A Multi-Stage, Multi-Modal Deep Learning Framework for Automated Lumbar Spine Degenerative Classification from MRI

Siddharth Panditrao
Dept. of Data Science & Eng.
Manipal University Jaipur
Jaipur, India
panditrao.sid@gmail.com

Devesh Nahar
Dept. of Data Science & Eng.
Manipal University Jaipur
Jaipur, India
devnahar23@gmail.com

Aashi Sharma
Dept. of Data Science & Eng.
Manipal University Jaipur
Jaipur, India
sharma.aashi2004@gmail.com

Contents

List of Figures

List of Tables

Abstract

This paper introduces a multi-stage deep learning framework designed for the automated classification of degenerative conditions affecting the lumbar spine, utilizing Magnetic Resonance Imaging (MRI). Recognizing the inherent complexity in assessing spinal pathologies, the proposed framework systematically decomposes the problem into distinct, sequential stages: initial coordinate generation for identifying relevant spinal levels and subsequent severity classification based on these locations. The coordinate generation process itself is further delineated into instance number prediction, which involves determining the most pertinent axial slice index using 3D Convolutional Neural Networks (CNNs) like ConvNeXt, and precise coordinate prediction, which refines the in-plane (x, y) locations using 2D CNNs such as ConvNeXt-base and EfficientNet-V2-L. Following successful localization, the second stage employs a Multiple Instance Learning (MIL) strategy integrated with a Bi-directional Long Short-Term Memory (Bi-LSTM) aggregator. This module processes feature representations extracted by 2D CNN encoders (ConvNeXt-small, EfficientNet-V2-s) from image patches meticulously cropped around the previously identified coordinates. A key aspect of this stage is the incorporation of specialized preprocessing steps, including condition-specific cropping algorithms and robust data augmentation techniques. Notably, these augmentations involve the deliberate introduction of random shifts to predicted coordinates and instance numbers, explicitly modeling and mitigating potential errors propagating from the initial localization stage. The framework incorporates several architectural innovations, including level-separated prediction heads for targeted outputs, dual regression and classification objectives for enhanced instance number prediction, the use of specialized pretraining datasets for coordinate models, and an attention-based Bi-LSTM MIL architecture employing an auxiliary loss function to guide the severity grading process. This comprehensive methodology aims to deliver accurate and automated assessments for common conditions like Spinal Canal Stenosis (SCS), Neuroforaminal Narrowing (NFN), and Spondylolisthesis (SS).

1 Introduction

Degenerative diseases impacting the lumbar spine, such as Spinal Canal Stenosis (SCS), Neuroforaminal Narrowing (NFN), and Spondylolisthesis (SS), represent a significant source of chronic pain and functional impairment globally. Magnetic Resonance Imaging (MRI) serves as the cornerstone diagnostic modality, providing unparalleled visualization of soft tissues, intervertebral discs, neural elements, and osseous structures within the spine. Despite its utility, the manual interpretation of lumbar spine MRI studies remains a challenging task. It demands considerable radiological expertise, is inherently time-consuming, and is prone to inter-observer variability, especially when quantifying the severity of degenerative findings which often exist on a continuous spectrum.

Recent advancements in deep learning have opened promising avenues for automating medical image analysis, potentially enhancing efficiency, improving diagnostic consistency, and aiding clinical decision-making. However, the application of these techniques to lumbar spine MRI is non-trivial. Challenges include the need to accurately identify specific vertebral levels within the complex spinal anatomy, localize subtle pathological changes, effectively process both 3D volumetric information and diagnostically critical 2D slice data, handle the common issue of class imbalance in severity grading (where milder forms are often more prevalent than severe ones), and ensure model robustness across diverse imaging protocols and patient populations.

To systematically address these complexities, this paper details a sophisticated multi-stage deep learning framework. Eschewing a monolithic end-to-end model, our approach segments the diagnostic workflow into logically sequential sub-tasks. The initial phase focuses on localization, explicitly divided into (1a) predicting the relevant axial slice index, often termed 'instance number', for each

lumbar intervertebral level, and (1b) subsequently predicting the precise in-plane (x, y) coordinates of relevant features within selected slices. Only after successful localization does the framework proceed to (2) classifying the severity of degenerative conditions. This classification stage analyzes image patches extracted based on the predicted coordinates. This decomposition facilitates the development of highly specialized models optimized for each sub-task, enabling the use of appropriate data modalities (e.g., 3D volumes for initial slice identification, 2D slices for detailed coordinate and severity analysis) and tailored network architectures (e.g., 3D CNNs, 2D CNNs, Recurrent Neural Networks, Multiple Instance Learning).

2 Methodology

The core of our proposed system is a multi-stage pipeline, conceptually depicted in Figure 1. This design comprises three fundamental types of predictive models working in concert: instance number prediction models, coordinate prediction models, and severity prediction models. Notably, Stage 1, responsible for localization, is explicitly bifurcated into instance number prediction followed by coordinate prediction.

