Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Cláudio Luiz Leite Júnior

Multi-Pathology Segmentation in Lumbar Spine MRI: A Comparative Deep Learning Approach

São Carlos
2025

Cláudio Luiz Leite Júnior

Multi-Pathology Segmentation in Lumbar Spine MRI: A Comparative Deep Learning Approach

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Visão Computacional

Orientador: Jurandy Gomes de Almeida Junior

São Carlos
2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Relatório de Defesa de Dissertação
Candidato: Cláudio Luiz Leite Júnior.

Aos 29/10/2025, às 14:00, realizou-se na Universidade Federal de São Carlos, nas formas e termos do Regimento Interno do Programa de Pós-Graduação em Ciência da Computação, a defesa de dissertação de mestrado sob o título: Multi-Pathology Segmentation in Lumbar Spine MRI: A Comparative Deep Learning Approach, apresentada pelo candidato Cláudio Luiz Leite Júnior.. Ao final dos trabalhos, a banca examinadora reuniu-se em sessão reservada para o julgamento, tendo os membros chegado ao seguinte resultado:

Participantes da Banca	**Função**	**Instituição**	**Resultado**	**Resultado Final**
Prof. Dr. Jurandy Gomes de Almeida Junior	Presidente	UFSCar	Aprovado	Aprovado
Prof. Dr. Cesar Henrique Comin	Titular	UFSCar	Aprovado	
Prof. Dr. Fábio Augusto Menocci Cappabianco	Titular	UNIFESP	Aprovado	

Parecer da Comissão Julgadora*:

O candidato apresentou o trabalho e respondeu as perguntas adequadamente.

Encerrada a sessão reservada, o presidente informou ao público presente o resultado. Nada mais havendo a tratar, a sessão foi encerrada e, para constar, eu, Ivan R Silva, representante do Programa de Pós-Graduação em Ciência da Computação, lavrei o presente relatório, assinado por mim e pelos membros da banca examinadora.

Prof. Dr. Jurandy Gomes de Almeida Junior

Representante do PPG: Ivan R Silva

Prof. Dr. Cesar Henrique Comin

Prof. Dr. Fábio Augusto Menocci Cappabianco

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Jurandy Gomes de Almeida Junior, Cesar Henrique Comin, Fábio Augusto Menocci Cappabianco e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Prof. Dr. Jurandy Gomes de Almeida Junior

(X) Não houve alteração no título ( ) Houve alteração no título. O novo título passa a ser:

Observações:
a) Se o candidato for reprovado por algum dos membros, o preenchimento do parecer é obrigatório.
b) Para gozar dos direitos do título de Mestre ou Doutor em Ciência da Computação, o candidato ainda precisa ter sua dissertação ou tese homologada pelo Conselho de Pós-Graduação da UFSCar.

Este trabalho é dedicado à minha mãe,
que me ensinou o valor em sonhar.

Agradecimentos

Agradecimentos à minha esposa, família e amigos por estarem comigo neste e outros tantos momentos nesta vida. E ao meu orientador, Jurandy Almeida, por toda paciência, condução e aprendizado.

"O correr da vida embrulha tudo,
a vida é assim: esquenta e esfria,
aperta e daí afrouxa,
sossega e depois desinquieta.
O que ela quer da gente é coragem."
(Guimarães Rosa)

Abstract

Low back pain is a leading cause of disability worldwide, and Magnetic Resonance Imaging (MRI) of the lumbar spine is fundamental for its diagnosis. However, the manual analysis of these images is a time-consuming and subjective process, and existing computational methods are often limited to a single pathology or rely on complex, multi-stage pipelines. A central challenge, not yet fully addressed, is the simultaneous occurrence of multiple pathologies in the same anatomical structure—a common clinical scenario that current models fail to model effectively. This dissertation addresses this gap by proposing and validating a robust methodology for the automated segmentation of multiple co-existing pathologies in lumbar intervertebral discs. The work was structured in two complementary phases. The first established the methodological underpinnings of this research. It consisted of a rigorous empirical study evaluating five deep learning architectures and four loss functions to determine the most effective approaches for the fundamental task of segmenting the vertebrae and intervertebral discs. This step ensured that the subsequent investigation into pathologies was built upon a robust and validated base. The second phase, the main contribution of this work, systematically investigated three distinct strategies for multi-pathology segmentation: (i) **binary class segmentation**, a baseline that treats each pathology independently; (ii) **multi-class segmentation**, mapping 70 disease combinations to unique classes (non-overlapping masks); and (iii) **multi-label segmentation**, which uses binary channels to explicitly model the coexistence of multiple diagnoses (overlapping masks). Our results, derived from over 200 training pipelines, demonstrate that the multi-label approach, especially when implemented with the V-Net and Swin UNETR architectures, achieves diagnostic accuracy comparable to the baseline while offering significantly superior computational efficiency. By developing a unified framework that integrates the precise spatial localization of symptomatic areas with the classification of multiple concurrent diseases, this work establishes a practical and efficient guideline for future research and clinical applications in the automated diagnosis of spinal pathologies.

Keywords: Medical Image Segmentation. Deep Learning. Lumbar Spine MRI. Multi-Label Learning. Multi-Pathology Diagnosis.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Low back pain is a leading cause of disability worldwide, affecting the quality of life of millions and imposing a significant burden on healthcare systems. It is recognized as the most common cause of long-term disability, with estimates suggesting that approximately 80% of people are affected by it at some point in their lives (PALMER et al., 2000). This public health challenge is significantly exacerbated by the global population's aging demographic, a trend quantified by the United Nations Department of Economic and Social Affairs, which reported that the number of people aged 60 and over nearly tripled between 1950 and 2000 (MCNICOLL, 2002). This demographic shift presents a unique set of clinical challenges, as eloquently summarized by (FEHLINGS et al., 2015):

> "The elderly population represents a unique challenge to healthcare systems and to spine clinicians, as this age group is associated with multiple medical comorbidities, reduced bone density and osteoporosis, reduced mobility, spinal degeneration and deformities, poor balance and a greater propensity for falls."

In this complex clinical scenario, Magnetic Resonance Imaging (MRI) of the lumbar spine stands as a cornerstone for accurate diagnosis. However, its efficacy faces limitations; its manual interpretation is time-consuming and subjective, and despite a high rate of appropriate referrals, studies show that only a small fraction of these scans effectively contributes to clinical decision-making (GRAAF et al., 2024), highlighting a critical need for more efficient and robust analysis tools.

In this context, Computer Vision driven by Deep Learning (DL) has emerged as a powerful tool to support the analysis of medical images, with vast potential to optimize clinical workflows and assist in medical diagnoses (CASTIGLIONI et al., 2021). A fundamental pillar for any automated diagnostic system targeting spinal pathologies is the

ability to first reliably identify the anatomical structures of interest. The precise segmentation of vertebrae and intervertebral discs serves as the foundational layer upon which subsequent, more complex analyses can be reliably built. Therefore, to ensure that our investigation into pathologies rests on a robust and accurate foundation, a thorough evaluation of state-of-the-art anatomical segmentation methods is a crucial prerequisite and motivating component of this work.

Moving beyond this foundational step, a more significant clinical challenge lies in the automated identification and segmentation of the pathologies themselves. A central difficulty in this domain is the clinical reality of pathological co-occurrence, where a single intervertebral disc may simultaneously present with multiple conditions, such as a herniation and disc narrowing. For instance, a statistical analysis performed on the dataset used in this work, known as SPIDER (*SPIne segmentation: Discs, vERtebrae and spinal canal*) (GRAAF et al., 2024), reveals that 97.25% of the patients present some form of anomaly overlap in their lumbar intervertebral discs, underscoring the critical need for models capable of handling this complexity. Most current computational methods are not designed to handle this semantic overlap effectively. Existing DL-based research efforts have largely followed two distinct paths. The first group involves approaches that support diagnosis by segmenting symptomatic areas for a single disease at a time, shown in Figure 1. Notable examples include the work of (ALTUN; ALTUN; ALKAN, 2023), who used a segmentation approach to identify areas of Lumbar Spinal Stenosis, and (QIAN et al., 2024), who focused on segmenting regions of Disc Herniation. The second group includes approaches that perform radiological classification for multiple diseases, often employing multi-stage pipelines, shown in Figure 2. For instance, (LAMA et al., 2022) integrated radiomic features to optimize a CNN for classification, while (WINDSOR et al., 2022) used a series of neural networks to first detect vertebrae and discs, and then classify cropped regions using a Residual Network (ResNet).

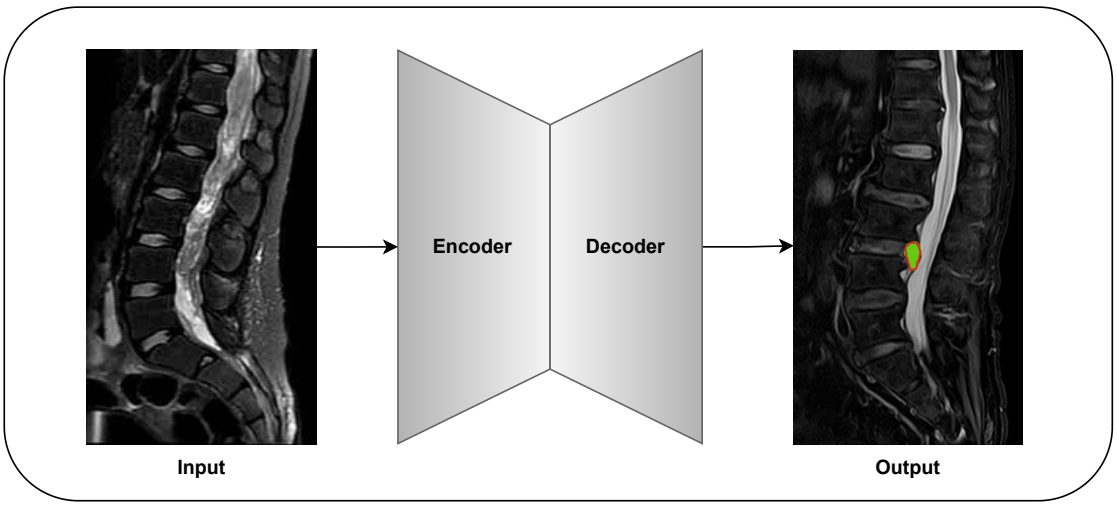
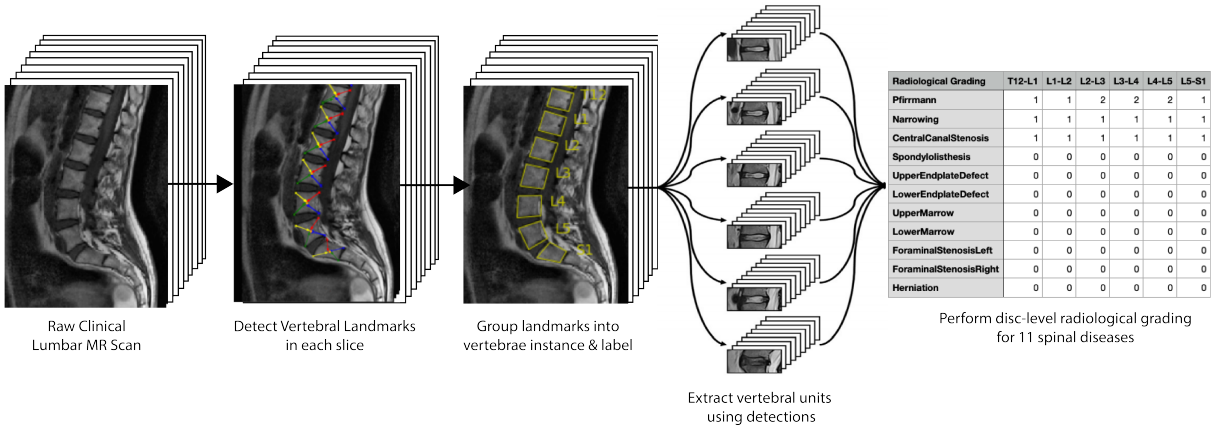


Figure 1 – Segmentation of Disc Herniation on Sagittal MRI.



Radiological Grading	T12-L1	L1-L2	L2-L3	L3-L4	L4-L5	L5-S1
Pfirrmann	1	1	2	2	2	1
Narrowing	1	1	1	1	1	1
CentralCanalStenosis	1	1	1	1	1	1
Spondylolisthesis	0	0	0	0	0	0
UpperEndplateDefect	0	0	0	0	0	0
LowerEndplateDefect	0	0	0	0	0	0
UpperMarrow	0	0	0	0	0	0
LowerMarrow	0	0	0	0	0	0
ForaminalStenosisLeft	0	0	0	0	0	0
ForaminalStenosisRight	0	0	0	0	0	0
Herniation	0	0	0	0	0	0

Figure 2 – Radiological classification steps by (WINDSOR et al., 2022).

While effective in their specific tasks, both approaches possess significant limitations. The former is constrained to a single-condition diagnosis, ignoring the frequent clinical scenario of co-occurrence, while the latter often relies on complex, computationally inefficient processes that lack precise spatial localization of the findings. This gap in the literature raises a central research question: *Is it possible to develop a unified deep learning framework that combines the precise spatial indication of symptomatic areas with the radiological classification of multiple, co-occurring pathologies in lumbar intervertebral discs?*

Motivated by this question, this dissertation proposes to investigate deep learning strategies capable of indicating symptomatic regions in lumbar MRIs while simultaneously classifying them into multiple disease categories, shown in Figure 3. The development of such an integrated solution requires overcoming three main challenges: (*i*) the inherent difficulty of modeling anatomical regions affected by more than one distinct pathology; (*ii*) the limitation of available medical datasets, which often provide annotations for only a single disease per image; and (*iii*) the technical complexity of adapting deep learning methods, often designed for single-label tasks, to perform multi-label semantic segmentation in a complex clinical context. Addressing these challenges would not only advance diagnostic support for lumbar spine pathologies but also contribute to the broader field of deep learning in medical imaging.

1.1 Objectives

The main objective of this dissertation is to investigate and validate a deep learning framework that integrates the capacity for indicating symptomatic areas with the radiological classification of multiple anomalies or diseases, aiming to support the analysis and diagnosis of lumbar spine magnetic resonance images. To achieve this main objective, the research was structured into two sequential and complementary specific objectives:

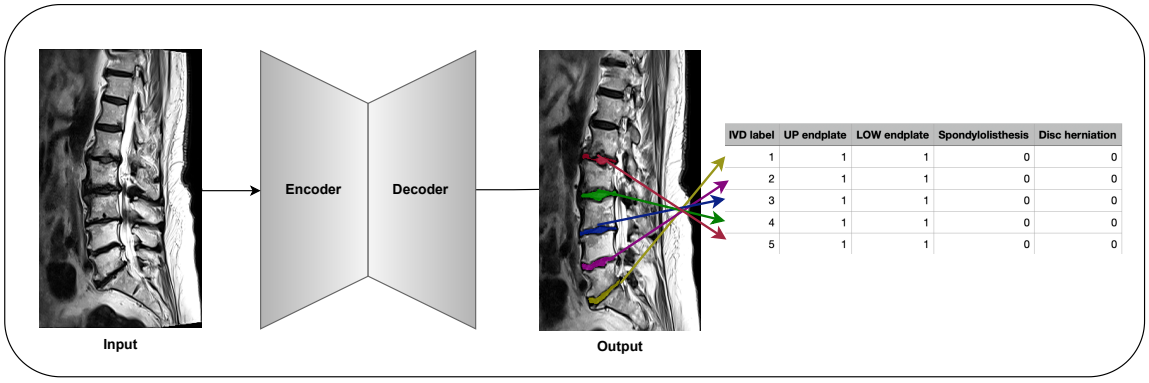


Figure 3 – Proposed process of segmenting multiple anomalies or diseases.

1. To establish a robust methodological foundation by systematically evaluating state-of-the-art architectures and loss functions to determine the most effective approaches for the fundamental task of the precise spatial indication of vertebrae and intervertebral discs.

2. Building upon the established foundation, to develop and validate a methodology for the precise spatial indication of symptomatic areas with the radiological classification of multiple, co-occurring pathologies, investigating different modeling strategies to identify the one that offers the best trade-off between diagnostic accuracy and computational efficiency when dealing with diagnostic overlap.

1.2 Hypotheses and Research Questions

This work is guided by the central hypothesis that it is possible to address the complex clinical reality of pathology co-occurrence through a deep learning approach that provides a precise indication of areas affected by multiple pathologies, overcoming the efficiency and scope limitations of current approaches that treat diseases in isolation or in multi-stage processes.

For the validation of this hypothesis and the achievement of the proposed objectives, the investigation was broken down into the following specific questions:

1. Which combination of deep learning architectures and loss functions presents the most robust performance for the foundational task of anatomical indication of the spine's structures, establishing a methodological foundation for pathological analysis?

2. In what way can the clinical challenge of pathology co-occurrence be computationally translated into a deep learning approach that allows a single model to learn to indicate and classify multiple conditions simultaneously?

3. Can approaches that handle the coexistence of multiple pathologies indeed offer a superior balance between diagnostic accuracy and computational efficiency when compared to strategies that treat pathologies independently or as exclusive combinations?

1.3 Contributions

The main contributions of this dissertation are:

- ❏ An extensive evaluation of methods for segmenting the lumbar spine. We provide an extensive evaluation of five state-of-the-art architectures (U-Net 3D, V-Net, UNETR, Swin UNETR, U-Net with ResNet-50) and four widely-used loss functions on both **semantic** and **instance segmentation**, establishing the robust performance ceiling that provides the necessary groundwork for our primary investigation.

- ❏ A novel methodology for addressing the multi-diagnosis problem as a semantic segmentation task. We propose and evaluate three strategies: (i) **binary class segmentation**, a baseline that treats each pathology independently; (ii) **multi-class segmentation**, which maps 70 possible disease combinations to unique, exclusive classes; and (iii) **multi-label segmentation**, an approach that uses dedicated binary channels for each disease to explicitly model the coexistence of multiple diagnoses at the same location.

- ❏ A systematic benchmark demonstrating that our proposed multi-label segmentation strategy offers a highly effective trade-off between accuracy and computational efficiency. Our results, derived from over 200 training pipelines, indicate that this approach achieves performance statistically comparable to the resource-intensive binary class strategy, thus establishing a practical and efficient guideline for future research and clinical applications.

1.4 Methodological Approach

The methodology of this dissertation was structured into two sequential and complementary phases, designed to build a robust foundation before addressing the core research problem. This progressive approach ensures that the conclusions regarding multi-pathology segmentation are based on a thorough understanding of the most effective tools for the fundamental task of anatomical lumbar spine segmentation.

