Spinal Canal Stenosis CAD

Shijia Zhang[1] and Ankush Jindal[2]

[1,2]Department of Biomedical Informatics & Data Science, Johns Hopkins University, Baltimore, MD

1. Introduction

According to the World Health Organization, low back pain is the leading cause of disability worldwide, affecting 619 million people in 2020, a 60% increase over 1990. Among the various pathologies that contribute to low back pain, spondylosis encompasses a spectrum of degenerative spine conditions that significantly affect patient mobility and quality of life. These conditions include intervertebral disc degeneration, neural foraminal narrowing, and various forms of spinal stenosis, all of which can result in nerve compression and associated symptoms.

Magnetic resonance imaging (MRI) has emerged as the gold standard for evaluating lumbar spine pathology, offering superior soft tissue contrast and multiplanar imaging capabilities. MRI enables detailed visualization of Neural structures and their relationship to surrounding tissues, Disc morphology and Presence and extent of stenosis.

In this project, we explore the role of MRI in detecting spinal canal stenosis by leveraging RSNA Lumbar Spine Degenerative Classification AI Challenge 2024. Accurate diagnosis and classification through MRI interpretation directly influences treatment strategy, surgical planning and prognosis assessment.

2. Dataset

2.1 Data Source

The study utilized the RSNA Lumbar Spine Degenerative Classification AI Challenge 2024 dataset (www.kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification), comprising imaging data collected from eight clinical sites across five continents. The dataset consists of 9,753 Axial T2-weighted magnetic resonance images acquired in DICOM format from 1,974 unique patients. These images capture the lumbar spine region, specifically focusing on five intervertebral disc levels from L1/l2 to L5/S1. The severity for

spinal canal stenosis on each disc level was graded on a three-point scale (Normal/Mild, Moderate and Severe).

2.2 Data Organization and Splitting

2.2.1 Cohort Division

The dataset was partitioned into three distinct cohorts while maintaining patient-level separation to prevent data leakage and ensuring each split had a natural distribution of severity grades (fig 2):

Table 1: Dataset Distribution Across Training, Validation, and Test Sets

Subset	Number of Patients	Approximate Ratio (%)
Training	1,284	65
Validation	296	15
Test	394	20
Total	**1,974**	**100**

Figure 1: L1/L2 vertebrae showing Spinal Canal Stenosis

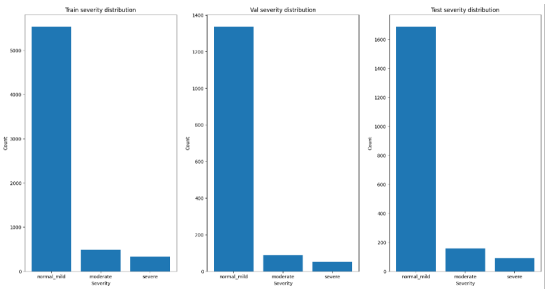


Figure 2: Distribution of data across Train, Val and Test splits

3. Methods

Our methodology leverages a hybrid architecture that effectively combines pre-trained vision models with level-specific embeddings to achieve accurate stenosis classification across different vertebral levels. We implemented six distinct backbone networks: Vision Transformer (ViT-B/16) [1], Swin Transformer (Swin-B) [2], BEiT [3], EfficientNet V2-M [4], ResNet-152 [5], and ConvNeXt Base [6]. Each of these models was pre-trained on

ImageNet-1K. To adapt these models for our specific task, we modified their architectures by removing the original classification heads while preserving their feature extraction capabilities, replacing the final classification layers with identity mappings to obtain feature vectors ranging from 768 to 2048 dimensions, depending on the specific architecture.

A key innovation in our approach is the integration of vertebral level information through a dedicated embedding module. This component maps each spinal level (L1/L2 through L5/S1) to a 256-dimensional embedding vector, which is then concatenated with the backbone's feature vector to provide crucial level-specific context. The combined features are processed through a classification head implemented as a multi-layer perceptron, comprising three main stages with decreasing dimensionality ($512 \rightarrow 256 \rightarrow 3$), each followed by layer normalization, GELU activation, and dropout layers for regularization.

We employed the AdamW optimizer with a learning rate of 1e-4 and weight decay of 0.02, processing the data in batches of 32 images over 30 epochs. Our transfer learning strategy was implemented in three phases: initial freezing of the backbone weights to preserve the pre-trained knowledge, selective unfreezing of the last 32 layers to allow fine-tuning, and full training of both the level embedding module and classification head. To address class imbalance in the severity grades, we utilized Cross-Entropy Loss with appropriate class weights.