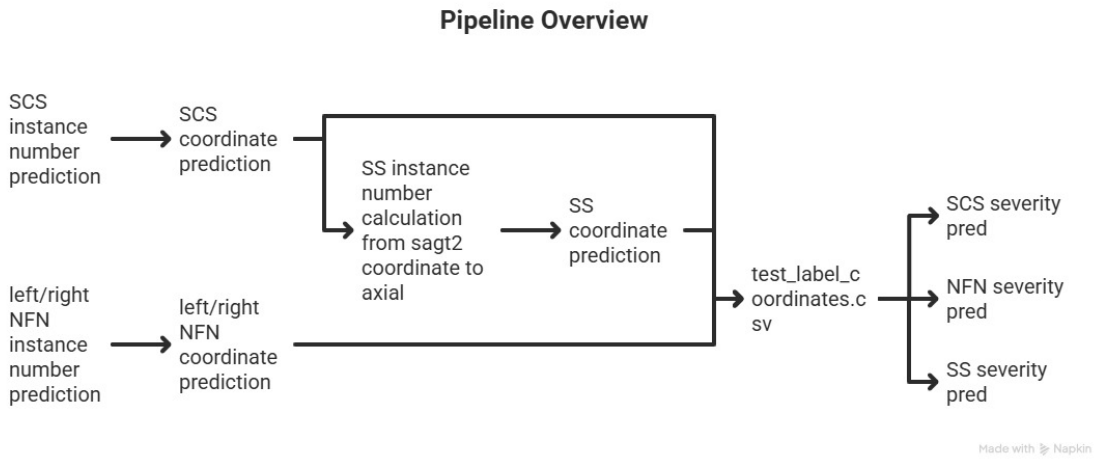

Figure 1: Overview of the Multi-Stage Processing Pipeline.

2.1 Stage 1: Coordinate Generation

The primary objective of this initial stage is to generate accurate spatial coordinates for each relevant spinal level, producing localization data (`test_label_coordinates.csv`) essential for the subsequent severity analysis.

2.1.1 Stage 1a: Sagittal Instance Number Prediction

The first step within localization focuses on identifying the most diagnostically relevant axial slice index (z-coordinate, referred to as 'instance number') for each intervertebral disc level (from L1/L2 down to L5/S1) by analyzing sagittal MRI volumes.

Input Data Preparation: Sagittal MRI volumes (e.g., T2-weighted sequences often preferred for SCS visualization) serve as input. Preprocessing involves ensuring correct slice order using DICOM

metadata, normalizing pixel intensities to a standard range (e.g., [0, 1]) to ensure consistent signal representation, and applying padding along the depth (slice) dimension to achieve a uniform input shape (e.g., 32 slices), facilitating batch processing. An example of a preprocessed input volume is shown in Figure 2.

Model Architecture: A 3D Convolutional Neural Network, specifically adopting the ConvNeXt architecture adapted for 3D inputs, forms the core of this sub-stage. This model utilizes a shared 3D ConvNeXt encoder to extract volumetric features, followed by distinct, level-separated prediction heads. Each head is dedicated to predicting the instance number for a specific lumbar level (L1/L2, L2/L3, etc.), allowing specialization. The architecture is depicted in Figure 3.

Training Strategy: To enhance prediction robustness, a dual-task learning approach is employed. Two sets of prediction heads are trained concurrently:

1. *Classification Task:* Treats the instance number as a discrete classification problem (predicting one of the 32 possible slice indices). The corresponding heads output logits of shape `(batch_size, 32)` for each level, trained using a standard Cross-Entropy Loss function.

2. *Regression Task:* Predicts the instance number as a continuous variable, specifically the normalized slice index z'. For regularization and potentially improved localization context, this task also predicts auxiliary normalized in-plane coordinates (x', y'). The regression heads output vectors of shape `(batch_size, 3)` per level. Coordinate normalization is applied as $(x', y', z') = (x/\text{width}, y/\text{height}, z/32)$ to stabilize training. An L1 Loss function is used, with the primary focus on minimizing the error for the z' component.

Prediction Ensembling: To leverage the strengths of both approaches and improve final accuracy, predictions from multiple models trained using 5-fold cross-validation (for both classification and regression tasks) are aggregated. The final instance number prediction for each spinal level is determined by calculating the median of the predicted instance numbers from all ensemble members.

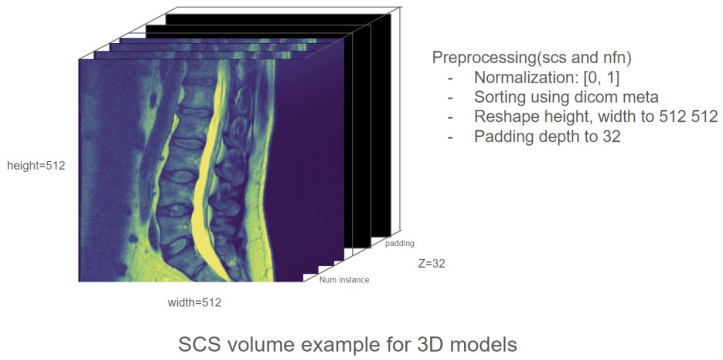

Figure 2: Exemplar Preprocessed 3D Sagittal Volume for Instance Number Prediction.

