Multi-View and Multimodal Radiological Grading Using Spinal MRIs

Robin Y. Park, Rhydian Windsor, Amir Jamaludin, and Andrew Zisserman

Visual Geometry Group, Department of Engineering Science,
University of Oxford, Oxford, UK
robinpark@robots.ox.ac.uk

Abstract. This paper proposes a transformer-based model that encodes MRI volumes from multiple sequences and anatomical views of the spine, and predicts multiple spinal gradings. The transformer ingests slice-wise 2D embeddings and a learnable class token to capture the relationships between the slice-wise embeddings. The method is applied to predict fine-grained radiological gradings of spinal stenosis conditions (spinal canal stenosis, right and left neural foraminal narrowing, and right and left subarticular stenosis) using T1-weighted, T2-weighted and STIR sequences in sagittal and axial views. Experiments show that our joint multi-view, multimodal model outperforms task-specific baselines trained on individual modalities or views across all tasks.

Keywords: Multimodal learning · Radiological grading · Stenosis

1 Introduction

Lumbar spinal stenosis affects over 100 million people worldwide and is a prevailing cause of disability in older adults [8, 11]. Developing accurate and rapid methods for diagnosing this condition can have significant impacts, both at the individual and societal level [2, 19, 23]. The current gold standard method for diagnosing stenosis and other spinal degenerative diseases is through multimodal, multi-view Magnetic Resonance Imaging (MRI) studies [13, 25]. By acquiring multiple scans with varying sequence parameters and views, clinicians can gather information on a wide range of different tissue characteristics that, in combination with expert judgment, allows for a more comprehensive diagnosis.

Given the need for accurate and rapid diagnosis of these images, several groups have investigated automating this process using deep learning (e.g. [12, 21]). Typically, this is achieved through multi-stage pipelines which first detect and label the vertebrae in each scan and then pass volumes surrounding each detected vertebra or intervertebral unit to a downstream classification network that grades disease severity.

A key challenge for such methods is making best use of the multiple streams of data provided by different scan types. Since incorporating more views and sequences requires more training data and complicated pre-processing, existing published pipelines typically focus on a single view (e.g. SpineNet [21]) or MR

pulse sequence (e.g. DeepSPINE [12]). However, this risks missing key information from an omitted scan type.

This paper describes an automated pipeline for diagnosing degenerative conditions that considers all information in a typical clinical multi-view, multi-sequence MR study to predict five stenosis gradings on the 2024 RSNA Lumbar Spine Degenerative Classification Challenge. Through careful design of the method used to extract vertebral volumes from axial sequences and adopting a transformer-based architecture to fuse information from multimodal input scans while remaining flexible to missing data, our method achieves high performance on all grading tasks, obtaining a macro average Area Under the Receiver Operating Characteristic (AUROC) curve of 0.961 for spinal canal stenosis, 0.886 for left neural foraminal narrowing, 0.875 for right neural foraminal narrowing, 0.904 for left subarticular stenosis and 0.909 for right subarticular stenosis.

In summary, our contributions are: (1) an end-to-end pipeline to provide fine-grained stenosis gradings, including the extraction of both sagittal and axial MR volumes; (2) a novel model architecture that effectively integrates MR volumes of different modalities and views; and (3) demonstrating that a joint model that incorporates all available views and modalities outperforms task-specific models trained on single sequences. The model code and weights will be open-sourced.

1.1 Related Work

There are several existing models that perform level-wise labelling, volume extraction and grading of spinal conditions. SpineNet [21] provides automated gradings of stenosis alongside other tasks (e.g. Pfirrmann grading and disc narrowing), relying solely on T2-weighted (T2w) sagittal sequences to predict gradings. DeepSPINE [12] incorporates axial views for stenosis gradings but uses only T2w sequences and 3D convolutions to extract features from volumes, which assumes a constant slice thickness across images (often not the case in clinical practice) and cannot be used to operate on 2D images. Context-Aware Spinal Transformer (CAST) [20, 21] uses slice-wise 2D encodings and can perform inference on multiple modalities and vertebral levels simultaneously, but has only been used with sagittal sequences. In recent years, many other transformer-based models have been used to successfully detect a variety of spinal conditions from medical images [9, 14, 15, 16, 18, 20, 22, 24].

