TECNOLOGIA SETÚBAL
ESCOLA SUPERIOR
POLITÉCNICO SETÚBAL

JACKSON SIEBEN

Automatic Detection of Lumbar Spinal Stenosis in Computed Magnetic Resonance Imaging

Thesis Report submitted as a partial requirement for the Master's degree in Software Engineering.

ADVISOR
Prof. Dr. Miguel López
CO-ADVISOR
Prof. Dr. Jorge Aikes Junior
SUPERVISOR
Dr. Nuno Cristino

November 2025

JACKSON SIEBEN

Automatic Detection of Lumbar Spinal Stenosis in Computed Magnetic Resonance Imaging

Examination Committee

Chair: Prof. José António Moinhos Cordeiro, ESTSetúbal/IPS

Member and Supervisor: Prof. Miguel López, ESTSetúbal/IPS

External Member: Prof. Augusto Marques Ferreira da Silva, DETI/ UA

November 2025

Resumo

A Estenose Lombar do Canal Vertebral (Lumbar Spinal Stenosis (LSS)) é uma condição degenerativa comum caracterizada pelo estreitamento do canal vertebral, que pode provocar dor, limitação funcional e redução da qualidade de vida. Apesar da disponibilidade de técnicas avançadas de imagiologia, como a Ressonância Magnética (RM), o diagnóstico e a classificação da gravidade da LSS continuam a ser desafiantes devido à interpretação subjetiva, à variabilidade interobservador e à ausência de critérios de avaliação padronizados. Estes desafios evidenciam a necessidade de ferramentas automáticas e objetivas que apoiem a tomada de decisão clínica.

Esta dissertação propõe uma framework de visão computacional e aprendizagem profunda para a deteção e classificação automática da LSS em imagens sagitais de RM lombar. A abordagem segue duas etapas: primeiro, as vértebras são localizadas através de um modelo de deteção You Only Look Once Version 8 (YOLOv8); em seguida, são extraídas regiões de interesse ao nível dos discos intervertebrais e classificadas com um modelo baseado em Swin Transformer em três categorias: Normal/Leve, Moderada e Severa. O pré-processamento incluiu filtragem, redimensionamento, realce de contraste e anotação manual das vértebras. O desenvolvimento e a validação foram realizados com o dataset multi-institucional da Radiological Society of North America (RSNA) 2024, que contém estudos diversificados de RM lombar com anotações padronizadas de gravidade.

A avaliação experimental demonstrou que o modelo de deteção YOLOv8 alcançou 97,83% de precisão, 98,03% de revocação e um F1-score de 97,93% na localização de vértebras. O classificador baseado em Swin Transformer obteve um F1-score ponderado de 77,05% e um F1-score macro de 67,98%, enquanto a framework integrada atingiu um F1-score ponderado de 77,15%. Adicionalmente, o sistema demonstrou viabilidade clínica ao processar cada estudo em média em 1,64 segundos, confirmando o seu potencial de integração nos fluxos de trabalho radiológicos.

Os resultados confirmam que métodos baseados em Inteligência Artificial (IA) podem aumentar a precisão e a eficiência no diagnóstico da LSS. São ainda discutidas limitações, como a diversidade restrita do dataset e a necessidade de validação clínica prospetiva, bem como perspetivas futuras, incluindo o uso de imagens multi-planares, datasets mais extensos e arquiteturas avançadas baseadas em transformadores.

Keywords: Estenose Lombar do Canal Vertebral; Ressonância Magnética (RM); Aprendizagem Profunda; YOLOv8; Swin Transformer; Redes Neuronais Convolucionais (CNN); Deteção de Objetos; Classificação de Imagem Médica; Apoio à Decisão Clínica; Inteligência Artificial em Saúde

Abstract

LSS is a common degenerative spinal condition characterized by narrowing of the spinal canal, which can result in pain, functional impairment, and a diminished quality of life. Despite the availability of advanced imaging techniques, such as Magnetic Resonance Imaging (MRI), the diagnosis and severity classification of LSS remain challenging due to subjective interpretation, inter-observer variability, and the absence of standardized assessment criteria. These challenges underscore the need for automated and objective tools to support clinical decision-making.

This thesis proposes a computer vision and deep learning framework for the automated detection and classification of LSS in sagittal lumbar MRI scans. The framework adopts a two-stage approach: first, vertebrae are localized using a YOLOv8 object detection model, and second, regions of interest at the intervertebral disc level are extracted and classified by a Swin Transformer-based model into three severity categories: Normal/Mild, Moderate, and Severe. Data preprocessing steps included filtering, resizing, contrast enhancement, and manual annotation of vertebrae to ensure high-quality inputs. The framework was developed and validated using the RSNA 2024 multi-institutional dataset, which contains diverse lumbar spine MRI studies with standardized severity labels.

Experimental evaluation demonstrated that the YOLOv8 detection model achieved high performance, with 97.83% precision, 98.03% recall, and an F1-score of 97.93% in vertebra localization. The Swin Transformer-based classifier achieved a weighted F1-score of 77.05% and a macro F1-score of 67.98%, while the integrated framework yielded a weighted F1-score of 77.15%. Additionally, the system demonstrated clinical feasibility by processing each study in an average of 1.64 seconds, supporting its potential for integration into diagnostic workflows.

The findings of this study confirm that Artificial Intelligence (AI)-driven methods can enhance diagnostic accuracy and efficiency in LSS assessment. Limitations, such as restricted dataset diversity and the need for prospective clinical validation, are discussed, along with future research directions, including the use of multi-plane imaging, larger annotated datasets, and advanced transformer-based architectures.

Keywords: Lumbar Spinal Stenosis (LSS); Magnetic Resonance Imaging (MRI); Deep Learning; YOLOv8; Swin Transformer; Convolutional Neural Networks (CNN); Object Detection; Medical Image Classification; Clinical Decision Support; Artificial Intelligence in Healthcare

Contents

List of Figures . . . . . **V**

List of Tables . . . . . **VI**

Acronyms . . . . . **VIII**

1 Introduction **1**

1.1 Motivation 2

1.2 Objectives 3

1.3 Scientific and Technical Contributions 3

1.4 Structure of the Thesis 4

2 State of The Art **6**

2.1 Lumbar Spinal Stenosis (LSS) 6

2.1.1 Clinical Background 6

2.1.2 Imaging Techniques for LSS 7

2.2 Artificial Intelligence (AI) 9

2.2.1 Artificial Neural Network (ANN) 9

2.2.2 Deep Learning (DL) 11

2.2.3 Convolutional Neural Networks (CNN) 12

2.2.4 You Only Look Once (YOLO) 15

2.2.5 Emerging AI Techniques in Computer Vision 18

3 Materials and Methods **21**

3.1 Dataset Description 21

3.2 Clinical Collaboration 22

3.3 Proposed Method 23

3.3.1 Data Preparation 23

3.3.2 Model Development 25

3.3.3 Model Evaluation 26

3.4 Final Method 27

3.4.1 Detection Data Preparation 28

3.4.2 Detection Model Development 29
3.4.3 Classification Data Preparation 30
3.4.4 Classification Model Development 32
3.4.5 Model Evaluation 34
3.4.6 Framework Evaluation 35
3.5 Development Environment and Computational Resources 36

4 Results and Discussion 38
4.1 Proposed Method Results and Analysis 38
4.1.1 Experimental Setup 38
4.1.2 Experiments Analysis 39
4.2 Detection Model Results and Analysis 43
4.3 Classification Results and Analysis 45
4.4 Framework Results and Analysis 47
4.5 Discussion of Key Findings and Limitations 49
4.6 Comparison with Literature 50

5 Conclusions and Future Work 52
5.1 Conclusions 52
5.2 Recommendations for Future Research 53

6 Bibliographic Reference 54

List of Figures

1 Axial plane of a lumbar vertebra (Bohinski, 2021). 7
2 Sagittal plane of Magnetic Resonance Imaging (MRI) (RSNA, 2024). . . . 8
3 Artificial neuron example. 10
4 Fully connected multilayer networks example (Bre et al., 2018). 11
5 Comparison of Machine Learning (ML) and Deep Learning (DL) workflows. 11
6 Transfer learning workflow example. 13
7 Convolutional operation example (Goodfellow et al., 2016). 14
8 Region Of Interest (ROI) pooling diagram (Fu and Wang, 2024). 14
9 Convolutional Neural Networks (CNN) layers example (Phung and Rhee, 2019). 15
10 You Only Look Once (YOLO) workflow Liu et al. (2020). 16
11 Timeline of YOLO versions Terven and Cordova-Esparza (2023). 16
12 You Only Look Once Version 8 (YOLOv8) Architecture (RangeKing, 2023). 17
13 Vision Transformer (ViT) model overview (Dosovitskiy et al., 2020). 19
14 Flowchart of a Generative Adversarial Network (GAN) architecture (Skandarani et al., 2023) 19
15 Faster Region-based Convolutional Neural Network (Faster R-CNN) workflow example (Ren et al., 2015). 20

16 Study MRI scans from Radiological Society of North America (RSNA) dataset. 22
17 Overall proposed method workflow. 23
18 Preprocessing stage workflow. 24
19 Generate bounding box stage workflow. 24
20 Overall final method workflow. 28
21 Preprocessing stage workflow. 29
22 Example of an annotated image. 30
23 Spinal curvature assessment via polynomial fitting of vertebra centers. . . . 31
24 Schematic overview of ROI extraction between two vertebrae L4/L5. 32
25 Example of a cropped ROI between two lumbar vertebrae. 32
26 Example of a shifted window approach in the Swin Transformer Liu et al. (2021). 33
27 Illustration of Intersection over Union (IoU) in Object Detection. 34

28 Training loss curves for Run f_rcnn_1. 40
29 Training loss curves for Run f_rcnn_2. 41
30 Training loss curves for Run f_rcnn_3. 42
31 Normalized confusion matrix across vertebral levels. 43
32 Training loss curves. 44
33 Representative detection result. 44
34 Normalized confusion matrix across severity classes for the Swin Transformer. 46
35 Training and validation loss curves for the Swin Transformer. 46
36 Example of a misclassified case by the Swin Transformer. 47

List of Tables

3.1 Dataset filtering steps and class distribution. 29

4.1 Experiments Configurations 39
4.2 Quantitative evaluation metrics of the Faster R-CNN runs. 39
4.3 Overall performance metrics. 43
4.4 Overall performance metrics. 45
4.5 Per-class performance metrics of the framework. 48
4.6 Intervertebral-level metrics. 48
4.7 Average computational time of the proposed framework. 48
4.8 Performance comparison with related studies on automated Lumbar Spinal Stenosis (LSS) classification. 51

Acronyms

AI Artificial Intelligence.

ANN Artificial Neural Network.

AP Average Precision.

AUROC Area Under the Receiver Operating Characteristic.

BCEL Binary Cross-Entropy Loss.

CADx Computer-Aided Diagnosis.

CLAHE Contrast Limited Adaptive Histogram Equalization.

CNN Convolutional Neural Networks.

COCO Common Objects in Context.

CPC Contrastive Predictive Coding.

CSP Cross Stage Partial.

CT Computed Tomography.

CV Computer Vision.

CVAT Computer Vision Annotation Tool.

DFL Distribution Focal Loss.

DL Deep Learning.

DL-based Deep Learning-Based.

EST Escola Superior de Tecnologia.

EU European Union.

FAIR Facebook AI Research.

Faster R-CNN Faster Region-based Convolutional Neural Network.

FN False Negatives.

FNR False Negative Rate.

FP False Positives.

FPN Feature Pyramid Network.

FPS Frames Per Second.

GAN Generative Adversarial Network.

GDPR General Data Protection Regulation.

IoU Intersection over Union.

IPS Instituto Politécnico de Setúbal.

IRB Institutional Review Board.

LSS Lumbar Spinal Stenosis.

mAP Mean Average Precision.

MHSAM Multi-Headed Self-Attention Module.

ML Machine Learning.

MRI Magnetic Resonance Imaging.

NLP Natural Language Processing.

R-CNN Region-based Convolutional Neural Network.

ResNet Residual Network.

ResNet-101-FPN ResNet-101 Feature Pyramid Network.

ResNet-50-FPN ResNet-50 Feature Pyramid Network.

ROI Region Of Interest.

RPN Region Proposal Network.

RSNA Radiological Society of North America.

SDT Self-Determination Theory.

SiLU Sigmoid Linear Unit.

SPPF Spatial Pyramid Pooling with Fused Features.

ToI Theory of Influence.

TP True Positives.

VIS-MAE Visualization and Segmentation Masked AutoEncoder.

ViT Vision Transformer.

VS Code Visual Studio Code.

YOLO You Only Look Once.

YOLOv8 You Only Look Once Version 8.

Chapter 1

Introduction

Lumbar Spinal Stenosis (LSS) is a medical condition characterized by the narrowing of the spinal canal, often leading to compression of the spinal cord and nerve roots (Yabuki et al., 2013). It is characterized by degenerative changes in the lumbar spine and can result in symptoms such as lumbar pain, pain radiating from the lumbar region to the lower extremities, and neurogenic claudication. The diagnosis of LSS typically involves a combination of clinical evaluation and imaging studies, with Magnetic Resonance Imaging (MRI) being a key diagnostic tool (Tomkins-Lane et al., 2020).

Despite the availability of diagnostic tools, evaluating the severity of LSS continues to present challenges due to the absence of a single, universally accepted scale. As noted by Abou-Al-Shaar et al. (2018), there is substantial variability in the methodologies utilized to assess the severity of LSS, which contributes to inconsistencies in outcome measures across studies. Similarly, Tumko et al. (2024) highlights the necessity for more objective and standardized methods to classify the severity of LSS, particularly through the implementation of advanced imaging techniques.

To overcome these limitations, innovative solutions are being explored, with Artificial Intelligence (AI) and Machine Learning (ML) emerging as promising tools. For instance, Convolutional Neural Networks (CNN) have demonstrated high diagnostic accuracy, with some models achieving an Area Under the Receiver Operating Characteristic (AUROC) of over 90% in detecting LSS from radiographs and MRI scans (Bogdanovic et al., 2024).

1.1 Motivation

The motivation for this research stems from the need to improve the image diagnostic process for LSS, ultimately enhancing patient outcomes and reducing healthcare costs.

Lumbar Spinal Stenosis (LSS) represents a debilitating condition associated with a substantial reduction in quality of life, affecting 9% in the general population and up to 47% in people over 60 years of age (Tomkins-Lane et al., 2016). Diagnosis LSS remains challenging due to the dependence on subjective clinical evaluations and qualitative imaging interpretations. The variability in radiologists' assessments of spinal canal narrowing on MRI scans often leads to inconsistent diagnoses and treatment plans. This problem is exacerbated by the lack of standardized criteria for visual evaluations (Lønne et al., 2014). These inconsistencies can result in delayed or inappropriate interventions, further compromising patient outcomes (Schepper et al., 2015).

The integration of AI has been increasingly adopted in medical imaging, particularly in data-intensive domains like radiology, to improve efficiency and accuracy in diagnostic workflows (Ahmad et al., 2021). AI models, such as CNN, have shown high precision in tasks like segmentation and classification, reducing inter-reader variability and improving diagnostic consistency. The van der Graaf et al. (2024b) study demonstrated success in spinal imaging tasks such as vertebral segmentation and disc herniation detection, however is a notable lack of AI-driven tools specifically designed for the detection and severity classification of LSS. This gap presents a critical opportunity to leverage AI technologies, offering clinicians a standardized, reproducible framework for assessing LSS severity. Such advancements could transform patient care by enabling earlier, more accurate diagnoses, which may reduce long-term disability and associated healthcare burdens (Tumko et al., 2024).

1.2 Objectives

This thesis aims to develop and evaluate a Computer Vision (CV) and Deep Learning-Based (DL-based) framework for the automated identification and classification of Lumbar Spinal Stenosis (LSS) from MRI scans. The framework is designed to assist radiologists in improving diagnostic accuracy and efficiency. Specifically, this research pursues the following overall objectives:

Explore the State of the Art: Identify and analyze the current state of the art in Computer Vision (CV) and Deep Learning (DL) techniques applied to medical imaging, with a focus on their use in detecting and diagnosing lumbar spinal stenosis from MRI scans.

Develop Algorithms for Automatic Detection and Classification: Design and implement deep learning-based methods to enable the automatic detection of LSS in MRI scans. This includes:

- To identify and segment key anatomical structures in MRI images, particularly those associated with LSS;
- To reduce the workload and time demands on medical specialists by automating diagnostic tasks and enabling a more efficient workflow;
- To evaluate the proposed methods in terms of accuracy, sensitivity, and specificity, compared to expert diagnoses, to ensure clinical reliability and minimize subjectivity.

Prototype a Clinical Decision Support System: Develop a proof-of-concept prototype for a clinical decision support system aimed at assisting radiologists and healthcare professionals. This system provided automated analysis of MRI scans and offered a second opinion to improve diagnostic accuracy and support decision-making.

Evaluate the Framework's Effectiveness: Conduct initial testing to assess the performance, usability, and reliability of the proposed framework. This evaluation involved comparing automated results to expert interpretations, with the goal of validating its potential to enhance diagnostic processes and improve patient outcomes.

1.3 Scientific and Technical Contributions

This thesis makes the following computational and software engineering contributions in the context of medical imaging and Lumbar Spinal Stenosis (LSS) diagnosis:

- **Developed an Automated Deep Learning Framework:** A CV- and DL-based framework was created for the automated detection, classification, and segmentation of LSS from MRI scans. This framework enhanced diagnostic accuracy and efficiency by reducing reliance on subjective interpretations, minimizing human error, and streamlining radiologists' workflows. Additionally, a proof-of-concept prototype for

a clinical decision support system was developed, offering automated analysis and second-opinion functionality to improve decision-making and patient outcomes.

- **Reproducible LSS Severity Classification:** A more objective and reproducible method for classifying LSS severity was introduced, addressing the lack of a universally accepted scale. This standardization improved the comparison of research findings and supported the development of standardized treatment protocols.
- **Validated AI-Driven Tools in Medical Imaging:** The effectiveness of AI-driven tools in LSS diagnosis was rigorously evaluated. The framework's accuracy, sensitivity, and specificity were assessed against expert diagnoses, ensuring clinical reliability and demonstrating its potential for real-world application.

1.4 Structure of the Thesis

- **Chapter 1 (Introduction)**

  Introduces the research problem, focusing on the challenges in diagnosing and classifying Lumbar Spinal Stenosis (LSS) and the need for more objective and efficient solutions. Outlines the motivation behind the study, the research objectives, and the expected contributions. Provides an overview of the clinical and technical context of LSS diagnosis, highlighting limitations of current methods and the potential of Artificial Intelligence (AI) and Deep Learning (DL) to address these challenges.