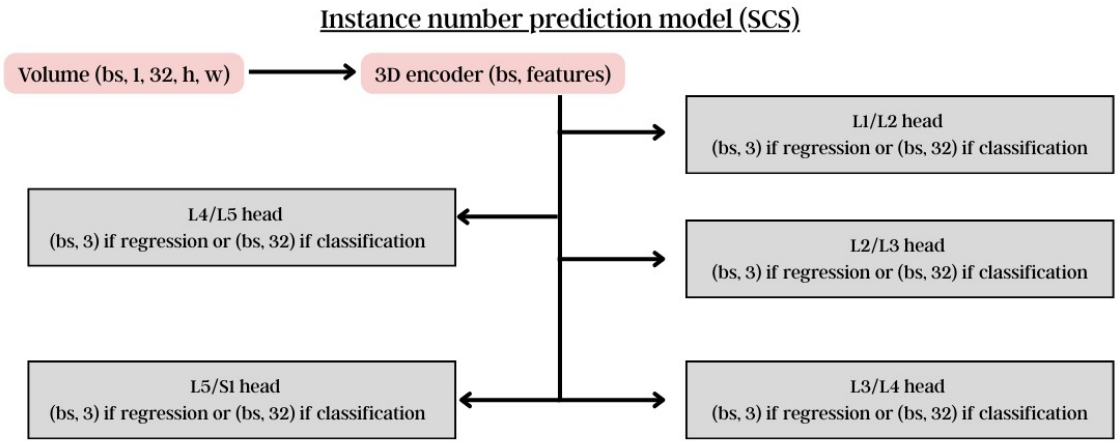

Figure 3: Architecture for 3D Instance Number Prediction using 3D ConvNeXt.

2.1.2 Stage 1b: Sagittal Coordinate Prediction

Once the most representative sagittal slice (based on the median instance number from Stage 1a) is identified for each level, this sub-stage predicts the precise in-plane (x, y) coordinates of the target feature within that slice.

Input Data: The input is the selected 2D sagittal slice for a given level. It is typically converted into a 3-channel image format (e.g., by channel duplication or stacking adjacent slices) to match the input requirements of standard pre-trained models, resized to a fixed resolution (e.g., 512x512 pixels), and normalized.

Model Architecture: This sub-stage utilizes 2D CNN architectures. A robust 2D CNN encoder (exploring options like ConvNeXt-base and EfficientNet-V2-L) extracts features from the input slice. Similar to Stage 1a, level-separated regression heads are appended to the encoder, each dedicated to predicting the (x, y) coordinates for a specific spinal level. The general structure is illustrated in Figure 4.

Training and Pretraining Strategy: The models are trained purely as a regression task to predict normalized coordinates $(x', y') = (x/\text{width}, y/\text{height})$ for each designated level, minimizing an L1 Loss function. Notably, performance was enhanced by employing a two-step pretraining strategy: models were first pre-trained on a large, publicly available dataset containing lumbar spine annotations before being fine-tuned on the primary task dataset. This domain-specific pretraining proved more effective than relying solely on standard ImageNet pre-trained weights.

Prediction Ensembling: To further boost accuracy and stability, predictions from the different 2D encoder models (ConvNeXt-base and EfficientNet-v2-l), each trained using cross-validation, are combined through simple averaging (mean) to yield the final sagittal (x, y) coordinate predictions.

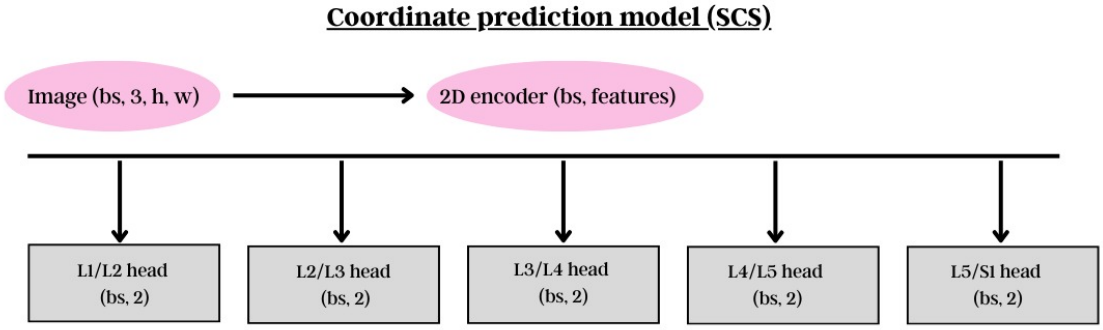

Figure 4: Architecture for 2D Sagittal Coordinate Prediction.