Phase 1: Foundational Analysis of Anatomical Segmentation (Article 1) The starting point of this research, detailed in the first article of this collection, was to establish a rigorous benchmark for the task of anatomical segmentation. Before investigating the

complex task of identifying pathologies, it was essential to determine which deep learning architectures and loss functions were most effective for the fundamental task of delineating the structures of interest: the vertebrae and intervertebral discs. In this phase, we conducted a comprehensive empirical study comparing five state-of-the-art architectures (U-Net 3D, V-Net, U-Net with ResNet-50, UNETR, and Swin UNETR) with four distinct loss functions (Dice, Generalized Dice, Dice + CE, and Dice + Focal). The evaluation was conducted on two critical problem formulations:

- ❏ Semantic Segmentation: Classifying each voxel simply as vertebra, disc, or background.
- ❏ Instance Segmentation: A more challenging task that requires the individual identification of each anatomical structure (e.g., distinguishing the L4 vertebra from L5).

The results of this phase, obtained from the public SPIDER dataset, not only identified the most robust models (V-Net and Swin UNETR) but also revealed the limitations of certain combinations, providing a crucial technical and empirical foundation for the subsequent phase of the research.

Phase 2: Investigation of Strategies for Multi-Pathology Segmentation (Article 2) Building on the insights from Phase 1, the second and primary phase of this research, presented in the second article, directly addresses the dissertation's central question: the development of a unified method for the segmentation of multiple, co-occurring pathologies. To investigate this problem, we proposed and systematically evaluated three conceptually distinct segmentation strategies, specifically designed to handle diagnostic overlap:

- ❏ Binary Class Segmentation: A baseline approach where a separate model is trained independently for each individual pathology.
- ❏ Multi-Class Segmentation: A strategy that reframes the problem by treating each unique combination of diseases (70 total combinations) as an exclusive, non-overlapping class.
- ❏ Multi-Label Segmentation: The central approach of our hypothesis, which uses a single model with multiple output channels, allowing for the prediction of several overlapping pathologies at the same location.

This comparative evaluation was performed using the most promising architectures and loss functions identified in Phase 1, applied to the more complex challenge of multi-pathology diagnosis on the same SPIDER dataset. The goal of this phase was to determine which strategy offers the best trade-off between diagnostic accuracy and computational efficiency, thereby validating our hypothesis that the multi-label approach is best suited to model the clinical reality of pathological comorbidities in the lumbar spine.

1.5 Organization of the Work

This dissertation is structured as a collection of articles, following the guidelines of the Postgraduate Program in Computer Science (PPGCC) at the Federal University of São Carlos (UFSCar). Following this introduction, the work is organized into two central chapters, each corresponding to a scientific paper, followed by the conclusion.

It is relevant to highlight that the central chapters of this dissertation are structured as scientific articles, validating the contributions of this work through peer review in relevant conferences. The article that constitutes **Chapter 3** has been accepted for publication in the 28th Iberoamerican Congress on Pattern Recognition (CIARP 2025). The main paper, presented in **Chapter 4**, has been accepted for publication in the 21th International Conference on Computer Vision Theory and Applications (VISAPP 2026).

Chapter 2 provides the theoretical background for this dissertation, covering the basic concepts of Deep Learning and medical image segmentation necessary to understand the subsequent chapters.

Chapter 3 presents the first article, which establishes the methodological foundation for this work.

Chapter 4 contains the second article, which represents the main contribution of this dissertation by directly addressing the challenge of multi-pathology segmentation.

Finally, **Chapter 5** presents the general conclusions of the work, synthesizing the results from both articles and discussing the implications for future research in the field.

Chapter 2

Basic Concepts

This chapter presents the basic concepts necessary for the development of this research, specifically addressing the principles of Magnetic Resonance Imaging (MRI), Image Segmentation and the Deep Learning architectures applied to medical image analysis. This content is directly established to support the main objective of this dissertation, which is to investigate and validate a unified framework for the precise spatial indication and radiological classification of multiple co-occurring pathologies in the lumbar spine. Regarding the general methodology of this work, this chapter is inserted as the theoretical foundation that underpins both the Foundational Analysis of Anatomical Segmentation (Phase 1) and the Investigation of Strategies for Multi-Pathology Segmentation (Phase 2), providing the essential technical background for the experiments detailed in the subsequent chapters.

2.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) has a wide range of applications in healthcare, both in clinical and research environments, for visualizing anatomical structures and evaluating or diagnosing human body tissues. It is based on an advanced imaging technique that utilizes magnetic fields and radiofrequency waves to generate detailed images of the internal structures of the human body, as detailed in (SERAI, 2022). Unlike other imaging modalities such as radiography or Computed Tomography (CT), which use ionizing radiation, as cited in (LEPAGE; GORE, 2004), MRI is based on the physics of atomic nuclei and their interactions with magnetic fields. This offers excellent soft tissue contrast, allowing for the differentiation between various tissue types, such as muscle, fat, organs, and blood vessels. This method is particularly useful for visualizing tissues such as

the brain, muscles, and the spinal column, offering high resolution and contrast between different tissue types.

MRIs can be T1-weighted or T2-weighted, which correspond to longitudinal and transverse relaxation times, respectively, or proton density, with each offering distinct information regarding tissue properties. As mentioned in (BIDHULT et al., 2016), T1-weighted images provide good anatomical contrast and are useful for highlighting normal anatomy, whereas T2-weighted images are frequently used to identify anomalies, areas of edema, or inflammation due to the higher contrast between different tissue structures.

Figure 4 – T1 and T2-weighted Sagittal Spine MRIs

Source: SPIDER Dataset, GRAAF et al.

MRIs can be acquired in different anatomical planes, including axial, sagittal, and coronal, providing a comprehensive view of anatomy and pathology. Axial images divide the body into horizontal slices providing cross-sections; sagittal images offer a lateral perspective, dividing the body into left and right halves; and coronal images allow for frontal visualization, dividing the body into anterior and posterior parts, as mentioned in (SERAI, 2022). The combination of these planes allows for a complete evaluation of the spine, including the visualization of vertebral bodies, intervertebral discs, the spinal cord, and surrounding structures. Figure 4 presents a spine MRI in the sagittal plane, weighted in T1 and T2 and Figure 5 show a MRIs of dorsal spine sagittal, coronal, and axial weighted in T2.

For the spine, MRI is a tool that can be highly useful for diagnosing a variety of conditions, including disc herniation, spinal stenosis, tumors, infections, and degenerative diseases. As cited in (HUTCHINS et al., 2022), MRI is superior to Computed Tomography

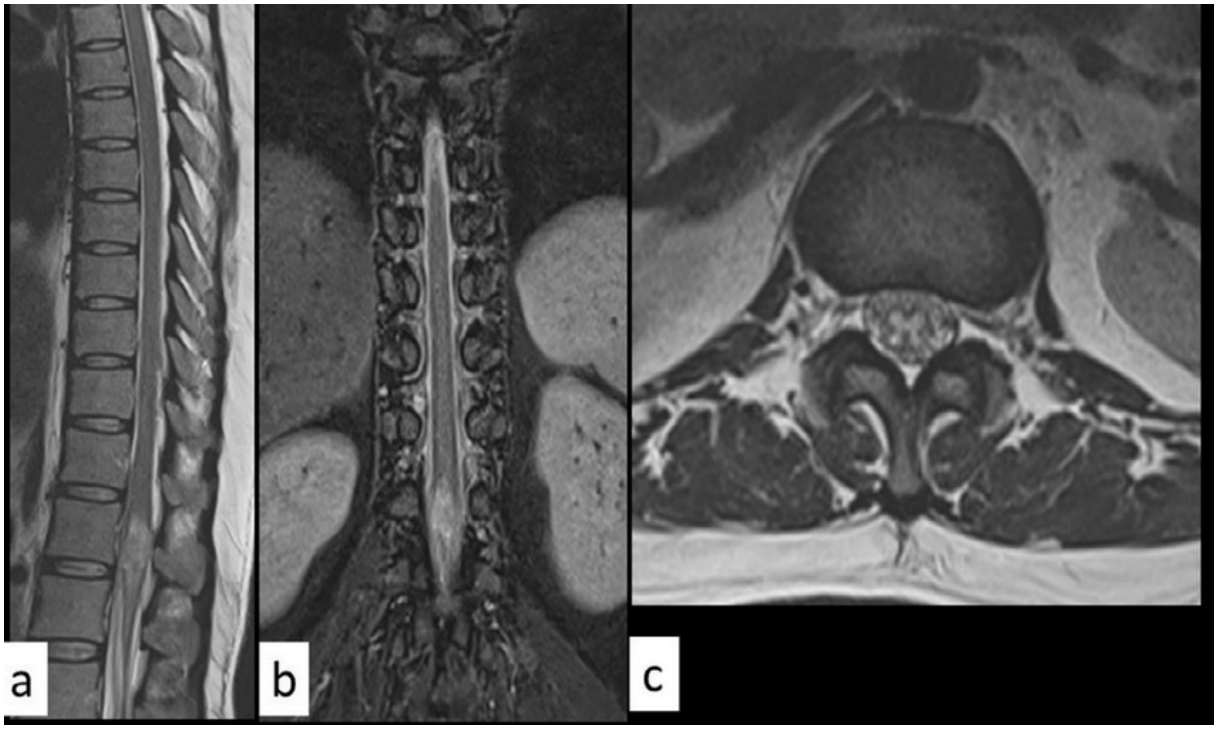


Figure 5 – MRI of dorsal spine sagittal (a), coronal (b), and axial (c). BHOI et al.

in evaluating certain spinal anomalies, as it allows for better visualization of soft tissues that are frequently involved in nerve root compression. Furthermore, MRI does not expose patients to radiation, making it a safer option. The ability of MRI to visualize the spinal cord and nerve roots in detail makes it particularly useful for evaluating spinal cord compression and other neurological conditions.

2.2 Image Segmentation

Image segmentation is a fundamental technique in computer vision, playing an essential role in the analysis, interpretation, and understanding of visual data. This process consists of partitioning an image into multiple segments or meaningful regions by grouping pixels with similar visual characteristics or those that correspond to objects and structures of interest within a real-world domain (MINAEE et al., 2021).

Formally, segmentation can be modeled as a function $S : I \rightarrow L$, where I represents the image space and L is a finite set of labels. For each pixel $p \in I$, the function assigns a label $l \in L$, such that pixels sharing the same label possess common visual or semantic properties.

In the context of medical imaging, segmentation is a critical step for computer-aided diagnosis, surgical planning, and radiotherapy (DESPOTOVIĆ; GOOSSENS; PHILIPS, 2015). In particular, the segmentation of spinal Magnetic Resonance Imaging (MRI) presents significant challenges due to anatomical complexity, variability between acquisition protocols, and low contrast among soft tissues (KOLARIK et al., 2019). The

automation of this process, driven by 3D Convolutional Neural Networks (CNNs), has demonstrated remarkable advancements in precision and clinical efficiency, allowing for the assertive identification of vertebral bodies and intervertebral discs with minimal human intervention.

Current literature classifies segmentation approaches into three main categories, based on the level of granularity and the distinction of instances: semantic segmentation, instance segmentation, and panoptic segmentation.

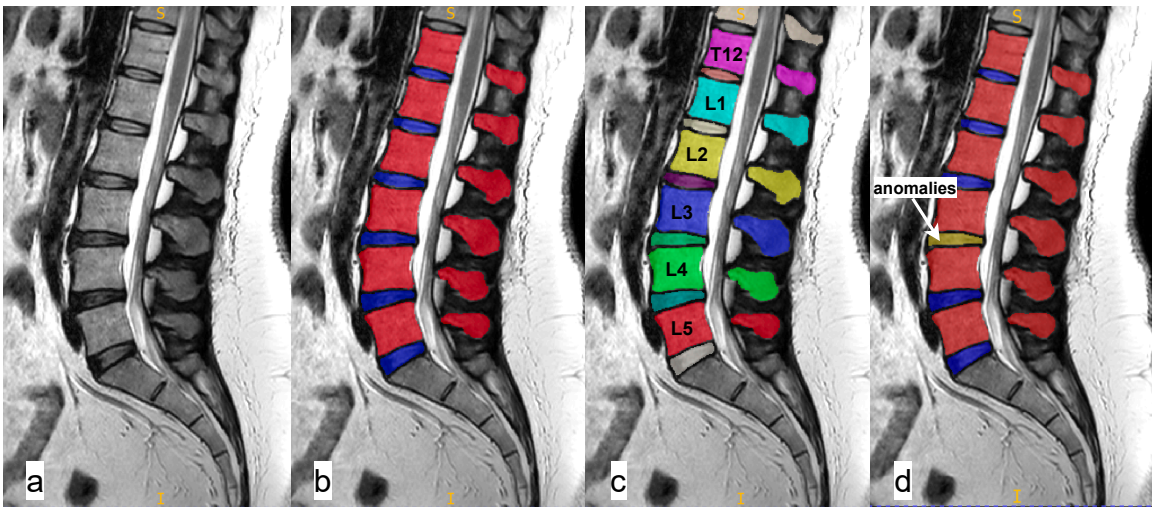


Figure 6 – MRI (a), Semantic Segmentation (b), Instance Segmentation (c) and Panoptic Segmentation (d).

In semantic segmentation, as shown in Figure 6(b), the objective is to assign a class label to every pixel in the image, without distinguishing between different instances of the same category (TAGHANAKI et al., 2021). The result is a dense map where semantically coherent regions (e.g., "intervertebral disc" or "vertebra") are identified, reflecting the global context of the scene.

In the medical field, this approach is widely used to delineate organs and lesions. Encoder-decoder architectures, such as U-Net and its variants, have established themselves as the gold standard for this task in MRI, due to their ability to preserve spatial details through skip connections (ANBUDEVI; SUGANTHI, 2022). However, their main limitation lies in the inability to separate adjacent objects of the same class, treating them as a single connected region.

Instance segmentation, illustrated in Figure 6(c), extends the previous concept by requiring not only the correct classification of pixels but also the individualization of each distinct object (HAFIZ; BHAT, 2020). Unlike the semantic approach, where all vertebrae would receive the same label, instance segmentation assigns a unique identifier to each vertebra (e.g., L1, L2, L3), integrating object detection tasks with pixel-level segmentation.

This distinction is crucial in biomedical scenarios where the count and individual morphology of structures are relevant for diagnosis (LIU et al., 2020). Modern methods frequently employ proposal-based approaches, such as Mask R-CNN, or proposal-free methods that group pixel embeddings to delineate individual instances.

Panoptic segmentation, represented in Figure 6(d), proposes a unification of the previous tasks, providing a holistic understanding of the scene. According to (ELHARROUSS et al., 2021), this approach assigns to each pixel both a class label and an instance identifier, encompassing both countable classes ("things", such as cells or vertebrae) and amorphous or background regions ("stuff", such as adipose tissue or image background).

The panoptic segmentation utilizes metrics such as Panoptic Quality (PQ) to jointly evaluate recognition quality and boundary precision. In the context of spinal imaging, this approach is particularly promising for providing a complete map that includes both tissue classification (semantic) and the precise identification of each vertebral level (instance), handling scenarios with ambiguous boundaries and overlapping objects more effectively.

2.3 Deep Learning

Deep Learning (DL) constitutes a class of machine learning techniques that utilizes artificial neural networks with multiple processing layers to model complex data and high-level abstractions. As discussed by LeCun, Bengio e Hinton (2015), this approach has proven particularly effective in analyzing grid-structured data, such as medical images.

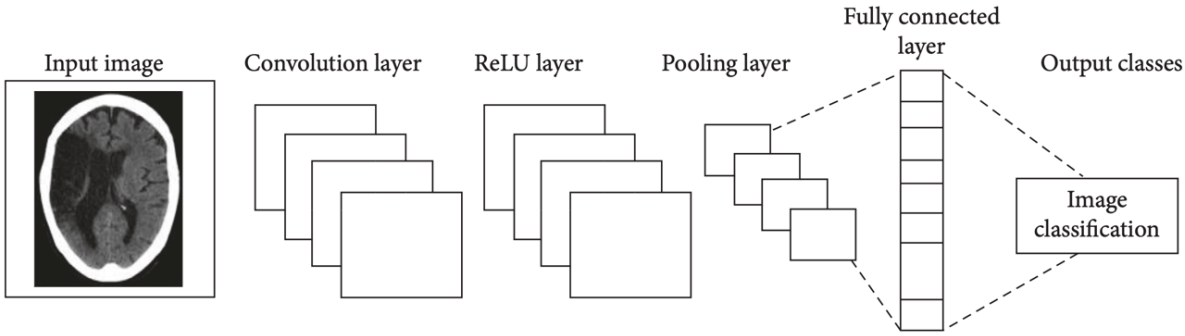


Figure 7 – Basic structure of a CNN for classification (TCHITO et al., 2021).

Among DL architectures, Convolutional Neural Networks (CNNs) have established themselves as one of the most prominent approaches for computer vision tasks. A CNN is composed of a sequence of layers that transform the input volume into an output through differentiable operations. The basic structure of a CNN for classification problem as shown in Figure 7, typically includes three main types of layers:

- **Convolutional Layers**: These constitute the foundation of the network. In these layers, a set of filters (or kernels) is applied over the input via dot product operations, producing feature maps that highlight local patterns such as edges and textures.

Following convolution, a non-linear activation function, such as the Rectified Linear Unit (ReLU), is commonly applied to enable the network to learn complex functions.

- **Pooling Layers**: These layers serve to reduce the spatial dimensionality of the feature maps, thereby decreasing the number of parameters and computational cost, in addition to introducing invariance to small translations. The most common operations are max pooling and average pooling.

- **Fully Connected Layers**: In the final layers of classification networks, the extracted features are flattened and passed through neurons connected to all activations of the previous layer, consolidating the information for the final decision.

In the field of medical imaging, the use of deep neural networks has surpassed classical methods and become the dominant approach for interpreting examinations such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) (SHEN; WU; SUK, 2017). However, while the standard CNN structure (Figure 7) is highly effective for image-level classification, it is insufficient for segmentation tasks that require dense, pixel-wise predictions. To address this, Encoder-Decoder architectures have become the most popular choice (MINAEE et al., 2021), as exemplified by the U-Net architecture shown in Figure 8.