Our approach operates on scans from multiple views (sagittal and axial) and modalities (T1w, T2w and STIR) to train a multi-task transformer-based model for stenosis grading. This extends the capabilities of SpineNetv2 [21] and DeepSPINE [12] by allowing inference from axial and non-T2w MRI volumes, using an adaptation of the CAST architecture [20, 21] designed to learn a joint representation of a disc across an arbitrary number of MRI volumes and slices within each volume. While all three models mentioned have been trained to predict stenosis, our tasks require a finer discernment between stenosis types than these models can detect due to the inclusion of subarticular stenosis, which occurs in the region between the central canal and the neural foramen. Our architecture is much simpler than CAST and is described in detail in section 2.2.

Several other groups have published work using data from the 2024 RSNA Lumbar Spine Degenerative Classification Challenge [5, 7, 10]. However, none of these papers have a vertebral detection or labelling step, instead opting to provide whole images as inputs. [7] collapses the distinct stenosis conditions into a single classification task. [5] treats the annotation points as a detection task, rather than predicting the radiological gradings. [10] reports condition-specific AUROC using unimodal task-specific models, to which our results are compared.

2 Model Architecture

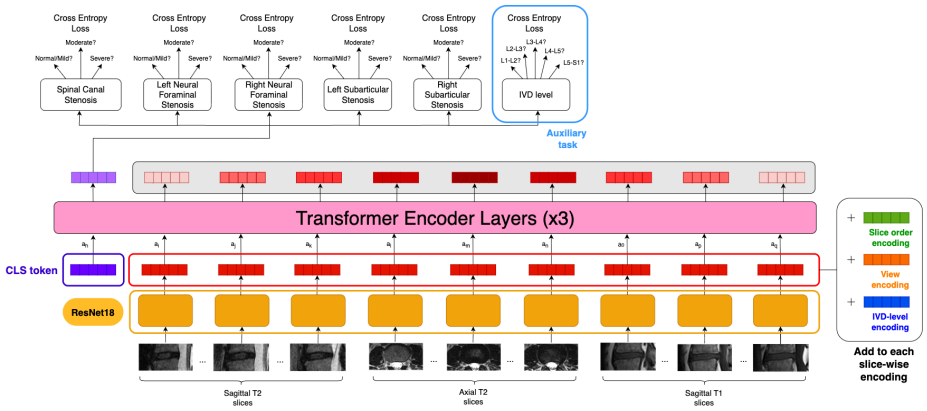

Fig. 1. Model architecture: Transformer-based model operating on multiple MRI sequences of a single intervertebral disc (IVD) for fine-grained stenosis prediction. A global class token (CLS) is prepended to slice-wise 2D ResNet18 embeddings generated using a shared encoder, then simultaneously passed through three transformer encoder layers. The CLS token attends to and aggregates information during training. The CLS token output from the transformer is used to predict the severity of each classification task using separate linear heads. While we show T1w sagittal, T2w sagittal and T2w axial sequences, our data often include STIR scans in lieu of T2w sagittal sequences. The model can operate on an arbitrary number of input sequences, modalities, and slices within each volume.

We propose a transformer-based architecture for the analysis of multi-sequence and multi-view MR volumes (see Figure 1). The model computes slice-wise embeddings from multiple views and sequences of an intervertebral disc (IVD) volume, which are jointly processed to capture both spatial context and cross-modal interactions. A global (CLS) token is appended to slice-wise embeddings from all modalities and then simultaneously passed through a three-layer transformer encoder to model contextual dependencies across slices. The CLS token output from the transformer encoder is subsequently passed through linear classification heads for the multi-class radiological grading tasks. The model can handle

varying types of input sequences in both training and inference. We explain each module in greater depth below, and then contrast the model with previous work.