- **Chapter 2 (State of The Art)**

  Reviews the relevant literature and technological advancements in medical imaging and AI-driven diagnostics. Provides an overview of existing Computer Vision (CV) and Deep Learning (DL) techniques applied to spinal imaging, with a focus on LSS detection and classification. Examines current challenges, gaps in the literature, and the potential of LSS to improve diagnostic accuracy and standardization.

- **Chapter 3 (Materials and Methods)**

  Details the design and development of the proposed Computer Vision (CV) and Deep Learning-Based (DL-based) framework for LSS diagnosis. Explains the methodology, including data collection, preprocessing, model selection, and training processes. Describes the development of a proof-of-concept prototype for a clinical decision support system, outlining its architecture, functionality, and integration with existing diagnostic workflows.

- **Chapter 4 (Results and Discussion)**

Presents the evaluation of the methods proposed for spinal stenosis classification and detection. First discusses the limitations observed in the initial proposed approach and then details the development and results of the final method, highlighting performance improvements and clinical relevance.

- **Chapter 5 (Conclusions and Future Work)**

  Summarizes the key contributions of the study, reflects on the limitations of the proposed methods, and outlines directions for future research and potential improvements.

Chapter 2

State of The Art

Reviews the clinical and technological foundations of Lumbar Spinal Stenosis (LSS) diagnosis, starting with clinical background, epidemiology, and current imaging modalities (Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and X-rays). Explores advancements in Artificial Intelligence (AI), focusing on deep learning techniques such as Convolutional Neural Networks (CNN) and emerging Computer Vision (CV) methodologies including vision transformers and Generative Adversarial Network (GAN). Analyzes their application to automating LSS detection, classification, and severity assessment. By integrating clinical challenges with AI progress, evaluates how these technologies address diagnostic variability, data scarcity, and the lack of standardized protocols, establishing the foundation for an automated framework to enhance diagnostic precision in lumbar spinal stenosis.

2.1 Lumbar Spinal Stenosis (LSS)

LSS is a degenerative condition characterized by the narrowing of the spinal canal or intervertebral foramina, leading to compression of the spinal cord and nerve roots. This narrowing is most commonly caused by age-related degenerative changes in the lumbar spine, including intervertebral disc degeneration, facet joint hypertrophy, and thickening of the ligamentum flavum (Katz and Harris, 2008). Less frequently, LSS may result from congenital abnormalities, trauma, or spondylolisthesis (Wu et al., 2024).

Figure 1 illustrates an axial view of a lumbar vertebra, comparing a normal spinal canal on the left with one affected by stenosis on the right. The left side shows a normal spinal canal and nerve root canals, which are unobstructed and provide sufficient space for the spinal cord and nerves. In contrast, the right side shows the pathological changes associated with lumbar spinal stenosis, where the spinal canal narrows due to facet joint hypertrophy, thickening of the ligaments, disc bulging, and bone spur formation. These degenerative changes compress the spinal cord and nerve roots, leading to inflammation and swelling of the affected nerves.

2.1.1 Clinical Background

The prevalence of LSS demonstrates a significant increase in individuals aged 60 years and above (Deyo et al., 2010), which accounts for its predominant occurrence among older adults (Ciol et al., 1996). Clinically, LSS is characterized by lower back pain and radiating leg pain, which may manifest unilaterally or bilaterally (Fritz et al., 1998). A hallmark symptom of LSS is neurogenic claudication, presenting as pain and weakness in the lower extremities during walking or standing, with symptoms typically alleviated by sitting or forward flexion (Watters et al., 2008). In rare cases, severe manifestations may include bladder or bowel dysfunction (Ishimoto et al., 2012).

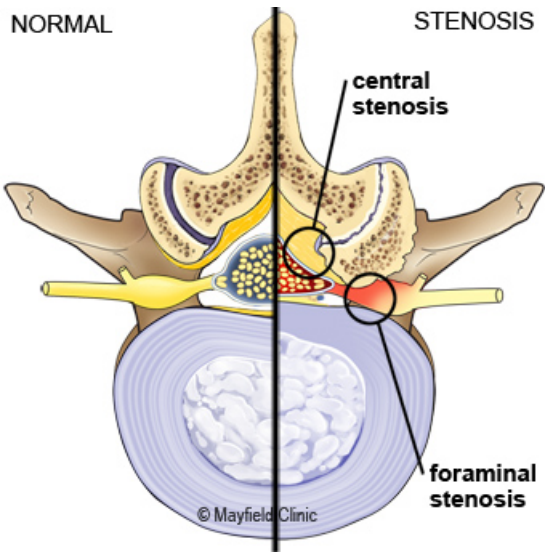

Figure 1: Axial plane of a lumbar vertebra (Bohinski, 2021).

The diagnosis of LSS is primarily established through clinical evaluation, which includes a detailed patient history and physical examination. Key diagnostic criteria include the presence of neurogenic claudication and relief of symptoms with forward flexion. It is crucial to differentiate LSS from other conditions with similar presentations, such as vascular claudication, which is associated with peripheral artery disease. Vascular claudication is typically relieved by rest rather than forward flexion and may be accompanied by diminished peripheral pulses and skin changes (Fritz et al., 1998). While imaging studies, such as MRI, are employed to confirm the diagnosis of LSS and evaluate the extent of spinal canal narrowing, it is noted that radiographic findings do not always correlate with the severity of clinical symptoms (Kalichman et al., 2009).

2.1.2 Imaging Techniques for LSS

Imaging plays a critical role in the diagnosis and management of LSS, offering detailed visualization of the spinal canal and surrounding structures. The most commonly utilized imaging modalities include X-rays, CT, and MRI (Steurer et al., 2011). Each technique possesses distinct strengths and limitations, with MRI being widely regarded as the gold standard for diagnosing LSS (Schepper et al., 2015).

X-rays are frequently employed as the initial imaging modality in the evaluation of LSS due to their widespread availability and cost-effectiveness (Kalichman et al., 2009). They provide essential information regarding spinal alignment, degenerative changes such as osteophytes and disc space narrowing, and conditions like spondylolisthesis (Cheung et al., 2014). However, X-rays are limited in their ability to visualize soft tissues, including the spinal cord, nerve roots, and intervertebral discs, rendering them insufficient for confirming the diagnosis of LSS. Despite these limitations, X-rays remain a valuable tool for initial assessment and for ruling out other spinal pathologies (Ciol et al., 1996).

CT scans offer detailed imaging of bony structures, making them particularly useful

for identifying structural causes of spinal stenosis, such as facet joint hypertrophy and osteophytes. CT is often utilized as an alternative to MRI in patients who cannot undergo MRI due to contraindications, such as the presence of certain implants or severe claustrophobia. CT myelography, which involves the injection of contrast dye into the spinal canal, can provide additional insights into spinal canal compression and is particularly valuable when MRI is not feasible. However, CT exposes patients to ionizing radiation and is less effective than MRI in detecting soft tissue abnormalities (Schepper et al., 2015).

MRI provides superior soft-tissue contrast, enabling detailed visualization of the spinal cord, nerve roots, intervertebral discs, and ligamentum flavum. Thickening of the ligamentum flavum reduces canal volume, compressing nerve roots and causing inflammation (Watters et al., 2008). It further permits integrated assessment of stenotic severity, disc pathology, and neural element compromise, supplying critical information for diagnosis and treatment planning (Kim et al., 2015).

Figure 2 shows a sagittal plane MRI image demonstrating lumbar spinal pathology, including anterior disc degeneration and hypertrophy of the posterior ligamentum flavum. These changes contribute to severe spinal canal stenosis at the L2-L3 and L4-L5 levels, with milder narrowing observed at L3-L4. The stenosis is most pronounced at L4-L5, where the anteroposterior diameter of the spinal canal is markedly reduced, consistent with advanced LSS.

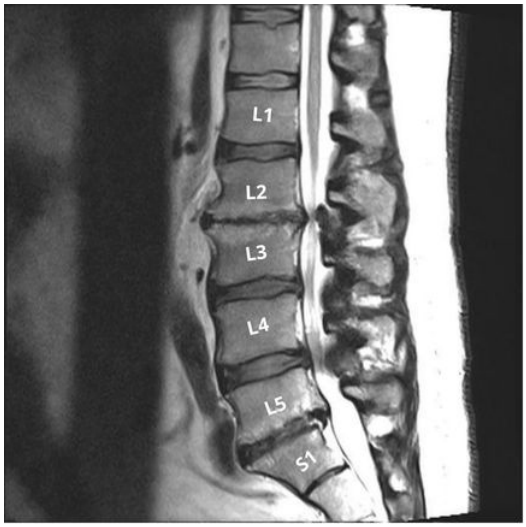

Figure 2: Sagittal plane of MRI (RSNA, 2024).

MRI is non-invasive, does not involve ionizing radiation, and can accurately differentiate LSS from other conditions with similar symptoms, such as tumors or infections (Cheung et al., 2014). Even with the disadvantages, MRI remains the most reliable and widely used imaging modality for LSS, offering unparalleled insights into spinal pathology (Moses et al., 2015).

2.2 Artificial Intelligence (AI)

AI encompasses computational methodologies that enable machines to perform tasks requiring human-like reasoning, such as pattern recognition, decision-making, and predictive analysis (Wang et al., 2023b). In medical imaging, AI has revolutionized diagnostic workflows by automating complex analyses of high-dimensional data, including MRI and CT scans (Karthik et al., 2024). These systems excel at identifying subtle anatomical and pathological features—such as tissue abnormalities or structural deformities—with high precision, achieving diagnostic accuracies of up to 90% in disease detection and demonstrating high sensitivity in detecting conditions like lung nodules or diabetic retinopathy (Aggarwal et al., 2021). By minimizing subjectivity and inter-observer variability, AI enhances diagnostic consistency, a critical advantage in conditions where qualitative assessments dominate clinical practice. For example, AI tools like icolung® quantify interstitial lung disease progression in systemic sclerosis patients, correlating imaging biomarkers with pulmonary function tests to reduce variability in radiological assessments (Guiot et al., 2025).

The evolution of AI in medical imaging has been propelled by advancements in computational power, algorithmic innovation, and the availability of large annotated datasets. Early AI applications in the 1960s–1980s focused on rule-based systems like Computer-Aided Diagnosis (CADx) for basic image processing tasks such as chest X-ray and mammography analysis (Buaka and Moid, 2024). However, the advent of Machine Learning (ML) and Deep Learning (DL) has enabled more sophisticated analyses, including automated segmentation, classification, and predictive modeling (Liu et al., 2024).

2.2.1 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) are computational models designed to simulate the operational principles of biological neural systems. These systems aim to execute computational tasks with enhanced efficiency compared to conventional architectures. ANNs represent efficient computational frameworks whose foundational premise derives from biological neural networks. In such architectures, neurons are interconnected through synaptic connections, each assigned a synaptic weight that encodes information related to input signals (Ansari, 2020).

Inspired by biological neural networks in both structural design and activation mechanisms, the artificial neuron was first conceptualized by McCulloch and Pitts (1943) as a binary computational unit with inputs, outputs, and a fixed activation threshold, establishing the foundational mathematical model for neural computation. Haykin (2001) describes this computational unit as a critical component for information processing within neural networks, outlining three fundamental components of an artificial neuron:

1. A set of synaptic connections, each characterized by a specific weight or strength. A key distinction from biological neural synapses lies in their operational range: artificial synapses may assume positive or negative values, whereas biological synapses predominantly operate through positive excitatory mechanisms;

2. A summation mechanism that calculates the weighted sum of input signals, performed by aggregating the products of input values and their assigned synaptic weights;

3. A nonlinear activation function, also termed a restrictive function, employed to confine the amplitude of the output signal to a finite permissible range. This ensures normalization of the output signal and introduces critical nonlinearity to the computational process.

The schematic representation of an artificial neuron, illustrated in Figure 3, comprises input signals denoted as x_1, x_2, ..., x_n. Synaptic weights associated with the neuron k are represented by W_{k1}, W_{k2}, ..., W_{kn}. A bias term B_k, introduced externally to the summation operator, modulates the net input to the activation function by either elevating or suppressing its value. The linear combiner output U_k, derived from the weighted sum of inputs and bias, is subsequently propagated through the activation function, represented by $\varphi(.)$. The final output of the neuron, Y_k, corresponds to the post-activation value.

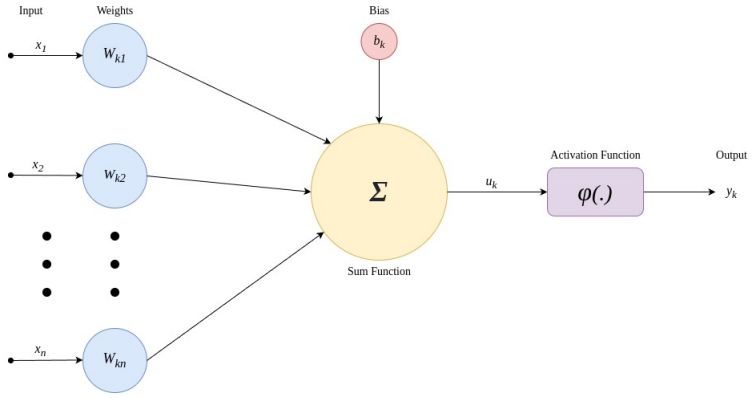

Figure 3: Artificial neuron example.

Such networks are typically organized into multilayer architectures, comprising an input layer for data transmission without direct processing, hidden layers responsible for computational transformations in function approximation, and an output layer delivering processed results (Livshin, 2019). Neurons within hidden and output layers modulate incoming signals through adjustable synaptic weights, while the input layer serve as an interface to propagate unaltered signal components (Hristev, 1990).

ANN may incorporate an indefinite number of neurons organized into interlinked layers, with the input layer responsible for representing the dataset and initial conditions, typically excluded from subsequent layers, as in applications such as image processing, their output may correspond directly to individual pixel values (Vasilev et al., 2019). In fully connected multilayer networks, as demonstrated in Figure 4, each neuron within a given layer is interconnected to all neurons in adjacent layers, resulting in high computational costs. However, the presence of repetitive patterns in data such as images enables the implementation of smaller detectors, trained to identify these recurrences across distinct regions, thereby optimizing computational efficiency without sacrificing model performance (Teoh, 2023).

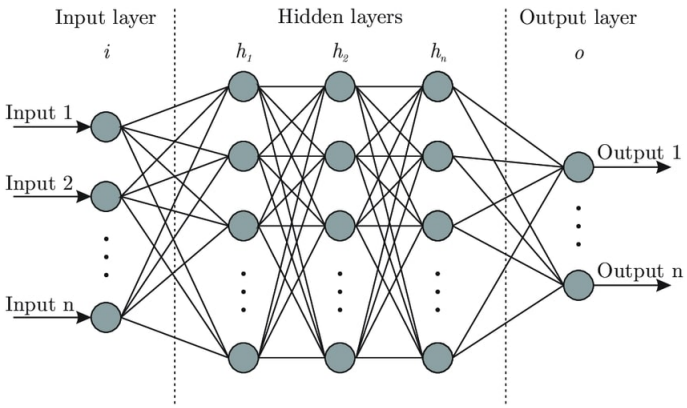

Figure 4: Fully connected multilayer networks example (Bre et al., 2018).

2.2.2 Deep Learning (DL)

DL, a specialized subset of ML, utilizes hierarchical architectures of ANN to autonomously extract high-level features from raw data through successive nonlinear transformations. Unlike traditional ML models, which depend on manual feature engineering, DL algorithms iteratively optimize their internal representations during training, as illustrated in Figure 5. This capability enables the identification of complex patterns in medical imaging data, including spatial hierarchies in volumetric scans and temporal dependencies in dynamic imaging sequences (Lecun et al., 2015).

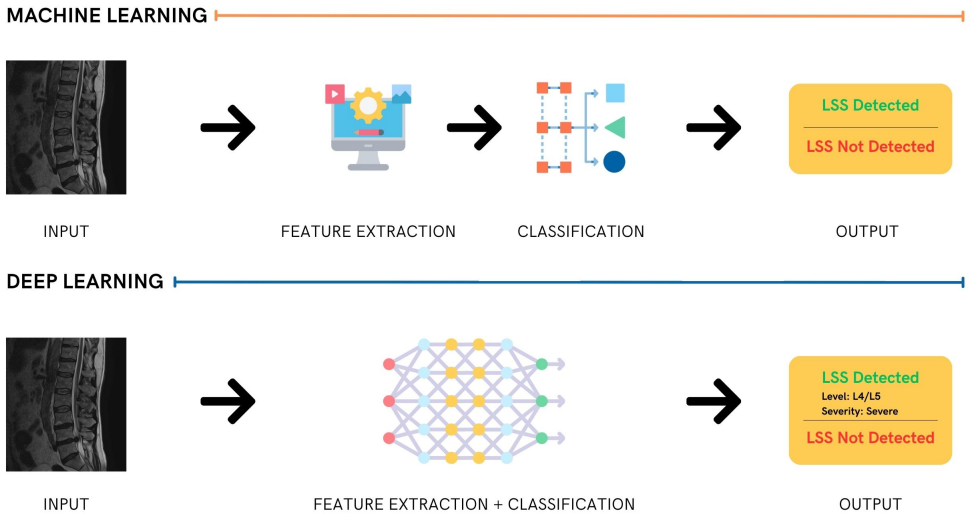

Figure 5: Comparison of ML and DL workflows.

This capacity is particularly transformative in clinical scenarios requiring the integration of multifactorial diagnostic criteria, such as differentiating malignant tumors from benign lesions based on heterogeneous imaging textures or predicting disease progression via longitudinal imaging biomarkers. For instance, Ronneberger et al. (2015) demonstrated

that DL architectures such as U-Net have significantly advanced biomedical image segmentation, enabling precise anatomical localization. Similarly, Goodfellow et al. (2014) highlighted that synthetic image generation using GAN facilitates the augmentation of datasets for rare pathologies, addressing challenges related to data scarcity in medical imaging research.

The efficacy of DL in medical imaging arises from its scalability with complex datasets and capacity to generalize across diverse imaging modalities. However, its performance depends on access to large-scale annotated datasets and rigorously optimized training protocols to minimize overfitting in high-dimensional parameter spaces. Techniques such as transfer learning—leveraging pre-trained models on non-medical datasets and fine-tuning them for domain-specific tasks—have proven effective in addressing data scarcity. For example, Esteva et al. (2017) achieved dermatologist-level accuracy in skin cancer classification using a CNN trained on 129,450 clinical images, demonstrating the utility of this approach. Attention mechanisms, which focus computational resources on diagnostically salient regions, further mitigate these challenges while improving model interpretability.

The Figure 6 illustrates the principle of transfer learning in DL. A base model pre-trained on a large-scale dataset transfers its learned weights, network parameters, to initialize a custom model designed for a specialized task. During adaptation, frozen layers retain the transferred weights to preserve generic feature extraction capabilities, while fine-tuned layers are retrained on task-specific data to adapt hierarchical representations. The fine-tuning process may involve hyperparameter adjustments, including learning rate reduction, epoch configuration, or selective layer unfreezing, to optimize convergence on the target dataset. By reusing and refining pre-trained weights, transfer learning reduces computational overhead and alleviates data scarcity, thereby enabling robust performance.