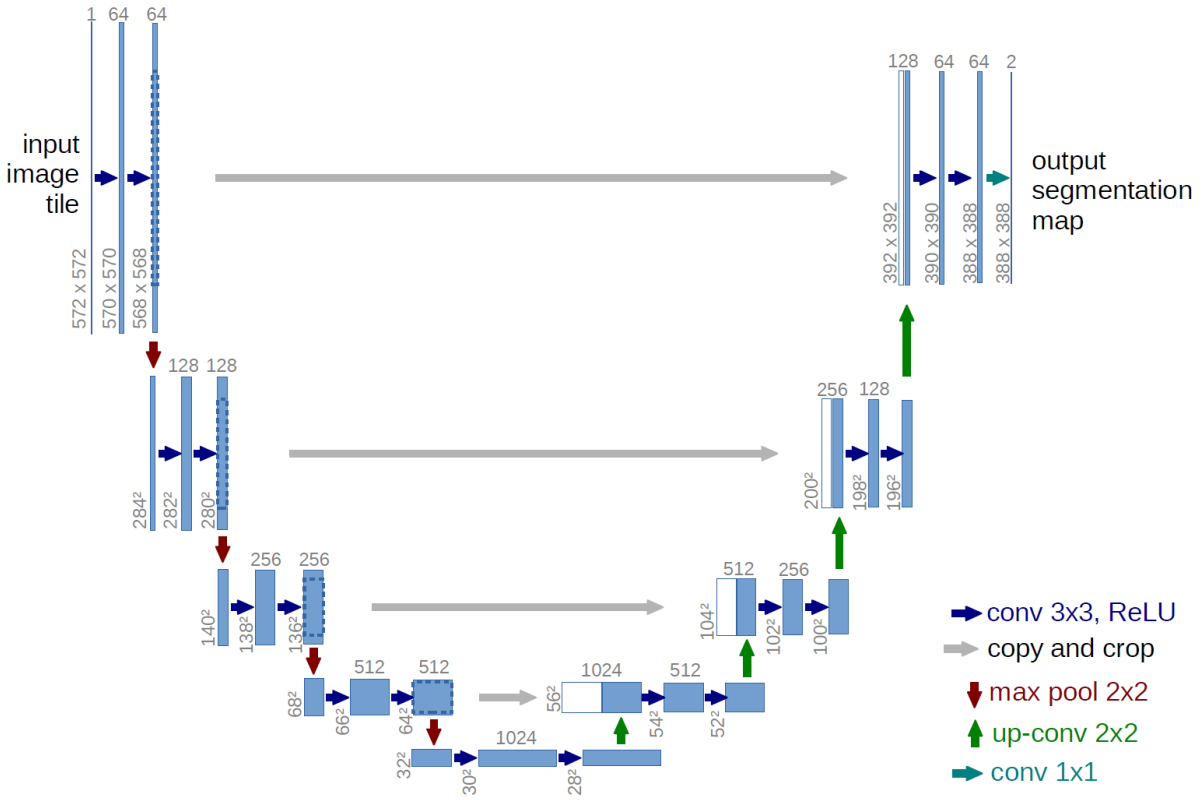


Figure 8 – Basic structure of a U-Net for segmentation. (RONNEBERGER; FISCHER; BROX, 2015).

Encoder-Decoder models are designed with two main pathways. The **Encoder** (contracting path) progressively reduces the spatial resolution of the data while increasing the depth of feature channels, capturing the global semantic context of the image (MINAEE et al., 2021). While essential for understanding "what" is in the image, this downsampling process tends to discard fine spatial details and edge information.

To mitigate this loss, the **Decoder** (expanding path) is responsible for recovering the original spatial resolution through upsampling operations, reconstructing the dense segmentation map. However, simple reconstruction from low-resolution features is often insufficient to delineate precise boundaries, a critical requirement in medical diagnostics (ZHOU et al., 2020).

In this context, the U-Net architecture (RONNEBERGER; FISCHER; BROX, 2015) introduced a fundamental concept that became the basis for most modern medical segmentation models: **skip connections**. The U-Net connects the contracting path to the expanding path by copying high-resolution feature maps from the encoder and concatenating them directly to the corresponding maps in the decoder.

This "copy and concatenate" strategy allows the network to combine deep semantic information (from the bottom of the "U") with texture and localization details preserved in the early layers of the encoder. The popularity of "U-Net-like" architectures is due to their robust ability to recover fine anatomical details, even in scenarios with limited datasets (AZAD et al., 2022). More recent variations, such as U-Net++ (ZHOU et al., 2019), propose redesigning these connections with nested and dense pathways to bridge the semantic gap between encoder and decoder features.

Despite the success of CNNs, they possess an intrinsic limitation: the local receptive field of convolution operations makes it difficult to model long-range dependencies within the image. To address this issue, recent architectures have integrated Transformer-based attention mechanisms into segmentation models.

In these hybrid approaches, such as UNETR (HATAMIZADEH et al., 2022) and Swin UNETR (HATAMIZADEH et al., 2021) (discussed in Chapters 3 and 4), the Transformer often assumes the role of the encoder. By processing the image as a sequence of patches, the Transformer is capable of learning global relationships between different anatomical regions from the very first layers. These global representations are then fed into a decoder (typically CNN-based, U-Net style) to reconstruct the segmentation with local precision, uniting the best of Transformers global modeling with the spatial accuracy of convolutions.

2.4 Evaluation Metrics

The quantitative evaluation of medical image segmentation methods is a critical step in validating the clinical efficacy of proposed models. In a supervised paradigm, the fundamental objective is to measure the degree of agreement between the segmentation

predicted by the computational model and the reference segmentation (ground truth), usually annotated by human experts (MüLLER; SOTO-REY; KRAMER, 2022).

Given the anatomical complexity and variability of lumbar spine pathologies, no single metric is capable of capturing all nuances of model performance. Therefore, the literature recommends the use of a set of complementary metrics that evaluate different aspects, such as volumetric overlap, contour precision, and instance detection (MINAEE et al., 2021).

(a) Ground truth mask (b) Predict mask (c) Overlap masks

Figure 9 – Visual representation of the confusion matrix components in segmentation. Adapted from (KUKIL, 2022).

The basis for most evaluation metrics is the confusion matrix, which categorizes predictions into four possible outcomes. In the context of semantic segmentation (pixel-based), each image unit (pixel in 2D or voxel in 3D) is treated as an independent classification instance (TAHA; HANBURY, 2015). Considering a class of interest (as shown in Figure 9) relative to the background, we define:

- ❏ **True Positive (TP):** Image unit correctly classified as belonging to the structure of interest.
- ❏ **False Positive (FP):** Image unit incorrectly classified as the structure of interest (over-segmentation).
- ❏ **True Negative (TN):** Image unit correctly classified as background.
- ❏ **False Negative (FN):** Image unit belonging to the structure of interest that were incorrectly classified as background (under-segmentation).

From these components, fundamental metrics are derived to evaluate classifier performance at the voxel level:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Although **Accuracy** is an intuitive metric, it is often inadequate for medical segmentation due to severe class imbalance, where the background occupies the vast majority

of the volume. This imbalance can inflate the result even in models that fail to segment the lesion (MüLLER; SOTO-REY; KRAMER, 2022). To mitigate this, **Precision** and **Recall** (also known as Sensitivity) are used:

$$\textbf{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Precision quantifies the purity of the prediction, penalizing over-segmentation, while Recall measures the model's ability to detect the entire structure, penalizing under-segmentation.

For segmentation tasks, overlap-based metrics are considered the gold standard. The **Dice Similarity Coefficient (DSC)** is the most widely adopted metric in the medical imaging literature (TAHA; HANBURY, 2015). It is defined as the harmonic mean of Precision and Recall, offering a robust measure of spatial agreement:

$$\textbf{DSC} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{4}$$

The DSC value ranges from 0 (no overlap) to 1 (perfect overlap). A related metric is the *Jaccard Index*, or **Intersection over Union (IoU)**, defined as:

$$\textbf{IoU} = \frac{TP}{TP + FP + FN} \tag{5}$$

Both DSC and IoU are particularly useful in scenarios with class imbalance, as they ignore the TN (background) component, focusing exclusively on the region of interest.

In multi-class or multi-label scenarios, where multiple pathologies or anatomical structures are evaluated simultaneously, it is common to use the *Macro-average* strategy. In this approach, the metric (e.g., DSC) is calculated individually for each class, and then the arithmetic mean is computed. This ensures that classes with smaller volumes (e.g., a small hernia) have the same weight in the final evaluation as voluminous classes (e.g., a vertebral body), preventing performance on larger structures from masking failures in critical pathologies.

While pixel-based metrics evaluate the quality of volumetric delineation and example-based metrics focus on the correct detection of entities and the consistency of predictions in multi-diagnostic scenarios.

In the specific context of *multi-label learning*, where a single voxel can be associated with multiple pathologies simultaneously (concurrent diagnosis), example-based metrics are essential. Unlike the traditional approach that aggregates results by class, example-based metrics evaluate prediction quality for each data instance individually (in this work, each voxel or region of interest) and then calculate the average over all instances (ZHANG; ZHOU, 2013). For instance, calculates the Dice similarity between the set of predicted

labels and the set of ground truth labels for each voxel, penalizing models that fail to capture the exact co-occurrence of diseases at that specific location. This evaluation granularity allows validating whether the model understands the semantic complexity of pathology overlap, going beyond simple independent volumetric segmentation.

Chapter 3

A Comprehensive Evaluation for Lumbar Spine Segmentation

This chapter presents the first article that constitutes this dissertation. Its content corresponds to the first phase of the methodology adopted in this research, focused on establishing a robust empirical foundation for the analysis of lumbar spine images.

The following paper has been accepted for publication and will be presented at the 28th Iberoamerican Congress on Pattern Recognition (CIARP 2025).

- **Title:** A Comprehensive Evaluation of Deep Learning Architectures and Loss Functions for Lumbar Spine Segmentation in MRI.
- **Authors:** Claudio Leite, Samuel Felipe dos Santos, and Jurandy Almeida.
- **Status:** Accepted for publication in the 28th Iberoamerican Congress on Pattern Recognition - CIARP 2025 (Qualis=A4).

Before addressing the main objective of this dissertation—the segmentation of multiple, co-occurring pathologies—it was crucial to determine which deep learning architectures and loss functions are most effective for the preceding task of anatomical segmentation. This chapter, therefore, presents a comprehensive comparative study of five state-of-the-art architectures and four distinct loss functions, evaluated on the tasks of semantic and instance segmentation of vertebrae and intervertebral discs . The results presented here not only establish a performance benchmark but also provide the technical justification for the selection of methods used in the following chapter, ensuring that the investigation into pathologies is built upon a solid and well-validated methodological foundation.

The remainder of this chapter is organized as follows.

Section 3.1 presents the abstract. Section 3.2 provides the introduction. Section 3.3 discusses the related work. Section 3.4 describes the methodology. Section 3.5 reports and analyzes the experimental results. Finally, Section 3.6 offers the conclusions.

3.1 Abstract

The aging global population has led to an increased prevalence of spinal pathologies, making the accurate analysis of lumbar spine Magnetic Resonance Imaging (MRI) a critical clinical task. Deep learning-based segmentation of vertebrae and intervertebral discs offers a promising avenue for automating and improving this analysis. This paper presents a comprehensive empirical study comparing the performance of five state-of-the-art deep neural network architectures (U-Net 3D, V-Net, UNETR, Swin UNETR, and U-Net with ResNet-50) in combination with four widely-used loss functions (Dice, Generalized Dice, Dice + CE, and Dice + Focal). We evaluate these 20 configurations on the public SPIDER dataset across two distinct and clinically relevant tasks: semantic segmentation (classifying voxels as vertebra, disc, or background) and instance segmentation (additionally identifying individual anatomical structures). Our findings indicate that while most models perform well on semantic segmentation, with V-Net and Swin UNETR showing superior performance, instance segmentation proves to be a significantly more challenging task that better differentiates model capabilities. In this more complex scenario, V-Net and Swin UNETR again demonstrate the most robust and accurate results. This detailed analysis of results and trade-offs provides a valuable guide for the selection of appropriate models and loss functions for lumbar spine segmentation, clarifying the strengths and weaknesses of different approaches for this important clinical problem.

3.2 Introduction

The aging of the global population presents one of the greatest challenges of the modern healthcare systems. Data from the United Nations Department of Economic and Social Affairs (MCNICOLL, 2002) reveal that the number of people aged 60 and over nearly tripled between 1950 and 2000, increasing from 205 million to 606 million. This demographic shift exacerbates the prevalence of age-related conditions, such as spinal pathologies. Comorbidities, such as reduced bone density and an increase in spinal degeneration, that commonly afflict the elderly are partially responsible for low back pain being a leading causes of disability worldwide (FEHLINGS et al., 2015).

Magnetic Resonance Imaging (MRI) of the lumbar spine is a cornerstone for an accurate diagnosis, yet its manual interpretation is time-consuming, subjective, and prone to inter-observer variability. According to van der Graaf et al. (GRAAF et al., 2024),

despite a high rate of appropriate referrals, only a small fraction of these scans effectively contributes to clinical decision-making, highlighting the need for more efficient and robust analysis tools.

In this context, Computer Vision driven by Deep Learning techniques emerges as a promising tool to support the analysis of medical images. These approaches can optimize the interpretation of lumbar spine MRI scans by indicating and assisting in the identification of potential abnormalities. These are promising technologies that have vast array of applications, such as surgical planning, disease progression monitoring, risk modeling and stratification, personalized screening, and direct diagnostic support (CASTIGLIONI et al., 2021). A fundamental step towards automating this analysis is the precise segmentation of anatomical structures of interest, such as vertebrae and intervertebral discs. The accurate delineation of these structures serves as the foundation upon which subsequent analyses can be reliably built.

Despite numerous deep learning models for lumbar spine segmentation been proposed recently, a systematic comparison of different architectures and training strategies remains a critical need in the field. To address this gap, this paper presents a comprehensive empirical study into the segmentation of vertebrae and intervertebral discs from MRI scans on the public SPIDER dataset (GRAAF et al., 2024). The main contributions of this work are:

- ❏ An extensive evaluation of methods for segmenting the lumbar spine, including five state-of-the-art architectures (U-Net 3D, V-Net, UNETR, Swin UNETR, U-Net with ResNet-50) and four widely-used loss functions (Dice, Generalized Dice, Dice + CE, and Dice + Focal), totaling 20 model-loss combinations.

- ❏ An evaluation of the methods on two critical formulations of the segmentation problem: (i) **semantic segmentation**, which aims to classify each voxel as either vertebra or disc; and (ii) **instance segmentation**, which, in addition to classifying, seeks to individualize each anatomical structure, offering a nuanced view of their capabilities.

- ❏ A detailed analysis of results and trade-offs, which serves as a valuable guide for the scientific community, clarifying the strengths and weaknesses of different approaches for this relevant clinical problem.

The remainder of this paper is organized as follows. Section 3.3 presents related work. Section 3.4 describes our methodology. Section 3.5 reports and analyzes our experimental results. Finally, Section 3.6 offers our conclusions.

3.3 Related Work

The automated segmentation of spinal structures in MRI is a critical task for clinical diagnostics and has become a major focus of deep learning research. Methodologies based on Convolutional Neural Networks (CNNs) have proven particularly effective, with the U-Net architecture emerging as a dominant paradigm. A survey by Andrew et al. (ANDREW et al., 2020), for instance, compared various deep learning approaches and concluded that U-Net is highly promising due to its ability to produce accurate segmentations even with limited training data.

Building on this foundation, numerous studies have proposed optimizations to the U-Net framework. For example, Wang et al. (WANG et al., 2022) introduced an Attention U-Net variant that integrates residual modules and a multi-level attention mechanism to refine feature fusion, employing a hybrid Dice and Cross-Entropy loss to stabilize training. Similarly, Wang et al. (WANG; XIAO; TAN, 2023) explored U-Net++, an architecture that redesigns skip connections to bridge the semantic gap between the encoder and decoder paths, thereby improving segmentation quality. Notably, the very work that introduced the SPIDER dataset used in our study, van der Graaf et al. (GRAAF et al., 2024), established strong baselines with a U-Net-like iterative model (IIS) and the popular nnU-Net framework.

Beyond purely end-to-end deep learning, hybrid strategies have also been explored. Hille et al. (HILLE et al., 2018) presented a semi-automatic approach using a level-set method guided by appearance-based probability maps. While not an end-to-end solution, this work underscores the importance of robustness against challenges common in clinical practice, such as low image quality and high anisotropy.

Despite these valuable advancements, the existing literature presents two key limitations. First, to our knowledge, no single study has conducted an extensive empirical benchmark comparing the latest Transformer-based models (UNETR, Swin UNETR) against established CNN architectures (U-Net 3D, V-Net, U-Net with ResNet-50) using a standardized dataset and a varied set of loss functions. Second, performance analysis is often restricted to a single problem formulation, typically semantic segmentation.

Our study directly addresses these gaps. We provide a comprehensive benchmark of 20 model-loss combinations and, crucially, evaluate them on both semantic and instance segmentation tasks. This extensive evaluation provides a more complete understanding of each model's capabilities, offering the scientific community a guide to the trade-offs inherent in this clinically vital problem.

3.4 Methodology

This section details the comprehensive methodology employed in our comparative study. We describe the selected architectures and loss functions, the dataset and pre-processing steps, the evaluation tasks, the performance metrics, and the implementation details.

3.4.1 Model Architectures

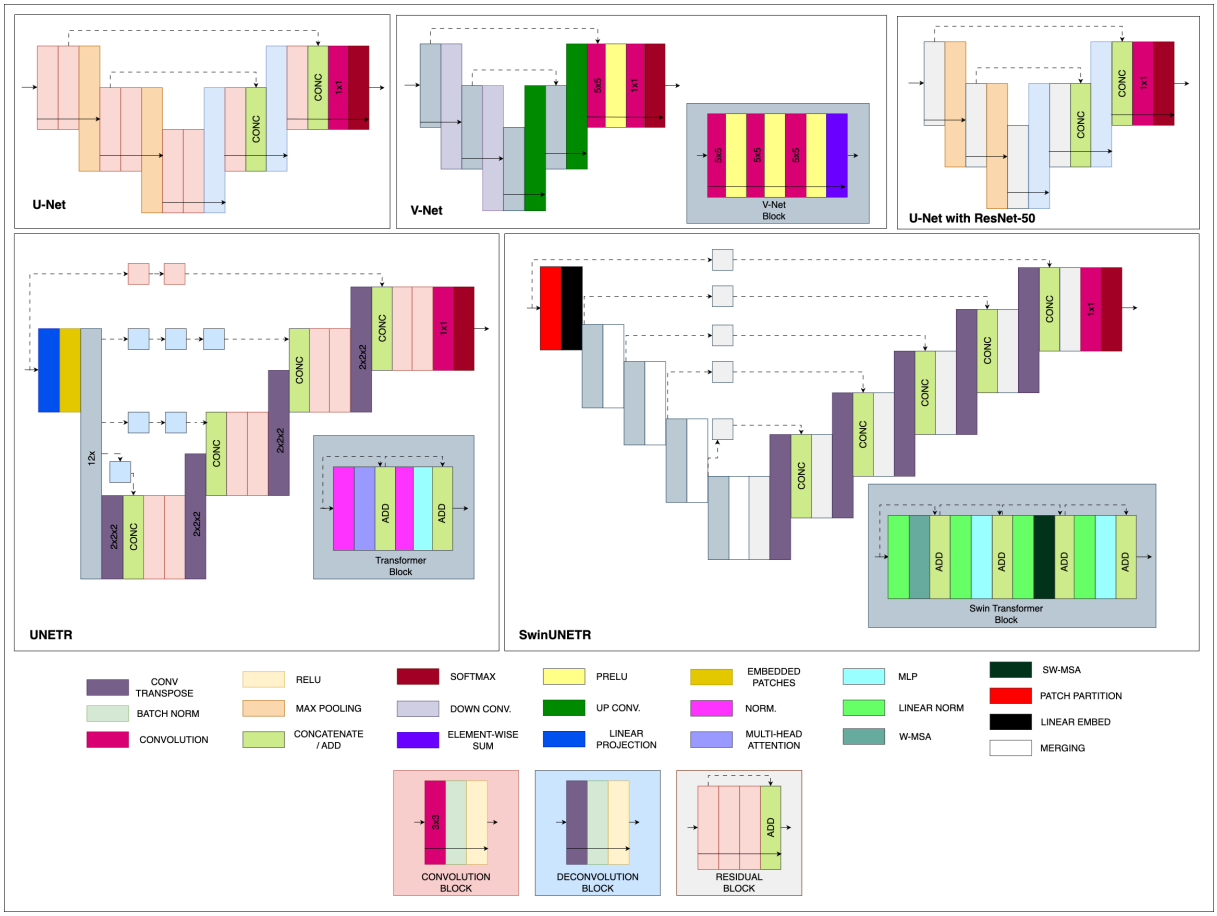


Figure 10 – Summary of the U-Net based architectures compared in this study.