2.1.3 Stage 1c: Axial Coordinate Generation

For conditions where axial views are critical (e.g., NFN, aspects of SS, subarticular stenosis relevant to SCS), a similar coordinate generation process is applied to axial MRI slices.

Methodology: The instance number prediction (identifying the relevant axial slices) adapts previously established techniques developed within the research community. Following slice identification, the coordinate prediction employs a 2D CNN architecture analogous to that described in Stage 1b, trained specifically to localize relevant features (e.g., foramina, facet joints, canal boundaries) on the axial plane.

2.2 Stage 2: Severity Prediction

With accurate localization coordinates established in Stage 1, the second major stage focuses on classifying the severity grade of the identified degenerative conditions. This stage leverages an attention-based Multiple Instance Learning (MIL) framework combined with sequence modeling using Bi-LSTMs.

2.2.1 Preprocessing and Data Augmentation for Severity Models

Careful preprocessing is essential to prepare the input for the severity classification models.

Coordinate-Guided Cropping: This is the cornerstone of the preprocessing strategy. Instead of processing entire slices, the models focus on smaller image patches extracted around the coordinates predicted in Stage 1.

- *Input Slice Selection:* For analyzing a specific level, a small stack of adjacent slices (e.g., 5 slices centered on the predicted instance number) is often considered as input instances for the MIL model. This provides local 3D context.
- *Cropping Ranges:* Patches are cropped using predefined pixel dimensions relative to the predicted (x, y) coordinate. These dimensions are carefully chosen based on the condition being

assessed and the MRI view, aiming to capture the relevant anatomy while excluding extraneous information. Specific ranges are detailed in Table 1. An example of a resulting cropped patch is shown in Figure 5.

- *Axial Cropping Specifics:* Cropping from axial views often involves additional logic. For instance, assessing features relevant to SCS might involve cropping around coordinates associated with left or right subarticular stenosis (potentially selected randomly during training for augmentation). Similarly, assessing SS might involve specific offsets relative to vertebral body coordinates to capture slippage. Examples visualizing these axial cropping strategies are provided in Figures 6 and 7.

Data Augmentation Strategy: Augmentation is applied strategically both before and after cropping:

1. *Pre-Cropping Augmentation (Simulating Stage 1 Errors):* This is identified as a critical step for building robustness against potential inaccuracies in the Stage 1 predictions. Before cropping, random shifts are deliberately introduced to the predicted (x, y) coordinates (e.g., uniformly sampled within ± 10 pixels) and the predicted instance number (z-coordinate, e.g., uniformly sampled within ± 2 slices). The probability of applying a shift to the instance number is directly informed by the measured error characteristics of the Stage 1a models (see Table 2), making the simulation more realistic.
2. *Post-Cropping Augmentation:* After the patches are cropped, standard image-level augmentations are applied to increase variability and prevent overfitting. These include techniques like `RandomBrightnessContrast` (adjusting brightness and contrast, p=0.25 probability) and `ShiftScaleRotate` (applying random shifts, scaling, and rotations within defined limits, p=0.5 probability).

For SCS				
Type	Left	Right	Upper	Lower
SagT2	96	32	40	40
Axial	96	96	96	96
For NFN				
Type	Left	Right	Upper	Lower
SagT1 (both L & R)	96	64	32	32
Axial (Right)	144	48	96	96
Axial (Left)	48	144	96	96
For SS				
Type	Left	Right	Upper	Lower
Axial (Right)	144	48	96	96
Axial (Left)	48	144	96	96

Table 1: Cropping Ranges (in pixels) Relative to Predicted (x, y) Coordinate for Different Conditions and Views.

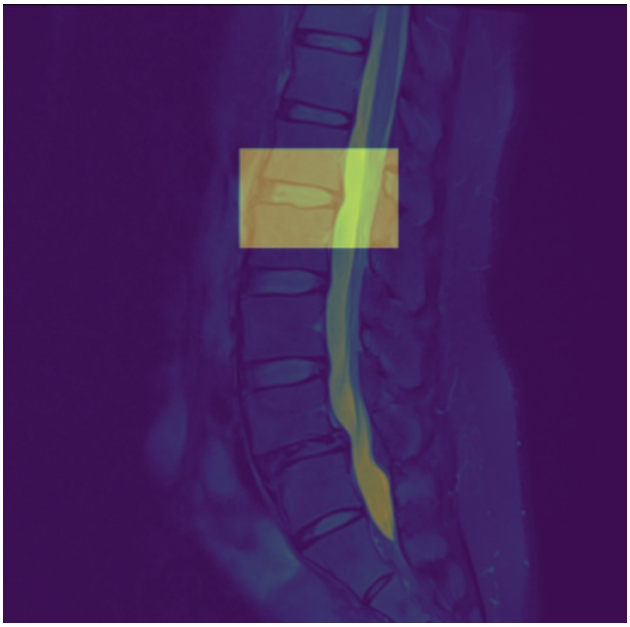

Figure 5: Example Cropped Sagittal T2 Patch for SCS Severity Prediction at L1/L2.