2.1 Model details

Slice encodings and class token. Each slice of an MRI volume is encoded using a 2D ResNet18 [6] encoder from scratch, adapted for single-channel images. The same encoder is shared across modalities. To each slice-wise encoding, additional embedding vectors are added: (i) slice order, which describes the slice index position within the volume (1 to 9), (ii) view type (axial or sagittal) and (iii) IVD level (L1-L2 to L5-S1), which are all calculated using linear layers applied to one-hot encodings. These added embeddings preserve key spatial and modality-specific information while the model jointly processes multimodal volumes. A global class token [3] is prepended to the slice-wise encodings to attend to and aggregate information across multimodal embeddings.

Transformer encoder layers. The class token and multimodal slice embeddings are simultaneously fed into a stack of three transformer encoder layers. Each encoder layer consists of multi-head self-attention followed by a position-wise feedforward network, with layer normalization and residual connections, which enables the model to capture spatial dependencies across slices, cross-modal interactions and learn contextualised representations of each slice.

Classification heads. The class token representation is passed through six separate classification heads: one for each condition (spinal canal stenosis, left/right neural foraminal narrowing and left/right subarticular stenosis) and an IVD level prediction head. Explicitly learning the IVD level should encourage the model to learn the associations between anatomical characteristics (which subtly vary across levels) and features of disease. The condition-specific heads generate probabilities corresponding to the grading classes: normal/mild, moderate and severe. The level prediction head, used for auxiliary loss, generates probabilities for each lumbar IVD level.

Dimensionality across modules. Each 2D MRI slice is embedded into a fixed-dimensional vector of size 512, resulting in an input tensor of shape $(B, M, S, 512)$, where B is the batch size, M is the number of modalities, and S is the number of slices per modality. The class tokens of shape $(B, 1, 512)$ are prepended to the flattened slice-wise embeddings $(B, M \times S, 512)$, and then passed into a three-layer transformer encoder, each with two attention heads. The resulting class token representation of shape $(B, 512)$ is then passed to each linear head for classification.

Loss function. Cross-entropy loss is computed for each condition. To address class imbalance, class weights inversely proportional to class frequencies in the training set are applied. The main classification loss is the average of the five condition-specific losses. An auxiliary loss to predict the IVD level is computed so that the model can learn the associations between the visual features and anatomical levels. The total loss is computed as a weighted sum of the main classification loss and the auxiliary loss, with weights of 0.8 and 0.2, respectively.

2.2 Design choices

In contrast with CAST [20, 21], slice-wise embeddings are not aggregated to the volume level before passing the embeddings through the transformer encoder layers. While [20] used multiple IVD levels as input to generate predictions at the whole spine level (e.g. spinal cancer), we provide volumes for a given IVD level to predict a level-specific grading. Since there are far fewer slices across the input volumes, they can be provided directly as tokens to the transformer layers, allowing the module to better learn localised features and inter-slice relationships that might be lost when taking volume-level embeddings as inputs. In addition, level information is encoded alongside visual features, and an auxiliary classification head is added for level prediction, which is intended to help the model better learn anatomical correlations with disease in downstream tasks.

3 Dataset

We use a publicly available dataset of anonymised MRI scans provided by the Radiological Society of North America (RSNA) for the 2024 Lumbar Spine Degenerative Classification Challenge [17]. The dataset includes 6,294 sequences (1,975 patients) from 12 contributing institutions from Asia, Europe, the Americas and Australia, with level-specific labels manually annotated by more than 50 clinicians. At each level from L1 to S1, spinal canal stenosis was annotated using T2w or STIR sagittal sequences, neural foraminal stenosis was graded using T1w sagittal sequences and subarticular stenosis was graded using T2w axial sequences. Alongside the gradings, the dataset included coordinates and the slice-wise index within each volume used to provide the annotation.

Table 1. Distribution of data by condition and severity: SCS stands for spinal canal stenosis, NFS for neural foraminal narrowing and SS for subarticular stenosis. 0 indicates Normal/Mild, 1 indicates Moderate and 2 indicates Severe gradings. Pat refers to number of patients. The number of IVDs vary across conditions for train and validation sets due to missing labels. The test set is limited to those IVDs with all valid labels and extracted IVDs across conditions.