Our comparative analysis includes five deep learning architectures, shown in Figure 10. They were selected for their prominence and diversity in medical image segmentation.

U-Net 3D. The U-Net (RONNEBERGER; FISCHER; BROX, 2015) architecture established a paradigm for medical image segmentation. The U-Net 3D (KERFOOT et al., 2018) extends this paradigm to handle volumetric data. Its architecture is characterized by a symmetric encoder-decoder structure. The encoder progressively downsamples the input volume through a series of convolutional and pooling layers to capture hierarchical contextual features. The decoder systematically upsamples these feature maps while merging them with high-resolution features from the corresponding encoder stage via skip

connections. These connections are crucial as they allow the network to combine high-level contextual information with fine-grained spatial details, enabling precise localization and delineation of anatomical structures.

V-Net. The V-Net (MILLETARI; NAVAB; AHMADI, 2016) is another prominent encoder-decoder network designed for 3D medical image segmentation. Its key contribution was the integration of residual connections within each stage of the network. These connections mitigate the vanishing gradient problem, enabling the training of deeper architectures capable of learning more complex feature representations.

ResNet-50. To leverage knowledge from large-scale natural image datasets, our study includes a hybrid model where the standard U-Net encoder is replaced by a ResNet-50 (WANG et al., 2023b) architecture. ResNet (Residual Network) architectures introduced the concept of residual learning, which allows for the effective training of networks that are substantially deeper than their predecessors.

UNETR. Breaking from purely convolutional designs, the UNETR (U-Net Transformer) (HATAMIZADEH et al., 2022) reformulates segmentation by integrating Transformers. It uses a Vision Transformer (ViT) as its encoder to model long-range global dependencies, a known limitation of CNNs. The input volume is divided into a sequence of patches, and the Transformer encoder learns a global contextual representation of the entire volume, which is then upsampled by a convolutional decoder.

Swin UNETR. The Swin UNETR (HATAMIZADEH et al., 2021) is an evolution of UNETR that uses the more efficient Swin Transformer as its encoder. Unlike ViT, which computes global self-attention, the Swin Transformer computes attention locally within non-overlapping windows and uses a shifted-window mechanism to learn global features. This significantly reduces computational complexity, making it highly scalable for high-resolution 3D images while combining the global modeling power of Transformers with the proven localization capabilities of a U-Net-style decoder.

3.4.2 Loss Functions

The choice of loss function is critical for optimizing segmentation models. We evaluate four distinct and widely-used loss functions. Let M be the number of classes, N be the number of voxels in the volume, $g_{i,j} \in \{0, 1\}$ be the ground truth label, and $p_{i,j} \in [0, 1]$ be the predicted probability for class i at voxel j.

Dice. Proposed alongside the V-Net (MILLETARI; NAVAB; AHMADI, 2016), the Dice Loss ($\mathcal{L}_{Dice}$) directly optimizes the Dice Similarity Coefficient, making it robust to class imbalance. It is defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_{i=1}^{M} \sum_{j=1}^{N} p_{i,j} g_{i,j}}{\sum_{i=1}^{M} \sum_{j=1}^{N} (p_{i,j} + g_{i,j})}. \tag{6}$$

Generalized Dice. The Generalized Dice Loss (GDL) (SUDRE et al., 2017) is a variant that assigns greater weight to smaller, rarer classes to prevent the model from

Table 1 – Details of MRI studies in the SPIDER dataset.

Hospital	Studies	T1	T2	Voxel Range (min–max) (mm)	Female (%)
UMC	41	39	39	(3.24×0.27×0.47) – (3.34×0.59×0.85)	55
RH1	43	43	37	(0.46×0.46×4.20) – (9.63×1.06×1.06)	58
RH2	44	24	44	(0.46×0.46×4.20) – (5.17×1.00×1.23)	59
OH	90	90	90	(3.15×0.24×0.24) – (3.39×0.83×1.02)	68
Total	**218**	**196**	**210**	(3.15×0.24×0.24) – (9.63×1.06×1.23)	**63**

ignoring them. The weight for each class, w_i, is the inverse of its volume. The loss, $\mathcal{L}_{\text{GDL}}$, is:

$$w_i = \frac{1}{(\sum_{j=1}^{N} g_{i,j})^2}, \tag{7}$$

$$\mathcal{L}_{\text{GDL}} = 1 - 2\frac{\sum_{i=1}^{M} w_i \sum_{j=1}^{N} p_{i,j} g_{i,j}}{\sum_{i=1}^{M} w_i \sum_{j=1}^{N} (p_{i,j} + g_{i,j})}. \tag{8}$$

Dice + CE. We also evaluate a hybrid loss, named Dice + CE Loss, combining Dice Loss (Equation 6) with multi-class Cross-Entropy Loss ($\mathcal{L}_{CE}$), which focuses on voxel-level classification accuracy. The loss, $\mathcal{L}_{CE}$, is defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{M} g_{i,j} \log(p_{i,j}). \tag{9}$$

Dice + Focal. Another relevant hybrid loss, named Dice + Focal Loss, combines Dice Loss (Equation 6) with Focal Loss ($\mathcal{L}_{\text{Focal}}$) (LIN et al., 2017), which modifies Cross-Entropy to focus training on hard-to-classify examples by down-weighting the loss assigned to well-classified examples. $\mathcal{L}_{FL}$ is defined as:

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{M} \alpha (1 - p_{i,j})^{\gamma} g_{i,j} \log(p_{i,j}) \tag{10}$$

where $\gamma > 0$ is a focusing parameter (typically 2) and α is a weighting factor.

3.4.3 Dataset

This study utilizes the public SPIDER (SPine Imaging Diagnostic Extended Resource) dataset (GRAAF et al., 2024), a comprehensive resource for advancing spinal imaging research. 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 multi-center and demographic diversity of SPIDER is fundamental for training and validating robust and generalizable models.

The cornerstone of this work lies in the segmentation annotations provided by SPIDER, which serve as our ground truth. The segmentation masks for all visible vertebrae (excluding the sacrum) and intervertebral discs were generated through a rigorous manual and iterative process, which was conducted by a medical trainee under the supervision

of a medical imaging specialist and an experienced musculoskeletal radiologist. It involved training a segmentation algorithm on a subset of the data, generating preliminary segmentations for the remaining data, and subsequent manual review and correction by experts. This refinement cycle ensures the high quality and precision of the anatomical delineations.

3.4.4 Data Transforms

The intrinsic heterogeneity of the SPIDER dataset, characterized by varying MRI dimensions, resolutions, and alignments, necessitates a robust strategy to provide uniform input for the deep learning models. To address this, we employed a random patch-based sampling strategy directly within the training pipeline. From each full-volume MRI, patches of a uniform spatial dimension of (32, 192, 192) voxels were randomly extracted during training. This patch size was deliberately chosen to meet two criteria: first, to satisfy the architectural constraints of Transformer-based models like UNETR and Swin UNETR, which require input dimensions to be multiples of their patch mechanisms. Second, it establishes a practical trade-off between capturing sufficient anatomical context for the learning task and maintaining computational feasibility. This dynamic sampling approach ensures that all models are trained on consistent input dimensions while being exposed to diverse regions of the anatomy.

3.4.5 Segmentation Tasks

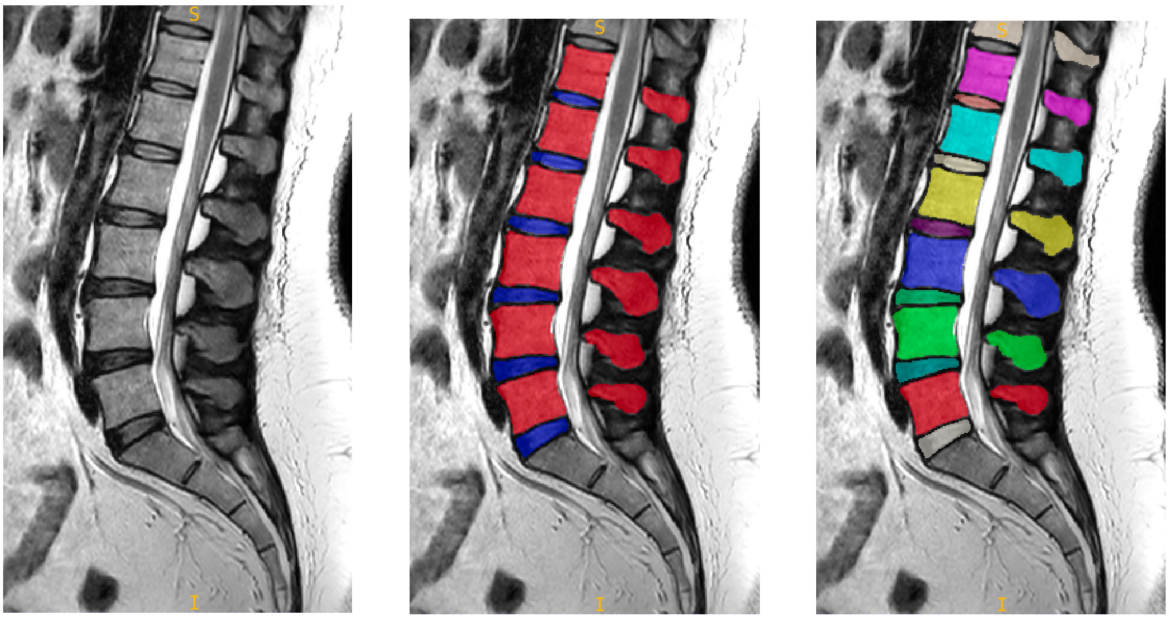
We evaluate all models on two distinct segmentation tasks, shown in Figure 11.

Semantic Segmentation. The first task is semantic segmentation, where the objective is to classify each voxel as one of three classes: vertebra, intervertebral disc, or background. All individual vertebra instances are unified under a single class label, and all disc instances under another.

Instance Segmentation. The second, more complex task is instance segmentation. This requires the individual identification and delineation of each anatomical structure. A unique identifier is assigned to each vertebra (e.g., L1, L2, L3, ...) and each intervertebral disc (e.g., L1-L2, L2-L3, L3-L4, ...), distinguishing different instances of anatomical structures within the same category.

3.4.6 Performance Metrics

To provide a comprehensive evaluation, we use three standard metrics: Precision, Recall, and Dice Similarity Coefficient (DSC). These metrics takes into account the amount of voxels correctly (i.e. True Positives - TP) and incorrectly (i.e. False Positives - FP) predicted as belonging to the class of interest, and the amount of voxels correctly (i.e. True



(a) MRI (b) Semantic Segmentation (c) Instance Segmentation

Figure 11 – Examples of an input image and its corresponding ground truth for the two segmentation tasks.

Negatives - TN) and incorrectly (i.e. False Negatives - FN) predicted as not belonging to it.

Precision is the ratio of voxels correctly predicted as positive to all voxels predicted as positive, penalizing the model for incorrectly classifying negatives as positives. In contrast, Recall is the ratio of voxels correctly predicted as positive to all positive voxels, penalizing the model for incorrectly classifying positives as negatives. Formally, they are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad ; \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{11}$$

The DSC metric is the harmonic mean of the Precision and Recall metrics:

$$\text{DSC} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{12}$$

The final reported score for each metric is the **macro-average**, where the metric is calculated for each class independently and then averaged. This ensures that all classes, regardless of their size, contribute equally to the final score.

3.4.7 Implementation Details

To ensure a fair and reproducible comparison, all experiments followed a standardized protocol. We utilized only the T2-weighted MRI scans, partitioned into a training set of 168 scans (134 for training, 34 for validation) and a hold-out test set of 42 scans. The training pipeline transforms included intensity scaling followed by a random crop to extract patches of a uniform spatial dimension of (32, 192, 192) voxels.

To ensure reproducibility, all training runs were initialized with a fixed random seed of 42. The models were trained for a total of 500 epochs using the Adam optimizer with a learning rate of 1e-4 and a batch size of 8 and the validation process was executed every 2 epochs. In the evaluation on the test set, inference on the full-sized volumes was performed with a sliding window approach.

3.5 Results and Discussion

To evaluate the performance of the investigated approaches, we conducted a battery of experiments covering two distinct segmentation tasks: semantic and instance. Five model architectures (U-Net 3D, V-Net, UNETR, Swin UNETR, and U-Net with ResNet-50) were evaluated in combination with four distinct loss functions (Dice, Generalized Dice, Dice + CE, and Dice + Focal). The performance of each combination was measured using the macro-average of the Precision, Recall, and DSC metrics, ensuring a fair evaluation across classes.

Table 2 – Results for the Semantic Segmentation task, grouped by metric. The table shows scores for different architectures and loss functions.

Metric	**Loss**	**U-Net 3D**	**V-Net**	**ResNet-50**	**UNETR**	**Swin UNETR**
Precision	Dice	0.84	0.89	0.86	0.84	0.90
	Dice+Focal	0.83	0.89	0.85	0.82	0.89
	Gen. Dice	0.81	0.88	0.41	0.79	0.86
	Dice + CE	0.85	0.89	0.85	0.84	0.90
Recall	Dice	0.79	0.87	0.79	0.75	0.85
	Dice+Focal	0.81	0.88	0.82	0.81	0.86
	Gen. Dice	0.75	0.83	0.39	0.73	0.83
	Dice + CE	0.79	0.89	0.85	0.82	0.86
DSC	Dice	0.81	0.87	0.80	0.78	0.86
	Dice+Focal	0.81	0.88	0.83	0.80	0.87
	Gen. Dice	0.77	0.84	0.39	0.75	0.83
	Dice + CE	0.81	0.88	0.84	0.82	0.88

3.5.1 Semantic Segmentation

As shown in Table 2, our results for the semantic segmentation task reveal that most model-loss combinations achieve high performance across all metrics. This indicates that modern deep learning architectures are highly capable of learning the distinctive features of anatomical structures at a class level.

Notably, architectures leveraging residual connections (**V-Net**) or hierarchical Transformers (**Swin UNETR**), demonstrated superior performance, with Precision, Recall, and DSC scores often exceeding 0.83. In contrast, the **UNETR** and **U-Net 3D** models

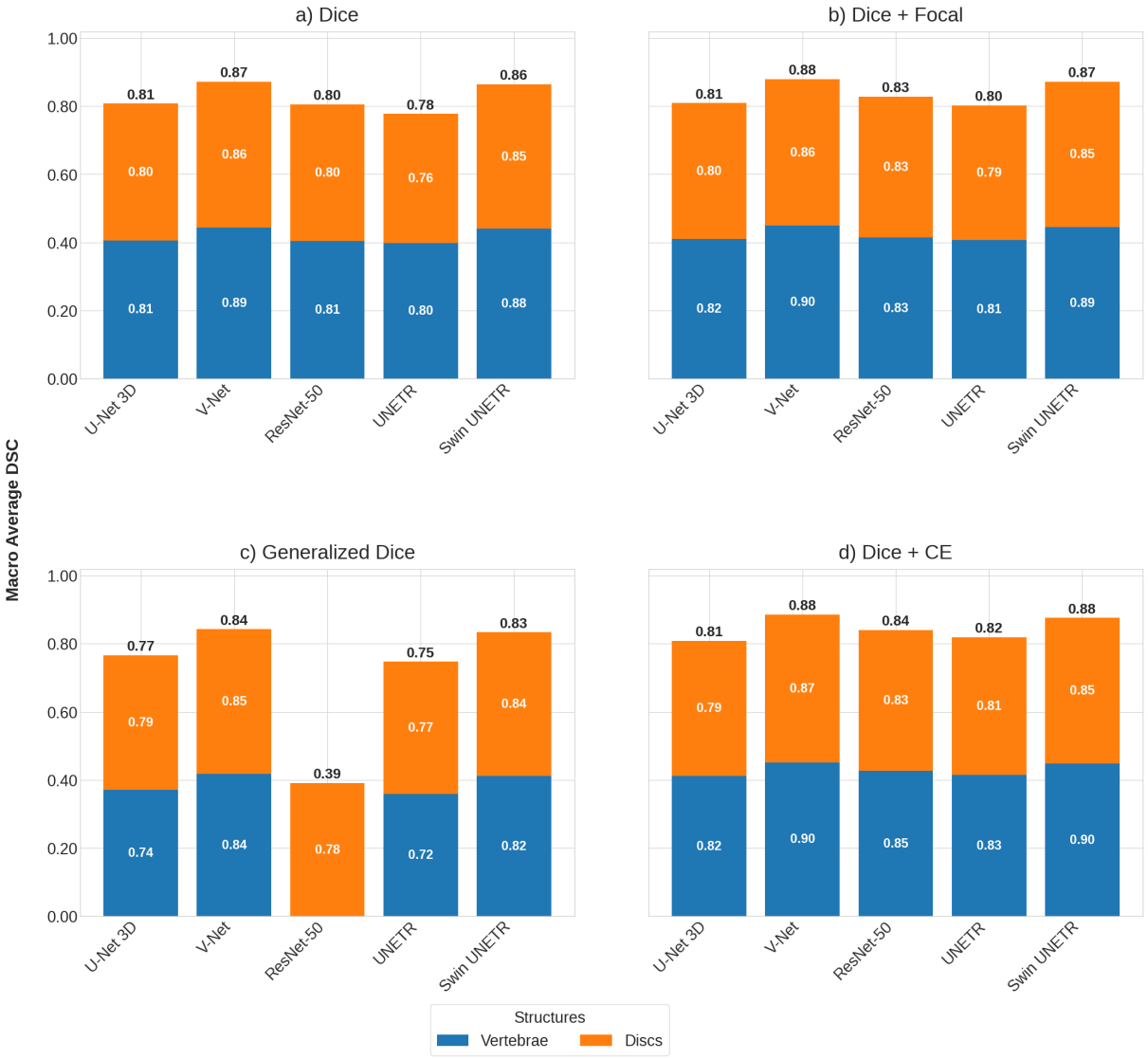


Figure 12 – Results for the semantic segmentation task. The chart displays the macro-average Dice Similarity Coefficient (DSC) for five different models, with each bar segmented by the "Vertebrae" and "Discs" classes.

obtained slightly lower results. The choice of loss function also revealed a clear performance hierarchy. The **Dice** and **Dice + CE** losses consistently provided the most robust and highest scores across the top-performing models. Conversely, the **Generalized Dice** loss consistently yielded the weakest performance for all architectures, indicating its lower effectiveness for this task.

