Multi-Pathology Segmentation of the Lumbar Spine

Claudio Leite
Federal University of São Carlos (UFSCar)

Jurandy Almeida
Federal University of São Carlos (UFSCar)

***Abstract*—The diagnosis of spinal pathologies is complex due to the frequent overlap of multiple diseases in the same anatomical location, a scenario that current segmentation or classification methods do not efficiently address. This work presents an empirical study on the segmentation of multiple overlapping pathologies, proposing and systematically comparing three strategies: (i) a baseline *binary class* approach using independent models; (ii) a *multi-class* approach mapping disease combinations to unique labels; and (iii) a *multi-label* approach using parallel channels to explicitly model co-occurrence. We evaluated over 300 training and inference pipelines, combining five neural network architectures and three loss functions. Our preliminary results show that the multi-label strategy significantly outperforms the other approaches in both accuracy and computational efficiency, establishing a promising direction for developing robust, scalable diagnostic tools.**

I. Introduction

Low back pain is a leading cause of disability worldwide [2], and Magnetic Resonance Imaging (MRI) of the lumbar spine is a cornerstone for its diagnosis [1]. However, a central challenge in the automated analysis of these images is the semantic overlap of multiple classes, where a single intervertebral disc may simultaneously present with conditions like a herniation and disc narrowing—a scenario most current methods are not designed to handle.

Current deep learning methods for lumbar spine analysis typically follow two paths: single-disease segmentation [4], [5] or complex multi-stage classification pipelines [6]. The former ignores pathological co-occurrence, while the latter are often computationally inefficient and difficult to generalize. Neither approach effectively handles the common clinical scenario of multiple, overlapping pathologies within a unified framework.

To address this gap, this paper presents an ongoing, comprehensive study on the segmentation of multiple pathologies in lumbar discs. We propose and compare three distinct strategies to manage diagnostic overlap. Our goal is to determine the most effective and efficient strategy for simultaneously delineating and diagnosing multiple co-occurring conditions.

The main contributions of this work in progress are:

- The proposition of three strategies for the multi-diagnosis problem: (i) a **binary class** approach treating each pathology independently; (ii) a **multi-class** approach mapping disease combinations to unique and exclusive classes; and (iii) a **multi-label** approach, which explicitly models the coexistence of diagnoses in distinct binary channels.
- A systematic comparative analysis with over 300 training and inference pipelines, evaluating five neural networks and three loss functions, establishing a reference benchmark for this complex clinical scenario.
- The demonstration that the proposed multi-label strategy offers a superior trade-off between accuracy and computational cost, achieving performance comparable to the significantly more expensive binary approach.

The rest of this work is organized as follows. Section II discusses related work. Section III presents our approaches to multi-pathology diagnosis. Section IV describes the experimental setup. Section V reports our results. Finally, Section VI offers our conclusions and directions for future work.

II. Related Work

The literature on spinal imaging with deep learning is extensive. One significant branch of research focuses on segmenting symptomatic areas for a single pathology, such as Lumbar Spinal Stenosis [4] or Disc Herniation [14], [15]. While effective for their specific tasks, these methods are inherently single-label and cannot address the frequent coexistence of multiple diseases.

A second line of work aims to diagnose multiple diseases but typically relies on complex, multi-stage pipelines that separate structure localization from classification [6], [16]. These fragmented approaches are often computationally expensive and difficult to generalize. Although some recent works have explored multi-label classification of spinal pathologies [17], [18], they operate on an image or patch level, failing to provide the precise spatial localization that semantic segmentation offers. Our work bridges this gap by proposing end-to-end segmentation frameworks explicitly designed to handle multi-pathology overlap.

III. Our Approaches

To address the challenge of multi-pathology diagnosis, we propose and systematically evaluate three distinct semantic segmentation strategies. These strategies—Binary Class, Multi-Class, and Multi-Label Segmentation—represent different conceptual frameworks for handling diagnostic overlap and are detailed in the following subsections.

A. Binary Class Segmentation

The first strategy (Figure 1), **Binary Class Segmentation**, serves as our baseline and decomposes the complex multi-diagnosis problem into a series of independent binary segmentation tasks. In this formulation, a separate segmentation model is trained specifically for each of the pathologies analyzed.