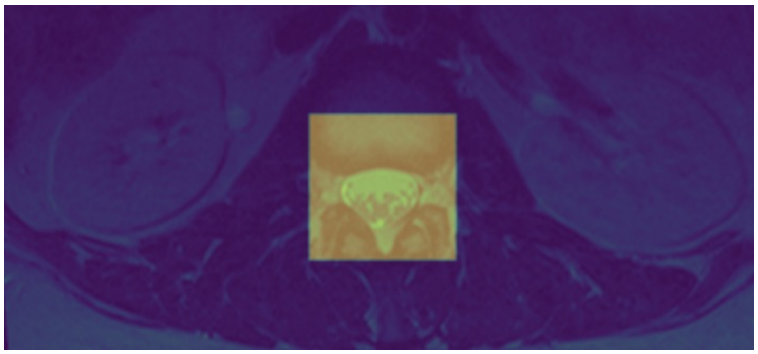

Figure 6: Visualization of Axial Cropping Strategy for SCS using adjusted SS coordinates.

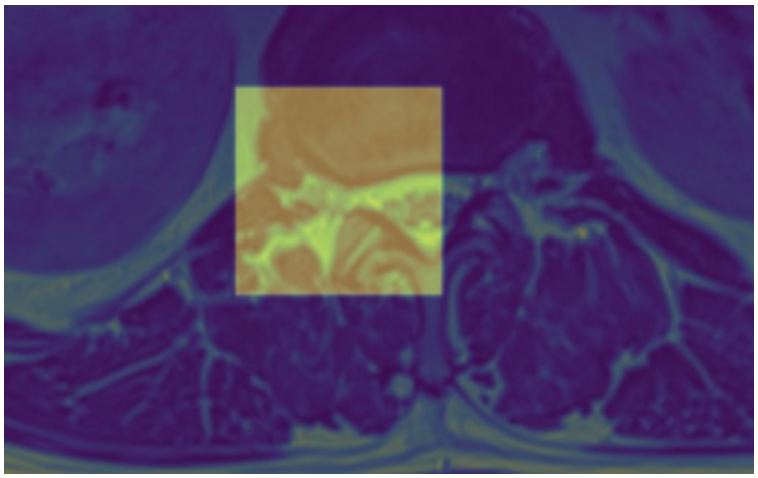

Figure 7: Visualization of Axial Cropping Strategy for SS Severity focusing on right subarticular region.

2.2.2 Model Architecture for Severity Prediction

The severity prediction stage utilizes an advanced architecture combining CNN feature extraction, sequential modeling with Bi-LSTM, and attention-based MIL.

Overall Structure: A shared 2D CNN encoder first processes each input instance (cropped patch) independently to extract a high-level feature vector. The sequence of feature vectors corresponding to the instances (e.g., adjacent slices) for a given spinal level is then fed into the Bi-LSTM MIL module for aggregation and final classification.

Encoder Backbone: Relatively lightweight yet powerful CNN architectures, specifically ConvNeXt-small and EfficientNet-V2-s, were chosen as the feature extractors after experiments indicated diminishing returns or potential overfitting with larger models in this setup.

Bi-LSTM MIL Module: This core component (detailed in Listing 1) is designed to aggregate information effectively from the sequence of instance features:

- It first employs a multi-layer Bi-directional LSTM to process the sequence of feature vectors, capturing contextual dependencies between adjacent slices or views.
- Two distinct attention mechanisms operate on the LSTM's output sequence:
  - The primary attention mechanism calculates weights to compute a weighted sum of the LSTM hidden states. This aggregated vector (`weighted_instances`) represents the entire bag of instances and serves as the input to the final classification head.
  - A parallel auxiliary attention mechanism generates separate attention scores (`aux_attn_scores`) for each instance in the sequence. These scores are *directly supervised* via an auxiliary loss function during training, potentially encouraging the model to focus on the most relevant instances (slices) for the severity decision.

Classification Head: A simple linear layer followed by a Softmax activation function maps the final aggregated feature vector (`weighted_instances`) to probability scores for each severity class.

The architectures tailored for SCS and SS severity prediction are visualized in Figures 8 and 9, respectively.

```python
import torch
import torch.nn as nn

class LSTMMIL(nn.Module):
    def __init__(self, input_dim):
        super(LSTMMIL, self).__init__()
        self.lstm = nn.LSTM(input_dim, input_dim // 2, num_layers=2,
                            batch_first=True, dropout=0.1, bidirectional=True)
        self.aux_attention = nn.Sequential( # For auxiliary loss
            nn.Tanh(),
            nn.Linear(input_dim, 1)
        )
        self.attention = nn.Sequential( # For primary weighted average
            nn.Tanh(),
            nn.Linear(input_dim, 1)
        )

    def forward(self, bags):
        # bags shape: (batch_size, num_instances, input_dim)
        bags_lstm, _ = self.lstm(bags)
        # aux_attn_scores shape: (batch_size, num_instances)
        aux_attn_scores = self.aux_attention(bags_lstm).squeeze(-1)
        # attn_scores shape: (batch_size, num_instances)
        attn_scores = self.attention(bags_lstm).squeeze(-1)
        # attn_weights shape: (batch_size, num_instances)
        attn_weights = torch.softmax(attn_scores, dim=-1)
        # weighted_instances shape: (batch_size, input_dim)
        weighted_instances = torch.bmm(attn_weights.unsqueeze(1), bags_lstm).
    squeeze(1)
        return weighted_instances, aux_attn_scores
```

Listing 1: Bi-LSTM MIL Module Implementation.