A per-class breakdown of the DSC scores (Figure 12) confirms the balanced, high performance of **V-Net** and **Swin UNETR**, while also revealing a critical failure. Notably, the combination of **U-Net with ResNet-50** and the **Generalized Dice** loss led to an almost complete failure in segmenting the "Vertebrae" class. We hypothesize this stems from a conflict between the GDL's heavy penalization of errors on the "Disc" class and the model's potentially suboptimal ImageNet pre-trained features. This combination likely guided the optimizer into a local minimum, prioritizing discs while ignoring verte-

brae. This issue was not observed in randomly initialized models, which are free from this potentially misleading inductive bias.

3.5.2 Instance Segmentation

Table 3 – Results for the Instance Segmentation task, grouped by metric. The table shows scores for different architectures and loss functions.

Metric	Loss	U-Net 3D	V-Net	ResNet-50	UNETR	Swin UNETR
Precision	Dice	0.47	0.72	0.55	0.47	0.58
	Dice+Focal	0.53	0.70	0.59	0.56	0.61
	Gen. Dice	0.17	0.09	0.10	0.08	0.08
	Dice + CE	0.56	0.73	0.68	0.57	0.66
Recall	Dice	0.44	0.70	0.54	0.44	0.57
	Dice+Focal	0.46	0.69	0.54	0.48	0.58
	Gen. Dice	0.13	0.12	0.11	0.09	0.10
	Dice + CE	0.47	0.70	0.60	0.50	0.58
DSC	Dice	0.43	0.69	0.52	0.43	0.55
	Dice+Focal	0.47	0.67	0.54	0.49	0.57
	Gen. Dice	0.13	0.10	0.10	0.08	0.08
	Dice + CE	0.49	0.70	0.61	0.51	0.59

In contrast, the instance segmentation task presents a significantly more challenging scenario, as evidenced by the drop in performance across all models (see Table 3). This task requires not only correct voxel classification but also the spatial disentanglement of adjacent structures (e.g., L1 vertebra vs. L2 vertebra). The resulting DSC scores, mostly falling in the 0.4 to 0.7 range, underscore this difficulty and serve as a much better differentiator of model capabilities.

In this more complex setting, the superiority of **V-Net** became even more pronounced. They consistently stood out as the top-performing architecture across all metrics, achieving Precision, Recall, and DSC scores above 0.65 in most cases. The performance gap to other models was significant. The SwinUNETR model, while strong in semantic segmentation, faced significant drops in Precision, Recall. The UNETR and U-Net 3D models delivered the weakest results in this task.

The influence of the loss function also proved more critical. While Dice and Dice + CE loss continued to favor the top-performing models, the Generalized Dice loss yielded the worst performance for every model by a significant margin. This comparative analysis highlights that instance segmentation imposes challenges that clearly distinguish the effectiveness and spatial generalization capabilities of each architecture and loss function.

A granular, per-instance analysis (see Figure 13) reveals further complexities. While the patch-based sampling strategy may contribute to this by limiting the available anatomical context, we hypothesize that it is also a primary factor in the catastrophic performance collapse observed with the **Generalized Dice** loss, evident by its failure to learn multiple

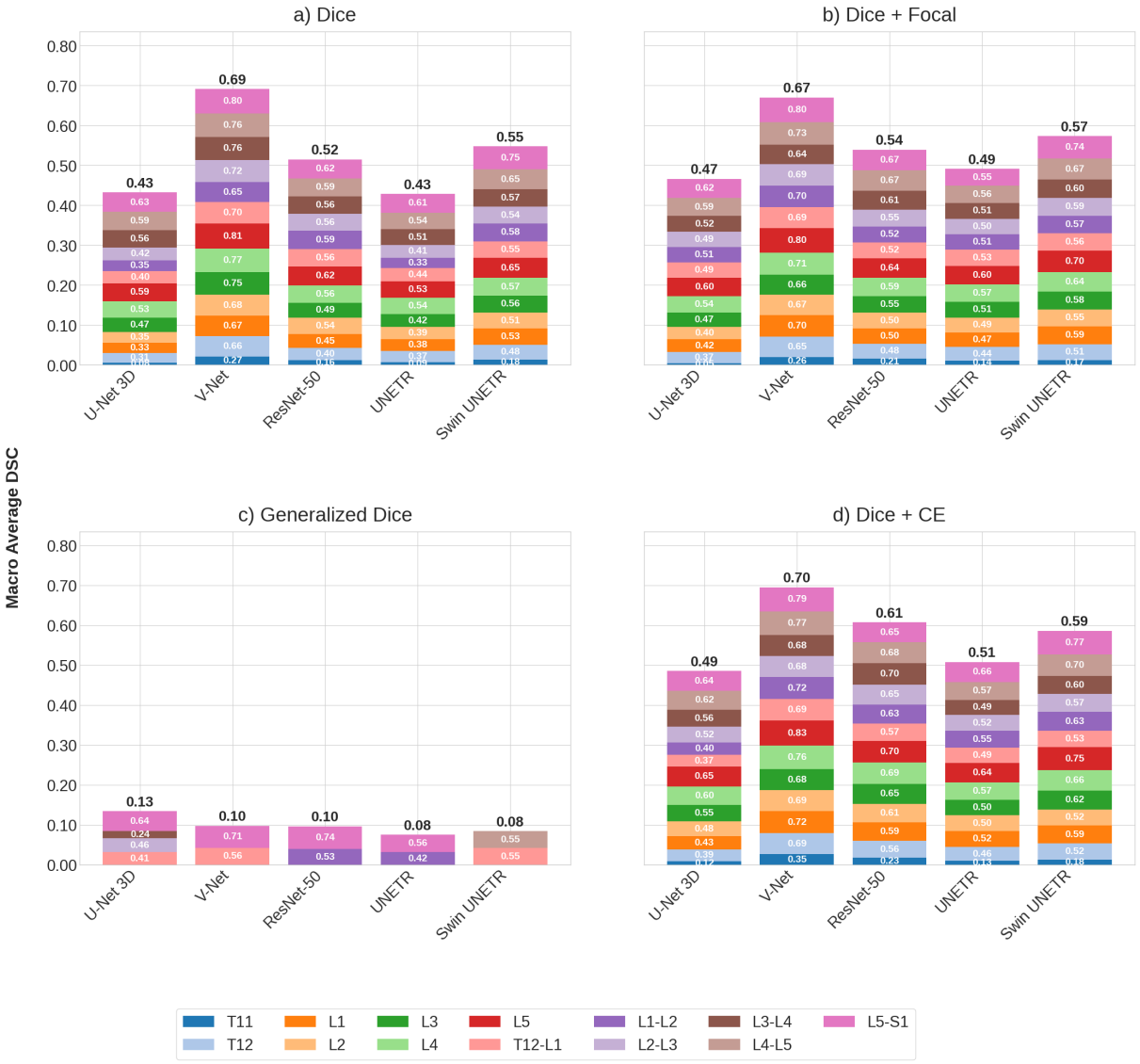


Figure 13 – Results for the instance segmentation task. The chart displays the macro-average Dice Similarity Coefficient (DSC) for five different models, with each bar segmented by the performance on each of the 13 anatomical classes, which include seven vertebrae (T11–L5) and six intervertebral discs (T12-L1–L5-S1).

instance classes. Our assumption is that this random cropping introduces high variability in the relative volumes of each instance within the training patches. Probably, this variance destabilizes the GDL's weighting mechanism, which is inherently sensitive to class volume, resulting in an erratic and ineffective training signal for this specific task.

3.6 Conclusion

This work presented a comprehensive empirical study, evaluating five state-of-the-art deep learning architectures in combination with four loss functions for lumbar spine segmentation in magnetic resonance images. We conducted our evaluation using the public

SPIDER dataset on two clinically relevant tasks: semantic and instance segmentation.

Our results reveal a clear distinction in the challenge posed by each task. In semantic segmentation, most models demonstrated high performance, with DSC scores consistently above 0.8. In contrast, instance segmentation, which requires the individualization of each anatomical structure, proved to be a substantially greater challenge. In this more complex scenario, which served as a crucial differentiator of model capabilities, the V-Net and Swin UNETR architectures emerged as the most robust and accurate. The choice of loss function also proved to be more impactful in this scenario, with the Dice and Dice + CE losses achieving good results with the best models, while the Generalized Dice loss achieved the worst performance for all models.

The main implication of this study is the need to evaluate segmentation models in scenarios that reflect real-world clinical complexity. Semantic segmentation, while useful, can mask a model's deficiencies in delineating adjacent structures of the same class. The success of V-Net and Swin UNETR suggests that both refined convolutional strategies and hierarchical Transformer-based approaches are promising paths for this task.

Despite its comprehensiveness, our study has limitations that must be considered. In addition to focusing on a single public dataset (SPIDER) and T2-weighted sequences, it is crucial to reflect on the impact of data transforms. The random patch-based sampling strategy, while a standard and necessary approach for training on large volumes, introduced a significant technical challenge that profoundly impacted our results. Particularly in instance segmentation, this methodology creates a direct conflict with the formulation of the **Generalized Dice** loss. Because the randomly sampled patches frequently do not contain instances of all anatomical classes, the GDL's weighting mechanism, which is designed to operate on the assumption that all classes are present, becomes unstable. This technical incompatibility directly explains the catastrophic drop in performance observed with this loss function, as the model receives an erratic and unreliable training signal, a problem rooted in the interaction between the sampling strategy and the mathematical design of the loss function.

Therefore, future work should not only validate these findings on multiple datasets but also explore more sophisticated processing strategies that preserve global context, such as whole-volume analysis with models that support variable-sized inputs and additionally include measures of variability in the results (e.g., confidence intervals or interquartile ranges). In summary, this study provides a valuable guide for researchers and practitioners, establishing a clear benchmark on the performance of models and loss functions, and emphasizing the importance of instance segmentation as the true test of robustness for automating lumbar spine analysis.

Chapter 4

Multi-Pathology Segmentation in Lumbar Spine MRI: A Comparative Deep Learning Approach

This chapter presents the second and main article that constitutes this dissertation, representing the culmination of the proposed investigation. Its content corresponds to the second and final stage of the overall research methodology.

- **Title:** Multi-Pathology Segmentation in Lumbar Spine MRI: A Comparative Deep Learning Approach.
- **Authors:** Claudio Leite, Samuel Felipe dos Santos, and Jurandy Almeida.
- **Status:** Accepted for publication in the 21th International Conference on Computer Vision Theory and Applications - VISAPP 2026 (Qualis=A3).

Building upon the methodological foundation established in the previous chapter, this work advances to the central objective of this dissertation: to develop and validate a robust methodology for the precise spatial indication of symptomatic areas with the radiological classification of multiple, co-occurring pathologies in lumbar spine MRIs. To achieve this, the chapter proposes and systematically evaluates three distinct strategies for handling diagnostic overlap: binary class, multi-class, and multi-label segmentation. By comparing these approaches in a benchmark of over 200 training pipelines, this study seeks to answer the main research question and validate the hypothesis that the multi-label formulation offers a superior trade-off between accuracy and computational efficiency. The results

presented here, therefore, aim to establish a practical and efficient guideline for future research and clinical applications in the automated diagnosis of spinal pathologies.

The remainder of this chapter is organized as follows.

Section 4.1 presents the abstract. Section 4.2 provides the introduction. Section 4.3 discusses the related work. Section 4.4 describes our proposed approaches. Section 4.5 details the experimental methodology. Section 4.6 reports and analyzes the experimental results. Finally, Section 4.7 offers our conclusions.

4.1 Abstract

Low back pain is a leading cause of disability worldwide. Magnetic Resonance Imaging (MRI) is a cornerstone for diagnosis, yet deep learning methods are needed to overcome limitations such as diagnostic overlap, where a single anatomical location presents with multiple pathologies. This paper presents a comprehensive empirical study on the segmentation of multiple pathologies in lumbar intervertebral discs. We systematically compare three distinct strategies for handling diagnostic overlap: (i) **binary class segmentation**, a baseline that treats each pathology independently; (ii) **multi-class segmentation**, mapping 70 disease combinations to unique classes (non-overlapping masks); and (iii) **multi-label segmentation**, which uses binary channels to explicitly model the coexistence of multiple diagnoses (overlapping masks). These strategies are evaluated across five state-of-the-art architectures and four loss functions, encompassing over 200 distinct training pipelines. Our results demonstrate that the proposed multi-label segmentation strategy achieves a superior trade-off between accuracy and computational efficiency, outperforming the costly binary class approach and establishing a practical guideline for future research.

4.2 Introduction

Low back pain is a leading cause of disability worldwide, affecting the quality of life of millions and imposing a significant burden on healthcare systems (PALMER et al., 2000). Magnetic Resonance Imaging (MRI) of the lumbar spine is a cornerstone for accurate diagnosis, yet its efficacy faces limitations. Despite a high rate of appropriate referrals, only a small fraction of these scans effectively contributes to clinical decision-making (PALMER et al., 2000). In this context, deep learning has emerged as a powerful tool to support the analysis of medical images, with the potential to optimize screening, risk stratification, and diagnostic support (CASTIGLIONI et al., 2021).

A central challenge in this domain is the semantic overlap of multiple classes within the same pixel or voxel. For instance, a single intervertebral disc may simultaneously

present with conditions like a herniation and disc narrowing — a scenario that most current methods are not designed to handle.

Current research efforts to address these limitations have largely followed two distinct paths. The first focuses on the segmentation of symptomatic areas for a single disease, such as segmenting regions of possible lumbar spinal stenosis (ALTUN; ALTUN; ALKAN, 2023) or identifying possible herniated discs (QIAN et al., 2024). The second performs radiological classification for multiple diseases, often employing multi-stage pipelines that combine different methods, such as integrating radiomic features or using neural networks in series to first detect anatomical structures and then classify cropped regions of interest (WINDSOR et al., 2022).

While effective, both methods have significant limitations. The first is constrained to a single-condition diagnosis, ignoring pathological co-occurrence. The second often relies on complex, sequential processes that are computationally inefficient and difficult to generalize.

To address these gaps, this paper presents a comprehensive empirical study on the segmentation of multiple, overlapping pathologies in lumbar intervertebral discs. The core of our work lies in the systematic proposal and comparison of three distinct strategies to manage diagnostic overlap. Based on these strategies, we conduct an extensive comparative analysis involving five state-of-the-art deep neural networks and four loss functions. Our goal is to determine the most effective and efficient strategy for simultaneously delineating regions of interest and classifying them into multiple diagnostic categories.

The main contributions of this work are:

- ❏ A novel application framework for addressing the multi-diagnosis problem as a semantic segmentation task. We propose and evaluate three strategies: (i) **binary class segmentation**, a baseline that treats each pathology independently; (ii) **multi-class segmentation**, which maps 70 possible disease combinations to unique, exclusive classes (i.e., non-overlapping masks); and (iii) **multi-label segmentation**, an approach that uses binary channels, one for each disease, to explicitly model the coexistence of multiple diagnoses at a same location (i.e., overlapping masks).

- ❏ A systematic benchmark evaluating about 200 training pipelines, considering five neural network architectures, four loss functions, and ten different models (i.e., one for multi-class, one for multi-label, and eight for binary class). This study establishes a valuable reference point for the community regarding the performance of different approaches in this complex clinical scenario.

- ❏ A demonstration that our proposed multi-label segmentation strategy offers a superior trade-off between accuracy and computational efficiency. Our results indicate

that this approach achieves performance comparable to the costly binary class segmentation strategy, thus establishing a practical and efficient guideline for future research and applications in this field.

The rest of this paper is organized as follows. Section 4.3 discusses related work. Section 4.4 presents our approaches to multi-pathology diagnosis. Section 4.5 describes the experimental setup. Section 4.6 reports our results. Finally, Section 4.7 offers our conclusions and directions for future work.

4.3 Related Work

The literature on the analysis of spinal imaging using deep learning is extensive. For this study, related works were organized into two main approaches: (i) those that perform the segmentation of symptomatic areas for a single disease, and (ii) those that employ radiological classification, often in multiple stages, for the diagnosis of multiple diseases.

Diagnostic support via indication of symptomatic areas. A significant branch of research that indicates symptomatic areas of a specific pathology to support diagnosis. For Lumbar Spinal Stenosis (LSS), for instance, Altun et al. (ALTUN; ALTUN; ALKAN, 2023) employed an LSS-UNET on axial images, while Li et al. (LI et al., 2021) used MANet, a U-Net enhanced with multi-scale attention. Other works (LAIWALLA et al., 2023; HOHENHAUS et al., 2024) have also proposed variations of the U-Net to segment and quantify spinal canal compromise in cases of stenosis.

Similarly, Disc Herniation has been a common target. Mbarki et al. (MBARKI et al., 2020) utilized a U-Net with a VGG-16 encoder, and Qian et al. (QIAN et al., 2024) proposed a modified U-Net with ResNet-18 and an attention mechanism for the same task. While effective in their respective tasks, a fundamental limitation of these methods is that they are inherently designed for a single condition, i.e., they deal with only one pathology at a time, ignoring the frequent coexistence of multiple diseases.

Diagnostic support via radiological classification of diseases or anomalies. Conversely, research efforts have focused on the diagnosis of multiple diseases, but this is often achieved through complex, multi-stage classification pipelines that separate structure localization from disease classification. One common line of work involves feature extraction followed by a classifier. Some approaches (BEULAH; SHARMILA; PRAMOD, 2018; BEULAH; SHARMILA; PRAMOD, 2022) applied algorithms like Expectation-Maximization (EM) for segmentation and then used image descriptors (HOG, Gabor) with an SVM to classify disc bulge and different types of disc degeneration. Likewise, Alsmirat et al. (ALSMIRAT et al., 2022) used SIFT to extract regions of interest and an AlexNet for the classification of herniated discs. Another, more modern, line of work uses end-to-end neural networks, but still in a sequential manner. SpineNetV2 (WINDSOR et al., 2022) uses three distinct networks in series: one to detect vertebral bodies, another to

classify the vertebrae, and a third to refine the predictions, before a final ResNet classifies the disc crop. Similar approaches used a Faster R-CNN for localization and a ResNet for classification (PAN et al., 2021), and employed a Mask R-CNN for localization followed by multiple CNNs for classification (CHEN et al., 2024).