Let $X \in R^{D \times H \times W}$ be an input Magnetic Resonance Imaging (MRI) volume and $\mathcal{P} = \{p_1, \ldots, p_n\}$ be the set of n pathologies. For each pathology $p_k \in \mathcal{P}$, a corresponding

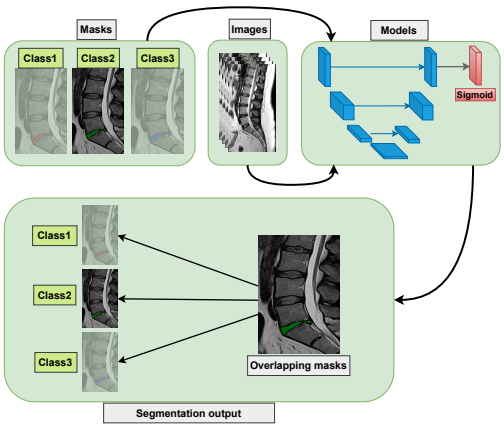


Fig. 1: Binary Class Segmentation.

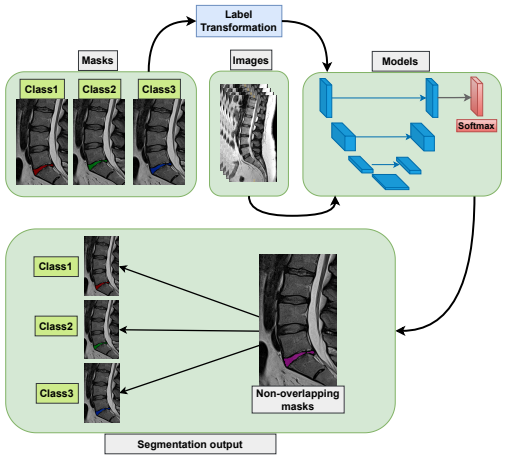


Fig. 2: Multi-Class Segmentation.

binary ground truth mask $Y_k \in \{0,1\}^{D\times H\times W}$ indicates its location. The objective is to learn a set of n independent models $\{\mathcal{M}_k\}_{k=1}^{n}$, where each model $\mathcal{M}_k : R^{D\times H\times W} \rightarrow [0,1]^{D\times H\times W}$ is trained to predict the mask Y_k. The final diagnosis for the volume X is the collection of all n individual predictions and can be expressed as:

$$\hat{Y}_k = \mathcal{M}_k(X), \quad \text{for } k = 1,\ldots,n \tag{1}$$

where each predicted mask $\hat{Y}_k$ is generated by a model $\mathcal{M}_k$.

The main advantage of this strategy is its simplicity and focus. By training a model for a single task (e.g., segmenting only Disc Herniation), the network can specialize in learning the unique visual features of that condition. This can lead to high performance for each individual disease and establishes a robust performance ceiling against which the other, more complex strategies can be compared.

However, the drawback of this approach lies in its high computational cost and resource inefficiency. The need to train, validate, and store n independent models multiplies the experimentation time and hardware demand. Furthermore, in a clinical scenario, it would be necessary to run all these models on a single patient's MRI to obtain a complete diagnosis, making the process slow and cumbersome. This strategy also fails to learn any correlations that may exist between different diseases, treating each as an isolated event.

B. Multi-Class Segmentation

The second strategy (Figure 2), **Multi-Class Segmentation**, reformulates the multi-diagnosis problem into a non-overlapping multi-class segmentation task. The premise is to create a unique class identifier for every possible combination of pathologies that can occur in a specific intervertebral disc.

Let $\mathcal{C}_{obs} \subseteq 2^{\mathcal{P}}$ be the set of m unique pathology combinations observed in the dataset. We create a new ground truth mask $Y' \in \{0,1,\ldots,m\}^{D\times H\times W}$. For each intervertebral disc, all its voxels are assigned a single integer label $c \in \{1,\ldots,m\}$ that uniquely identifies the specific combination of pathologies present, with $c = 0$ for the background. The goal is to learn a single model $\mathcal{M} : R^{D\times H\times W} \rightarrow R^{(m+1)\times D\times H\times W}$ that performs standard multi-class over these fused labels. This operation can be expressed as:

$$\hat{Y}' = \underset{c}{\text{argmax}}(\mathcal{M}(X)) \tag{2}$$

where the model $\mathcal{M}$ outputs a probability distribution over the $m+1$ classes for each voxel, and the final prediction $\hat{Y}'$ is the class with the maximum probability.