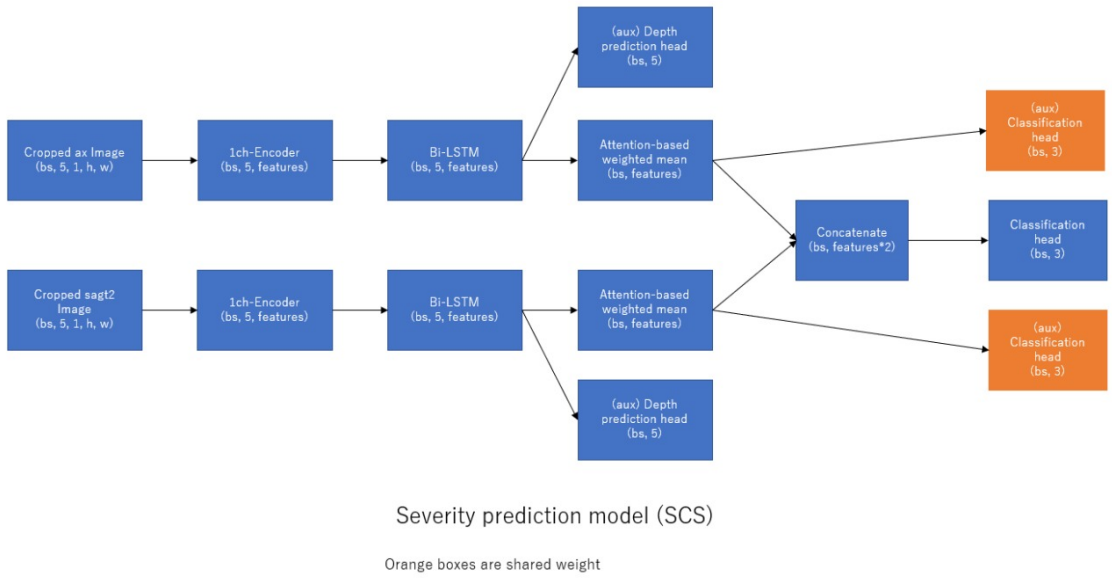

Figure 8: Severity Prediction Model Architecture for SCS.

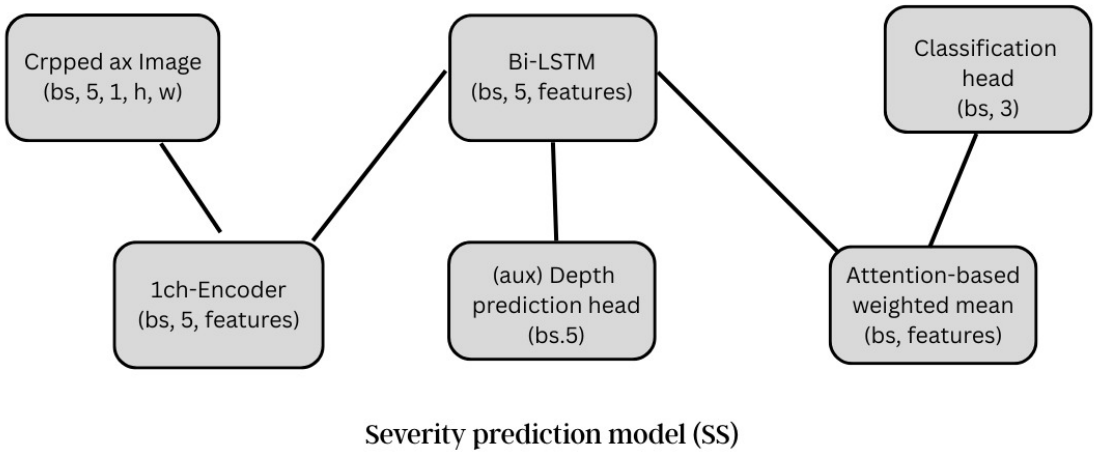

Figure 9: Severity Prediction Model Architecture for SS. Structure similar to SCS model, adapted for SS task.

2.2.3 Training and Ensembling for Severity

The severity models are trained using specific strategies:

Loss Function: A composite loss function is employed, combining the primary Cross-Entropy loss calculated on the final severity predictions with an auxiliary Cross-Entropy loss applied directly to the `aux_attn_scores` generated by the auxiliary attention head. This auxiliary loss encourages the model to assign meaningful attention weights during training.