		IVDs														
		SCS			**NFN (L)**			**NFN (R)**			**SS (L)**			**SS (R)**		
Split	**Pat**	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**
Train	1552	6710	757	371	6009	1424	313	6024	1383	295	5203	1431	722	5221	1421	723
Val	194	837	74	48	747	177	45	749	180	38	652	180	38	647	184	86
Test	194	729	69	38	661	146	29	652	152	32	600	157	79	585	158	93

Annotations are provided for five different definitions of stenosis at three levels of severity (normal/mild, moderate and severe): spinal canal stenosis, left and right neural foraminal stenosis and left and right subarticular stenosis. Figure 2 shows examples of each condition at various levels.

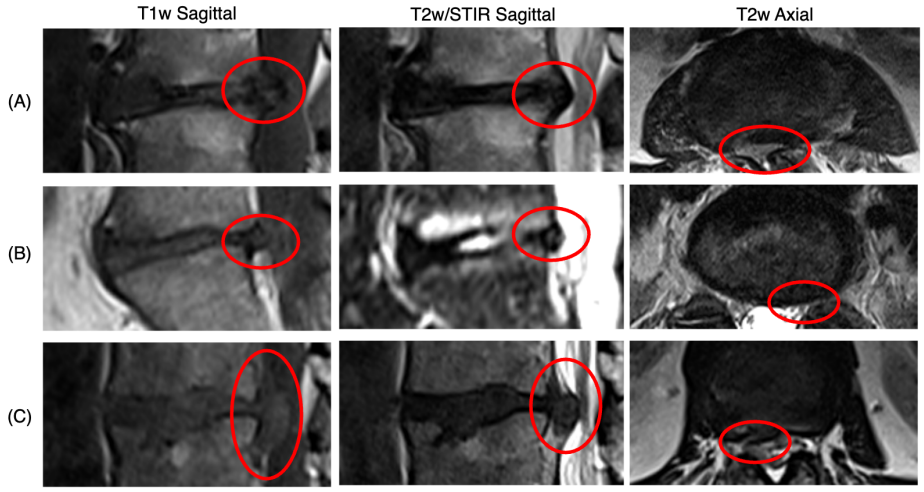

Fig. 2. IVDs with Stenosis: (A) shows T1w sagittal, STIR sagittal and T2w axial scans of an L2-L3 IVD with severe spinal canal stenosis. (B) shows T1w sagittal, T2w sagittal, and T2w axial scans of an L5-S1 IVD with severe left neural foraminal stenosis. (C) shows T1w sagittal, T2w sagittal, and T2w axial scans of an L1-L2 IVD with severe right subarticular stenosis. Stenosis is marked in red in each image.

Multi-label stratified sampling is used to split the data into train, validation and test sets, ensuring that IVDs from a single patient are contained in a single split. Table 1 shows the distribution of data splits by condition and severity.

4 Detection and Implementation Details

4.1 IVD Extraction and Labelling

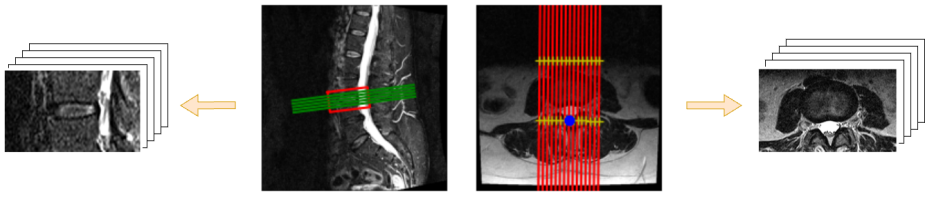

Fig. 3. Example sagittal and axial IVDs: SpineNetV2 is used to detect sagittal IVDs. Red box on the sagittal scan shows the detected region of interest (ROI) for the L3-L4 level, and green lines depict axial scans that intersect this region. The axial scan shows intersecting sagittal slices as red lines and intersections with the anterior and posterior of the ROI in yellow. The blue dot identifies the midpoint of the posterior intersection, which is used as the reference point to extract the axial IVDs.