The main advantage of this approach is its conceptual simplicity, allowing standard deep learning architectures to be trained end-to-end. However, its primary drawback is the combinatorial explosion in the number of classes, which increases model complexity and exacerbates class imbalance, as many pathology combinations are extremely rare. In this study, the $m = 70$ classes observed in the dataset were used.

C. Multi-Label Segmentation

The final and most flexible strategy is **Multi-Label Segmentation** (Figure 3), which directly addresses the problem of coexisting pathologies. In this formulation, each disease is treated as an independent, binary segmentation channel, allowing for multiple, overlapping predictions.

Let the ground truth be a tensor $Y \in \{0,1\}^{n\times D\times H\times W}$, where each channel Y_k is the binary mask for pathology $p_k \in \mathcal{P}$. The objective is to learn a single model $\mathcal{M} : R^{D\times H\times W} \rightarrow [0,1]^{n\times D\times H\times W}$ that takes an MRI volume and outputs a tensor of n probability maps, one for each pathology. The operation for this approach can be expressed as:

$$\{\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n\} = \mathcal{M}(X) \tag{3}$$

where a single model $\mathcal{M}$ simultaneously generates a set of n distinct prediction masks.

The primary advantage of this strategy is its flexibility and efficiency. By treating each disease as an independent task within a single model, it can generalize to pathology combinations not seen during training and is far more computationally efficient than the Binary Class strategy. Additionally,

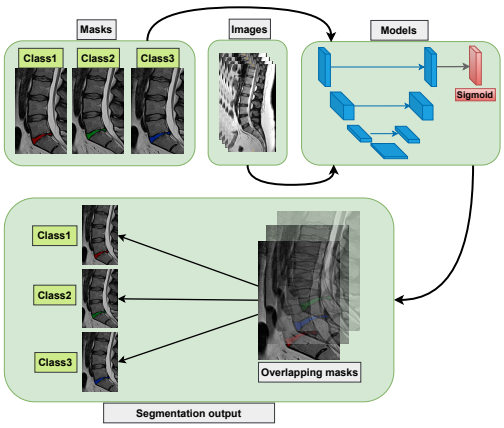


Fig. 3: Multi-Label Segmentation.

the model can learn shared features that may be relevant to multiple diseases, potentially improving overall performance.

Its complexity, however, lies in the training process, as it requires optimizing n tasks simultaneously. This demands rigorous gradient management and careful selection of the loss function to ensure balanced learning across all tasks.

IV. Experimental Setup

Preliminary experiments were conducted on the public SPIDER (SPine Imaging Diagnostic Extended Resource) dataset [3]. It contains 447 T1 and T2-weighted MRI series from 257 patients, collected from four different hospitals in the Netherlands, with patient ages ranging from 18 to 95 years. The dataset is enriched with two types of expert-validated annotations: precise anatomical segmentations of vertebrae, intervertebral discs, and the spinal canal, alongside detailed radiological classifications for a spectrum of degenerative changes, including disc herniation, spondylolisthesis, and Modic changes. This simultaneous availability of segmentation masks and multi-pathology labels provides the essential ground truth required for our investigation into automated multi-pathology diagnosis.

To evaluate the proposed strategies, we selected five state-of-the-art architectures for medical image segmentation, spanning convolutional (CNN), hybrid, and Transformer-based models. The CNN architectures include the **U-Net 3D** [8], an extension of U-Net for volumetric data, and the **V-Net** [9], which integrates residual connections for deeper networks. As a hybrid approach, we used a **U-Net with a ResNet-50** [11] encoder for more powerful feature extraction. Finally, we explored Transformer-based models: the **UNETR** [7], which uses a Transformer as an encoder to capture long-range spatial dependencies, and its evolution, the **Swin UNETR** [10], which employs the more efficient attention of the Swin Transformer to model global relationships. To optimize model training, we evaluated three distinct loss functions tailored for this task: the standard Dice loss [9], a hybrid Dice+Focal loss [12], and the Generalized Dice loss [13].

All experiments followed a standardized protocol using only T2-weighted MRI scans. The dataset was partitioned into training (134 scans), validation (34 scans), and test (42 scans) sets. To ensure reproducibility, all runs were initialized with a fixed random seed (42). For preprocessing, all MRI volumes were intensity-scaled and resized to $32\times192\times192$ voxels using center cropping or zero-padding. All models were trained for 500 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 8. Performance was measured using the Dice Similarity Coefficient (DSC). We report both the macro-average across classes, which assess performance at the class level; and the example-based average across voxels, which assess performance at the instance level [19].

