, M.L., et al.: Global, regional, and national burden of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *The Lancet Rheumatology* **5**(6), e316–e329 (2023)
5. Herzog, R., Elgort, D.R., Flanders, A.E., Moley, P.J.: Variability in diagnostic error rates of 10 MRI centers performing lumbar spine MRI examinations on the same patient within a 3-week period. *The Spine Journal* **17**(4), 554–561 (2017)
6. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. *Artificial Intelligence* **172**(16–17), 1897–1916 (2008)
7. Jamaludin, A., Kadir, T., Zisserman, A.: SpineNet: Automated classification and evidence visualization in spinal MRIs. *Medical Image Analysis* **41**, 63–73 (2017)
8. Kowlagi, N., Nguyen, H.H., McSweeney, T., Saarakkala, S., Määtä, J., Karppinen, J., Tiulpin, A.: A stronger baseline for automatic Pfirrmann grading of lumbar spine MRI using deep learning. In: ISBI. pp. 1–5 (2023)
9. Lee, G.Y., Lee, J.W., Choi, H.S., Oh, K.J., Kang, H.S.: A new grading system of lumbar central canal stenosis on MRI: an easy and reliable method. *Skeletal Radiology* **40**(8), 1033–1039 (2011)
10. Lewandrowski, K.U., de Carvalho, P.S.T., Calderaro, A.L., Dos Santos, T.S., de Lima E Silva, M.S., de Carvalho, P., J., Yeung, A.: Outcomes with transforaminal endoscopic versus percutaneous laser decompression for contained lumbar herniated disc: a survival analysis of treatment benefit. *Journal of Spine Surgery* **6**(Suppl 1), S84–S99 (2020)

11. Li, M.D., Chang, K., Bearce, B., Chang, C.Y., Huang, A.J., Campbell, J.P., Brown, J.M., Singh, P., et al.: Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *npj Digital Medicine* **3**(1) (2020)
12. Lu, J.T., Pedemonte, S., Bizzo, B., Doyle, S., Andriole, K.P., Michalski, M.H., Gonzalez, R.G., Pomerantz, S.R.: Deep spine: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. In: *MLHC*. pp. 403–419. PMLR (2018)
13. Minin, A., Leonova, O., Krutko, A., Elgaeva, E., Antonets, D., Shtokalo, D., Tsepilov, Y.: SpineScan: a deep learning model for lumbar spine MRI annotation and Pfirrmann grading assessment. *European Spine Journal* (2025)
14. Modic, M.T., Steinberg, P.M., Ross, J.S., Masaryk, T.J., Carter, J.R.: Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. *Radiology* **166**(1), 193–199 (1988)
15. Natalia, F., Meidia, H., Afriliana, N., Young, J.C., Yunus, R.E., Al-Jumaily, M., Al-Kafri, A., Sudirman, S.: Automated measurement of anteroposterior diameter and foraminal widths in MRI images for lumbar spinal stenosis diagnosis. *PLOS ONE* **15**(11), e0241309 (2020)
16. Neubert, A., Fripp, J., Engstrom, C., Walker, D., Weber, M.A., Schwarz, R., Crozier, S.: Three-dimensional morphological and signal intensity features for detection of intervertebral disc degeneration from magnetic resonance images. *Journal of the American Medical Informatics Association* **20**(6), 1082–1090 (2013)
17. Parikh, D., Grauman, K.: Relative attributes. In: *ICCV*. p. 503–510 (2011)
18. Pedregosa, F., Bach, F.R., Gramfort, A.: On the consistency of ordinal regression methods. *J. Mach. Learn. Res.* **18**, 55:1–55:35 (2014)
19. Pedregosa, F., Cauvet, E., Varoquaux, G., Pallier, C., Thirion, B., Gramfort, A.: Learning to rank from medical imaging data. In: *International Workshop on Machine Learning in Medical Imaging*. pp. 234–241. Springer (2012)
20. Pfirrmann, C.W., Metzdorf, A., Zanetti, M., Hodler, J., Boos, N.: Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine* **26**(17), 1873–1878 (2001)
21. Saffari, S.E., Löve, , Fredrikson, M., Smedby, O.: Regression models for analyzing radiological visual grading studies – an empirical comparison. *BMC Medical Imaging* **15**(1) (2015)
22. Shi, X., Cao, W., Raschka, S.: Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications* **26**(3), 941–955 (2023)
23. Souiri, Y., Noury, E., Adeli, E., Mousavi, P., Tizhoosh, H.R.: Deep relative attributes. In: *ACCV*. pp. 118–133. Springer (2016)
24. Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: Context-aware transformers for spinal cancer detection and radiological grading. In: *MICCAI*. pp. 271–281 (2022)
25. Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: SpineNetV2: Automated detection, labelling and radiological grading of clinical MR scans (2022), <https://arxiv.org/abs/2205.01683>
26. Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: Automated detection, labelling and radiological grading of clinical spinal MRIs. *Scientific Reports* **14**(1) (2024)
27. Zheng, H.D., Sun, Y.L., et al.: Deep learning-based high-accuracy quantitation for lumbar intervertebral disc degeneration from MRI. *Nature Communications* **13**, 841 (2022)