SpineNetV2 [21] is used to automatically detect the vertebrae and extract IVDs in T1w and T2w/STIR sagittal scans. Each IVD is of dimension $9 \times 112 \times 224$ (slice $\times$ height $\times$ width).

SpineNetV2 [21] does not have the capability to extract axial IVD volumes. However, since every pixel in a DICOM may be projected to 3D space relative to the patient [1], the T2w/STIR sagittal IVD regions of interest (ROI) from the same patient are used to extract axial volumes. All axial slices that intersect the posterior of the sagittal ROI are included in the extracted axial volumes. The midpoint of the axial ROI is defined as the point on the IVD posterior that lies halfway between the first and last sagittal slices that intersect the axial scan, which is an approximate sagittal midpoint of the spine. The width of the extracted axial volume is defined as twice the lateral depth of the sagittal slices. The height of the extracted volume is chosen such that the aspect ratio of the extracted patch is 2:1 to get IVD volumes of the same dimension as sagittal IVDs ($9 \times 112 \times 224$). Figure 3 shows example sagittal and axial sequences with the sagittal ROI and intersections between the two sequences visualised.

4.2 Training Details

All models were trained for 200 epochs using the AdamW optimiser with a learning rate scheduler which linearly warms up from 1e-6 to 1e-5 over 10 iterations, then switches to a cosine annealing schedule with periodic restarts every 30 iterations. Augmentations were applied in training to improve robustness. Each volume was randomly shifted by up to 2 slices, flipped with 50% probability (mirrored left-to-right within each slice for axial, flipping slice order for sagittal) with corresponding left and right labels swapped, and perturbed by an intensity offset sampled from [-0.1, 0.1]. Random translations of up to 22 pixels horizontally and 12 pixels vertically were applied, along with scaling in the range [0.9, 1.1]. Each image was rotated in-plane by a random angle between -20 and 20 degrees. The final model was selected based on the highest macro-averaged AUROC achieved over 10 consecutive validation epochs. The same random seed is set across Python, NumPy, and PyTorch with CUDA to ensure reproducibility.

5 Experimental Results for Radiological Grading

Model performance on radiological grading was evaluated using the Area Under the Receiver Operating Characteristic (AUROC) curve. As each task is multi-class, One-vs-Rest (OvR) AUROC was used for each class, and the macro average AUROC across classes was computed as a measure of overall performance.

The performance of our multimodal, multi-task method was compared against unimodal, single-task models trained only on the sequence that was used for the annotation of the condition (e.g. T1w sagittal for left neural foraminal narrowing). The unimodal models use attention-aggregated slice embeddings as a direct input to the classification layers for grading. We also compared against multimodal, multi-task results with modality drop out (T2w/STIR sagittal & T2w axial, sagittal sequences only, T1w sagittal & T2w axial).

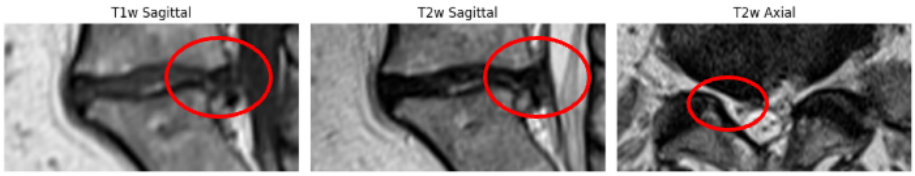

Fig. 4. Classification example: Slices from T1w sagittal, T2w sagittal and T2w axial sequences of an L5-S1 IVD, successfully used by the model to provide a moderate right subarticular stenosis grading.

In addition, we compare against the best condition-specific AUROCs reported in [10], which used a single sequence for single task predictions without a labelling or extraction step. While [10] uses the same data from RSNA, the data splits (training, validation and test) are likely to be different.