Although these methods can classify multiple conditions, they often rely on a fragmented pipeline (detection, cropping, classification) and involve many parameters. More recently, Chen et al. (CHEN et al., 2025) and Wang et al. (WANG et al., 2023a) have made advances in multi-label classification, but they still operate within an image or patch classification paradigm, rather than providing a precise spatial localization of the pathology through the indication of symptomatic areas.

Existing methods are often limited to single-disease segmentation or rely on fragmented, multi-stage classification pipelines. Neither approach adequately addresses the common clinical scenario of multiple, co-occurring pathologies within the same anatomical structure. In this context, this work aims to overcome these limitations by proposing a unified framework that integrates the spatial localization of symptomatic areas with the radiological classification of multiple concurrent diseases. Our objective is to support the diagnosis of coexisting conditions in intervertebral discs, thereby advancing the state-of-the-art in the analysis of regional spinal pathologies.

4.4 Our Approaches

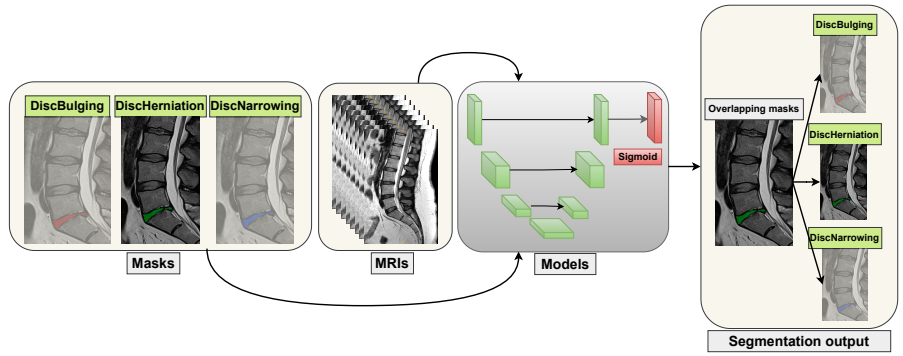
To address the challenge of multi-pathology diagnosis, we propose and systematically evaluate three distinct semantic segmentation strategies, as shown in Figure 14. These strategies — Binary Class, Multi-Class, and Multi-Label — represent different conceptual frameworks for handling diagnostic overlap and are detailed in the following subsections.

4.4.1 Binary Class Segmentation

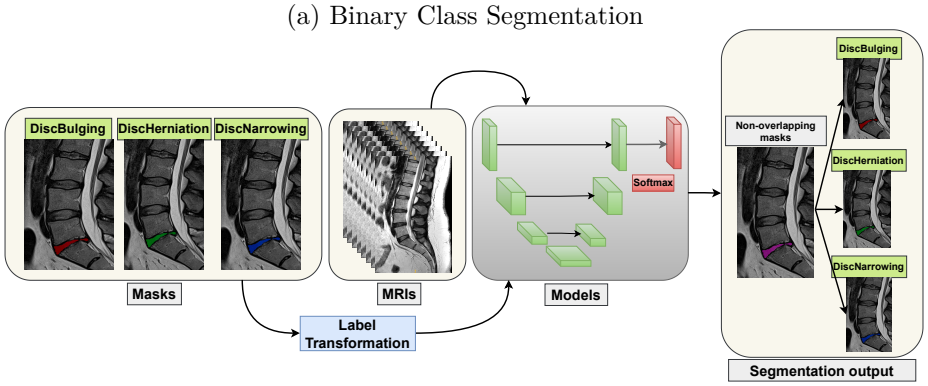
The first strategy (Figure 14a), **Binary Class Segmentation**, serves as a baseline that decomposes the complex multi-diagnosis problem into a series of independent binary segmentation tasks. In this formulation, a separate segmentation model is trained specifically for each of the pathologies analyzed.

Let $X \in \mathbb{R}^{D\times H\times W}$ be an input 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 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 $\mathcal{M}_k : \mathbb{R}^{D\times H\times W} \to [0,1]^{D\times H\times W}$ 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:

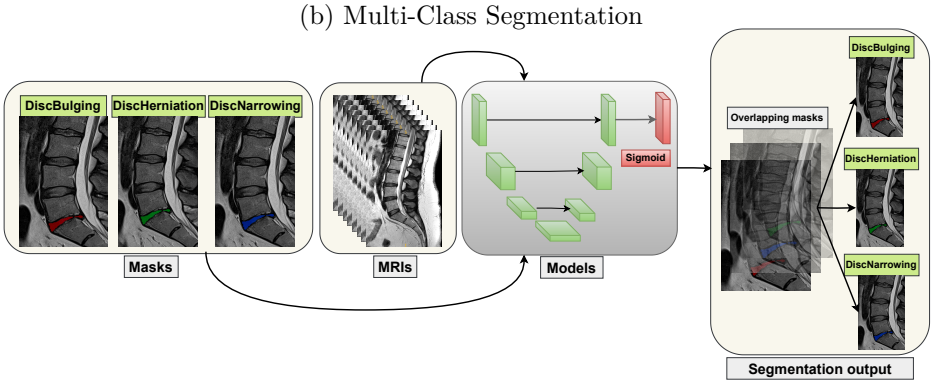
$$\hat{Y}_k = \mathcal{M}_k(X), \quad \text{for } k = 1, \ldots, n \tag{13}$$



(a) Binary Class Segmentation



(b) Multi-Class Segmentation



(c) Multi-Label Segmentation

Figure 14 – Illustration of the three segmentation strategies for multi-pathology diagnosis. (a) Binary Class trains n separate models. (b) Multi-Class trains one model on m unique combination classes. (c) Multi-Label trains one model with n output channels.

where each model $\mathcal{M}_k$ predict a mask $\hat{Y}_k$.

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 is its high computational cost and resource inefficiency. The need to train, validate, and store n independent models multiplies the experimentation time and the demand for hardware resources. In addition, 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 correlations that may exist between diseases, treating each as an isolated event.

4.4.2 Multi-class Segmentation

The second strategy (Figure 14b), **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 unique pathology combinations observed in the training dataset, such that $|\mathcal{C}_{obs}| = m$ and $2^{\mathcal{P}}$ denotes the power set of $\mathcal{P}$, i.e., the complete collection of all possible subsets of $\mathcal{P}$. 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 background. The objective is to learn a single model $\mathcal{M} : \mathbb{R}^{D \times H \times W} \to \mathbb{R}^{(m+1) \times D \times H \times W}$ that performs standard multi-class segmentation over these fused labels. This operation can be expressed as:

$$\hat{Y}' = \underset{c}{\text{argmax}}(\mathcal{M}(X)) \tag{14}$$

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 most probable class.

The main advantage of this approach is its conceptual simplicity, serving as a baseline for comparison that can handle multiple pathologies with a single model. 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.

4.4.3 Multi-label Segmentation

The final and most flexible strategy is **Multi-Label Segmentation** (Figure 14c), 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 $\mathcal{M} : \mathbb{R}^{D \times H \times W} \to [0, 1]^{n \times D \times H \times W}$ 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:

$$\{\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n\} = \mathcal{M}(X) \tag{15}$$

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. In addition, the model can also learn shared features that may be relevant to multiple diseases, potentially improving overall performance. However, its complexity lies in the training process, as it requires optimizing n tasks simultaneously.

4.5 Experimental Setup

This section details the comprehensive experimental setup used to evaluate our proposed strategies, covering the dataset, data preprocessing, model architectures, loss functions, and the performance metrics employed.

4.5.1 Model Architectures

To evaluate the performance of the proposed strategies, we selected five deep learning architectures for medical image segmentation. The choice of these models was based on their relevance and representation in the state-of-the-art and their prior validation for the segmentation of anatomical structures (LEITE; SANTOS; ALMEIDA, 2025). This selection includes approaches based on Convolutional Neural Networks (CNNs), Transformer-based models, and hybrid architectures.

4.5.1.1 U-Net 3D

The U-Net 3D (KERFOOT et al., 2018) is a direct extension of the U-Net (RONNEBERGER; FISCHER; BROX, 2015) architecture for processing volumetric data and is characterized by a symmetric encoder-decoder structure, with a contracting path to capture image context and an expanding path to enable precise localization of segmented features.

4.5.1.2 V-Net

Like U-Net 3D, V-Net (MILLETARI; NAVAB; AHMADI, 2016) is another prominent CNN architecture for volumetric segmentation. Its key contribution is the integration of residual connections within each stage of the network, which mitigate the vanishing gradient problem, enabling the training of deeper architectures capable of learning more complex feature representations.

4.5.1.3 ResNet-50

This hybrid approach combines the effectiveness of two canonical architectures: ResNet and U-Net. In this model, the U-Net's encoder is replaced by a ResNet-50 architecture (HE et al., 2016), which serves as a powerful feature extractor.

The decoder, in turn, follows the symmetric expanding path of the U-Net (RONNEBERGER; FISCHER; BROX, 2015), using the features extracted by the encoder at multiple scales to reconstruct a precise segmentation mask.

4.5.1.4 UNETR

Inspired by success of Transformers, UNETR (HATAMIZADEH et al., 2022) reformulates the task of volumetric segmentation as a sequence-to-sequence prediction problem. In this architecture, a Transformer is used as the encoder to learn sequence representations of the input volume, effectively capturing long-range spatial dependencies — a known limitation of CNNs. While maintaining the "U-shaped" design, the Transformer encoder is directly connected to a CNN-based decoder via skip connections at different resolutions, combining global information capture with local detail reconstruction.

4.5.1.5 Swin UNETR

The Swin UNETR (HATAMIZADEH et al., 2021) is an evolution of UNETR that uses the more efficient Swin Transformer as its encoder. Unlike the Transformer's global self-attention, the Swin Transformer computes attention locally within non-overlapping windows and uses a shifted-window mechanism to learn global features. This approach is designed to more effectively model long-range relationships, making it useful for segmenting structures with variable sizes and shapes.

4.5.2 Loss Functions

To find the most robust optimization for our three strategies, we evaluated four loss functions chosen for their properties in handling class imbalance and optimizing overlap metrics.

4.5.2.1 Dice

Originally proposed by Milletari et al. (MILLETARI; NAVAB; AHMADI, 2016), the Dice loss directly optimizes the Dice Similarity Coefficient (DSC). It is inherently robust to strong class imbalance, making it a powerful baseline. It is defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{k=1}^{C}\sum_{v=1}^{N_v}\hat{Y}_{k,v}Y_{k,v}}{\sum_{k=1}^{C}\sum_{v=1}^{N_v}(\hat{Y}_{k,v} + Y_{k,v})} \tag{16}$$

where C is the number of segmentation channels (e.g., n for multi-label, $m+1$ for multi-class), N_v is the total number of voxels, $Y_{k,v}$ is the ground truth for class k at voxel v, and $\hat{Y}_{k,v}$ is the corresponding predicted probability.

4.5.2.2 Dice + Focal

This is a weighted sum of the Dice loss and the Focal loss (LIN et al., 2017), which reshapes the standard cross-entropy loss to focus training on hard-to-classify examples. It is defined as:

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{N_v}\sum_{v=1}^{N_v}\sum_{k=1}^{C}\alpha(1-\hat{Y}_{k,v})^{\gamma}Y_{k,v}\log(\hat{Y}_{k,v}) \tag{17}$$

where $\gamma > 0$ is a focusing parameter (typically 2) and α is a weighting factor. The final loss is a weighted sum, $\mathcal{L}_{\text{DiceFocal}} = \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \lambda_{\text{Focal}}\mathcal{L}_{\text{Focal}}$, where λ_{Dice} and λ_{Focal} are weighting parameters for each term, used with default values of 1.0 for both.

4.5.2.3 Generalized Dice

An evolution of the Dice loss, the Generalized Dice Loss (GDL) (SUDRE et al., 2017) is designed for severe class imbalance. It introduces a class-specific weight, w_k, that is inversely proportional to the class region's volume, ensuring that smaller classes are not ignored during training. The weight and loss are defined as:

$$w_k = \frac{1}{(\sum_{v=1}^{N_v} Y_{k,v})^2} \tag{18}$$

$$\mathcal{L}_{\text{GDL}} = 1 - 2\frac{\sum_{k=1}^{C} w_k \sum_{v=1}^{N_v}\hat{Y}_{k,v}Y_{k,v}}{\sum_{k=1}^{C} w_k \sum_{v=1}^{N_v}(\hat{Y}_{k,v}+Y_{k,v})} \tag{19}$$

4.5.2.4 Dice + CE

This hybrid loss is a summation of the Dice loss and the standard Cross-Entropy (CE) loss (MAO; MOHRI; ZHONG, 2023). This combination leverages the complementary strengths of both functions. While the Dice loss is robust to severe class imbalance by optimizing the overlap metric directly, the CE loss operates at the voxel-level, penalizing individual misclassifications. This composite loss ensures both high fidelity for region-based overlap and fine-grained accuracy at the pixel level, leading to more stable optimization. The Cross-Entropy component is defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N_v}\sum v = 1^{N_v}\sum_{k=1}^{C} Y_{k,v}\log(\hat{Y}_{k,v}) \tag{20}$$

This combination is formulated as $\mathcal{L}_{\text{DiceCE}} = \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}}$, where λ_{Dice} and λ_{CE} are weighting factors that control the relative importance of the region-based and voxel-level components, respectively. In this work, both factors were set to 1.0.

4.5.3 Dataset

This study utilizes the public SPIDER (SPine Imaging Diagnostic Extended Resource) dataset (GRAAF et al., 2024), a comprehensive resource for advancing spinal imaging research. 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 multi-center and demographic diversity of SPIDER is fundamental for training and validating robust and generalizable models.

The dataset is enriched with two types of high-quality manual annotations. First, it includes anatomical segmentations of all visible vertebrae (excluding the sacrum), intervertebral discs, and the spinal canal. They were generated through a meticulous process conducted by a medical trainee under the supervision of a medical imaging specialist and an experienced musculoskeletal radiologist. To ensure accuracy, an iterative, semi-automatic approach was employed: an initial algorithm was trained on a small, manually segmented subset and then used to generate preliminary segmentations for the remaining data. Then, these segmentations were reviewed and manually corrected by experts, with the corrections being reincorporated to refine the algorithm in a continuous cycle, resulting in a robust and detailed segmentation ground truth. Second, the dataset includes radiological classifications for each intervertebral disc level, performed by an experienced radiologist. These classifications assess the presence and severity of a spectrum of degenerative changes, including: Modic changes (types I, II, or III), Low Endplate, Up Endplate, Spondylolisthesis, Disc herniation, Disc narrowing, Disc bulging, and Pfirrmann grade (from 1 to 5).

The simultaneous availability of precise segmentation masks and detailed multi-pathology classifications is a key feature of the SPIDER dataset. This unique combination enables our investigation, as it allows for the creation and evaluation of ground truths for multi-pathology diagnosis.

4.5.4 Data Preprocessing

The SPIDER dataset exhibits intrinsic heterogeneity typical of multi-center clinical data, with MRI scans varying in dimensions, spatial resolution, and alignment. This causes anatomical structures, such as vertebral bodies and intervertebral discs, to appear in different positions across the volumes, posing a significant challenge for deep models, which operate more robustly and efficiently with uniform input data.

To mitigate this variability, we employed a label-guided random cropping strategy. During the training phase, patches with a fixed spatial dimension of $32 \times 192 \times 192$ voxels were randomly extracted from the full volumes. This process was guided by the ground truth annotations to ensure each patch contained semantically relevant information. The sampling was strategically configured to draw patches centered on regions of interest

(positive samples) and background areas (negative samples) in a 1:2 ratio. This approach guarantees that approximately 66.7% of the sampled patches are centered on background regions, compelling the model to learn discriminative features from challenging contexts rather than focusing only on trivial foreground examples. This normalization procedure not only ensures input data consistency for all models in our comparative study but also serves as an effective data augmentation technique, enhancing model robustness by exposing it to varied anatomical contexts.

Table 4 – Results for DSC, Precision and Recall for the three proposed strategies across different models and loss functions. Values are presented as (Binary class / Multi-class / Multi-label). Best score per class is in bold.

Table 5 – DSC

Metric	Loss	U-Net 3D	V-Net	ResNet-50	UNETR	Swin UNETR
Macro-average	Dice	0.34/0.22/0.38	0.38/0.08/0.40	0.37/0.22/0.39	0.35/0.21/0.37	0.39/0.25/**0.42**
	Dice + Focal	0.32/0.28/0.35	0.37/0.29/0.41	0.37/0.22/0.40	0.34/0.18/0.36	0.38/0.27/**0.42**
	Generalized Dice	0.34/0.01/0.32	0.33/0.01/0.34	0.38/0.04/0.31	0.35/0.23/0.32	**0.39**/0.20/0.35
	Dice + CE	0.34/0.24/0.36	0.37/0.34/**0.41**	0.36/0.24/0.38	0.34/0.07/0.36	0.39/0.23/**0.41**
Example-based	Dice	0.26/0.29/0.35	0.06/0.07/0.37	0.36/0.30/0.39	0.26/0.29/0.34	0.34/0.32/**0.43**
	Dice + Focal	0.23/0.32/0.36	0.33/0.39/0.41	0.35/0.30/0.41	0.25/0.24/0.31	0.33/0.36/**0.43**
	Generalized Dice	0.27/0.01/0.33	0.00/0.00/0.40	0.35/0.03/0.39	0.26/0.23/0.33	0.35/0.25/**0.41**
	Dice + CE	0.26/0.27/0.33	0.31/**0.43**/0.41	0.35/0.35/0.37	0.25/0.19/0.33	0.35/0.34/0.41

Table 6 – Precision

Metric	Loss	U-Net 3D	V-Net	ResNet-50	UNETR	Swin UNETR
Macro-average	Dice	0.34/0.29/0.36	0.36/0.07/0.39	**0.40**/0.24/**0.40**	0.35/0.27/0.38	**0.40**/0.30/**0.40**
	Dice + Focal	0.35/0.32/0.38	0.38/0.35/0.40	0.41/0.24/**0.42**	0.34/0.30/0.40	0.39/0.32/0.39
	Generalized Dice	0.36/0.01/0.33	0.31/0.00/0.36	**0.41**/0.02/0.40	0.36/0.26/0.33	**0.41**/0.33/0.40
	Dice + CE	0.34/0.40/0.35	0.38/**0.43**/0.39	0.41/0.39/0.38	0.36/0.12/0.39	0.41/0.39/0.39
Example-based	Dice	0.27/0.38/0.31	0.05/0.08/0.29	0.36/0.39/0.35	0.26/0.37/0.29	0.33/**0.40**/0.37
	Dice + Focal	0.25/0.37/0.35	0.35/**0.44**/0.35	0.35/0.40/0.37	0.25/0.39/0.27	0.32/0.42/0.37
	Generalized Dice	0.28/0.01/0.36	0.00/0.00/0.43	0.35/0.03/**0.44**	0.27/0.28/0.33	0.35/0.29/0.43
	Dice + CE	0.28/0.33/0.28	0.32/0.46/0.33	0.36/**0.48**/0.30	0.25/0.39/0.28	0.35/**0.48**/0.34

Table 7 – Recall

Metric	Loss	U-Net 3D	V-Net	ResNet-50	UNETR	Swin UNETR
Macro-average	Dice	0.40/0.22/0.47	**0.51**/0.33/0.52	0.41/0.25/0.46	0.42/0.21/0.46	0.45/0.25/0.50
	Dice + Focal	0.38/0.30/0.39	0.44/0.30/**0.51**	0.41/0.26/0.47	0.43/0.16/0.41	0.46/0.29/**0.51**
	Generalized Dice	0.39/**0.58**/0.37	0.55/0.08/0.41	0.43/0.42/0.32	0.39/0.24/0.38	0.44/0.19/0.38
	Dice + CE	0.41/0.19/0.45	0.44/0.33/**0.53**	0.40/0.21/0.48	0.41/0.06/0.42	0.43/0.20/0.52
Example-based	Dice	0.31/0.28/0.49	0.08/0.06/**0.61**	0.43/0.28/0.53	0.31/0.27/0.50	0.43/0.32/0.60
	Dice + Focal	0.27/0.35/0.45	0.37/0.38/0.58	0.41/0.28/0.54	0.30/0.21/0.45	0.41/0.35/**0.60**
	Generalized Dice	0.31/0.01/0.37	0.01/0.01/0.44	0.42/0.05/0.40	0.31/0.25/0.39	0.43/0.27/**0.46**
	Dice + CE	0.31/0.28/0.49	0.36/0.45/**0.62**	0.40/0.33/0.55	0.30/0.14/0.49	0.41/0.31/0.60

Table 8 – DSC results per class for the three proposed strategies across different models and loss functions. Values are presented as (Binary class / Multi-class / Multi-label). Best score per class is in bold.