Self-supervised learning frameworks, such as Contrastive Predictive Coding (CPC), mitigate reliance on labeled data by exploiting inherent structures in high-dimensional signals through contrastive loss and autoregressive modeling, as demonstrated by van den Oord et al. (2018) across domains including speech, images, text, and reinforcement learning. These frameworks encode latent representations designed to maximize mutual information between contextual observations. Parallel advancements include contrastive methods like SimCLR, proposed by Chen et al. (2020), which enhances representation learning through comprehensive data augmentations such as random cropping and color distortion, coupled with nonlinear projection heads. This approach achieves state-of-the-art performance in unsupervised and semi-supervised learning paradigms by refining feature discriminability in embedding spaces.

2.2.3 Convolutional Neural Networks (CNN)

A CNN is an approach based on training and extracting significant features from input images, functioning similarly to a biological neural network. This capability enables the CNN to identify portions of an image and recognize distinct features that remain consistent even in the presence of transformations such as rotation, scaling, and translation (Teoh, 2023). The core operation of the CNN is convolution, which involves two one-dimensional functions of a real-valued argument and can be represented by an asterisk, as shown in Equation 2.1. In this context, the function f represents the input data, while

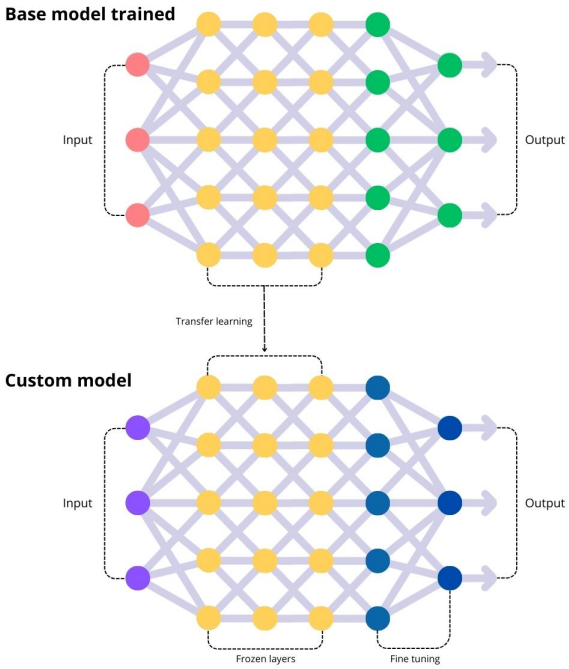

Figure 6: Transfer learning workflow example.

the argument g is known as the kernel, whose output results in a feature map (Goodfellow et al., 2016).

$$s(t) = (f * g)(t) \tag{2.1}$$

The Figure 7 from Goodfellow et al. (2016) illustrates the convolutional operation applied to a two-dimensional input array, such as a grayscale image. A small kernel or filter, represented as a weight matrix, is systematically slid across the input tensor with a predefined stride, performing element-wise multiplication between the kernel values and the corresponding local region of the input at each position. The products are aggregated through summation to produce a single scalar value in the output feature map, a process mathematically defined as discrete convolution. This spatial invariance is emphasized by the kernel's shared parameters, which detect localized features—edges, textures, or patterns—regardless of their position.

Pooling is a downsampling operation in CNNs that reduces spatial dimensions while retaining critical features. Common types include max pooling, which selects maximum values from sub-regions, and average pooling, which computes regional averages (LeCun et al., 1989). Region Of Interest (ROI) Pooling is a specialized technique used to standardize variable-sized candidate regions into fixed-size feature maps. Dividing each region proposal into a grid of bins, applies max pooling within each bin, and concatenates the results to preserve spatial hierarchies, as exemplified in Figure 8. Unlike standard pooling, which operates on entire images, ROI Pooling focuses on regions extracted from feature maps. This enables efficient training and inference by maintaining critical geological fea-

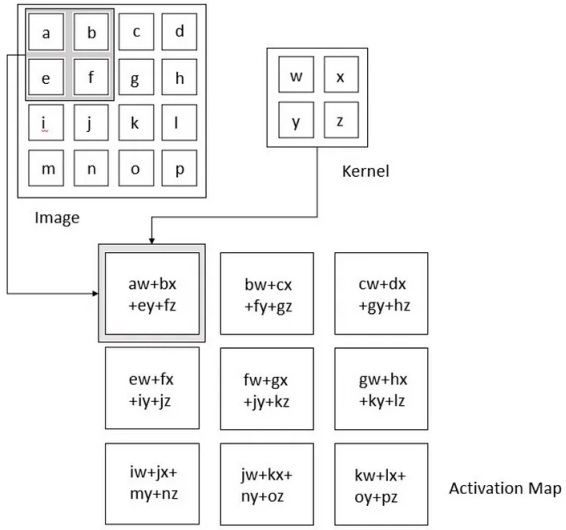

Figure 7: Convolutional operation example (Goodfellow et al., 2016).

tures, such as rock textures or layer boundaries, in variable-sized regions (Girshick, 2015).

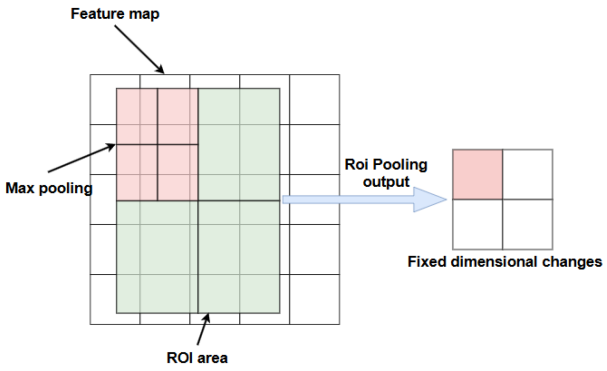

Figure 8: ROI pooling diagram (Fu and Wang, 2024).

The layers of a CNN play a crucial role in its architecture, consisting of filters, also referred to as kernels or feature detectors, which are applied to all regions of the input data. Each filter can be understood as a set of weights that are adjusted during training (Vasilev et al., 2019). According to Teoh (2023), the most common CNN architecture comprises four main layers, as illustrated in Figure 9. The convolutional layer extracts essential features through operations, while the pooling layer enhances efficiency by reducing the matrix size. The convolutional and pooling layers form the network's filter, whereas the classifier is composed of fully connected layers.

Recent studies highlight the efficacy of CNN s in the diagnosis of LSS. For example, Bogdanovic et al. (2024) developed a fully automated CNN model to measure spinal stenosis on MRI, achieving diagnostic accuracy comparable to expert radiologists. Similarly, Tumko et al. (2024) proposed a neural network framework for LSS classification, leveraging multi-planar MRI inputs to capture axial and sagittal views of spinal compres-

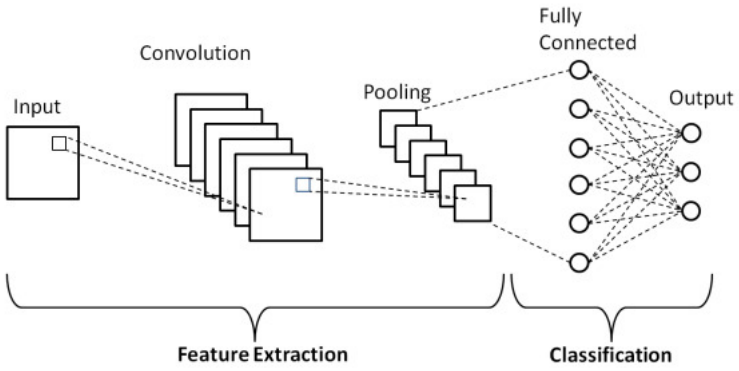

Figure 9: CNN layers example (Phung and Rhee, 2019).

sion. Challenges persist, however, including the need for large, diverse datasets to ensure model generalizability and address class imbalance in stenosis severity grades.

2.2.4 You Only Look Once (YOLO)

Object detectors based on CNN are widely used in recommendation systems. For instance, a system that detects empty parking spots using urban surveillance cameras is typically powered by models that are often slow and have low accuracy. Improving the accuracy of real-time object detectors is extremely important for managing autonomous processes and reducing dependence on human resources (Bochkovskiy et al., 2020).

In order to improve accuracy and reduce processing time in real-time object detection, Redmon et al. (2015) developed a framework called You Only Look Once (YOLO). The name refers to the fact that the framework looks at the image only once to predict what objects are present and where they are located, essentially framing object detection as a regression problem. YOLO works in a straightforward prediction-based manner, where a single CNN simultaneously predicts bounding boxes and the class probabilities for those boxes. Training is performed directly on full images, which enhances detection performance.

The main idea behind YOLO is to divide the input image into an SxS grid and perform detections in each of the grid cells. In each cell, YOLO predicts B bounding boxes along with a confidence score for each box. The confidence score is computed using Equation 2.2, indicating whether an object is present in the cell. The higher the confidence, the thicker the box border. Furthermore, each cell predicts the probability distribution over the C possible classes for the object. Each cell results in $5 + C$ values, including the coordinate pair *(x,y)*, the box dimensions *(w,h)*, the confidence score, and the class C. Based on these predictions, the final image is generated with bounding boxes around the detected objects, as illustrated in Figure 10 (Liu et al., 2020).

$$Confidence = Pr(Object) \cdot IoU(Gt, pred) \tag{2.2}$$

Over the years, YOLO has undergone several updates, improvements, and architec-

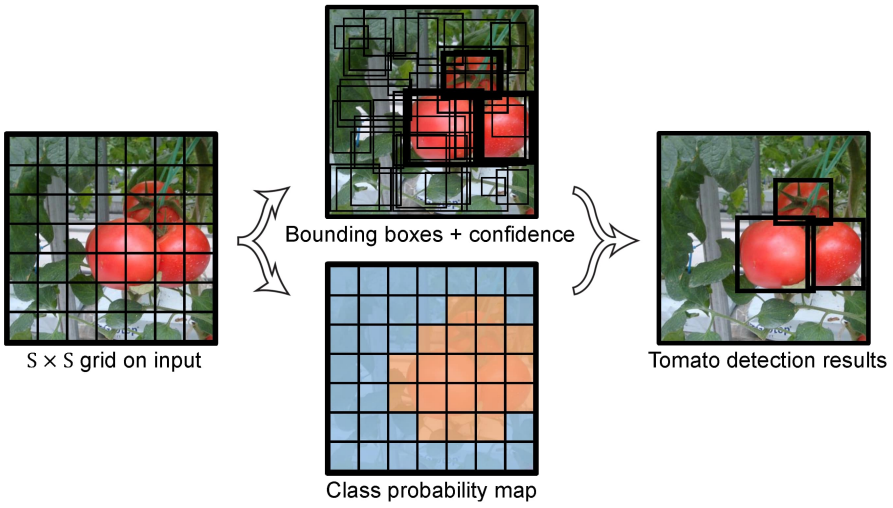

Figure 10: You Only Look Once (YOLO) workflow Liu et al. (2020).

tural changes. Some of these versions were developed by the original creators with the ongoing goal of enhancing detection performance and improving training, Mean Average Precision (mAP), and Frames Per Second (FPS). Additionally, other versions were created by different contributors, such as YOLOv4, developed by Bochkovskiy et al. (2020), and YOLOv7, developed by Wang et al. (2022). Figure 11 shows the evolution of the YOLO versions over time.

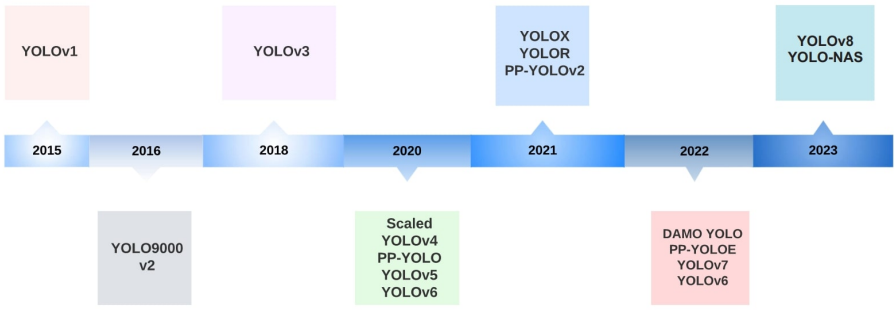

Figure 11: Timeline of YOLO versions Terven and Cordova-Esparza (2023).

A recent version of the YOLO family is You Only Look Once Version 8 (YOLOv8), the architecture of You Only Look Once Version 8 (YOLOv8) can be subdivided into two distinct parts, as illustrated in Figure 12. The first part, called the backbone, is a modified version of the CSPDarknet53 architecture. This part consists of a 53-layer convolutional network and employs cross-stage connections to improve the flow between different layers. The second part is referred to as the head, and it is composed of several convolutional layers followed by fully connected layers. These layers are responsible for predicting bounding boxes, objectness scores, and class probabilities for objects detected in the image (Mehra, 2023).

The convolutional layers in YOLOv8 are Conv2D layers, meaning they are two-

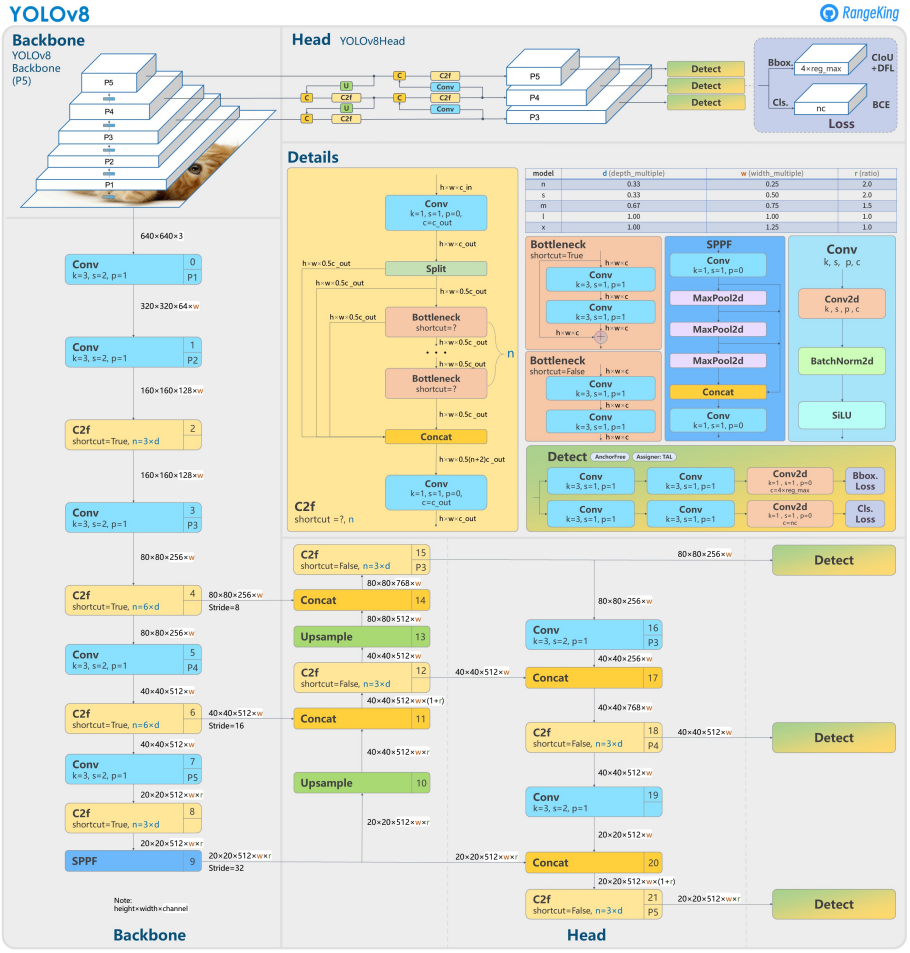

Figure 12: YOLOv8 Architecture (RangeKing, 2023).

dimensional convolutional layers. Essentially, the kernel or filter of the layer slides over the 2D input data, performing element-wise multiplication. The output is the sum of these results in a single pixel. Additionally, YOLOv8 includes the *BatchNorm2D* layer, a ANN layer that normalizes activations between the 2D layers. The activation function used is Sigmoid Linear Unit (SiLU) (Databricks, 2023).

This version of YOLO begins with a convolutional layer in the backbone that uses a 3×3 filter. Unlike some previous versions, which featured Cross Stage Partial (CSP) Bottleneck modules with 3 convolutional layers—known as C3 modules—the YOLOv8 implementation uses C2f modules, which are CSP Bottleneck modules with 2 convolutional layers. These offer increased speed compared to the standard C2. The Bottleneck consists of two 3×3 convolutional layers with residual connections (Buhl, 2023).

Detection in YOLOv8 is performed using a decoupled head, which consists of two separate branches: one for object classification and another for bounding box prediction through regression. It uses different loss functions: for the classification task, it employs the Binary Cross-Entropy Loss (BCEL), and for bounding box prediction, it uses the Distribution Focal Loss (DFL) and CIoU Loss. YOLOv8 is an anchor-free detection model, meaning it predicts the center of an object directly, making the detection process

more flexible and efficient (Wang et al., 2023a).

The YOLOv8 architecture also includes parameters for the convolutional layers, called Conv, such as k for kernel size, s for stride, p for padding, and c for the number of channels. The backbone concludes with a technique called Spatial Pyramid Pooling with Fused Features (SPPF), which splits the output of the layer into several sub-regions. The head also includes Concat layers, used to concatenate data from different sources, and Upsample layers, which increase the spatial resolution of the data. Finally, the architecture ends with detection layers, known as detect, composed of parallel Conv and Conv2D layers that compute the loss and generate the bounding boxes for the detected objects.

2.2.5 Emerging AI Techniques in Computer Vision

Recent advancements in CV have expanded the capabilities of AI in medical imaging, offering novel solutions to challenges such as data scarcity, diagnostic variability, and anatomical complexity inherent to LSS diagnosis (Esteva et al., 2017). CNN enhanced with self-attention mechanisms have demonstrated significant improvements in feature extraction and global contextual analysis. For instance, hybrid architectures like the Multi-Headed Self-Attention Module (MHSAM) refine CNN-based models by dynamically weighting diagnostically salient regions, enabling holistic analysis of spinal structures across MRI slices (Lin et al., 2024).

Transformers are self-attention-based architectures that have emerged as the predominant model in Natural Language Processing (NLP), as demonstrated by Vaswani et al. (2017). Their computational efficiency and scalability enable the training of models with unprecedented scale, exceeding 100 billion parameters (Brown et al., 2020). In CV, CNN architectures remain dominant; however, inspired by the success of Transformers in NLP, Dosovitskiy et al. (2020) introduced the Vision Transformer (ViT), which applies a pure Transformer architecture directly to image recognition tasks.