V. Current State of the Research

In this stage of the research, we have processed and evaluated over 300 training and inference pipelines to systematically compare the three proposed strategies. The analysis of the obtained results has allowed us to draw important conclusions about the feasibility of each approach.

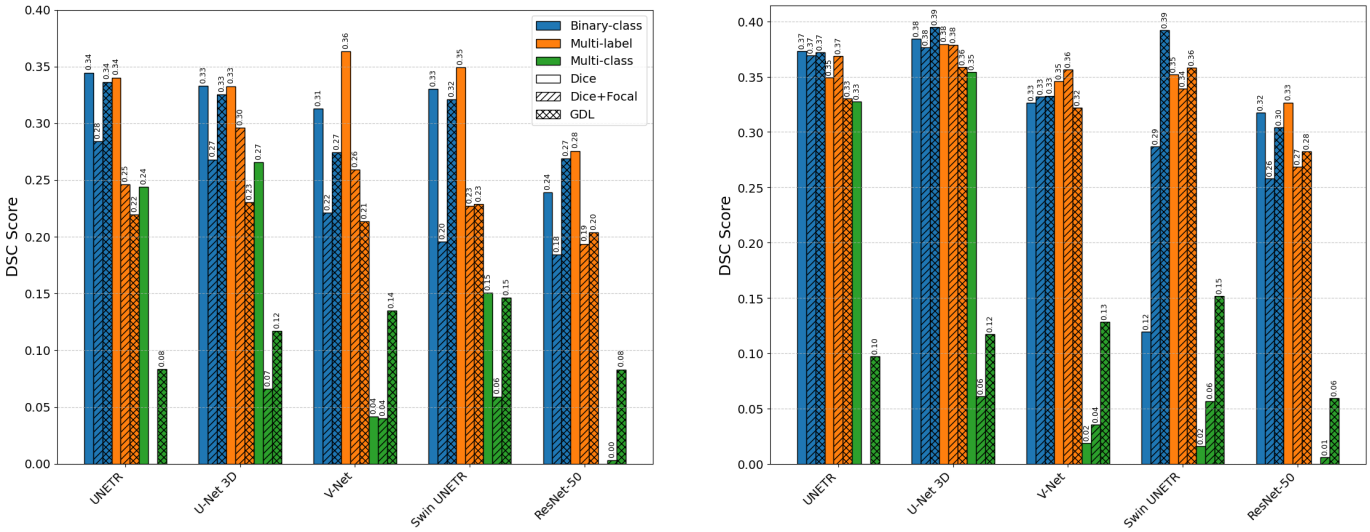
A. Preliminary Results and Discussion

Figure 4 provides a detailed comparison between the Binary Class, Multi-Class, and Multi-Label strategies across all architecture-loss combinations. The Multi-Class strategy proved to be ineffective. Due to the extreme class imbalance created by mapping 70 unique pathology combinations to distinct classes, the macro-average DSC scores were consistently below 0.15 for nearly all models. The poor performance confirms that this approach is not viable for this problem. Consequently, it was excluded from further comparison.

The results in Figure 4(a) show a clear and significant finding: the **Multi-Label strategy consistently outperforms the Binary Class baseline**. The best Multi-Label configurations (V-Net and Swin UNETR with Dice loss) achieved a macro-average DSC of $\mathbf{0.36}$ and $\mathbf{0.35}$. In contrast, the best-performing Binary Class model (UNETR with Dice/GDL) only reached a score of 0.34. This strongly suggests that training a single, unified model capable of learning shared features across pathologies is more effective than training specialized, isolated models.

The example-based DSC results in Figure 4(b) measure the model's ability to predict the correct set of labels for each voxel. Here, the performance gap narrows. The best Binary Class models achieved a score of $\mathbf{0.39}$, while the best Multi-Label model was just behind at $\mathbf{0.38}$. This indicates that while the Multi-Label approach is superior at identifying pathologies at a macro level, the specialized single models remain competitive at identifying the precise combination of labels at a given location.