Table 9 – Dice

Classe	**U-Net 3D**	**V-Net**	**ResNet-50**	**UNETR**	**Swin UNETR**
Modic	0.32 / 0.16 / 0.38	0.36 / 0.03 / 0.40	0.36 / 0.14 / 0.38	0.32 / 0.14 / 0.39	0.37 / 0.21 / **0.41**
UP endplate	0.31 / 0.17 / 0.42	0.37 / 0.23 / 0.44	0.38 / 0.15 / 0.43	0.35 / 0.12 / 0.43	0.40 / 0.21 / **0.45**
LOW endplate	0.31 / 0.16 / 0.41	0.39 / 0.03 / 0.44	0.33 / 0.15 / 0.41	0.34 / 0.12 / 0.42	0.40 / 0.20 / **0.45**
Spondylolisthesis	0.01 / 0.00 / **0.04**	0.01 / 0.00 / **0.04**	0.01 / 0.00 / **0.04**	0.01 / 0.00 / 0.03	0.02 / 0.00 / **0.04**
Disc herniation	0.11 / 0.00 / 0.10	0.04 / 0.00 / 0.11	0.09 / 0.01 / 0.10	0.12 / 0.00 / 0.10	0.12 / 0.00 / **0.14**
Disc narrowing	0.42 / 0.28 / 0.44	0.50 / 0.09 / 0.48	0.48 / 0.30 / 0.47	0.45 / 0.29 / 0.44	0.46 / 0.34 / **0.53**
Disc bulging	0.51 / 0.40 / 0.54	0.57 / 0.08 / 0.59	0.55 / 0.41 / 0.56	0.53 / 0.43 / 0.53	0.59 / 0.41 / **0.60**
Pfirrman grade	0.70 / 0.60 / 0.67	**0.76** / 0.14 / 0.72	0.74 / 0.60 / 0.69	0.68 / 0.58 / 0.65	0.74 / 0.61 / 0.75
Overall	0.34 / 0.22 / 0.38	0.38 / 0.08 / 0.40	0.37 / 0.22 / 0.39	0.35 / 0.21 / 0.37	0.39 / 0.25 / **0.42**

Table 10 – Dice + Focal

Classe	**U-Net 3D**	**V-Net**	**ResNet-50**	**UNETR**	**Swin UNETR**
Modic	0.29 / 0.27 / 0.35	0.34 / 0.26 / 0.40	0.35 / 0.18 / 0.38	0.31 / 0.16 / 0.38	0.37 / 0.24 / **0.41**
UP endplate	0.32 / 0.30 / 0.37	0.34 / 0.29 / 0.43	0.38 / 0.18 / 0.42	0.34 / 0.14 / 0.41	0.40 / 0.23 / **0.45**
LOW endplate	0.30 / 0.29 / 0.35	0.30 / 0.30 / 0.43	0.36 / 0.17 / 0.41	0.35 / 0.13 / 0.39	0.38 / 0.21 / **0.44**
Spondylolisthesis	0.00 / 0.00 / 0.04	0.01 / 0.00 / **0.05**	0.01 / 0.00 / **0.05**	0.01 / 0.00 / 0.04	0.01 / 0.00 / 0.04
Disc herniation	0.09 / 0.00 / 0.08	0.08 / 0.01 / 0.12	0.08 / 0.00 / 0.13	0.12 / 0.00 / 0.10	0.11 / 0.00 / **0.14**
Disc narrowing	0.41 / 0.36 / 0.42	**0.55** / 0.31 / 0.50	0.48 / 0.23 / 0.52	0.40 / 0.18 / 0.42	0.50 / 0.38 / 0.51
Disc bulging	0.49 / 0.43 / 0.54	0.56 / 0.51 / 0.59	0.57 / 0.40 / 0.58	0.51 / 0.26 / 0.51	0.57 / 0.44 / **0.60**
Pfirrman grade	0.68 / 0.61 / 0.66	**0.77** / 0.67 / 0.73	0.70 / 0.60 / 0.71	0.69 / 0.57 / 0.60	0.73 / 0.66 / 0.75
Overall	0.32 / 0.28 / 0.35	0.37 / 0.29 / 0.41	0.37 / 0.22 / 0.40	0.34 / 0.18 / 0.36	0.38 / 0.27 / **0.42**

Table 11 – Generalized Dice

Classe	**U-Net 3D**	**V-Net**	**ResNet-50**	**UNETR**	**Swin UNETR**
Modic	0.31 / 0.01 / 0.30	0.35 / 0.00 / 0.29	0.37 / 0.04 / 0.26	0.31 / 0.24 / 0.32	**0.39** / 0.21 / 0.34
UP endplate	0.31 / 0.01 / 0.33	**0.39** / 0.00 / 0.34	0.37 / 0.05 / 0.27	0.34 / 0.24 / 0.36	**0.39** / 0.20 / 0.36
LOW endplate	0.32 / 0.01 / 0.33	0.01 / 0.01 / 0.34	0.35 / 0.05 / 0.27	0.33 / 0.24 / 0.36	**0.37** / 0.20 / 0.35
Spondylolisthesis	0.02 / 0.00 / 0.00	0.00 / 0.00 / 0.00	0.02 / 0.00 / 0.00	0.02 / 0.00 / 0.00	**0.03** / 0.00 / 0.00
Disc herniation	0.12 / 0.00 / 0.00	0.01 / 0.00 / 0.00	0.13 / 0.00 / 0.00	**0.15** / 0.00 / 0.02	0.12 / 0.00 / 0.00
Disc narrowing	0.42 / 0.01 / 0.44	**0.52** / 0.01 / 0.45	0.49 / 0.05 / 0.46	0.44 / 0.28 / 0.41	0.49 / 0.27 / 0.46
Disc bulging	0.51 / 0.02 / 0.52	**0.60** / 0.01 / 0.58	0.56 / 0.07 / 0.56	0.50 / 0.33 / 0.51	0.59 / 0.27 / 0.57
Pfirrman grade	0.69 / 0.02 / 0.63	**0.76** / 0.01 / 0.69	0.74 / 0.09 / 0.68	0.68 / 0.49 / 0.60	0.74 / 0.47 / 0.71
Overall	0.34 / 0.01 / 0.32	0.33 / 0.01 / 0.34	0.38 / 0.04 / 0.31	0.35 / 0.23 / 0.32	**0.39** / 0.20 / 0.35

Table 12 – DICE + CE

Classe	**U-Net 3D**	**V-Net**	**ResNet-50**	**UNETR**	**Swin UNETR**
Modic	0.30 / 0.21 / 0.35	0.33 / 0.31 / 0.39	0.35 / 0.18 / 0.38	0.33 / 0.00 / 0.37	0.37 / 0.17 / **0.40**
UP endplate	0.32 / 0.22 / 0.40	0.33 / 0.32 / **0.45**	0.35 / 0.18 / 0.42	0.35 / 0.00 / 0.42	0.40 / 0.17 / **0.45**
LOW endplate	0.30 / 0.21 / 0.38	0.35 / 0.33 / **0.45**	0.36 / 0.19 / 0.41	0.34 / 0.00 / 0.40	0.36 / 0.17 / **0.45**
Spondylolisthesis	0.00 / 0.00 / **0.04**	0.01 / 0.00 / **0.04**	0.01 / 0.00 / **0.04**	0.01 / 0.00 / 0.03	0.02 / 0.00 / **0.04**
Disc herniation	0.11 / 0.00 / 0.08	0.05 / 0.02 / 0.10	0.08 / 0.00 / 0.10	**0.13** / 0.00 / 0.09	0.11 / 0.00 / 0.11
Disc narrowing	0.44 / 0.26 / 0.42	**0.54** / 0.47 / 0.48	0.48 / 0.25 / 0.46	0.40 / 0.00 / 0.43	0.53 / 0.23 / 0.49
Disc bulging	0.50 / 0.42 / 0.53	0.58 / 0.54 / **0.60**	0.52 / 0.42 / 0.56	0.52 / 0.00 / 0.53	0.55 / 0.40 / 0.58
Pfirrman grade	0.72 / 0.56 / 0.64	**0.77** / 0.70 / 0.74	0.72 / 0.67 / 0.68	0.67 / 0.54 / 0.64	0.74 / 0.67 / 0.74
Overall	0.34 / 0.24 / 0.36	0.37 / 0.34 / **0.41**	0.36 / 0.24 / 0.38	0.34 / 0.07 / 0.36	0.39 / 0.23 / **0.41**

4.5.5 Performance Metrics

We used two families of standard metrics to analyze performance: macro-average metrics, which assess performance at the class level; and example-based metrics, which assess performance at the instance level.

4.5.5.1 Macro-average Metrics

These metrics are computed per segmentation class (or channel) k. Let Y_k be the ground truth binary mask for class k, and $\hat{Y}_k$ be the corresponding binarized prediction mask from the model (with probabilities thresholded at 0.5). The **Precision** and **Recall** for class k measure the fraction of positive predictions that are correct and the fraction of actual positives that are correctly identified, respectively,

$$\text{Precision}_k = \frac{\sum_v (\hat{Y}_{k,v} \cdot Y_{k,v})}{\sum_v \hat{Y}_{k,v}}, \text{Recall}_k = \frac{\sum_v (\hat{Y}_{k,v} \cdot Y_{k,v})}{\sum_v Y_{k,v}}. \tag{21}$$

The **Dice Similarity Coefficient (DSC)** for class k is the harmonic mean of Precision and Recall:

$$\text{DSC}_k = \frac{2 \sum_v (\hat{Y}_{k,v} \cdot Y_{k,v})}{\sum_v \hat{Y}_{k,v} + \sum_v Y_{k,v}}. \tag{22}$$

To ensure all C classes contribute equally to the final evaluation, we report the **macro-average** of these metrics, calculated by averaging the independent scores for each class, e.g., $\text{DSC}_\text{M} = \frac{1}{C} \sum_{k=1}^{C} \text{DSC}_k$.

4.5.5.2 Example-based Metrics

Macro-average metrics assess aggregate performance but do not capture a model's ability to predict the correct set of labels for each instance (ZHANG; ZHOU, 2013).

For this, we use example-based metrics, treating each voxel as an independent instance.

Let $\mathcal{P}_v = \{k | Y_{k,v} = 1\}$ be the set of true pathology labels for a voxel v, and let $\hat{\mathcal{P}}_v = \{k | \hat{Y}_{k,v} \geq \tau\}$ be the set of predicted labels for that same voxel (with a threshold $\tau = 0.5$). The example-based ($\text{Precision}_\text{E}$) and ($\text{Recall}_\text{E}$) compute the proportion of correctly predicted labels and correctly identified ground-truth labels, respectively, averaged over all voxels,

$$\text{Precision}_\text{E} = \frac{1}{N_v} \sum_{v=1}^{N_v} \frac{|\mathcal{P}_v \cap \hat{\mathcal{P}}_v|}{|\hat{\mathcal{P}}_v|}, \text{Recall}_\text{E} = \frac{1}{N_v} \sum_{v=1}^{N_v} \frac{|\mathcal{P}_v \cap \hat{\mathcal{P}}_v|}{|\mathcal{P}_v|}. \tag{23}$$

To balance these, we compute the example-based (DSC_E), which is the harmonic mean for each voxel individually before averaging (ZHANG; ZHOU, 2013):

$$\text{DSC}_\text{E} = \frac{1}{N_v} \sum_{v=1}^{N_v} \frac{2 \times |\mathcal{P}_v \cap \hat{\mathcal{P}}_v|}{|\mathcal{P}_v| + |\hat{\mathcal{P}}_v|} \tag{24}$$

These metrics provide a granular view of performance, penalizing models that fail to capture the correct combination of co-occurring pathologies at a specific location.

4.5.6 Implementation Details

All experiments were conducted following a standardized protocol. We used only the T2-weighted MRI scans. The dataset was partitioned into a training set of 168 scans (80% or 134 for training, 20% or 34 for validation) and a hold-out test set of 42 scans. For preprocessing, all input volumes underwent intensity scaling and were subsequently random cropped to a spatial dimension of $32 \times 192 \times 192$ voxels.

To ensure reproducibility, all training runs were initialized with a fixed random seed of 42. The models were trained for a total of 500 epochs using the Adam optimizer with a learning rate of 1e-4 and a batch size of 8 and the validation process was executed every 2 epochs. In the evaluation on the test set, inference on the full-sized volumes was performed with a sliding window approach.

4.6 Experimental Results

This section compares the performance of our proposed strategies. We first conduct an overall analysis of the methods with multiple performance metrics, followed by a per-class analysis of the results, and concluding with an evaluation of the computational cost of the evaluated strategies.

Table 4 presents the macro-average and example-based metrics of DSC, Precision, and Recall for every combination of architecture, loss function, and segmentation strategy.

The Multi-Label strategy demonstrates a clear performance advantage over the other two approaches. As we can see in Table 5, its best configuration (Swin UNETR with Dice or Dice + Focal) yielded a macro-average DSC of 0.42, outperforming the best Binary Class model (Swin UNETR), which achieved a score of 0.39. The Multi-Class strategy performed worst compared to the other approaches, reaching a macro-average DSC of 0.34 with its best configuration (V-Net with Dice + CE). This poor performance is attributable to the extreme class imbalance and data sparsity inherent in this formulation. The superior performance of the Multi-Label strategy indicates that training a single, unified model that can learn shared features across pathologies is more effective than training separate, isolated models.

For example-based DSC, which measures the model's ability to predict the correct *set* of labels for each voxel, the best Multi-Label results (Swin UNETR with Dice or Dice + Focal) achieved an example-based DSC of 0.43, tying with the best configuration for the Multi-Class strategy (V-Net with Dice + CE), that obtained the same performance. In this setting, the Binary Class strategy performed worst, with its best configuration (Swin UNETR with Generalized Dice or Dice + CE) scoring 0.35. These results demonstrates that the Multi-Label strategy is not only superior at identifying pathologies at a macro level, but is also effective at correctly identifying the full set of labels at a given location.

Tables 6 and 7 reveal the trade-off between recall and precision. The top-performing models, such as

Swin UNETR with Dice and Dice + Focal loss for the Multi-Label strategy, show a strong balance between macro-average recall (0.50 and 0.51, respectively) and precision (0.40 and 0.39), indicating they are not biased towards over- or under-segmenting.

After evaluating the overall performance of the models with multiple metrics, we conducted a break-down of the per-class DSC for each of the 8 pathologies across all architectures, loss functions, and segmentation strategies. These results are presented in Table 8, and highlight the varying difficulty of segmenting different conditions. Classes, such as Pfirrman grade and Disc bulging, consistently achieved the highest DSC. For example, in the Multi-Label setting, Swin UNETR with Dice + Focal loss reached DSC of 0.75 and 0.60, respectively. While pathologies such as Spondylolisthesis and Disc herniation, proved extremely challenging for all strategies, with DSC often between zero and 0.14, likely due to significant class imbalance.

Finally, we evaluated the computational cost of each segmentation strategy and architecture, quantifying the total number of parameters (in millions) and the computational complexity in Giga Floating-point Operations Per Second (GFLOPs). A detailed comparison is presented in Table 13.

Table 13 – Comparison of computational cost (parameters and GFLOPs) for each strategy and architecture. The cost for the Binary Class strategy reflects the total for all 8 models.

Architecture	Params (M)	GFLOPs
Binary class (8 models)		
U-Net 3D	15.84	151.20
V-Net	364.72	6760.64
ResNet-50	541.36	1401.60
UNETR	740.96	1760.48
Swin UNETR	495.92	7047.20
Multi-class (1 model)		
U-Net 3D	2.04	159.54
V-Net	45.87	1508.46
ResNet-50	67.70	245.52
UNETR	92.62	222.66
Swin UNETR	61.99	888.70
Multi-label (1 model)		
U-Net 3D	1.98	33.16
V-Net	45.62	911.36
ResNet-50	67.68	182.32
UNETR	92.62	220.32
Swin UNETR	61.99	881.68

A critical distinction in this analysis is the practical requirement of the Binary class

(one-vs-all) strategy. To segment all 8 target classes, this approach requires the deployment of 8 independent models. Consequently, the total computational cost reported for this strategy represents the sum of all 8 models.

In sharp contrast, the Multi-class and Multi-label strategies operate as single, unified models capable of handling all classes simultaneously. This fundamental difference results in a dramatic reduction in computational overhead. For instance, the Binary class strategy using Swin UNETR totals 495.92 M parameters and 7047.20 GFLOPs, whereas the Multi-label strategy implemented with Swin UNETR requires only 61.99 M parameters and 881.68 GFLOPs, an approximate 8-fold reduction.

The experimental results strongly support the efficacy of the proposed **Multi-Label segmentation strategy**. It not only achieves superior overall performance but does so with a single model, making it vastly more computationally efficient and practical for clinical application than the standard Binary Class strategy. The ability of the multi-label models to learn shared representations appears to be a key factor in their success.