The ViT splits an image into fixed-size patches, as illustrated on Figure 13, linearly embeds them into tokens, and processes the sequence with a standard Transformer encoder. Positional embeddings are added to retain spatial information, and a learnable `[class]` token aggregates global features for classification. Unlike CNNs, the ViT lacks inherent inductive biases, such as locality and translation equivariance, requiring large-scale pretraining to achieve state-of-the-art performance. When transferred to downstream tasks, ViT matches or surpasses CNNs while using fewer computational resources, demonstrating that Transformers can excel in vision without convolutional operations (Dosovitskiy et al., 2020).

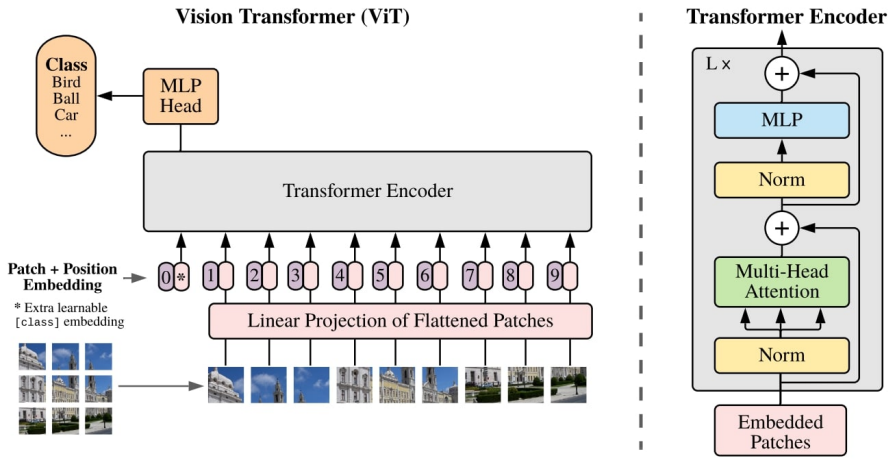

Figure 13: ViT model overview (Dosovitskiy et al., 2020).

In parallel, GANs have emerged as a transformative tool in medical imaging, particularly in addressing the challenge of limited annotated datasets. GANs consist of two neural networks: a generator and a discriminator, as shown in Figure 14. The generator creates synthetic data that mimics real data, while the discriminator evaluates the quality of generateed data, classifying them as real or synthetic generated data. This adversarial process enables the generation of high-quality synthetic images, which can significantly augment limited training datasets—a critical advantage in medical imaging, where annotated Low-Sample-Size datasets are often scarce (Rusanov et al., 2022).

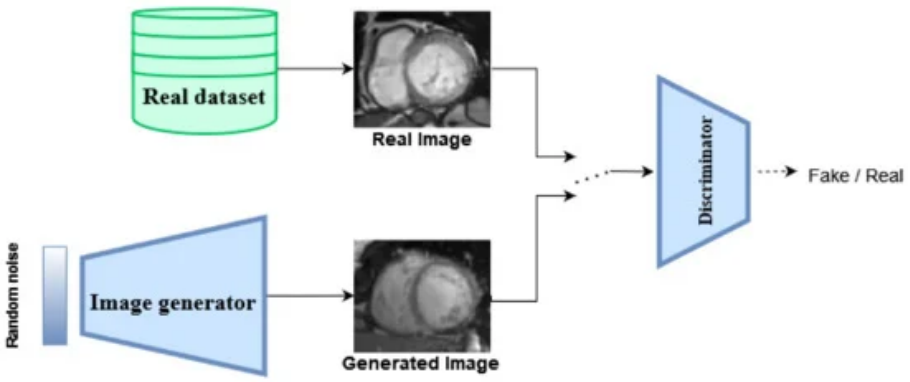

Figure 14: Flowchart of a GAN architecture (Skandarani et al., 2023) .

Faster Region-based Convolutional Neural Network (Faster R-CNN) is favored over single-stage detectors like YOLO due to its higher accuracy, particularly in applications like medical image analysis. As shown in Figure 15, this two-stage approach first uses a Region Proposal Network (RPN) to generate potential ROI by sliding over the feature map and proposing candidate bounding boxes. These initial proposals are then refined and classified in the second stage by a Faster R-CNN detection network, which performs both classification and bounding box regression. This process allows Faster R-CNN to

achieve precise localization of abnormalities in medical images, ensuring high detection accuracy (Ma et al., 2023).

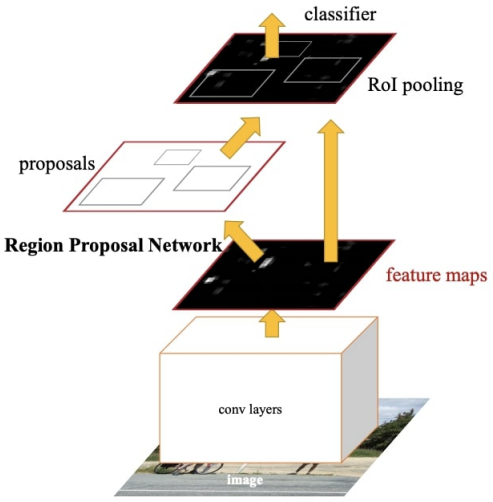

Figure 15: Faster R-CNN workflow example (Ren et al., 2015).

Chapter 3

Materials and Methods

Based on the conducted study on LSS, ANN, and their associated concepts and methodologies, it is essential to establish the process that was adopted to achieve the established objectives. Therefore, this chapter presents the materials and methods that were employed throughout the development of the framework.

3.1 Dataset Description

The Radiological Society of North America (RSNA) 2024 Kaggle competition[1] dataset is the primary source for training the AI model. The dataset is designed to identify degenerative lumbar spine conditions in MRI scans. It contains multi-institutional data, compiled from eight medical institutions across five continents, providing a diverse and robust set of MRI images of the lumbar spine. The dataset includes five distinct lumbar spine degenerative conditions:

- Left Neural Foraminal Narrowing
- Right Neural Foraminal Narrowing
- Left Subarticular Stenosis
- Right Subarticular Stenosis
- Spinal Canal Stenosis

Each condition is evaluated at five levels of the intervertebral disc (L1 / L2, L2 / L3, L3 / L4, L4 / L5 and L5 / S1), with severity scores classified into three levels: Normal / Mild, Moderate, and Severe, supplemented by spatial coordinates delineating the anatomical center of the pathological region. The dataset consist with three different plane MRI images: Axial T2, Sagittal T1 and Sagittal T2.

The dataset is structured based on studies and series, with each study corresponding to an individual patient's MRI examination. Multiple series may be associated with each study, each representing distinct sequences or views of the patient's spinal condition. The dataset comprises 1,975 studies and 6,294 series. Each series consists of between 5 and 192 sequential images, as is typical in MRI studies, with the images representing slices of the lumbar spine. In total, the dataset contains 147,218 DICOM files. Figure 16 presents an example study from the dataset, with all view planes corresponding to the same patient and extracted from the center slice of the MRI volume. Panel (a) shows an axial T1-weighted image; panel (b) a sagittal T1-weighted image; and panel (c) a sagittal T2-weighted image.

[1] `https://www.kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification`

Figure 16: Study MRI scans from RSNA dataset.

3.2 Clinical Collaboration

A formal interdisciplinary collaboration was established between the Instituto Politécnico de Setúbal (IPS) and Hospital da Luz Setúbal, a tertiary healthcare institution specializing in neurosurgical care, to advance translational research in LSS. The Neurosurgery Department at Hospital da Luz Setúbal maintains a clinical focus on spinal pathology, with expertise in the diagnosis and management of degenerative spinal disorders. This department actively pursues research initiatives integrating multimodal gait analysis, advanced neuroimaging, and computational analytics into standard clinical workflows to develop predictive frameworks for personalized therapeutic decision-making.

The heterogeneity in methodological approaches to assessing LSS severity—a critical factor influencing treatment outcomes—has been identified as a source of variability in cross-study comparability. To mitigate these limitations, this collaboration prioritized the implementation of standardized, imaging-based diagnostic protocols augmented by CV and DL algorithms. These computational methodologies aim to quantify pathoanatomical features, such as dural sac cross-sectional area, foraminal dimensions, and correlate them with clinical symptomatology, thereby enabling objective stratification of disease severity.

The partnership operationalizes a synergistic integration of technical and clinical expertise to ensure translational applicability of the proposed framework within existing diagnostic workflows. For this thesis, no patient data from Hospital da Luz Setúbal were accessed or used; model development and experiments were conducted exclusively on publicly available, de-identified datasets, and no protected health information was handled. Any future studies that involve hospital data will require secure, GDPR-compliant pipelines (including DICOM anonymization and pseudonymization), restricted access for authorized personnel only, and prior approval by the Hospital da Luz Setúbal Institutional Review Board (IRB), in accordance with the Declaration of Helsinki.

By leveraging IPS's computational resources and the hospital's ongoing clinical expertise, this initiative seeks to establish a robust, clinically informed predictive framework and evaluation criteria, with prospective validation on institutional data contingent on future approvals. The resultant framework is anticipated to contribute to evidence-based standardization in LSS management, aligning with the broader objectives of precision medicine in neurosurgical care.

3.3 Proposed Method

During the initial phases of this research, an early methodological approach was developed and implemented based on preliminary insights and assumptions. This initial method is presented in this section to document the reasoning, design, and limitations that ultimately led to a revised strategy. While this proposed method did not yield optimal results or fully meet the study objectives, it played a critical role in shaping the final approach. The inclusion of both methods in this thesis aims to provide a transparent account of the research process and highlight the iterative nature of methodological development.

The workflow presented in Figure 17 systematically delineates the study methodology, beginning with the ingestion of the dataset and progressing through a structured data preparation phase, which comprises the filtering, preprocessing, and generation of the bounding box. Following data preparation, model development is conducted through a sequential process involving transfer learning, hyperparameter definition, preliminary testing, and full-scale training with fine-tuning. The final phase involves a comprehensive evaluation of the model.

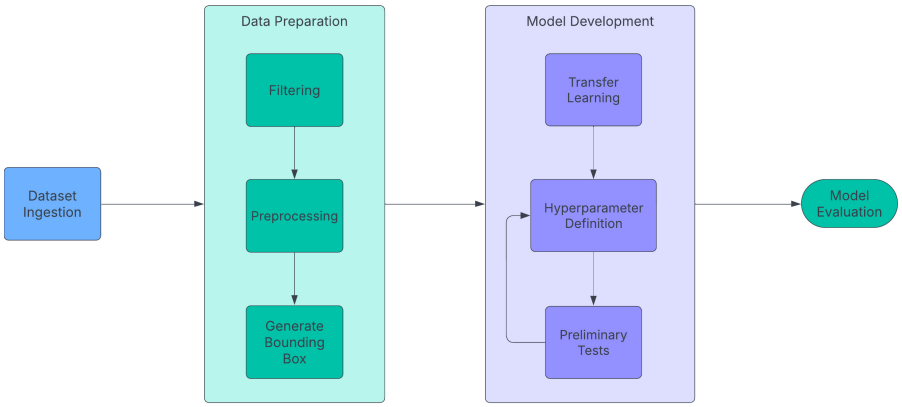

Figure 17: Overall proposed method workflow.

3.3.1 Data Preparation

For this study, a subset of the RSNA dataset was utilized to train the model, specifically focusing on sagittal T2-weighted MRI images of patients diagnosed with spinal canal stenosis. The selection of T2-weighted sequences was based on their established efficacy in visualizing spinal canal narrowing, intervertebral disc degeneration, and other key features associated with LSS. These sequences were particularly advantageous due to their high contrast and sensitivity in detecting soft tissue abnormalities, which were critical for accurate diagnosis.

The dataset was filtered to include only studies meeting the predefined inclusion criteria, resulting in 1,973 studies with 1 series each. Each series consisted of between 8 and 29 sequential images, culminating in a training corpus of 33,554 DICOM files. These images

formed the basis for training the DL model developed for the detection and classification of LSS.

As shown in Figure 18, the preprocessing stage began with resizing images to a resolution of 512×512 pixels to ensure standardization. Subsequently, contrast enhancement was applied using Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve visualization and enhance the distinguishability of relevant features.

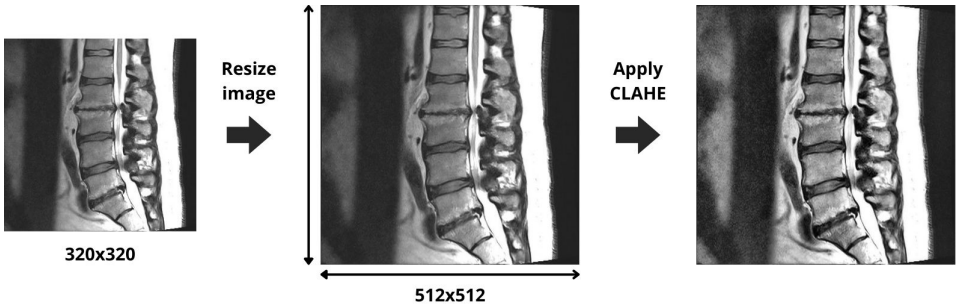

Figure 18: Preprocessing stage workflow.

Figure 19 illustrates the final step, where bounding boxes were generated using the spatial coordinates (X, Y) provided in the RSNA dataset. Each coordinate defined the center of a rectangular region measuring 70×50 pixels, ensuring that the pathological region was fully encompassed for model training. Each bounding box was labeled according to the severity classification annotated in the dataset. This process resulted in a fully annotated and structured dataset, optimized for the model development phase.

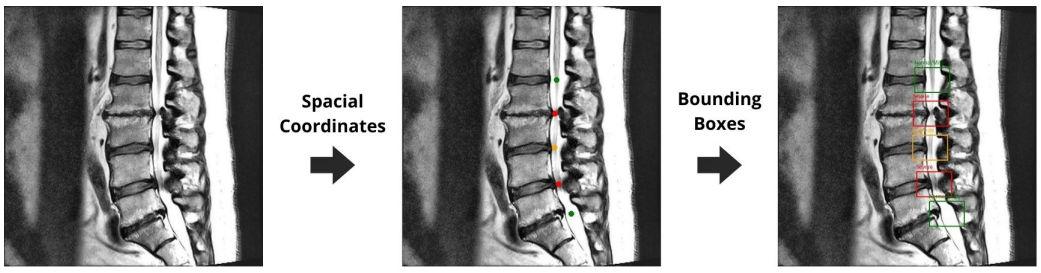

Figure 19: Generate bounding box stage workflow.

The dataset was partitioned into three subsets: 70% for training, 20% for validation, and 10% for testing. This division was selected to balance effective model learning, hyperparameter optimization, and unbiased performance assessment. Allocating the majority of the data to training exposed the model to sufficient examples to learn meaningful features and patterns. The validation subset was used during training to monitor performance, guide hyperparameter tuning, and mitigate overfitting through mechanisms such as early stopping and model selection. The test subset remained completely isolated from the training process and was reserved exclusively for the final evaluation of generalization on unseen data.

3.3.2 Model Development

The first stage involves using a pre-trained model provided by the Detectron2[2] framework, an open-source library developed by Facebook AI Research (FAIR) for object detection and segmentation tasks. The development and training of the model were conducted using Python[3], leveraging the computational resources of Google Colab Pro[4]. This cloud-based platform was selected for its high-performance GPU/TPU support, such as NVIDIA A100, V100, which accelerates the training of deep neural networks while minimizing hardware constraints.

Detectron2 was selected for its modular design, computational efficiency, and robust support for advanced architectures like Faster R-CNN, which are critical for handling the nuanced complexities of medical imaging (Wu et al., 2022).

The selected pre-trained model is a Faster R-CNN with a ResNet-50 Feature Pyramid Network (ResNet-50-FPN) backbone, which is an advanced DL model for object detection. It is an extension of the previous Region-based Convolutional Neural Network (R-CNN) architecture, and its key contribution is the introduction of a RPN, which significantly improves the speed and performance of object detection (Ren et al., 2015).

The model is initialized with pre-trained ImageNet weights, leveraging transfer learning to accelerate convergence and improve generalization. This approach is particularly effective in medical domains, where pre-training on large natural image datasets has been shown to enhance performance even with small target datasets (Tajbakhsh et al., 2016).

The framework employs a ResNet-101 Feature Pyramid Network (ResNet-101-FPN) backbone to power the Faster R-CNN model. ResNet-101's residual learning architecture mitigates the vanishing gradient problem in deep networks, enabling robust hierarchical feature extraction (He et al., 2016)—a critical capability for capturing the fine-grained details of spinal anatomy. The network was fine-tuned from pre-trained weights to adapt its feature representations to the spinal stenosis detection task. Hyperparameters and optimization settings were determined through empirical testing, with particular consideration given to the dataset's pronounced class imbalance.

The training configuration was designed to effectively adapt the pre-trained ResNet-101-FPN backbone to the specific characteristics of the spinal stenosis dataset. A batch size of 16 balanced GPU memory usage with gradient stability, while the AdamW optimizer combined adaptive learning rates with weight decay to improve generalization. The base learning rate of 0.0001 enabled stable fine-tuning without overwriting pre-trained features. Training was performed for 250000 iterations to ensure convergence for the complex detection task. To address the underrepresentation of Moderate and Severe cases, class weights of [4.0, 0.5, 4.0] were applied for the `Moderate`, `Normal/Mild`, and `Severe` classes respectively, increasing the loss contribution from minority classes and reducing that from the majority class (`Normal/Mild`).

Modifying these parameters after unfreezing the layers is crucial for tailoring the pre-trained model to the new task. Each task may have different requirements regarding data

[2] `https://ai.meta.com/tools/detectron2/`
[3] `https://www.python.org/`
[4] `https://colab.google/`

representations, model structure, and learning rates. Customizing these parameters helps the model adapt to the new task, learn relevant features, and improve overall performance (He et al., 2016).

3.3.3 Model Evaluation

The model evaluation is an important process, in this study some metrics were used to evaluate and help to translate the results of the model. The evaluation of performance in lumbar spinal stenosis localization necessitates the use of rigorously established metrics that quantify both spatial accuracy and consistency.

Precision quantifies the reliability of a model's positive predictions by measuring the proportion of correctly identified positive instances among all instances classified as positive. This metric is expressed in Equation 3.1, True Positives (TP) represent cases where the model correctly identifies a positive instance, while False Positives (FP) denote negative instances erroneously classified as positive. A high precision value indicates that the model minimizes incorrect positive predictions, ensuring confidence in its affirmative classifications (Faceli, 2011).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.1}$$

Recall evaluates a model's ability to capture all relevant positive instances within a dataset, thereby reflecting its sensitivity to true positives. It is defined in Equation 3.2, False Negatives (FN) correspond to positive instances that the model fails to detect. A high recall value signifies robust coverage of true positives, which is critical in applications where missing positive cases carries significant risks (Faceli, 2011). Unlike precision, recall prioritizes comprehensive identification of positives, often at the cost of increased false alarms.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.2}$$

The F1-score harmonizes precision and recall into a single metric by computing their harmonic mean. Given aggregated precision *Precision* and recall *Recall* for a classification task, the F1-score is defined in Equation 3.3. The harmonic mean penalizes large disparities between precision and recall, ensuring that both metrics are balanced and neither dominates the overall performance evaluation (Géron, 2019).