Training Details: Experimental results suggested that relatively short training durations were optimal, for example, 7 epochs when using the ConvNeXt-small encoder and 14 epochs for the EfficientNet-V2-s encoder, potentially mitigating overfitting on the complex task.

Ensembling: For final inference, predictions generated by models utilizing different encoder backbones (ConvNeXt-small, EfficientNet-V2-s) and those trained across different folds during cross-validation are likely aggregated, commonly through averaging predicted probabilities, to produce a more robust final severity classification.

3 Experimental Setup & Results

3.1 Datasets and Implementation

The development and evaluation of this framework were conducted using the dataset provided for the RSNA 2024 Lumbar Spine Degenerative Classification challenge. The implementation relied heavily on the PyTorch deep learning library. As mentioned, external datasets and methods were leveraged for specific pretraining tasks (Stage 1b coordinate models) and axial instance number calculation (Stage 1c).

3.2 Stage 1a Performance

The accuracy of the sagittal instance number prediction models (Stage 1a), comparing the classification and regression approaches based on their prediction error relative to the ground truth slice

index, is presented in Table 2.

Model Type	Error = ±0	Error = ±1	Error = ±2	Error > ±2
Classification	71.08%	27.04%	1.43%	0.44%
Regression	67.48%	30.59%	1.61%	0.31%

Table 2: Sagittal T2 Instance Number Prediction Accuracy (SCS) - Distribution of Prediction Error relative to ground truth slice index.

Both methods achieve high precision, with the vast majority of predictions falling within one slice of the correct location. The final ensembled prediction using the median further enhances the reliability required for selecting the appropriate slice for Stage 1b coordinate prediction.

3.3 Stage 2 Performance and Ablations

Qualitative reports indicated that the architectural choices within the Stage 2 severity model were critical. The described MIL approach incorporating Bi-LSTM and dual attention mechanisms significantly outperformed baseline attempts using simpler 2.5D models (which process slices nearly independently). Specifically, adding the Bi-LSTM to model sequential dependencies between instance features before attention aggregation, along with the guiding effect of the auxiliary attention loss, reportedly yielded substantial improvements in evaluation metrics compared to a simpler attention-only MIL baseline. This underscores the importance of capturing inter-slice context and carefully guiding the attention process in the MIL framework for this task.

4 Discussion

This paper details a comprehensive framework that systematically addresses the challenges of automated lumbar spine MRI analysis through a multi-stage, multi-modal approach. The decomposition into localization (instance number, then coordinates) and subsequent severity classification allows for specialized model development and optimization.

Key Findings and Contributions: The study highlights several effective strategies:

- The explicit, sequential separation of Stage 1 into instance number prediction (using 3D context) and coordinate prediction (using 2D slices) is a beneficial architectural choice.
- Combining 3D inputs for initial slice identification and 2D inputs for fine-grained localization and patch-based severity analysis effectively leverages different aspects of the MRI data.
- A crucial innovation is the data augmentation strategy in Stage 2 that simulates potential errors from Stage 1 by randomly shifting coordinates and instance numbers, significantly enhancing the robustness of the severity classifier to real-world localization inaccuracies.
- The proposed Bi-LSTM + Attention MIL architecture, particularly with the inclusion of an auxiliary attention loss, proves effective for aggregating information from multiple cropped instances (slices) for accurate severity grading.
- Domain-specific pretraining provides a tangible advantage for the coordinate prediction subtask compared to standard ImageNet pretraining alone.

Limitations and Negative Results: The development process also identified approaches that were less effective within this specific implementation:

- Alternative sequence models like MAMBA or pure Self-Attention mechanisms did not surpass the performance of Bi-LSTM for aggregating instance features in the MIL module.
- Sharing parameters between the primary and auxiliary attention heads in the MIL component was found to be detrimental, suggesting distinct roles for these mechanisms.
- Empirical testing revealed that including certain MRI sequences for specific tasks did not improve performance (e.g., using SagT1 images for SCS classification, or SagT2 for NFN).
- Counterintuitively, larger CNN backbones or extended training durations led to degraded performance for the severity classification task, possibly due to overfitting on the complex patch-based inputs or challenges in optimizing larger models with the given setup.
- Standard Vision Transformer models did not outperform the selected CNN backbones in the severity classification experiments conducted here.