Among the model architectures, the Transformer-based **Swin UNETR** was a consistent top performer, particularly when paired with the **Dice** and **Dice + Focal** losses in the multi-label setting. The CNN-based V-Net was also capable of obtaining similar results when paired with the Dice + CE loss. This suggests that both advanced attention mechanisms and well-designed convolutional structures are effective for this complex task.

In summary, a Multi-Label strategy implemented with a Swin UNETR or optimized with a standard Dice or Dice + Focal losses, or with a V-Net optimized with the Dice + CE loss, represents the most effective and efficient approach for multi-pathology segmentation in lumbar spine MRI.

4.7 Conclusions

In this paper, we presented a comprehensive empirical study to address the segmentation of multiple, co-occurring pathologies in lumbar spine MRI. We systematically designed, implemented, and evaluated three distinct strategies — Binary Class, Multi-Class, and Multi-Label segmentation — across five deep neural networks and four loss functions.

Our results demonstrate that the **Multi-Label segmentation strategy offers a superior trade-off between accuracy and computational efficiency**. By training a single, unified model to predict all pathologies simultaneously, this approach achieved a higher macro-average DSC score than the resource-intensive Binary Class baseline, which requires training a separate model for each condition. We found that models like **Swin UNETR** with the **Dice** or **Dice + Focal losses** and **V-Net** with the **Dice + CE loss**, yielded the best performance in the multi-label framework.

This work establishes a valuable benchmark for multi-pathology medical image analysis and provides a clear, practical guideline for developing scalable diagnostic tools. Fu-

ture work will focus on including additional pathologies, exploring advanced multi-label loss functions, and analyzing model robustness regarding hyper-parameters such as the positive-negative sampling ratio and initialization variability. We also plan to enhance interpretability through XAI and qualitative results, like visualization of predictions, while collaborating with health experts to validate the clinical utility of our results and present clinical interpretations.

Chapter 5

Conclusion

The automated analysis of lumbar spine pathologies from magnetic resonance images represents a significant clinical challenge, especially given the frequent co-occurring of multiple pathologies or conditions in the same anatomical structure. Existing computational approaches were largely limited to single-disease segmentation or multi-stage classification pipelines, which are computationally expensive and lack precise spatial localization of the findings. Motivated by this gap, the central objective of this dissertation was to develop and validate a unified deep learning methodology capable of segmenting multiple pathologies simultaneously and efficiently in lumbar intervertebral discs.

To achieve this goal, a progressive methodology structured in two complementary phases was adopted. The first phase, presented in **Chapter 3**, was crucial to the success of the investigation as it established the technical and methodological robustness of the work. It consisted of a rigorous benchmark for the fundamental task of anatomical segmentation of vertebrae and intervertebral discs, which allowed for the objective identification of the most suitable models and loss functions. This foundational step ensured that the selection of tools for pathology analysis was not arbitrary but based on empirical evidence, thereby strengthening the validity of the subsequent conclusions.

The second phase, detailed in **Chapter 4**, directly addressed the core research question by systematically evaluating three strategies for multi-pathology segmentation: (i) **binary class segmentation**, a baseline that treats each pathology independently; (ii) **multi-class segmentation**, mapping 70 disease combinations to unique classes (non-overlapping masks); and (iii) **multi-label segmentation**, which uses binary channels to explicitly model the coexistence of multiple diagnoses (overlapping masks). The results conclusively demonstrated the ineffectiveness of the multi-class approach and, more importantly, revealed the efficacy of the multi-label strategy. By training a unified model, it

achieved performance metrics comparable to the binary class approach, but with vastly superior computational efficiency, validating it as the most practical approach for clinical deployment.

This dissertation, therefore, affirmatively answers its research questions. The direct comparison between the strategies demonstrated that the multi-label formulation is the most effective for computationally translating the clinical challenge of pathological co-occurrence. Furthermore, the results validate the central hypothesis that a unified approach handling the precise spatial indication of symptomatic areas with the radiological classification of multiple, co-occurring pathologies offers a superior trade-off between diagnostic accuracy and computational efficiency, requiring the training of a single model instead of multiple ones.

In summary, the main contribution of this work is the establishment of a valuable benchmark and a clear, practical guideline for the analysis of multiple pathologies in medical images. By demonstrating the feasibility and efficiency of a multi-label segmentation framework, this research offers a scalable and robust solution with the potential to optimize the diagnosis and clinical management of diseases of the lumbar spine.

5.1 Future Work

Despite the promising results, this research has limitations that open avenues for future investigations. Based on the findings and challenges encountered, we propose the following directions:

- ❑ **Integration of Anatomical and Pathological Models:** A promising next step is to cascade the models developed in this dissertation. The robust anatomical segmentation output (vertebrae and discs) from the first phase could be utilized as an attention mechanism or masking input for the second phase. This would effectively filter out background noise, allowing the multi-pathology model to focus learning exclusively on the relevant anatomical regions of interest.

- ❑ **Validation on Multiple Datasets and Organs:** Although the SPIDER dataset is diverse, validating the top-performing models on external datasets is crucial. Furthermore, to verify the versatility of the multi-label strategy in handling overlapping regions, future work should apply this methodology to other medical domains, such as brain tumor segmentation in the BraTS dataset (PATI et al., 2021), comparing our approach against the hierarchical strategies commonly used in the literature to handle tumor sub-region overlaps.

- ❑ **Parameter-Efficient Domain Adaptation:** A significant barrier to clinical deployment is the domain shift between scanners. Instead of computationally expensive full-model fine-tuning, future research should investigate Parameter-Efficient

Fine-Tuning (PEFT) techniques. Specifically, utilizing adapters such as Low-Rank Adaptation (LoRA) (HU et al., 2021) or Weight-Decomposed Low-Rank Adaptation (DoRA) (LIU et al., 2024) could allow the proposed multi-pathology base model to adapt to new clinical environments or specific rare pathologies with minimal computational overhead.

- ❑ **Robust Evaluation Metrics:** While the Dice Similarity Coefficient (DSC) is the standard metric in medical imaging, it can be biased towards the positive class in certain scenarios. As highlighted by Chicco and Jurman (CHICCO; JURMAN, 2020), metrics like the Matthews Correlation Coefficient (MCC) or the Jaccard Index (IoU) may provide a more truthful evaluation of binary classifications in imbalanced datasets. Future benchmarks should include these metrics to rigorously validate the model's reliability.

- ❑ **Exploration of Advanced Loss Functions:** Future work should focus on exploring sophisticated loss functions designed to better handle the trade-off between false positives and false negatives, such as the Tversky Loss (SALEHI; ERDOGMUS; GHOLIPOUR, 2017). By adjusting its hyperparameters, it is possible to impose higher penalties on false negatives, which is critical in medical diagnostic scenarios where missing a pathology is often more detrimental than a false alarm.

- ❑ **Open-World Segmentation:** The current paradigm operates in a "closed world", identifying only the pathologies for which it was explicitly trained. An advanced research direction would be to develop models capable of operating in an "open-world" scenario, where the system could not only segment known classes but also identify and flag the presence of previously unseen anomalies.

- ❑ **Preservation of Global Context:** The patch-based sampling strategy limits the anatomical context available to the model. Investigating architectures capable of analyzing whole volumes or strategies that preserve global context could mitigate failures, particularly in instance segmentation and the differentiation of rare pathologies.

- ❑ **Prospective Clinical Validation:** The ultimate goal is to integrate a tool like this into the clinical workflow. Prospective studies will be necessary to evaluate the model's real-world impact on diagnostic accuracy, radiologist efficiency, and, ultimately, patient outcomes.

Bibliography

ALSMIRAT, M. et al. Deep learning-based disk herniation computer aided diagnosis system from mri axial scans. **IEEE Access**, IEEE, v. 10, p. 32315–32323, 2022.

ALTUN, İ.; ALTUN, S.; ALKAN, A. Lss-unet: Lumbar spinal stenosis semantic segmentation using deep learning. **Multimedia Tools and Applications**, Springer, v. 82, n. 26, p. 41287–41305, 2023.

ANBUDEVI, M. K. a.; SUGANTHI, K. Review of semantic segmentation of medical images using modified architectures of unet. **Diagnostics**, MDPI, v. 12, n. 12, 2022.

ANDREW, J. et al. Spine magnetic resonance image segmentation using deep learning techniques. In: IEEE. **2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)**. [S.l.], 2020. p. 945–950.

AZAD, R. et al. Medical image segmentation review: The success of u-net. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 46, n. 1, p. 10076–10095, 2022.

BEULAH, A.; SHARMILA, T. S.; PRAMOD, V. Disc bulge diagnostic model in axial lumbar mr images using intervertebral disc descriptor (idd). **Multimedia Tools and Applications**, Springer, v. 77, n. 20, p. 27215–27230, 2018.

_____. Degenerative disc disease diagnosis from lumbar mr images using hybrid features. **The Visual Computer**, Springer, v. 38, n. 8, p. 2771–2783, 2022.

BHOI, S. et al. Atypical presentation of anti-mog ab disease. **Kansas Journal of Medicine**, v. 14, p. 310–313, 12 2021.

BIDHULT, S. et al. Validation of t1 and t2 algorithms for quantitative mri: performance by a vendor-independent software. **BMC Medical Imaging**, v. 16, n. 1, p. 1–8, 2016.

CASTIGLIONI, I. et al. Ai applications to medical images: From machine learning to deep learning. **Physica medica**, Elsevier, v. 83, p. 9–24, 2021.

CHEN, K. et al. Deep learning-based intelligent diagnosis of lumbar diseases with multi-angle view of intervertebral disc. **Mathematics**, MDPI, v. 12, n. 13, p. 2062, 2024.

CHEN, Y. et al. Deep learning-based computer-aided diagnostic system for lumbar degenerative diseases classification using mri. **Biomedical Signal Processing and Control**, Elsevier, v. 109, p. 108002, 2025.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. **BMC genomics**, BioMed Central, v. 21, n. 1, p. 1–13, 2020.

DESPOTOVIĆ, I.; GOOSSENS, B.; PHILIPS, W. Mri segmentation of the human brain: Challenges, methods, and applications. **Computational and Mathematical Methods in Medicine**, Hindawi, v. 2015, 2015.

ELHARROUSS, O. et al. Panoptic segmentation: A review. **arXiv preprint arXiv:2111.10250**, 2021.

FEHLINGS, M. G. et al. **The aging of the global population: the changing epidemiology of disease and spinal disorders**. [S.l.]: LWW, 2015. S1–S5 p.

GRAAF, J. W. van der et al. Lumbar spine segmentation in mr images: a dataset and a public benchmark. **Scientific Data**, Nature Publishing Group UK London, v. 11, n. 1, p. 264, 2024.

HAFIZ, A. M.; BHAT, G. M. A survey on instance segmentation: state of the art. **International Journal of Multimedia Information Retrieval**, Springer, v. 9, p. 171–189, 2020.

HATAMIZADEH, A. et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: SPRINGER. **International MICCAI brainlesion workshop**. [S.l.], 2021. p. 272–284.

_____. Unetr: Transformers for 3d medical image segmentation. In: **Proceedings of the IEEE/CVF winter conference on applications of computer vision**. [S.l.: s.n.], 2022. p. 574–584.

HE, K. et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

HILLE, G. et al. Vertebral body segmentation in wide range clinical routine spine mri data. **Computer methods and programs in biomedicine**, Elsevier, v. 155, p. 93–99, 2018.

HOHENHAUS, M. et al. Quantification of cervical spinal stenosis by automated 3d mri segmentation of spinal cord and cerebrospinal fluid space. **Spinal Cord**, Nature Publishing Group UK London, v. 62, n. 7, p. 371–377, 2024.

HU, E. J. et al. Lora: Low-rank adaptation of large language models. **arXiv preprint arXiv:2106.09685**, 2021.

HUTCHINS, J. et al. A systematic review of validated classification systems for cervical and lumbar spinal foraminal stenosis based on magnetic resonance imaging. **European Spine Journal**, v. 31, n. 6, p. 1358–1369, 2022.

KERFOOT, E. et al. Left-ventricle quantification using residual u-net. In: SPRINGER. **International workshop on statistical atlases and computational models of the heart**. [S.l.], 2018. p. 371–380.

KOLARIK, M. et al. 3d volumetric segmentation of the vertebral bodies and intervertebral discs from mri using deep learning. **Computer Methods and Programs in Biomedicine**, Elsevier, v. 182, p. 105051, 2019.

KUKIL. **Intersection over Union (IoU) in Object Detection and Segmentation**. 2022. LearnOpenCV. Acessado em: 30 dez. 2025. Disponível em: <https://learnopencv.com/intersection-over-union-iou-in-object-detection-and-segmentation/>.

LAIWALLA, A. N. et al. Lumbar spinal canal segmentation in cases with lumbar stenosis using deep-u-net ensembles. **World Neurosurgery**, Elsevier, v. 178, p. e135–e140, 2023.

LAMA, R. S. D. et al. Computer-aided diagnosis of vertebral compression fractures using convolutional neural networks and radiomics. **Journal of digital imaging**, Springer, v. 35, n. 3, p. 446–458, 2022.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LEITE, C.; SANTOS, S. F. dos; ALMEIDA, J. A comprehensive evaluation of deep learning architectures and loss functions for lumbar spine segmentation in mri. In: **CIARP**. [S.l.: s.n.], 2025. p. 1–14.

LEPAGE, M.; GORE, J. C. Contrast mechanisms in magnetic resonance imaging. In: **Journal of Physics: Conference Series**. [S.l.]: IOP Publishing, 2004. v. 3, n. 1, p. 214–221.

LI, H. et al. Automatic lumbar spinal mri image segmentation with a multi-scale attention network. **Neural Computing and Applications**, Springer, v. 33, n. 18, p. 11589–11602, 2021.

LIN, T.-Y. et al. Focal loss for dense object detection. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 2980–2988.

LIU, D. et al. Cell r-cnn v3: A novel panoptic paradigm for instance segmentation in biomedical images. **arXiv preprint arXiv:2002.06345**, 2020.

LIU, S.-Y. et al. Dora: Weight-decomposed low-rank adaptation. **arXiv preprint arXiv:2402.09353**, 2024.

MAO, A.; MOHRI, M.; ZHONG, Y. Cross-entropy loss functions: Theoretical analysis and applications. v. 202, p. 23928–23946, 2023.

MBARKI, W. et al. A novel method based on deep learning for herniated lumbar disc segmentation. In: IEEE. **2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)**. [S.l.], 2020. p. 394–399.

MCNICOLL, G. World population ageing 1950-2050. **Population and development Review**, John Wiley & Sons, Inc., v. 28, n. 4, p. 814–816, 2002.

MILLETARI, F.; NAVAB, N.; AHMADI, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: IEEE. **2016 fourth international conference on 3D vision (3DV)**. [S.l.], 2016. p. 565–571.

MINAEE, S. et al. Image segmentation using deep learning: A survey. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 44, n. 7, p. 3523–3542, 2021.

MüLLER, D.; SOTO-REY, I.; KRAMER, F. Towards a guideline for evaluation metrics in medical image segmentation. **BMC Research Notes**, v. 15, 2022.

PALMER, K. T. et al. Back pain in britain: comparison of two prevalence surveys at an interval of 10 years. **Bmj**, British Medical Journal Publishing Group, v. 320, n. 7249, p. 1577–1578, 2000.

PAN, Q. et al. Automatically diagnosing disk bulge and disk herniation with lumbar magnetic resonance images by using deep convolutional neural networks: method development study. **JMIR medical informatics**, JMIR Publications Inc., Toronto, Canada, v. 9, n. 5, p. e14755, 2021.

PATI, S. et al. The federated tumor segmentation (fets) challenge. **arXiv preprint arXiv:2105.05874**, 2021.

QIAN, J. et al. Lumbar disc herniation diagnosis using deep learning on mri. **Journal of Radiation Research and Applied Sciences**, Elsevier, v. 17, n. 3, p. 100988, 2024.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **International Conference on Medical image computing and computer-assisted intervention**. [S.l.], 2015. p. 234–241.

SALEHI, S. S. M.; ERDOGMUS, D.; GHOLIPOUR, A. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: SPRINGER. **International workshop on machine learning in medical imaging**. [S.l.], 2017. p. 379–387.

SERAI, S. D. Basics of magnetic resonance imaging and quantitative parameters t1, t2, t2*, t1rho and diffusion-weighted imaging. **Pediatric Radiology**, v. 52, n. 2, p. 217–227, 2022.

SHEN, D.; WU, G.; SUK, H.-I. Deep learning in medical image analysis. **Annual review of biomedical engineering**, Annual Reviews, v. 19, p. 221–248, 2017.

SUDRE, C. H. et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: SPRINGER. **International Workshop on Deep Learning in Medical Image Analysis**. [S.l.], 2017. p. 240–248.

TAGHANAKI, S. A. et al. Deep semantic segmentation of natural and medical images: a review. **Artificial Intelligence Review**, Springer, v. 54, p. 137–178, 2021.

TAHA, A.; HANBURY, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. **BMC Medical Imaging**, v. 15, 2015.

TCHITO, C. T. et al. Biomedical image classification in a big data architecture using machine learning algorithms. **Journal of Healthcare Engineering**, v. 2021, p. 1–11, 05 2021.

WANG, S. et al. Automatic segmentation of lumbar spine mri images based on improved attention u-net. **Computational Intelligence and Neuroscience**, Wiley Online Library, v. 2022, n. 1, p. 4259471, 2022.

WANG, Y. et al. Deep learning-driven diagnosis of multi-type vertebra diseases based on computed tomography images. **Quantitative Imaging in Medicine and Surgery**, v. 14, n. 1, p. 800, 2023.

WANG, Z. et al. Dice semimetric losses: Optimizing the dice score with soft labels. In: SPRINGER. **International conference on medical image computing and computer-assisted intervention**. [S.l.], 2023. p. 475–485.

WANG, Z.; XIAO, P.; TAN, H. Spinal magnetic resonance image segmentation based on u-net. **Journal of Radiation Research and Applied Sciences**, Elsevier, v. 16, n. 3, p. 100627, 2023.

WINDSOR, R. et al. Spinenetv2: automated detection, labelling and radiological grading of clinical mr scans. **arXiv preprint arXiv:2205.01683**, 2022.

ZHANG, M.-L.; ZHOU, Z.-H. A review on multi-label learning algorithms. **IEEE transactions on knowledge and data engineering**, IEEE, v. 26, n. 8, p. 1819–1837, 2013.

ZHOU, S. et al. High-resolution encoder–decoder networks for low-contrast medical image segmentation. **IEEE Transactions on Image Processing**, IEEE, v. 29, p. 461–475, 2020.

ZHOU, Z. et al. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. **IEEE Transactions on Medical Imaging**, IEEE, v. 39, n. 6, p. 1856–1867, 2019.