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.3}$$

The efficiency of a system is a critical factor to consider, with execution speed and runtime serving as key metrics for its assessment. The system's performance must meet predefined time constraints, which are primarily influenced by the underlying hardware configuration (Garcia-Garcia et al., 2017).

The clinical validity of the model was evaluated by comparing its predictions against annotations from a specialist physician, ensuring consistency with real-world diagnostic practices. Agreement metrics assessed how closely the model's outputs aligned with the specialist's clinical judgments, a critical benchmark for its potential use in healthcare workflows.

3.4 Final Method

A core scientific contribution of this work was the development of a revised and more effective approach, born from the insights and limitations encountered with the initially proposed method. As will be detailed in the Section 4.1, the initial single-stage model was hindered by the dataset's class imbalance and the inherent difficulty of performing detection and classification simultaneously. Therefore, the methodological pivot to a decoupled, two-stage framework is a key contribution that directly addresses these challenges. This final method, integrates improvements in data annotation and model design to successfully tackle the separate objectives of vertebra detection and spinal stenosis classification.

The workflow depicted in Figure 20 outlines the methodology employed in this study. It begins with dataset ingestion, followed by detection data preparation, which includes filtering, preprocessing and manual annotation of vertebrae. These annotations are then used to develop a vertebra detection model through transfer learning, hyperparameter tuning, and preliminary evaluation.

Simultaneously, the vertebra annotations are leveraged to extract ROI at the intervertebral disc level (e.g., L1/L2, L2/L3) for classification data preparation. The resulting dataset serves as input to the classification model, which is also developed using transfer learning, hyperparameter optimization, and initial testing. Both models are evaluated independently to assess their performance in vertebra detection and spinal stenosis classification tasks.

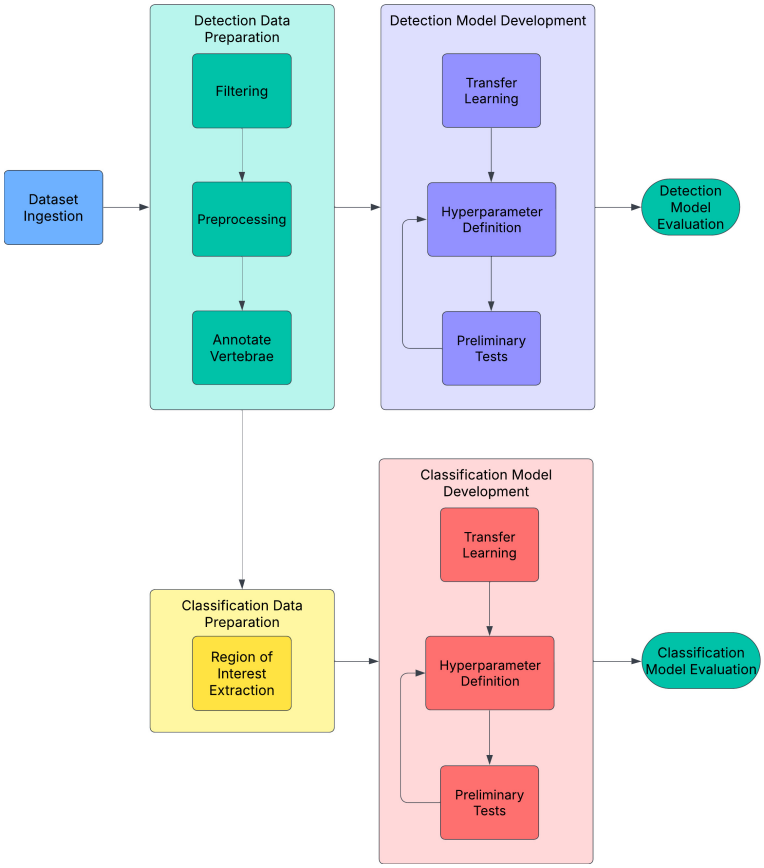

Figure 20: Overall final method workflow.

3.4.1 Detection Data Preparation

For the final method, the data preparation process was adapted from the initial approach described in Section 3.3.1, incorporating key modifications to better support vertebra detection. Table 3.1 summarizes the stepwise filtering of the dataset, starting from the original RSNA dataset and including the filtered subset introduced in Section 3.3.1.

In the final stage of dataset preparation, an additional filtering step was applied to exclude all series containing only Normal/Mild classifications. This process yielded a more balanced distribution across the severity classes and a total of 709 studies. Furthermore, only the central slice from each MRI series was retained, ensuring a clear and representative visualization of the spinal canal for model training. This selection resulted in a final dataset of 709 images.

As demonstrated in Figure 21, the pre-processing stage begins by resizing the images to a resolution of 640×640 pixels to ensure standardization. Subsequently, contrast enhancement is applied using CLAHE to improve visualization and enhance the distinguishability of relevant features. This resolution was chosen as a compromise between computational efficiency and detection accuracy, following the observations of the RA-YOLO study (Hao et al., 2023), which reports that larger input images can improve accuracy but significantly prolong training time, while smaller images accelerate training at the expense of accuracy.

Table 3.1: Dataset filtering steps and class distribution.

Filtering Step	Studies	Series	Images	Normal/Mild	Moderate	Severe
Original RSNA dataset	1,975	6,294	147,218	37,612	7,949	3,081
Filter T2 sagittal with stenosis diagnosis	1,973	1,973	33,554	8,535	730	468
Exclude series with only Normal/Mild labels	709	709	12,430	2,301	730	468
Select central MRI slice	709	709	709	2,301	730	468

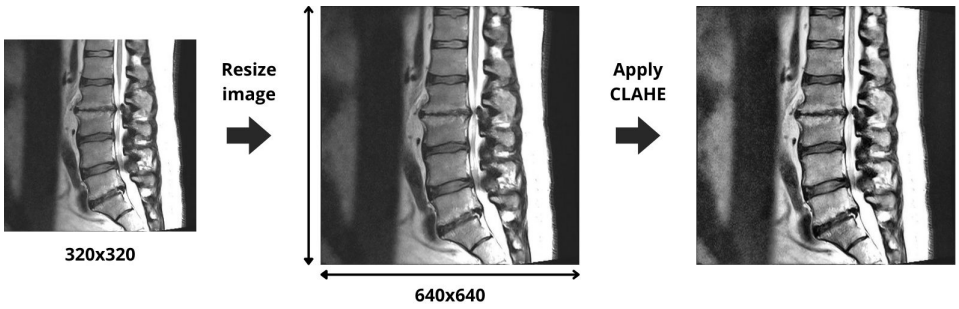

Figure 21: Preprocessing stage workflow.

Figure 22 illustrates the final step of the process, in which bounding boxes were annotated using the Computer Vision Annotation Tool (CVAT)[5], an open source, web-based platform developed by *Intel* for annotating image and video data in computer vision tasks, including object detection, classification, and segmentation. The annotated labels correspond to the vertebrae L1, L2, L3, L4, L5, and S1. This annotation process results in a fully structured and labeled dataset, optimized for the subsequent model development phase.

Finally, the annotated images and corresponding labels were organized into a dataset compatible with the YOLO Ultralytics detection format. To ensure a reliable evaluation protocol aligned with the full validation requirements of the framework, the data set was first divided by studies, assigning 80% of the studies for training and 20% for validation. This approach guarantees that images from the same study do not appear in both subsets, preventing data leakage and enhancing the generalization of the model. To facilitate reproducibility, the final dataset was uploaded to the Hugging Face Hub[6].

3.4.2 Detection Model Development

The selection of YOLOv8 for the vertebra detection stage was motivated by efficiency, architectural fit, and empirical evidence. As a single-stage detector, YOLOv8 is faster and less computationally intensive during both training and inference, aligning with the goal of an efficient, clinically practical workflow. Architecturally, its modern anchor-free design with a decoupled detection head provides greater flexibility across object sizes and

[5] `https://www.cvat.ai/`
[6] `https://huggingface.co`

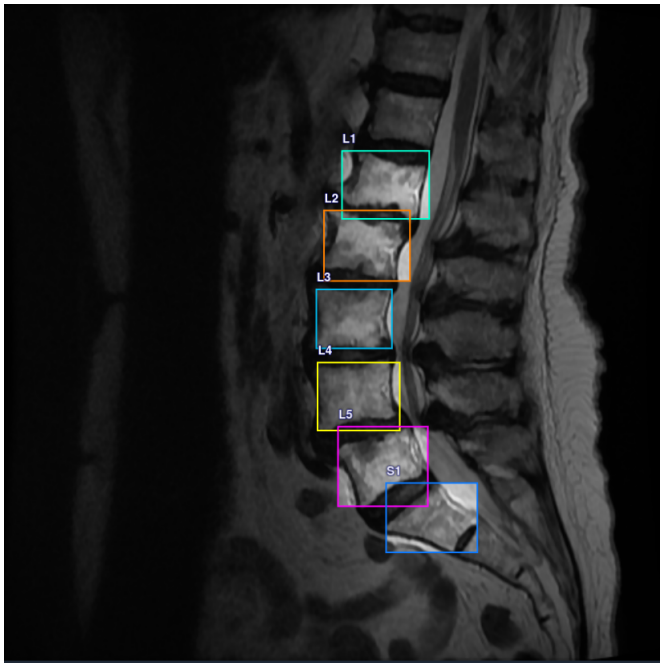

Figure 22: Example of an annotated image.

aspect ratios, which is well-suited to the consistent localization of vertebrae across diverse patient anatomies in sagittal lumbar MRI.

The first stage of the phase involved using a pre-trained model provided by the Ultralytics YOLOv8 framework[7]. Specifically, the `YOLOv8m` variant, the *medium* version, was selected as a balanced option between model complexity and performance, offering improved accuracy compared to the smaller variants while maintaining reasonable computational efficiency.

The model was fine-tuned on the dataset described in Section 3.4.1. Training was conducted for 200 epochs with an input image size of 640×640 pixels and a batch size of 16. The training process incorporated early stopping with a patience value of 20, meaning the training would halt if no improvement in validation performance was observed over 20 consecutive epochs. The dataset was provided in a custom YAML configuration file, and the training pipeline automatically handled data loading, augmentation, and evaluation.

3.4.3 Classification Data Preparation

The annotations described in Section 3.4.1 were used to construct the dataset for the vertebral level classification task. Each sample is generated by extracting a ROI corresponding to a specific intervertebral level (e.g., L1/L2, L2/L3). The cropping process begins with an assessment of the spine's curvature to determine whether it deviates laterally to the left or right. The centers of all annotated vertebrae are first extracted in normalized image coordinates. A second-degree polynomial is then fitted to these points,

[7] `https://yolov8.com`

with the vertical coordinate (y) defined as the independent variable and the horizontal coordinate (x) as the dependent variable. This fitting models the overall spinal alignment as a smooth curve extending from the uppermost to the lowermost vertebrae.

To quantify curvature, the fitted polynomial is compared with a straight-line approximation, and the sign of the quadratic coefficient is used to classify the direction of deviation: a positive coefficient indicates a leftward curvature, with the spinal canal positioned on the right side of the spine, whereas a negative coefficient indicates a rightward curvature, with the spinal canal positioned on the left side. This directional information is subsequently used to guide horizontal adjustments to the cropping position, ensuring that the ROI remains centered on the spinal canal even when lateral bending is present. An example of a leftward curvature with the canal positioned on the right is shown in Figure 23.

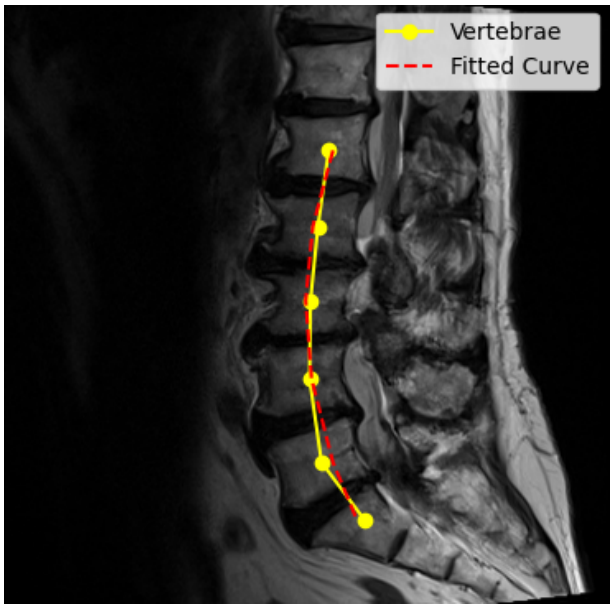

Figure 23: Spinal curvature assessment via polynomial fitting of vertebra centers.

Once the direction of curvature has been established, the centers of the two vertebrae defining the intervertebral level of interest are calculated, and their midpoint is determined. This midpoint serves as the initial reference for the crop center and is subsequently adjusted based on both the identified curvature and the anatomical level. Specifically, no offset is applied at the upper lumbar levels (L1–L2); a vertical shift is introduced at the lower lumbar level (L5–S1) to accommodate sacral inclination; and a lateral shift is applied at the mid-lumbar levels to account for alignment variations. The Euclidean distance between the two vertebral centers is then employed to define the side length of a square crop, thereby ensuring proportional coverage across anatomical scales. The final crop, centered on the adjusted midpoint, provides a standardized and anatomically relevant ROI for subsequent classification. Figure 24 presents a schematic illustration of the ROI extraction process.

An example of a resulting cropped image is shown in Figure 25. The cropped region is subsequently resized to a standardized resolution of 256×256 pixels, ensuring uniformity

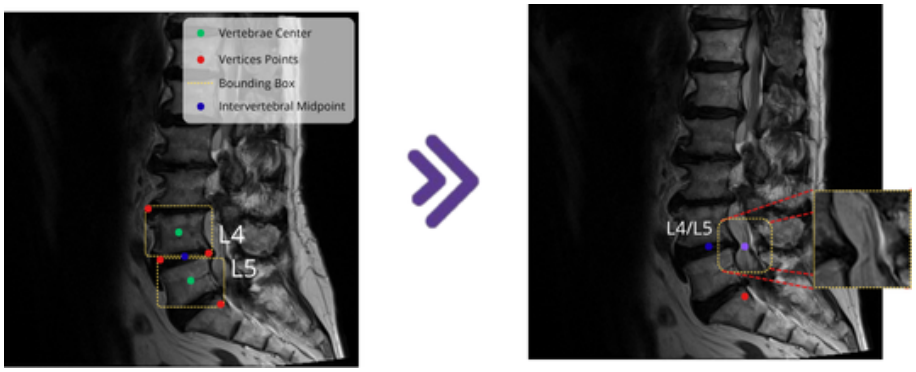

Figure 24: Schematic overview of ROI extraction between two vertebrae L4/L5.

for convolutional neural network training. This approach preserves anatomical relevance while reducing variability unrelated to the target classification task.

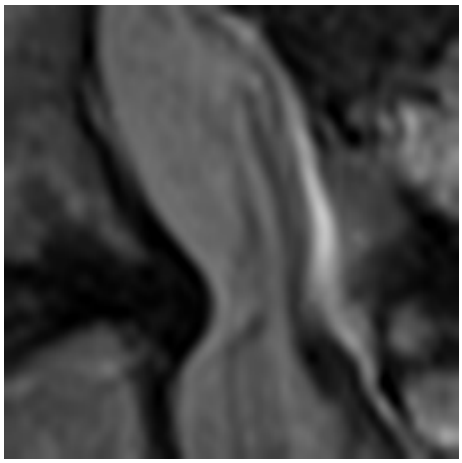

Figure 25: Example of a cropped ROI between two lumbar vertebrae.

The final dataset was formatted to be fully compatible with TensorFlow and convolutional architectures such as ResNet. To maintain consistency across experiments, the same study-based split described in Subsection 3.4.1 was applied. This ensures that the same studies are used in the training and validation subsets of both the detection and classification datasets, each adapted to its respective format. To facilitate reproducibility, the dataset was uploaded to the Hugging Face Hub.

3.4.4 Classification Model Development

For the classification task, a Swin Transformer architecture was chosen over a traditional CNN, such as ResNet. While CNNs excel at learning local features, we hypothesized that the Swin Transformer's self-attention mechanism would be more effective at capturing the subtle, long-range anatomical variations that define spinal stenosis. Unlike the fixed local view of CNNs, the transformer can model relationships between different parts

of an image, which is crucial for a nuanced task like severity grading from complex MRI scans.

Unlike traditional convolutional neural networks, the Swin Transformer processes images by partitioning them into non-overlapping windows and applying self-attention within each region. By periodically shifting these windows across layers, presented in Figure 26, the architecture achieves cross-window connections, enabling the model to integrate fine-grained local details with broader contextual information. Furthermore, its hierarchical design progressively merges image patches into larger representations, producing multi-scale feature maps that are particularly well-suited for analyzing the subtle anatomical variations present in medical images.

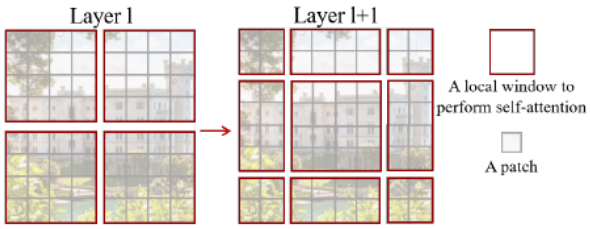

Figure 26: Example of a shifted window approach in the Swin Transformer Liu et al. (2021).

In this study, the base variant of the Swin Transformer with a patch size of 4×4 pixels and a window size of 7×7 was used. The model was initialized with weights pre-trained on the ImageNet dataset, providing a robust set of low- and mid-level visual features that facilitated transfer learning to the medical imaging domain. The classification head of the network was adapted to output three logits corresponding to the target severity classes: Normal/Mild, Moderate, and Severe.

The model was implemented in PyTorch and trained with a batch size of 32 over a maximum of 50 epochs. Optimization was performed using the AdamW optimizer with an initial learning rate of $3e-5$ and weight decay of $5e-4$. A cosine annealing learning rate schedule was applied to encourage smooth convergence, while early stopping with a patience of seven epochs prevented overfitting by halting training when validation performance plateaued. To address class imbalance, balanced class weights were computed from the training data and applied within a weighted cross-entropy loss function. The resulting class weights were 0.50 for Normal/Mild, 1.58 for Moderate, and 2.52 for Severe.