Implementation was carried out using the PyTorch framework, with backbone models initialized using pre-trained weights from either torchvision.models or the Hugging Face model repository. Data preprocessing followed standard practices, including resizing images to 224×224 pixels and normalizing using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). We also incorporated data augmentation techniques, including random horizontal flips and rotations, to enhance model generalization.

Model evaluation was conducted using a comprehensive set of metrics including accuracy, precision, recall, and F1-score, with confusion matrices generated for each severity grade. All evaluations were performed on held-out validation and test sets to ensure unbiased assessment of model performance.

4. Results

Our comprehensive evaluation of six deep learning architectures for spinal canal stenosis classification revealed notable differences in performance across various metrics. The BEiT [3] and Vision Transformer (ViT) [1] architectures demonstrated superior overall performance (accuracy 84.7%), while EfficientNet [4] showed relatively lower performance across most metrics (accuracy 76.8%).

The macro-precision metrics revealed that transformer-based models (BEiT and ViT) consistently outperformed traditional convolutional architectures. [Appendix Table 2] BEiT achieved the highest macro-precision of 0.581, followed closely by ViT at 0.578.

When considering class distribution through weighted metrics, weighted F1 scores showed consistent performance among the top models (BEiT: 0.865, ViT: 0.864). BEiT and ViT also demonstrated the highest ROC-AUC scores (0.855 and 0.856)[fig 4]. The superior performance of transformer-based architectures (BEiT and ViT) suggests that their self-attention mechanisms are particularly effective in capturing relevant features for the classification of spinal stenosis.

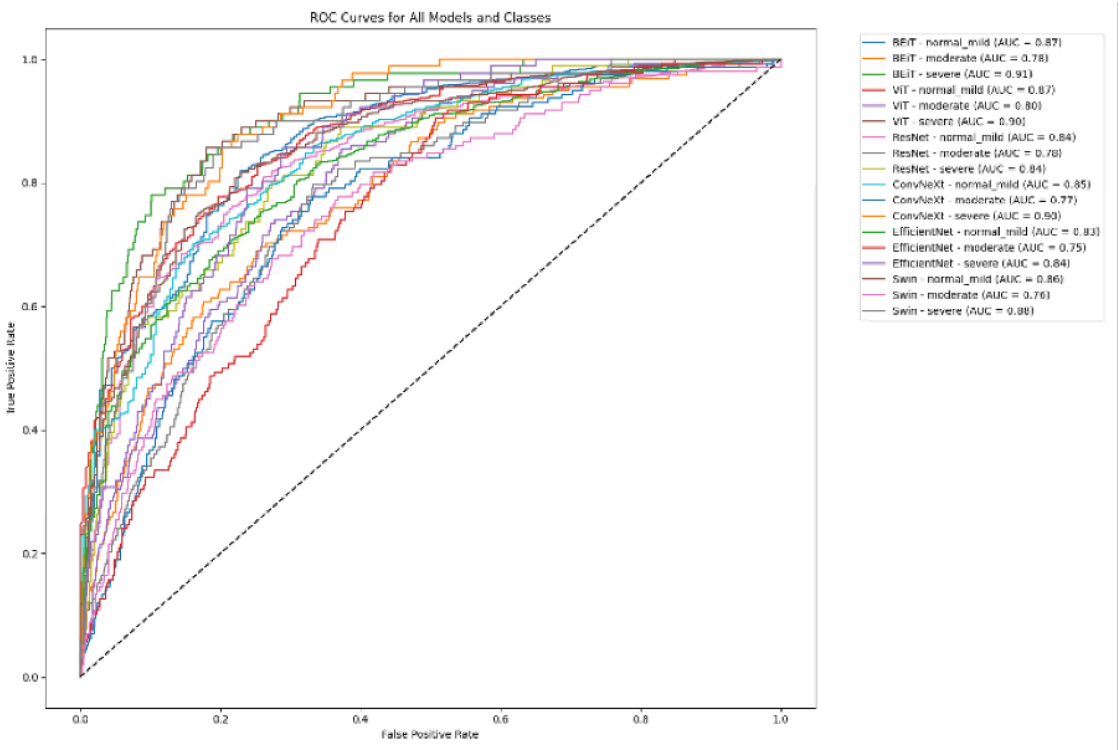


Figure 3: AUC-ROC curves comparing all models for each class

5. Conclusion

The comparative analysis of six different architectures has revealed that transformer-based models consistently outperform traditional convolutional neural networks in this specific medical imaging context. This superiority manifests not only in overall accuracy but also in the models' ability to handle the inherent complexities of spinal stenosis classification across different vertebral levels. The integration of level-specific embeddings with pre-trained vision models has proven to be an effective approach, suggesting that architectural innovations combining domain-specific knowledge with advanced deep learning techniques can yield substantial improvements in medical image analysis.