Our results confirm the efficacy of the **Multi-Label strategy**, which achieves superior performance with a single model, making it more computationally efficient and clinically practical than the Binary Class approach. Its success stems from the ability to learn shared representations. The **Swin UNETR** and



(a) Macro-average DSC

(b) Example-based DSC

Fig. 4: DSC results for the three proposed strategies across different models and loss functions.

V-Net architectures consistently performed best, especially with **Dice loss**, confirming our approach's robustness.

B. Next Steps

Based on these results, the research has advanced to a second experimental phase to refine and validate our conclusions. We are currently utilizing random crops of dimension $32 \times 192 \times 192$ instead of a fixed center crop for training. Furthermore, inference will be performed with a sliding window approach to generate full-size segmentation maps, aiming to improve the method's clinical applicability.

VI. Conclusion

In this work-in-progress, we presented a systematic evaluation of three deep learning strategies for the challenging task of multi-pathology segmentation in lumbar spine MRI. Our preliminary results indicate that a **Multi-Label segmentation strategy** is significantly more effective and efficient than treating each disease independently or mapping combinations to unique classes. Architectures like **Swin UNETR** and **V-Net**, optimized with a **Dice loss**, have shown the most promise.

This study establishes a strong foundation for future research. Our ongoing work involves exploring random crops for training and a sliding window-based inference, aiming to generate full-size segmentation maps of the scan and thus enhance the clinical applicability of our approach. The ultimate goal is to develop a validated, scalable, and clinically useful tool for automated spinal diagnosis.

Acknowledgment

This research was supported by FAPESP (grants 2023/17577-0 and 2024/22985-3) and CNPq (grants 315220/2023-6, 420442/2023-5, and 444982/2024-8).

References

[1] G. McNicoll, "World Population Ageing 1950-2050," *Popul. Dev. Rev.*, pp. 814-816, 2002.
[2] M.G. Fehlings et al., "The aging of the global population: the changing epidemiology of disease and spinal disorders," *Neurosurgery*, pp. S1-S5, 2015.
[3] J. W. van der Graaf et al., "Lumbar spine segmentation in MR images: a dataset and a public benchmark," Sci. Data, vol. 11, no. 264, 2024.
[4] İ. Altun et al., "LSS-UNET: Lumbar spinal stenosis semantic segmentation using deep learning," *Multimed. Tools Appl.*, 82:41287-41305, 2023.
[5] J. Qian et al., "Lumbar disc herniation diagnosis using deep learning on MRI," *J. Radiat. Res. Appl. Sci.*, 17(3):100988, 2024.
[6] R. Windsor et al., "SpineNetV2: automated detection, labelling and radiological grading of clinical MR scans," arXiv:2205.01683, 2022.
[7] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in *WACV*, pp. 1748-1758, 2022.
[8] E. Kerfoot et al., "Left-ventricle quantification using residual U-Net," in *STACOM*, LNCS 11395, pp. 371-380, 2019.
[9] F. Milletari et al., "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*, pp. 565-571, 2016.
[10] A. Hatamizadeh et al., "Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images," in *BrainLes*, LNCS 12962, pp. 272-284, 2022.
[11] K. He et al., "Deep residual learning for image recognition," in *CVPR*, pp. 770-778, 2016.
[12] T.-Y. Lin et al., "Focal loss for dense object detection," in *ICCV*, pp. 2980-2988, 2017.
[13] C. H. Sudre et al., "Generalised Dice Overlap as a deep learning loss function for highly unbalanced segmentations," in *DLMIA, MICCAI Wkshp.*, pp. 240-248, 2017.
[14] W. Mbarki et al., "A novel method based on deep learning for herniated lumbar disc segmentation," in *IC ASET*, pp. 394-399, 2020.
[15] J. Qian et al., "Lumbar disc herniation diagnosis using deep learning on MRI," *J. Radiat. Res. Appl. Sci.*, 17(3):100988, 2024.
[16] Q. Pan et al., "Automatically diagnosing disk bulge and disk herniation...: Method development study," *JMIR Med. Inform.*, 9(5):e14755, 2021.
[17] Y. Chen et al., "Deep learning-based computer-aided diagnostic system for lumbar degenerative diseases classification using MRI," *Biomed. Signal Process. Control*, 109:108002, 2025.
[18] Y. Wang et al., "Deep learning-driven diagnosis of multi-type vertebra diseases based on computed tomography images," *Quant. Imaging Med. Surg.*, 14(1):800, 2023.
[19] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, 26(8):1819-1837, 2013.