To enhance generalization, a set of carefully designed image augmentations was applied during training using PyTorch's online augmentation pipeline. These included random resized cropping with a scale factor between 0.8 and 1.0 of the original image size, horizontal flipping with a probability of 0.5, rotations up to 15°, and color jitter with brightness, contrast, and saturation each varied within $\pm 20\%$. In addition, random erasing was applied with a probability of 0.3. Importantly, all augmentations were designed to preserve anatomical plausibility to avoid introducing unrealistic variations. Furthermore, the Mixup technique (Zhang et al., 2018) was employed, where pairs of training samples were linearly combined at both the image and label level. This strategy encouraged smoother decision boundaries and reduced the risk of overfitting.

3.4.5 Model Evaluation

Both the detection and classification models were evaluated independently, following the same core methodology. The evaluation employed the set of carefully selected metrics defined in Section 3.3.3, Precision, Recall, and F1-Score, ensuring consistency with the initial approach. These metrics were chosen to verify that the models meet clinical expectations while maintaining high diagnostic relevance.

For classification tasks, the evaluation metrics are computed by directly comparing the predicted class labels with the corresponding ground-truth labels for each input instance. In this context, TPs refers to instances where the predicted label matches the ground truth. FPs are cases where the model predicts a class incorrectly, assigning a positive label to a negative instance. In contrast, FNs occurs when the model fails to identify a positive instance, incorrectly classifying it as negative. These definitions are applied at the instance level, making the calculation of Precision, Recall, and F1-Score straightforward and directly interpretable in terms of classification accuracy.

Intersection over Union (IoU) is a widely used metric to evaluate the spatial accuracy of object detection models. Measures the degree of overlap between a predicted bounding box and the corresponding ground-truth bounding box. Specifically, IoU is calculated as the area of the intersection between the predicted and ground truth boxes divided by the area of their union. An IoU value ranges from 0 - no overlap to 1 - perfect overlap, with higher values indicating better localization. In detection evaluation, a predicted box is typically considered a correct detection, also known as TP, only if its IoU with a ground truth box exceeds a predefined threshold, set at 0.5. This threshold ensures that the model not only identifies the correct object class but also localizes it with sufficient precision. Figure 27 illustrates the calculation IoU between two overlapping boxes.

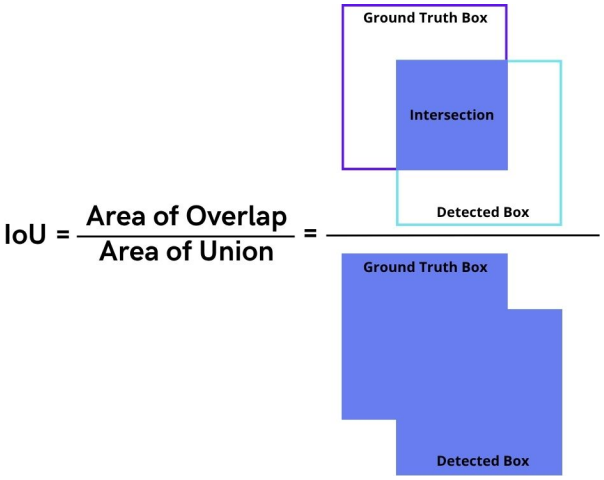

Figure 27: Illustration of IoU in Object Detection.

The evaluation of the detection model uses the same core performance metrics as the classification model - Precision, Recall, and F1-Score - but TPs, FPs, and FNs are defined using the IoU-based matching strategy. A detection is counted as a true positive only when it meets two criteria: the predicted class matches the ground truth, and the

predicted bounding box exceeds the IoU threshold with the corresponding ground truth box. Predictions that do not meet these conditions are counted as false positives, while unmatched ground-truth objects are considered false negatives. By incorporating IoU into the evaluation, these metrics reflect both the classification correctness and the spatial accuracy of the detection model.

3.4.6 Framework Evaluation

The evaluation of the framework, which integrates vertebra detection and stenosis classification into a unified pipeline, was conducted using a study-based, end-to-end protocol. This evaluation did not involve retraining or re-evaluating the individual models; rather, it relied on the outputs already obtained from the detection and classification stages described in Sections 3.4.2 and 3.4.4. The aim was to combine these results to assess the reliability and clinical relevance of the full system when considered as a system.

To ensure a rigorous and unbiased assessment, the framework evaluation was based on a held-out validation set comprising 20% of the available studies. This split was consistently applied at the study level across both tasks, preserving data independence and preventing leakage. For each study, the central sagittal T2-weighted MRI slice was used in accordance with the established preprocessing protocol.

The first level of evaluation was performed at the intervertebral scale. For each disc space (e.g. L2-L3), detection precision was checked by computing the IoU between predicted and ground-truth bounding boxes of the two adjacent vertebrae. If both vertebrae achieved an IoU of at least 0.5, the corresponding ROI was extracted and classified. The classification output was then compared to the ground-truth stenosis label, and metrics such as precision, recall, accuracy, and F1-score were computed across all valid levels.

Beyond the per-level analysis, results were aggregated at the study scale to provide a more clinically meaningful evaluation. For each study, the number of valid intervertebral levels (i.e., levels where both vertebrae were correctly detected and classification was performed) was recorded. These results were then consolidated to measure how well the framework performed across an entire case, enabling assessment of both partial and complete success scenarios. This aggregation gave a high-level perspective on the robustness and reliability of the system in practice.

By structuring the evaluation around both intervertebral levels and whole studies, the framework's effectiveness could be measured not only in terms of classification accuracy, but also in its capacity to consistently integrate detection and classification outputs. Finally, inference time per study was recorded to quantify the computational efficiency of the full pipeline from image input to final classification.

3.5 Development Environment and Computational Resources

The development and experimentation for this project were carried out using the Python[8] programming language, a widely adopted ecosystem in the machine learning and computer vision communities due to its extensive library support and ease of integration. All code was written and managed within the Visual Studio Code (VS Code)[9] integrated development environment, which provided an efficient and customizable workspace for debugging, version control, and task automation.

A number of open-source libraries and frameworks played a crucial role throughout the pipeline:

- **PyTorch**[10] was the primary deep learning framework used for model development, training, and inference. Its dynamic computational graph and high-level APIs facilitated rapid experimentation and fine-tuning.
- **OpenCV**[11] was employed for image manipulation tasks such as resizing, normalization and visualization of model outputs.
- **NumPy**[12] and **Pandas**[13] were essential for numerical operations, data preprocessing, and managing structured metadata associated with the MRI scans and corresponding annotations.

To manage dependencies and ensure reproducibility, the project environment was encapsulated using `virtualenv`, and all packages were version-locked using a requirements text file. Code and experimental configurations were tracked using Git and GitHub, supporting iterative development and collaboration.

Model training and evaluation were executed on the Vision supercomputer[14] located at the VISTA Lab, University of Évora. The Vision system is specifically designed for high-performance computing tasks in computer vision and artificial intelligence. It comprises two compute nodes and one management node, with each compute node featuring:

- **CPU:** Dual AMD EPYC 7742 (Rome architecture), providing a total of 128 physical cores per node
- **System Memory:** 1 TB of DDR4 RAM, supporting large-scale parallel data loading and in-memory preprocessing
- **GPUs:** 8 × NVIDIA A100 Tensor Core GPUs (40 GB HBM2 memory per GPU), enabling massive parallelism and accelerated tensor operations

[8] `https://www.python.org/`
[9] `https://code.visualstudio.com/`
[10] `https://pytorch.org`
[11] `https://opencv.org`
[12] `https://numpy.org`
[13] `https://pandas.pydata.org`
[14] `https://vision.uevora.pt`

- **Total GPU Memory:** 320 GB per node

The Vision infrastructure allowed for efficient execution of large-scale training experiments, particularly those involving 3D medical imaging data and class-imbalanced datasets. Jobs were scheduled and managed via SLURM, enabling optimized resource allocation and job queuing in a multi-user environment. In summary, the combination of a powerful software stack and access to advanced computational infrastructure significantly contributed to the scalability, reproducibility, and performance of the developed models.

Chapter 4

Results and Discussion

This chapter presents the experimental results and critical analysis of the proposed deep learning framework for the automated detection and classification of LSS in MRI scans. Building upon the methodology described in Chapter 3, the results are organized to address each component of the framework. Key evaluation metrics—including precision, recall, F1-score, and IoU are employed to assess both spatial accuracy and diagnostic reliability across independent validation datasets. The discussion situates these findings within clinical expectations, examines inherent limitations such as dataset heterogeneity and computational constraints, and compares the framework's performance with that reported in the existing literature. By integrating empirical evidence with both technical and clinical perspectives, this chapter demonstrates the framework's potential to contribute to the standardization of LSS diagnosis, while also outlining opportunities for further refinement.

4.1 Proposed Method Results and Analysis

This section presents the outcomes of the initial object detection experiments conducted using Faster R-CNN. The objective was to develop a unified model capable of detecting and classifying LSS in sagittal MRI slices, addressing three severity classes: Normal/Mild, Moderate, and Severe.

4.1.1 Experimental Setup

Several experiments were conducted to achieve the target performance in both detection and classification tasks. Among these, three representative training runs were selected to illustrate the incremental modifications made to the training strategy. The configurations of these experiments are summarized in Table 4.1. Each experiment is identified by a unique Run ID and includes details regarding the dataset version, the total number of images, class distribution, and key training hyperparameters. The dataset versions represent incremental refinements of the annotations and data composition, and their specific improvements will be discussed in detail later in this chapter.

Specifically, the table reports the dataset version used in each run, the number of LSS severity classes, the total number of images, and the number of annotated instances per class. Additionally, it includes the batch size and the number of training iterations employed in each configuration. Certain configurations are not presented in the table, as they were consistently applied across all experiments—namely, an image size of 512×512 pixels and an initial learning rate of $1e-4$.

Table 4.1: Experiments Configurations

Run ID	Dataset	Images	Annotations per Class			Batch Size	Iterations
			Normal/Mild	Moderate	Severe		
f_rcnn_1	v1.0	57041	144490	12809	8306	4	5000
f_rcnn_2	v1.1	46974	202286	17932	11628	4	5000
f_rcnn_3	v1.2	82038	196938	62685	40590	16	150000

4.1.2 Experiments Analysis

The development of the Faster R-CNN model followed an iterative refinement process, guided by systematic evaluation of performance limitations at each stage. Modifications were informed by both quantitative metrics and qualitative inspection of failure cases. Table 4.2 summarizes the key results for all three experimental runs.

Table 4.2: Quantitative evaluation metrics of the Faster R-CNN runs.

Run ID	Precision	Recall	F1-Score	Per-Class Precision		
				Normal/Mild	Moderate	Severe
f_rcnn_1	0.432%	11.200%	0.832%	1.296%	0%	0%
f_rcnn_2	0.511%	13.700%	0.984%	2.532%	0%	0%
f_rcnn_3	**4.699%**	**46.600%**	**8.537%**	**6.673%**	**4.048%**	**3.377%**

The initial experiment used dataset version v1.0, comprising 57041 images, with a batch size of 4 over 5000 training iterations. Quantitative results were unsatisfactory: overall precision, recall and F1 score in all classes were 0.432%, 11.2%, and 0.832%, respectively, as shown in Table 4.2. At the class level, the Severe and Moderate classes had 0% precision, while Normal/Mild achieved only 1.296% precision. Qualitative inspection of predictions revealed that the model consistently favored the majority class, failing to detect minority-class instances. This outcome was primarily attributed to the extreme class imbalance in the dataset, with Normal/Mild totaling 144490 instances versus 8306 Severe cases.

Training dynamics confirmed that the model converged smoothly, yet without meaningful generalization. The total loss decreased from 2.245 to 1.518 as presented on Figure 28, with the classification loss dropping from 1.417 to 0.715 and the RPN classification loss from 0.679 to 0.080. While the losses suggest stable training, they failed to translate into effective detection for minority classes, indicating that the neural network had overfitted to the majority class patterns and was insensitive to the underrepresented pathological features. These findings highlighted the need for more targeted interventions, such as augmentations or dataset curation, to address the imbalance and improve model generalization.

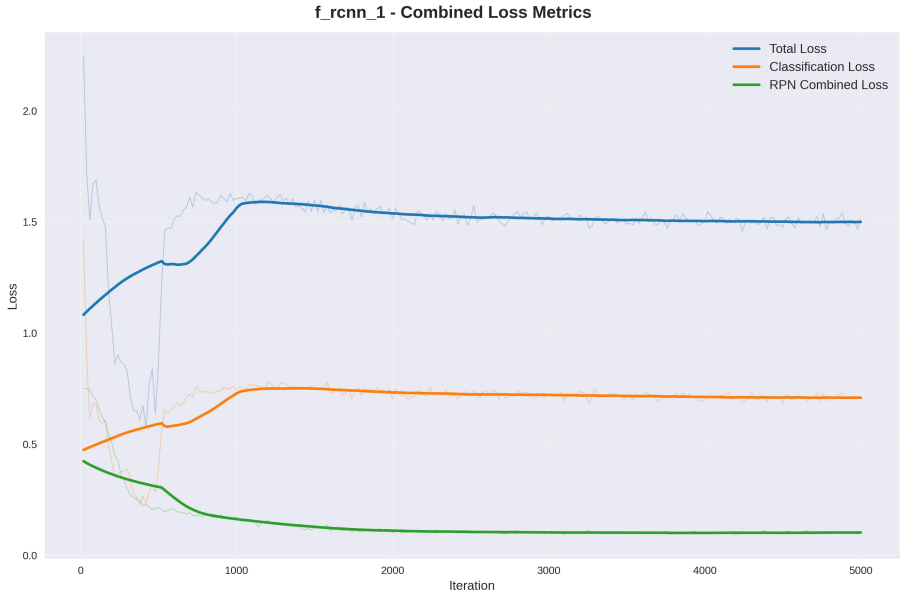

Figure 28: Training loss curves for Run f_rcnn_1.

To improve generalization in the second experiment, the dataset v1.0 training set was augmented using horizontal flips, small rotations, and brightness/contrast adjustments, expanding the dataset to 46974 images, generating the v1.1 dataset. The augmentations were specifically applied to increase the representation of minority classes, Moderate and Severe, aiming to mitigate the extreme class imbalance that had caused the unsatisfactory performance in the first run.

No changes were made to the model architecture or hyperparameters, the batch size and number of training iterations remained the same. Quantitative results showed modest improvement, the overall precision improved to 0.511%, recall improved to 13.7% and F1 Score improved to 0.984% compared to the previous run. At the class level, Normal/Mild precision increased to 2.532%, while Severe and Moderate classes still had 0% precision. Qualitative inspection indicated that the model slightly diversified its predictions but continued to favor the majority class, revealing that simple augmentations were not sufficient to fully correct the imbalance.

Training dynamics confirmed stable convergence over 5000 iterations. Figure 29 presents the training loss curves, showing that the total loss decreased from an initial 2.238 to 1.048, the classification loss decreased from 1.494 to 0.448 and the RPN classification loss from 0.663 to 0.014. Overall, while the neural network demonstrated improved convergence and slightly better generalization to Normal/Mild, it failed to detect minority classes, highlighting the need for further targeted interventions such as more extensive dataset curation and enhancing class distribution by augmentation, oversampling, or architectural modifications.

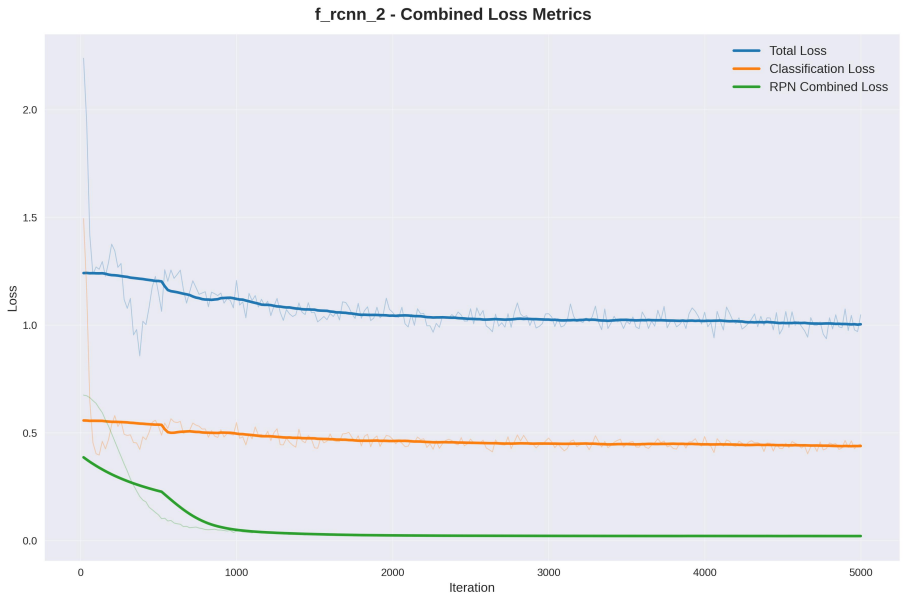

Figure 29: Training loss curves for Run f_rcnn_2.

The third experimental run was conducted using a newly curated dataset, version v1.2, derived from v1.1. A detailed inspection of the data revealed two major sources of bias and noise: (i) the initial and final slices of the MRI volumes often contained no relevant anatomical features, and (ii) several studies included only Normal/Mild annotations across all vertebrae, further exacerbating the imbalance. To address these issues, all studies containing exclusively Normal/Mild labels were removed, and 25% of the initial and final slices of each MRI were discarded, yielding 12430 useful images.

The curated dataset provided a cleaner and more focused representation of spinal canal stenosis severity. To further enhance minority class representation, the training set was expanded by a factor of nine through a diverse augmentation pipeline, including random rotations, motion blur, grid distortion, random shadow, elastic transformations, and brightness/contrast adjustments, resulting in 82038 images.

In addition to dataset curation, the training configuration was also modified to increase the model's learning capacity. The maximum number of iterations was tripled to 150000, allowing the network more opportunities to converge on meaningful patterns, while the batch size was increased to 16 in order to stabilize gradient updates and make more efficient use of available computational resources. These adjustments were intended to complement the improvements in dataset quality and provide a fairer assessment of the model's ability to generalize.

Quantitative results showed a substantial improvement compared to previous experiments, with an overall precision of 4.699%, recall of 46.6%, and F1-score of 8.537%. While these values remain very low in absolute terms, they mark the first time the model was able to detect all three severity classes. Specifically, Normal/Mild achieved 6.673% precision, Moderate 4.048%, and Severe 3.377%. This indicates that the curated v1.2 dataset partially mitigated the bias toward Normal/Mild present in earlier runs.

Training dynamics, as shown in Figure 30, were also markedly more stable compared to earlier runs. The total loss decreased smoothly from 1.581 to 0.348. Classification loss dropped from 0.861 to 0.184, while RPN classification loss converged to near-zero, from 0.694 to 0.005, indicating a highly consistent region proposal process. Overall, these training metrics confirmed that the longer optimization schedule, combined with a curated dataset, improved both numerical stability and generalization.