The high performance achieved by our models, particularly in weighted metrics that account for class imbalance, suggests potential clinical utility. Our approach could serve as a valuable tool for radiologists and clinicians, potentially:

1. Accelerating the diagnostic process

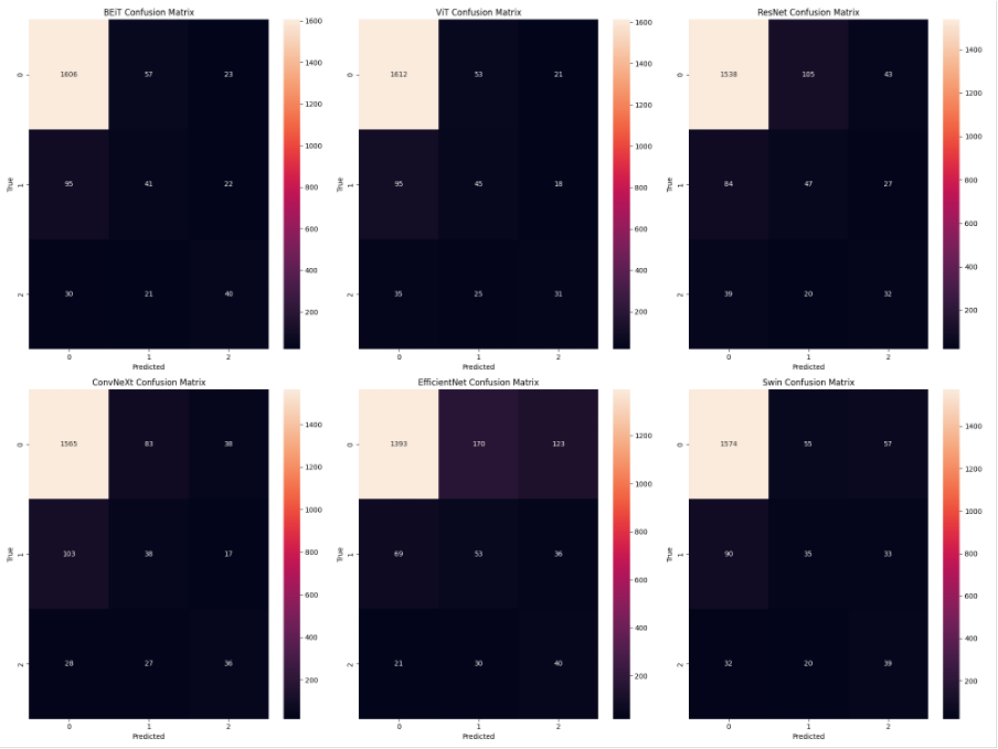


Figure 4: Confusion Matrix for all models

2. Providing consistent grading of stenosis severity
3. Supporting clinical decision-making
4. Reducing the workload in high-volume clinical settings

6. Limitations

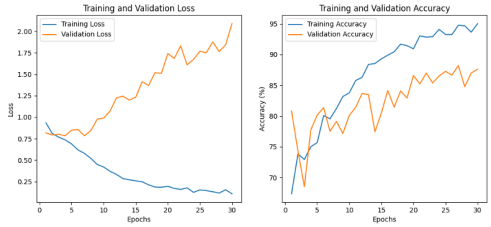
Despite the promising results, there are several limitations such as low macro-precision and recall scores for minority classes. Also, the current approach relies only on 2D axial images, potentially missing important spatial information from sagittal and coronal views. In future work, we aim to address these limitations by using synthetic data generation techniques to oversample minority classes and incorporating multi-view approaches.

7. Appendix

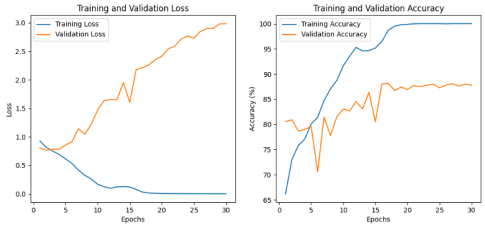
7.1 Training and Validation Curves

This appendix presents the training and validation loss and accuracy curves for all six models evaluated in this study.