Table 2. IVD-level grading performance on RSNA test data (n=194). SCS stands for spinal canal stenosis, NFN for neural foraminal narrowing and SS for subarticular stenosis. Row 1 reports the best-performing results from [10]. Rows 2 to 8 are our own models trained on our data splits. Rows 2 to 4 use attention-aggregated single-sequence embeddings without transformer layers to predict a single task (SCS for T2w/STIR sagittal, NFN for T1w sagittal, SS for T2w axial). The right and left tasks are trained separately for single-task models. Rows 5 to 7 use two of three modalities using the model architecture described in Section 2. Row 8 jointly learns across all sequences.

		Macro AUC				
	Method/Data	**SCS**	**NFN (L)**	**NFN (R)**	**SS (L)**	**SS (R)**
(1)	Limicia et al. [10]	0.747	0.752	0.763	0.843	0.845
(2)	T1w Sagittal	-	0.853	0.850	-	-
(3)	T2w/STIR Sagittal	0.958	-	-	-	-
(4)	T2w Axial	-	-	-	0.881	0.885
(5)	T1w Sag + T2w Ax	0.930	0.880	0.868	0.895	0.894
(6)	T2w/STIR Sag + T2w Ax	0.959	0.855	0.859	0.886	0.895
(7)	T1w + T2w/STIR Sag	0.958	0.877	**0.876**	0.894	0.897
(8)	All	**0.961**	**0.886**	0.875	**0.904**	**0.909**

5.1 Model performance

Table 2 shows the results by condition. Our final model (row 8) achieves the strongest performance across all unimodal, single-task models (rows 1-5) and combination of two modalities (rows 5-7) except right neural foraminal narrowing, with macro average AUROC of 0.961 for spinal canal stenosis, 0.886 for left neural foraminal narrowing, 0.875 for right neural foraminal narrowing, 0.904 for left subarticular stenosis and 0.909 for right subarticular stenosis. The dual

modal models (rows 5-7) also outperform the single modal models (rows 2-4) for all conditions except spinal canal stenosis. Figure 4 shows slices from T1w sagittal, T2w sagittal and T2w axial sequences of an L5-S1 IVD with moderate right subarticular stenosis, correctly graded by our model.

Table 3 gives class-specific OvR AUROC for our main results (row 8 in Table 2). Figure 5 shows the OvR ROC curves of the severe class for each condition. Across all conditions, the moderate class is the most difficult task, while the severe class is the best performing class.

Table 3. Class-specific performance on RSNA test data (n=194). Our main results from row 8 of Table 2 with One-vs-Rest (OvR) AUROC reported for each class. 0 indicates Normal/mild, 1 indicates Moderate and 2 indicates Severe gradings.

Condition	Avg	0	1	2
SCS	0.961	0.977	0.941	0.964
NFN (L)	0.886	0.890	0.832	0.937
NFN (R)	0.875	0.881	0.805	0.940
SS (L)	0.904	0.928	0.839	0.945
SS (R)	0.909	0.940	0.847	0.941

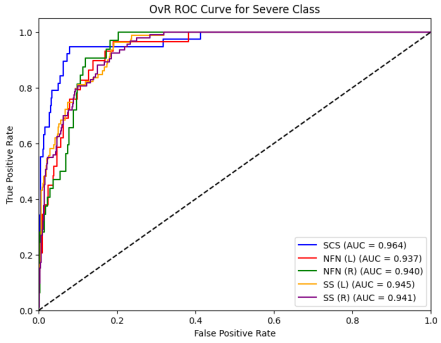

Fig. 5. One-vs-Rest (OvR) AUROC for the severe class for each condition.

5.2 Discussion of results

Our experiments show that training related tasks together using all available multimodal and multi-view inputs with multi-task objectives result in better classification performance than training unimodal, single-task models.