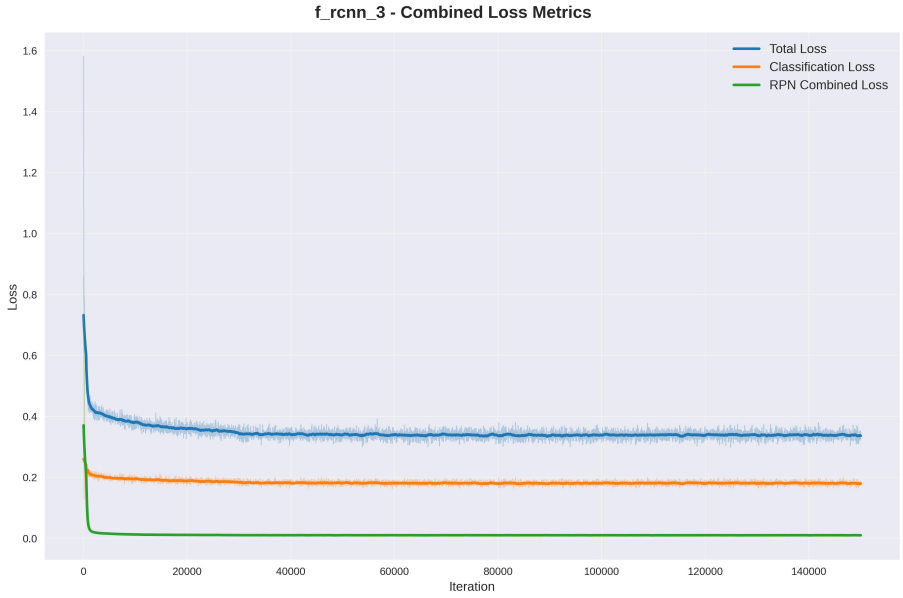

Figure 30: Training loss curves for Run f_rcnn_3.

Across the three experimental runs, Faster R-CNN demonstrated a clear but limited evolution. While early attempts completely collapsed to the majority class, the curated and augmented v1.2 dataset enabled the model to produce predictions across all severity classes for the first time. These results illustrate the importance of careful dataset design and the impact of targeted augmentations. However, even with several configurations tested beyond the three representative runs presented in this study, the model never achieved an average precision higher than 5%. This indicates that the data set lacked sufficient discriminative features for the task or that the chosen architecture and training pipeline were not well suited or were too complex to reliably detect and classify the severity of spinal stenosis in sagittal MRI slices.

In summary, although the Faster R-CNN experiments confirmed that the network could learn basic patterns, the overall performance remained far from acceptable for clinical or research deployment. These limitations motivated the design of a final method aimed at separately addressing detection and classification in a more direct and specialized manner.

4.2 Detection Model Results and Analysis

The YOLO-based detection model was evaluated for its ability to localize the lumbar vertebrae (L1-S1) and the corresponding spinal canal regions. Table 4.3 presents the overall performance metrics. The model achieved exceptionally strong results, with both precision and recall exceeding 97%. The resulting F1-score of 97.93% is not only a high value but also a key technical contribution of this work. Accurate vertebral localization is a fundamental prerequisite for any automated multi-stage diagnostic system, and this robust performance provides a reliable foundation for the subsequent classification stage, mitigating the risk of error propagation throughout the pipeline.

Table 4.3: Overall performance metrics.

Precision	**Recall**	**F1-Score**
97.831%	98.033%	97.928%

Figure 31 shows the normalized confusion matrix with per-level detection outcomes. Performance remained consistently high across all six vertebral levels. Slightly lower accuracy was observed at L1, reaching 96.06%, while S1 achieved the highest value of 98.92%. These results suggest that anatomical variability at the uppermost vertebra introduces additional challenges for detection, while performance in the central and lower regions of the lumbar spine is more stable.

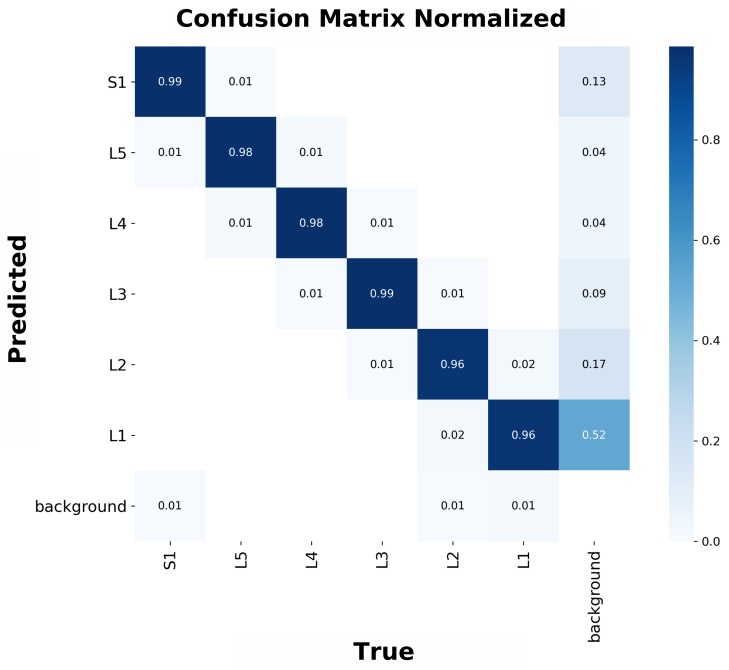

Figure 31: Normalized confusion matrix across vertebral levels.

The training dynamics were also examined to evaluate convergence and model stability. Figure 32 illustrates the evolution of the primary loss components: box regression,

classification, and distribution focal loss. The training process was configured for a maximum of 200 epochs with an early stopping patience of 20. Training concluded at epoch 123, with the optimal model identified at epoch 102. The loss values decreased steadily throughout training. Box regression started at 1.36 and reached 0.40, classification decreased from 2.68 to 0.23, and distribution focal loss declined from 1.44 to 0.86. These results confirm stable optimization and effective convergence.

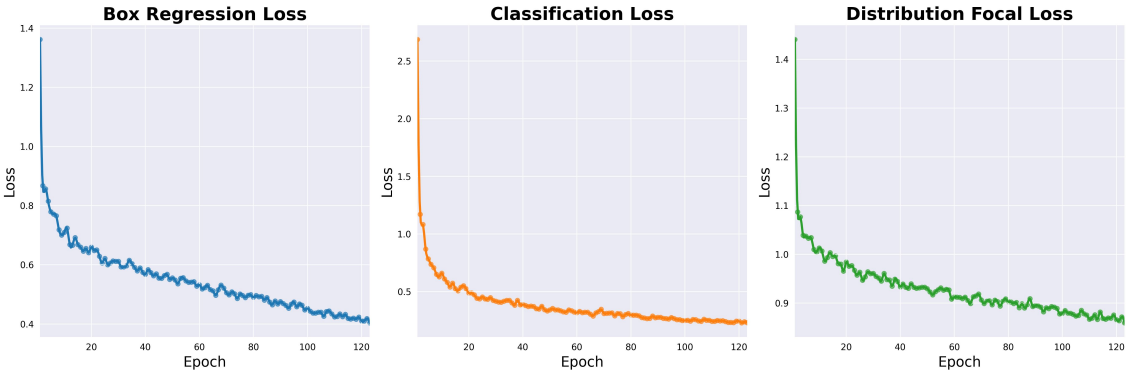

Figure 32: Training loss curves.

To complement the quantitative evaluation, Figure 33 presents a representative test image with predicted bounding boxes compared against ground-truth annotations. This visual example demonstrates the model's ability to precisely localize the vertebrae, showing strong agreement between predicted and reference labels.

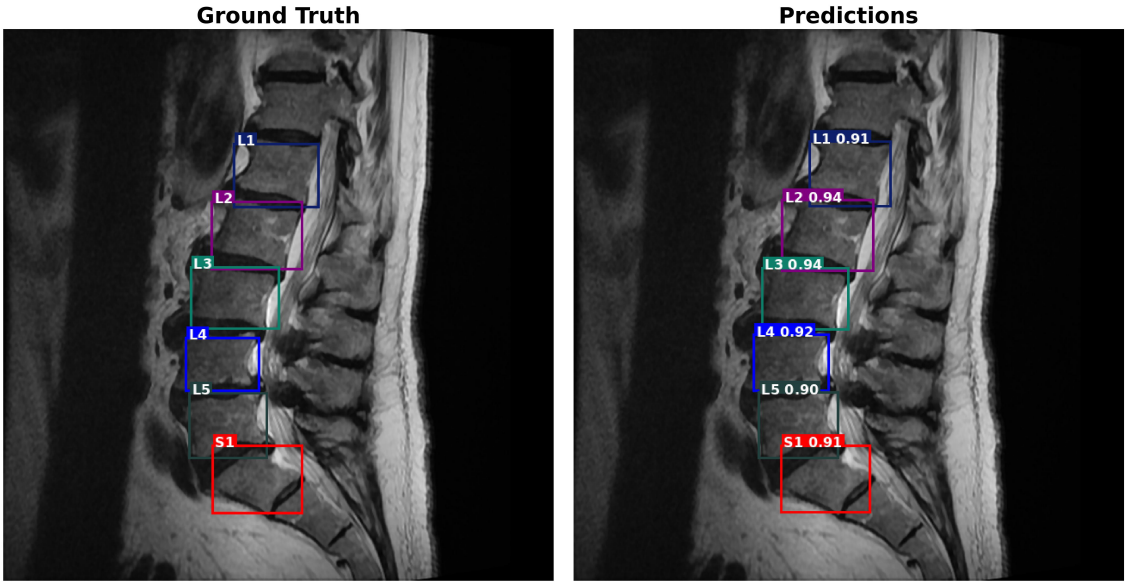

Figure 33: Representative detection result.

In summary, the detection model demonstrates excellent reliability in localizing individual lumbar vertebrae from L1 to S1. The consistently high detection rates across

all vertebral levels indicate that the model can accurately and robustly identify each anatomical region. This reliable vertebral-level detection provides a solid foundation for the subsequent classification stage, ensuring that the extracted regions correspond precisely to the intended vertebrae. Furthermore, the strong performance across both the upper and lower vertebrae highlights the model's ability to handle anatomical variability, which is critical for practical applications in clinical or research settings.

4.3 Classification Results and Analysis

Table 4.4 summarizes the overall performance of the Swin Transformer classifier using both macro and weighted averages. The weighted averages, which account for class imbalance by considering the proportion of instances in each class, indicated a precision of 78.65%, recall of 76.06%, and F1-score of 77.05%. Macro averages, which treat all classes equally, were 66.45% for precision, 70.25% for recall, and 67.98% for F1-score. The relatively small gap between macro and weighted averages suggests that the model performs more consistently across all classes.

Table 4.4: Overall performance metrics.

Macro Average			**Weighted Average**		
Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
66.450%	70.250%	67.980%	78.650%	76.060%	77.050%

Figure 34 presents the normalized confusion matrix for the three severity classes. The model achieved its highest accuracy on the Normal/Mild class, correctly predicting 83% of samples, while 14% were misclassified as Moderate and 3% as Severe. For the Moderate class, 56% of samples were correctly classified, 21% predicted as Normal/Mild, and 24% as Severe, reflecting some confusion with both neighboring classes. The Severe class exhibited improved recall, with 72% correctly identified, 24% misclassified as Moderate, and 4% as Normal/Mild. Overall, these results indicate that the model achieves a more balanced performance across severity levels, with improved ability to distinguish severe cases.

The observed confusion, particularly within the Moderate class, represents a critical finding. It likely reflects the inherent ambiguity of this diagnostic category in clinical practice. The visual features distinguishing a "severe" Normal/Mild case from an "early" Moderate case can be extremely subtle in a single 2D sagittal view. The model's difficulty in this regard suggests that information contained within a single MRI slice may be insufficient to reliably differentiate these borderline cases. This limitation highlights not only constraints of the model but also the potential shortcomings of relying on individual 2D slices for such a nuanced classification task.

Training dynamics, illustrated in Figure 35, show a steady decrease in training loss from 1.0127 at epoch 1 to 0.7005 at epoch 16. Validation loss decreased from 0.8168 to 0.6985 at epoch 15, which corresponds to the best-performing checkpoint of the 50 planned epochs. The convergence of both training and validation losses, along with the

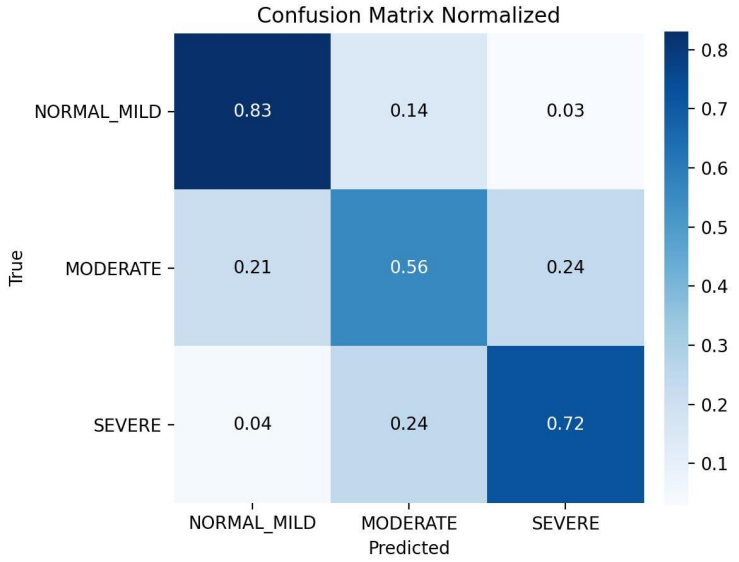

Figure 34: Normalized confusion matrix across severity classes for the Swin Transformer.

relatively small gap between them, demonstrates effective learning and generalization without significant overfitting.

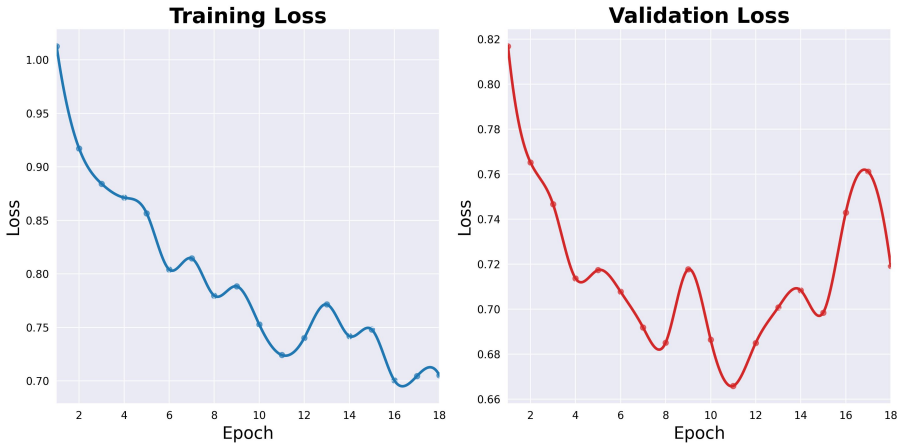

Figure 35: Training and validation loss curves for the Swin Transformer.

To illustrate the model's limitations, Figure 36 shows representative misclassified cases. Moderate and severe cases remain partially confused, reflecting overlapping features and subtle class boundaries. This indicates that incorporating additional contextual information, such as multi-slice or volumetric data, could help the model better distinguish borderline cases.

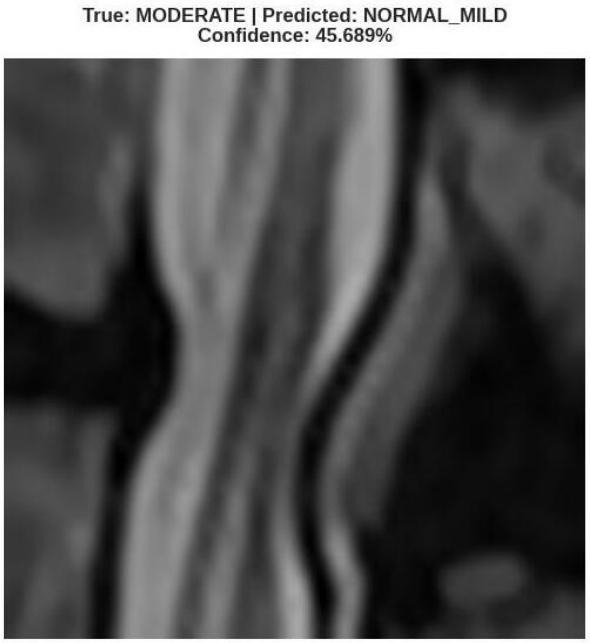

Figure 36: Example of a misclassified case by the Swin Transformer.

In summary, the Swin Transformer achieved strong classification performance, particularly in distinguishing Severe cases, while maintaining strong accuracy for Normal/Mild and reasonable performance for Moderate. While the model demonstrates the capacity to differentiate between severity levels, the distribution of classes and the subtle anatomical variations in MRI images continue to limit performance for less represented cases.

4.4 Framework Results and Analysis

The performance of the integrated vertebra detection and stenosis classification framework was evaluated at both the study level and the intervertebral level using the held-out validation set. All results are based on valid intervertebral levels, i.e., levels where both vertebrae were correctly detected and classified. The total number of valid levels analyzed was 692.

The overall performance of the framework, summarized by macro and weighted averages across all valid levels, is presented in Table 4.5. The macro-average F1-score of 67.316% reflects a balanced consideration of all classes, while the weighted average F1 of 77.151% accounts for the class distribution, highlighting the dominance of the majority Normal/Mild class.

Per-class performance metrics, shown in Table 4.5, further illustrate how the framework performs across individual stenosis severity levels. Normal/Mild achieves the highest F1-score, while Moderate and Severe classes are more challenging due to class imbalance and fewer examples. These results highlight the strengths and limitations of the pipeline at the class level.

Table 4.5: Per-class performance metrics of the framework.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Normal/Mild	89.410%	87.060%	88.220%	456
Moderate	48.950%	51.090%	50.000%	137
Severe	61.900%	65.660%	63.730%	99
Macro Average	66.757%	67.938%	67.316%	692
Weighted Average	77.468%	76.879%	77.151%	692

Performance at the intervertebral level is summarized in Table 4.6. Macro and weighted F1-scores are reported for each disc level, showing variability across the spine. In particular, the lower macro-F1 at L5_S1 reflects increased anatomical variability and fewer examples for the Severe class, while higher weighted F1 at some levels is influenced by the majority class presence.

Table 4.6: Intervertebral-level metrics.