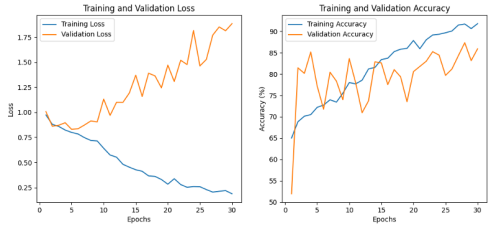
7.2 Detailed Performance Metrics



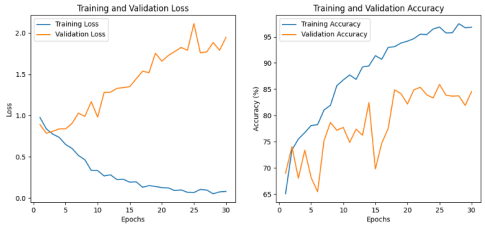
(a) BEiT



(b) ViT

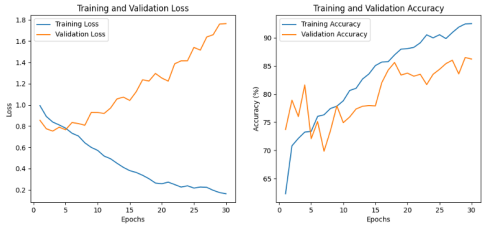


(c) Swin



(d) ResNet

(e) EfficientNet



(f) ConvNeXt

Figure 5: Training and validation curves showing loss and accuracy over epochs for all models. Each subplot displays the convergence behavior and learning dynamics of the respective architecture during the training process.

Metric	BEiT	ViT	ResNet	ConvNeXt	EfficientNet	Swin
Overall Performance						
Accuracy	**0.872**	**0.872**	0.836	0.847	0.768	0.852
Macro F1	**0.564**	0.549	0.512	0.523	0.471	0.516
Weighted F1	**0.865**	0.864	0.840	0.845	0.800	0.849
Macro ROC-AUC	0.855	**0.856**	0.818	0.840	0.808	0.836
Macro PR-AUC	**0.550**	0.541	0.490	0.524	0.473	0.507
Normal/Mild Class						
Precision	0.928	0.925	0.926	0.923	**0.939**	0.928
Recall	0.953	**0.956**	0.912	0.928	0.826	0.934
F1-Score	**0.940**	**0.940**	0.919	0.925	0.879	0.931
ROC-AUC	0.868	**0.871**	0.842	0.847	0.830	0.863
PR-AUC	0.975	**0.977**	0.970	0.971	0.969	0.976
Moderate Class						
Precision	0.345	**0.366**	0.273	0.257	0.209	0.318
Recall	0.259	0.285	0.297	0.241	**0.335**	0.222
F1-Score	0.296	**0.320**	0.285	0.248	0.258	0.261
ROC-AUC	0.785	**0.801**	0.775	0.770	0.749	0.761
PR-AUC	**0.300**	0.264	0.263	0.220	0.206	0.234
Severe Class						
Precision	**0.471**	0.443	0.314	0.396	0.201	0.302
Recall	**0.440**	0.341	0.352	0.396	**0.440**	0.429
F1-Score	**0.455**	0.385	0.332	0.396	0.276	0.355
ROC-AUC	**0.913**	0.895	0.838	0.904	0.845	0.883
PR-AUC	0.374	**0.383**	0.237	0.380	0.243	0.313

Table 2: Comprehensive performance metrics for all models across different evaluation criteria. Best performance for each metric is shown in bold.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, 'An image is worth 16x16 words: Transformers for image recognition at scale,' in *International Conference on Learning Representations*, 2021.

[2] Z. Liu, Y. Lin, Y. Cao *et al.*, 'Swin transformer: Hierarchical vision transformer using shifted windows,' in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[3] H. Bao, L. Dong, S. Piao & F. Wei, 'Beit: Bert pre-training of image transformers,' in *International Conference on Learning Representations*, 2022.

[4] M. Tan & Q. V. Le, 'Efficientnet: Rethinking model scaling for convolutional neural networks,' in *International Conference on Machine Learning*, 2019, pp. 6105–6114.

[5] K. He, X. Zhang, S. Ren & J. Sun, 'Deep residual learning for image recognition,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[6] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell & S. Xie, 'A convnet for the 2020s,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.