As an external baseline, we compare against the best results reported by [10], who also use the same dataset and report macro average AUROC for each condition. Our best method improves upon the baseline by 0.214 for spinal canal stenosis, 0.134 for left neural foraminal narrowing, 0.112 right neural foraminal narrowing, 0.061 for left subarticular stenosis and 0.064 for right subarticular stenosis. Neural foraminal narrowing is the most challenging task for our model overall. However, subarticular stenosis shows the least improvement from joint training when comparing against the baseline results. This may be due to fact that subarticular stenosis lies in the slim region between the central canal and neural foramen, which is more difficult to see in any other view sequence apart from axial scans. Across all five conditions, the moderate class is the most difficult for the model to distinguish. This is expected, as the extreme classes (normal/mild and severe) are most likely to be more clearly defined and have better inter-rater labelling consistency.

While it would have been ideal to compare against top performing models in the 2024 RSNA Lumbar Spine Degenerative Classification Challenge, the true test set used to evaluate entries for the competition is hidden; a subset of the training data provided is used to generate our evaluation set. As a result, the entries would have used the samples in our test split to train their models. Furthermore, the competition evaluation metric was a single weighted log-loss across all five conditions, rather than condition-specific metrics. We report AUROC separately for each task (OvR) and condition (macro average) for our main results to make them more interpretable and clinically relevant.

Note, we used AUROC rather than threshold-dependent metrics such as balanced accuracy or F1 due to poor calibration across multi-class probabilities. It has been demonstrated that the average confidence of neural networks is considerably higher than their accuracy [4]. This is exacerbated in cases where the data are highly unbalanced, as is often the case in medical domains, even when using class weights to adjust for imbalance. As a result, macro average and class-specific AUROC were used to evaluate performance.

6 Conclusion and extensions

This paper proposes a single model that can learn a joint embedding across all available sequences and view types to predict radiological gradings. Our joint model achieves better performance than single-task models trained with a single relevant sequence across all five spinal stenosis conditions.

One limitation of our work is that labelling and extracting axial volumes require a pre-defined ROI from a sagittal scan, which means that the method cannot be used in datasets without paired sagittal scans. In future work, our aim will be to perform detection directly within the axial scans and define an ROI without the need for a sagittal scan.

Beyond demonstrating superior performance, this work demonstrates a way to learn a joint, single representation for a given disc across modalities, preserving information such as slice order, view and IVD-level alongside the visual features. Having one model allows for a simpler deployment process with only one model to update and optimise, reducing maintenance overhead and runtime cost. Although we use stenosis gradings as an example application, the model can be adapted for other conditions or to jointly assess other imaging modalities, such as X-Ray or PET.

Acknowledgments. We are grateful to RSNA 2024 Lumbar Spine Degenerative Classification Kaggle challenge for providing data access for academic research. We are also grateful to our funders: EPSRC CDT in Health Data Science (EP/S02428X/1), Cancer Research UK via the EPSRC CDT in Autonomous Intelligent Machines and Systems (EP/S024050/1), EPSRC programme grant Visual AI (EP/T025872/1), and the Oxford Big Data Institute.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Atkinson, D.: Geometry in medical imaging: DICOM and NIfTI formats (2022)

[2] Battié, M.C., Jones, C.A., Schopflocher, D.P., Hu, R.W.: Health-related quality of life and comorbidities associated with lumbar spinal stenosis. The Spine Journal **12**(3), 189–195 (2012)

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[4] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)

[5] Guo, X., Malykhina, G.: TransCNN: Fusion of transformer and CNN for detection of lumbar degenerative spine lesions. In: 2025 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM). pp. 783–788. IEEE (2025)

[6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[7] Huynh, H.N., Ngoc, A.D.N., Nguyen, T.A.D., Phan, M.H., Nguyen, V.C., Tran, T.N.: Classification of lumbar spine degeneration using vision transformer with the RSNA dataset. In: Journal of Physics: Conference Series. vol. 2949, p. 012016. IOP Publishing (2025)

[8] Katz, J.N., Zimmerman, Z.E., Mass, H., Makhni, M.C.: Diagnosis and management of lumbar spinal stenosis: a review. Jama **327**(17), 1688–1699 (2022)