Detailed Future Work: While this framework demonstrates strong capabilities, several avenues for future research and development exist:

- *Enhanced 3D Context Modeling:* Explore more advanced 3D CNN or 3D Transformer architectures directly for severity classification, potentially capturing subtle through-plane features missed by the current 2.5D/MIL approach, while carefully managing computational costs.
- *Improved Interpretability:* Integrate techniques like Grad-CAM, attention map visualization, or concept-based explanations to provide clinicians with insights into the model's decision-making process, fostering trust and facilitating error analysis. Visualizing the learned attention weights from the MIL module could highlight which slices contribute most to the severity prediction.
- *Multi-Modal Data Integration:* Investigate methods to incorporate other relevant patient information, such as clinical history, symptoms, or demographic data available in Electronic Health Records (EHR), potentially creating more personalized and accurate risk assessments or predictions.
- *Semi-Supervised and Self-Supervised Learning:* Leverage the vast amounts of unlabeled MRI data often available in clinical archives. Techniques like contrastive learning, masked autoencoders, or consistency regularization could be used for pretraining models, potentially leading to more robust feature representations and reducing the reliance on large labeled datasets.
- *Ordinal Classification Approaches:* Explicitly model the ordinal nature of severity grades (e.g., Mild $<$ Moderate $<$ Severe) using specialized ordinal regression loss functions or network output structures, which might improve performance compared to treating severity as purely categorical.
- *Robustness Across Sites and Scanners:* Conduct extensive testing and potentially apply domain adaptation or generalization techniques to ensure the framework performs reliably across different hospitals, MRI scanners, and imaging protocols.
- *Prospective Clinical Validation:* The most critical next step involves rigorously evaluating the framework's performance in realistic clinical workflows through prospective studies, comparing

its diagnostic accuracy, efficiency, and impact on clinical decisions against current human expert performance.

5 Conclusion

The proposed multi-stage, multi-modal deep learning framework offers a detailed and demonstrably effective solution for the complex task of automated lumbar spine degenerative classification using MRI data. By meticulously addressing the distinct challenges of localization (through sequential instance number and coordinate prediction) and severity grading (via an advanced Bi-LSTM-Attention MIL architecture), the system achieves strong results. Key strengths include the hybrid use of 3D and 2D information, specialized model components, coordinate-guided analysis, and, critically, a data augmentation strategy designed to confer robustness against upstream localization errors. While computationally demanding in its entirety, the detailed methodology and insights gleaned from experimental ablations provide a valuable foundation and blueprint for the continued development and refinement of sophisticated AI tools aimed at enhancing the efficiency and consistency of spinal imaging analysis.

References

[1] Hoy, D., March, L., Brooks, P., et al. (2014). The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. *Annals of the Rheumatic Diseases*, 73(6), 968-974.

[2] Placeholder reference 2: Paper on MRI for lumbar spine diagnosis (e.g., Jarvik Deyo, 2002, Ann Intern Med).

[3] Placeholder reference 11: Paper on inter-observer variability in lumbar MRI reading (e.g., Carrino et al., 2009, Spine).

[4] Lee, Aric, et al. (2024). Applications of Artificial Intelligence and Machine Learning in Spine MRI. *Bioengineering*, 11(9), 894.

[5] Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60-88.

[6] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11976-11986).

[7] Tan, M., Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning (ICML)* (pp. 10096-10106). PMLR.

[8] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

[9] Schuster, M., Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673-2681.

[10] Ilse, Maximilian, Jakub Tomczak, and Max Welling. (2018). Attention-based deep multiple instance learning. In *International conference on machine learning (ICML)* (pp. 2127-2136). PMLR.

[11] Lin, Juncai, Honglai Zhang, and Hongcai Shang. (2024). Convolutional neural network incorporating multiple attention mechanisms for MRI classification of lumbar spinal stenosis. *Bioengineering*, 11(10), 1021.

[12] Wang, Zhiwei, et al. (2024). Accurate scoliosis vertebral landmark localization on x-ray images via shape-constrained multi-stage cascaded cnns. *Fundamental Research*, 4(6), 1657-1665.

[13] Vania, Malinda, and Deukhee Lee. (2021). Intervertebral disc instance segmentation using a multistage optimization mask-RCNN (MOM-RCNN). *Journal of Computational Design and Engineering*, 8(4), 1023-1036.

[14] Kim, Min-Jung, et al. (2021). Evaluation of a multi-stage convolutional neural network-based fully automated landmark identification system using cone-beam computed tomography-synthesized posteroanterior cephalometric images. *Korean Journal of Orthodontics*, 51(2), 77-85.

[15] Yilihamu, Elzat Elham-Yilizati, et al. (2025). Quantification and classification of lumbar disc herniation on axial magnetic resonance images using deep learning models. *La radiologia medica*, (Published online, volume/pages pending or use DOI if available).

[16] Savale, Ishita, et al. (2024). Lumbar Spinal Stenosis Degenerative Detection and Classification on MRI Images. *Preprint/Technical Report* (November 15, 2024). (Adjust source type as appropriate).

[17] Li, Haixing, et al. (2021). Automatic lumbar spinal MRI image segmentation with a multi-scale attention network. *Neural Computing and Applications*, 33, 11589-11602.

[18] Wang, Shuai, et al. (2022). Automatic segmentation of lumbar spine MRI images based on improved attention U-net. *Computational Intelligence and Neuroscience*, 2022, Article ID 4259471.

[19] Tabatabaei, Sadafossadat, Khosro Rezaee, and Min Zhu. (2023). Attention transformer mechanism and fusion-based deep learning architecture for MRI brain tumor classification system. *Biomedical Signal Processing and Control*, 86, 105119.