[9] Li, Y., Chen, J., Su, Z., Hai, J., Qiao, K., Qin, R., Lu, H., Yan, B.: Transformer-based diagnosis of nerve root compromise in MR imaging of lumbar spine. In: International Conference on Biomedical and Intelligent Systems (IC-BIS 2022). vol. 12458, pp. 534–540. SPIE (2022)

[10] Limicia, J.A., Chandra, W., Gunawan, A.A.S., Tedjasulaksana, J.J.: Comparative model evaluation of lightweight transformer-based and CNN-based architechture for degenerative lumbar spine classification. In: 2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS). pp. 1–6. IEEE (2025)

[11] Lin, S.I., Lin, R.M., Huang, L.W.: Disability in patients with degenerative lumbar spinal stenosis. Archives of physical medicine and rehabilitation **87**(9), 1250–1256 (2006)

[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: Machine Learning for Healthcare Conference. pp. 403–419. PMLR (2018)
[13] Lurie, J., Tomkins-Lane, C.: Management of lumbar spinal stenosis. BMJ **352** (2016)
[14] Park, J., Yang, J., Park, S., Kim, J.: Deep learning-based approaches for classifying foraminal stenosis using cervical spine radiographs. Electronics **12**(1), 195 (2022)
[15] Payne, D.L., Xu, X., Faraji, F., John, K., Pradas, K.F., Bernard, V.V., Bangiyev, L., Prasanna, P.: Automated detection of cervical spinal stenosis and cord compression via vision transformer and rules-based classification. American Journal of Neuroradiology **45**(4), 432–438 (2024)
[16] Qian, J., Su, G., Shu, X., Shen, K., Chen, B., Wang, X.: Lumbar disc herniation diagnosis using deep learning on MRI. Journal of Radiation Research and Applied Sciences **17**(3), 100988 (2024)
[17] Richards, T., Talbott, J., Ball, R., Colak, E., Flanders, A., Kitamura, F., Mongan, J., Prevedello, L., Vazirabad., M.: RSNA 2024 lumbar spine degenerative classification. https://kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification (2024), kaggle
[18] Stojšić, K., Miletić Rigo, D., Jurković, S.: Automated vertebral bone quality determination from T1-weighted lumbar spine mri data using a hybrid convolutional neural network–transformer neural network. Applied Sciences **14**(22), 10343 (2024)
[19] Waldrop, R., Cheng, J., Devin, C., McGirt, M., Fehlings, M., Berven, S.: The burden of spinal disorders in the elderly. Neurosurgery **77** (2015)
[20] Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: Context-aware transformers for spinal cancer detection and radiological grading. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 271–281. Springer (2022)
[21] Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: Automated detection, labelling and radiological grading of clinical spinal MRIs. Scientific Reports **14**(1), 14993 (2024)
[22] Yilihamu, E.E.Y., Zeng, F.S., Shang, J., Yang, J.T., Zhong, H., Feng, S.Q.: GPT4LFS (generative pre-trained transformer 4 omni for lumbar foramina stenosis): enhancing lumbar foraminal stenosis image classification through large multimodal models. The Spine Journal (2025)
[23] Zemedikun, D.T., Kigozi, J., Wynne-Jones, G., et al.: Healthcare resource utilisation and economic burden attributable to back pain in primary care: A matched case-control study in the united kingdom. British Journal of Pain **18**(2), 137–147 (2024). https://doi.org/10.1177/20494637231208364
[24] Zhao, M., Meng, N., Cheung, J.P.Y., Yu, C., Lu, P., Zhang, T.: SpineHRformer: a transformer-based deep learning model for automatic spine deformity assessment with prospective validation. Bioengineering **10**(11), 1333 (2023)
[25] Zileli, M., Crostelli, M., Grimaldi, M., Mazza, O., Anania, C., Fornari, M., Costa, F.: Natural course and diagnosis of lumbar spinal stenosis: WFNS spine committee recommendations. World Neurosurg X **7**, 100073 (February 2020)