Disc Level	**Macro F1**	**Weighted F1**	**Support**
L1_L2	47.513%	88.605%	417
L2_L3	57.366%	72.903%	426
L3_L4	65.617%	68.692%	426
L4_L5	66.872%	67.617%	426
L5_S1	42.860%	86.932%	423

The proposed framework achieves efficient end-to-end inference performance, as detailed in Table 4.7. On average, the vertebra detection step requires 0.345 seconds per study, while stenosis classification takes approximately 0.260 seconds per disc level, resulting in a mean classification time of 1.295 seconds per study. Consequently, the overall processing time per study amounts to 1.640 seconds, which highlights the computational efficiency of the framework and supports its suitability for near real-time clinical deployment.

Table 4.7: Average computational time of the proposed framework.

Metric	**Time (s)**
Mean detection time per study	0.345
Mean classification time per disc level	0.260
Mean classification time per study	1.295
Mean total processing time per study	1.640

Overall, the results show that the framework is effective in combining detection and classification to provide reliable study-level predictions, while also allowing identification of more challenging disc levels and stenosis classes. The combination of study-level and intervertebral-level evaluation ensures a comprehensive understanding of both clinical relevance and technical performance.

4.5 Discussion of Key Findings and Limitations

Having presented the quantitative performance metrics of the developed framework, this section now turns to a critical discussion of the key findings and their implications. The research evolved from an initial monolithic deep learning model into a refined two-stage framework for the automated analysis of LSS from MRI scans. This iterative development process not only highlighted the challenges inherent in this diagnostic task but also revealed effective strategies to mitigate them.

The initial attempt employed a single Faster R-CNN model to detect and classify LSS severity simultaneously. This approach proved ineffective, with average precision remaining below 5%. The poor performance was primarily attributable to the extreme class imbalance within the dataset, which caused the model to overfit the majority Normal/Mild class while failing to reliably identify Moderate and Severe cases. This outcome highlighted the difficulty of unifying detection and classification in a single architecture and motivated the development of a more specialized, decoupled pipeline.

The final two-stage framework successfully addressed these challenges by separating anatomical localization from pathological classification:

- High-Fidelity Vertebra Detection: The first stage, powered by a YOLOv8 model, achieved excellent results in localizing lumbar vertebrae (L1-S1). With a precision of 97.83%, recall of 98.03%, and F1-score of 97.93%, the model provided a highly reliable foundation for the subsequent analysis by ensuring accurate ROI extraction across all vertebral levels.
- Effective Stenosis Classification: Building on this reliable detection, the Swin Transformer model achieved a weighted average F1-score of 77.05% for severity classification. Performance was strongest for Normal/Mild cases with 83% accuracy and improved notably for Severe cases with 72% accuracy, representing a substantial gain compared with the initial approach.

When integrated, the end-to-end framework produced a weighted average F1-score of 77.15% across all valid intervertebral levels. Equally important, the system demonstrated computational efficiency, processing an entire study in an average of 1.640 seconds. This rapid inference supports the potential for near real-time clinical use, directly addressing the thesis objective of improving diagnostic efficiency.

A central strength of this work is the demonstration that a specialized, decoupled pipeline is a more effective strategy for LSS analysis than monolithic designs. By leveraging sagittal T2-weighted MRI sequences alone, the framework establishes a proof-of-concept for achieving standardized and reproducible severity classification. While axial acquisitions remain clinically important for comprehensive assessment, this sagittal-based approach highlights the potential for reducing dependence on time-consuming protocols in selected scenarios. Such a streamlined strategy may contribute to lowering diagnostic costs and accelerating clinical workflows, while providing radiologists with an objective and reproducible assessment to complement traditional interpretation.

Despite these strengths, two main limitations must be acknowledged. First, the persistent confusion between Moderate and Severe cases indicates that features extracted from individual sagittal slices may not fully capture the nuanced anatomical detail required to distinguish between these clinically critical categories. This limitation is not merely academic, in patient management, the distinction between Moderate and Severe stenosis often determines whether surgical intervention or conservative treatment is pursued. A model that cannot reliably make this distinction cannot yet be trusted as a standalone diagnostic tool, though it may still be valuable as an initial screening or second-opinion system.

Second, the framework is restricted to lumbar central canal stenosis and does not incorporate other relevant pathologies such as foraminal or lateral recess stenosis, limiting its comprehensiveness in real-world clinical settings. Finally, the dataset was derived from curated central slices of the RSNA 2024 competition, which constrains the generalizability of the results. Broader validation on larger, multi-sequence, and volumetric datasets will be essential to confirm robustness across diverse patient populations, imaging protocols, and slice orientations.

4.6 Comparison with Literature

This thesis aligns with a growing body of research emphasizing the value of AI-based tools in improving the efficiency and consistency of spinal MRI interpretation. Similar to the observations of Won et al. (2025) and Yilihamu et al. (2025), this work demonstrates that AI can accelerate image analysis and reduce the repetitive burden on radiologists. The rapid inference time of 1.640 seconds per study represents a tangible contribution to this broader goal of enhanced workflow efficiency.

A recurring challenge identified in the literature is class imbalance in LSS datasets, a factor that was also central to this study. Won et al. (2025) explicitly noted the limitations imposed by imbalanced data, while Mousavi (2024) described it as a key obstacle requiring approaches such as data augmentation and tailored loss functions. The difficulties encountered in the early Faster R-CNN experiments of this thesis corroborate these observations, underscoring the necessity of robust strategies to manage imbalance in medical imaging tasks.

A distinctive contribution of this research is its exclusive reliance on sagittal T2-weighted MRI scans for stenosis classification. In contrast, many prior studies—including those by Bogdanovic et al. and Won et al. (2025)—have relied primarily on axial T2 images. The present approach therefore challenges conventional reliance on axial views, aligning more closely with Baek and Chung (2023), who likewise demonstrated that sagittal sequences alone can support accurate classification. This divergence highlights the potential for more streamlined and resource-efficient diagnostic workflows.

The performance of the framework is summarized alongside comparable studies in Table 4.8, where all reported values correspond to the F1-score metric. This thesis achieved a weighted average F1-score of 77.15%, which is competitive with the F1-scores of 78.4% and 75.7% reported by Won et al. (2025) and slightly lower than the F1-score of 90.51%

reported by Mousavi (2024) and the F1-score of 69.84% reported by Baek and Chung (2023). The table provides a concise overview of the main metrics, input modalities, and reported outcomes, allowing for a direct comparison with existing approaches.

Table 4.8: Performance comparison with related studies on automated LSS classification.

Study	**Input Modality**	**Performance**
This thesis	Sagittal T2 MRI	77.15%
Won2025	Axial + Sagittal T2 MRI	78.4%, 75.7%
Mousavi2024	Sagittal T1/T2 + Axial T2	90.51%
Baek2023	Sagittal T2 MRI	69.84%

The differences in reported performance largely reflect methodological variations, for instance, Baek and Chung (2023) utilized a Swin Transformer architecture that relies on implicitly learned hierarchical image features through meta-learning, whereas conventional methods such as YOLACT or YOLOS often rely on more explicit, handcrafted feature extraction pipelines.

Similarly, Mousavi (2024) integrated multiple MRI modalities (Sagittal T1, Sagittal T2, and Axial T2), which likely provided richer information than the single-modality Sagittal T2 design adopted in this work. Together, these comparisons emphasize both the strengths of the current framework—competitive performance and efficiency—and areas where multi-modality or feature-based approaches may achieve higher metrics.

A principal innovation of this work is the development of a highly effective vertebral detection stage, setting it apart from related literature. For example, Mousavi (2024) reported unsatisfactory performance in their detection pipeline, which represented a critical bottleneck for their overall system. In contrast, the YOLOv8 model developed in this study achieved F1-scores exceeding 97%, providing a robust solution to the foundational localization problem and enabling more reliable and accurate downstream analysis.

In summary, this thesis contributes to the field by demonstrating a novel, fully decoupled pipeline that delivers competitive classification accuracy using a more efficient sagittal-only workflow. Its principal innovations lie in the highly accurate vertebral detection stage and the effective application of a Swin Transformer for implicit feature learning in LSS severity grading. These contributions not only confirm the feasibility of sagittal-only approaches but also provide a promising direction for developing practical, real-time AI tools to support spinal imaging in clinical settings.

Chapter 5

Conclusions and Future Work

This chapter summarizes the main findings of the research, highlights the contributions made, and acknowledges the limitations encountered. In addition, it provides recommendations for future research directions aimed at extending and strengthening the proposed framework for the automated analysis of LSS.

5.1 Conclusions

This thesis presents the successful development and validation of a novel, two-stage DL framework for the automated diagnosis of LSS from Sagittal T2-weighted MRI. The study demonstrated that a decoupled approach, wherein vertebral localization and stenosis classification are addressed as separate, specialized tasks, effectively mitigates the challenges of class imbalance and task complexity that rendered initial single-stage models ineffective.

A primary scientific contribution of this work is the design of a highly accurate and computationally efficient pipeline. The framework's first stage, employing a YOLOv8 model, achieved an F1-score exceeding 97%. The selection of YOLOv8 over more complex architectures like Faster R-CNN was a deliberate methodological decision, prioritized for its anchor-free design and rapid inference speed, which are critical for practical clinical workflows. This robust detection performance provided a reliable foundation for the subsequent classification stage.

Building upon this accurate localization, the classification model, based on the Swin Transformer architecture, achieved a weighted F1-score of 77.15%. This result affirms the viability of transformer-based methods for spinal pathology analysis, as their attention mechanisms are well-suited for capturing the subtle, long-range anatomical variations that define stenosis severity. A key finding is that this level of performance was achieved using single Sagittal MRIsequence. This outcome challenges the conventional reliance on axial sequences and suggests a potential for more streamlined and cost-effective imaging protocols. From the perspective of clinical adoption, the framework demonstrates high efficiency, processing an entire patient study in approximately 1.64 seconds, which supports its viability as a tool for near real-time diagnostic support.

Despite its strengths, the framework exhibits a critical limitation in its ability to consistently distinguish between moderate and severe stenosis. This classification ambiguity is clinically significant, as this distinction often dictates treatment pathways, including the decision between conservative management and surgical intervention. This weakness is likely attributable to the inherent ambiguity in 2D single-slice images, which may not capture the full three-dimensional complexity of the spinal canal required for the precise classification of borderline cases.

In summary, this research presents a successful proof-of-concept for a clinical decision-support system. By offering an objective and reproducible method for LSS assessment, the framework has the potential to enhance diagnostic efficiency, reduce radiologist workload,

and mitigate the inter-observer variability associated with subjective interpretation.

5.2 Recommendations for Future Research

In light of the findings and limitations of this study, the following directions are recommended for future research to advance the automated analysis of LSS:

1. *Incorporation of Volumetric Data*: A critical next step is the training of models on full three-dimensional volumetric MRI or multi-slice inputs. This approach would provide richer contextual information regarding spinal canal morphology, which is essential for reducing the ambiguity between moderate and severe cases and improving overall diagnostic accuracy.

2. *Expansion of Pathological Scope*: The current framework is limited to central canal stenosis. Subsequent research should extend the model to include other clinically relevant conditions, such as foraminal stenosis and subarticular stenosis, in order to create a more comprehensive diagnostic tool.

3. *Multi-modal and Multi-sequence Analysis*: The integration of additional MRI sequences, such as sagittal T1 or axial T2 scans, could provide complementary anatomical information and enhance classification performance. An investigation of data fusion techniques represents a promising direction for improving model robustness and accuracy.

4. *Clinical Validation on Diverse Datasets*: As this framework was developed using a curated dataset from a single competition source, rigorous validation on larger and more diverse datasets is essential for establishing clinical readiness. Such efforts must include data from multiple institutions, covering a wider range of patient demographics, scanner protocols, and pathologies to ensure generalizability.

By pursuing these research avenues, the contributions of this thesis may be extended toward the development of a clinically validated and comprehensive AI-assisted system for the diagnosis and management of lumbar spinal pathologies.

Bibliographic Reference

Hussam Abou-Al-Shaar, Owoicho Adogwa, and Ankit I. Mehta. Lumbar spinal stenosis: Objective measurement scales and ambulatory status. *Asian Spine Journal*, 12:765–774, 2018. ISSN 1976-1902. doi: 10.31616/ASJ.2018.12.4.765.

Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, Daniel S.W. Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian, and Ara Darzi. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digital Medicine 2021 4:1*, 4:1–23, 4 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00438-z.

Zubair Ahmad, Shabina Rahim, Maha Zubair, and Jamshid Abdul-Ghafar. Artificial intelligence (ai) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. a comprehensive review. *Diagnostic Pathology*, 16:1–16, 12 2021. ISSN 17461596. doi: 10.1186/s13000-021-01085-4.

Gustav Andreisek, Richard A. Deyo, Jeffrey G. Jarvik, Francois Porchet, Sebastian F.X. Winklhofer, and Johann Steurer. Consensus conference on core radiological parameters to describe lumbar stenosis - an initiative for structured reporting. *European Radiology*, 24:3224–3232, 11 2014. ISSN 14321084. doi: 10.1007/s00330-014-3346-z.

S. Ansari. *Building Computer Vision Applications Using Artificial Neural Networks: With Step-by-Step Examples in OpenCV and TensorFlow with Python*. Apress, 2020. ISBN 9781484258866.

Ji-Won Baek and Kyungyong Chung. Swin transformer-based object detection model using explainable meta-learning mining. *Applied Sciences*, 13(5), 2023. ISSN 2076-3417. doi: 10.3390/app13053213.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *CoRR*, abs/2004.10934, 2020. doi: 10.48550/arXiv.2004.10934.

Sanja Bogdanovic, Matthias Staib, Marco Schleiniger, Livio Steiner, Leonardo Schwarz, Christoph Germann, Reto Sutter, and Benjamin Fritz. Ai-based measurement of lumbar spinal stenosis on mri: External evaluation of a fully automated model. *Investigative radiology*, 59:656–666, 9 2024. ISSN 1536-0210. doi: 10.1097/RLI.0000000000001070.

Robert Bohinski. Spinal decompression: laminectomy & foraminotomy — mayfield brain & spine, 2021. URL `https://mayfieldclinic.com/pe-decompression.htm`.

Facundo Bre, Juan M. Gimenez, and Víctor D. Fachinotti. Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, 158:1429–1441, 1 2018. ISSN 0378-7788. doi: 10.1016/J.ENBUILD.2017.11.045.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,

Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020-December, 5 2020. ISSN 10495258. doi: 10.48550/arXiv.2005.14165.

Emefa Surprize Deborah Buaka and Md Zubab Ibne Moid. Ai and medical imaging technology: evolution, impacts, and economic insights. *Journal of Technology Transfer*, 49:2260–2272, 12 2024. ISSN 15737047. doi: 10.1007/s10961-024-10100-x.

Nikolaj Buhl. Yolo models for object detection explained [yolov8 updated], 2023. URL `https://encord.com/blog/yolo-object-detection-guide/`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-3:1575–1585, 2 2020. doi: 10.48550/arXiv.2002.05709.

Jason Pui Yin Cheung, Dino Samartzis, Hideki Shigematsu, and Kenneth Man Chee Cheung. Defining clinically relevant values for developmental spinal stenosis: A large-scale magnetic resonance imaging study. *Spine*, 39:1067–1076, 6 2014. ISSN 15281159. doi: 10.1097/BRS.0000000000000335.

Marcia A. Ciol, Richard A. Deyo, Eric Howell, and Suzanne Kreif. An assessment of surgery for spinal stenosis: Time trends, geographic variations, complications, and reoperations. *Journal of the American Geriatrics Society*, 44:285–290, 3 1996. ISSN 1532-5415. doi: 10.1111/J.1532-5415.1996.TB00915.X.

Databricks. Convolutional layer, 2023. URL `https://www.databricks.com/glossary/convolutional-layer`.

Richard A. Deyo, Sohail K. Mirza, Brook I. Martin, William Kreuter, David C. Goodman, and Jeffrey G. Jarvik. Trends, major medical complications, and charges associated with surgery for lumbar spinal stenosis in older adults. *JAMA*, 303:1259–1265, 4 2010. ISSN 0098-7484. doi: 10.1001/JAMA.2010.338.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR 2021 - 9th International Conference on Learning Representations*, 10 2020. doi: 10.48550/arXiv.2010.11929.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature 2017 542:7639*, 542:115–118, 1 2017. ISSN 1476-4687. doi: 10.1038/nature21056.

Katti Faceli. *Inteligência artificial : uma abordagem de aprendizado de máquina*. Grupo Gen - LTC, 2011. ISBN 9788521618805,8521618808,9788521620150,8521620152.

Julie M. Fritz, Anthony Delitto, William C. Welch, and Richard E. Erhard. Lumbar spinal stenosis: A review of current concepts in evaluation, management, and outcome measurements. *Archives of Physical Medicine and Rehabilitation*, 79:700–708, 6 1998. ISSN 0003-9993. doi: 10.1016/S0003-9993(98)90048-X.

Peng Fu and Jiyang Wang. Lithology identification based on improved faster r-cnn. *Minerals 2024, Vol. 14, Page 954*, 14:954, 9 2024. ISSN 2075-163X. doi: 10.3390/MIN14090954.

Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *TPAMI - Transactions on Pattern Analysis and Machine Intelligence*, 2017. doi: 10.48550/arXiv.1704.06857.

Yansong Ge, Yaoxing Lu, Cheng Ma, Benteng Lu, Erteng Ma, Yafei Zhang, and Fei Zhao. Effect of different interventions on lumbar spinal stenosis: A systematic evaluation and network meta-analysis. *World Neurosurgery*, 194:123459, 2 2025. ISSN 1878-8750. doi: 10.1016/J.WNEU.2024.11.042.

Ross Girshick. Fast r-cnn. *Neural Computation*, 2015. doi: 10.48550/arXiv.1504.08083.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2016. ISBN 9780262035613.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Science Robotics*, 3:2672–2680, 6 2014. ISSN 10495258. doi: 10.48550/arXiv.1406.2661.

Julien Guiot, Monique Henket, Fanny Gester, Béatrice André, Benoit Ernst, Anne-Noelle Frix, Dirk Smeets, Simon Van Eyndhoven, Katerina Antoniou, Lennart Conemans, Janine Gote-Schniering, Hans Slabbynck, Michael Kreuter, Jacobo Sellares, Ioannis Tomos, Guang Yang, Clio Ribbens, Renaud Louis, Vincent Cottin, Sara Tomassetti, Vanessa Smith, and Simon L. F. Walsh. Automated ai-based image analysis for quantification and prediction of interstitial lung disease in systemic sclerosis patients. *Respiratory Research 2025 26:1*, 26:1–9, 1 2025. ISSN 1465-993X. doi: 10.1186/S12931-025-03117-9.

Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 1 edition, 2019. ISBN 1492032646,9781492032649.

Andrew J. Haig, Michael E. Geisser, Henry C. Tong, Karen S.J. Yamakawa, Douglas J. Quint, Julian T. Hoff, Anthony Chiodo, Jennifer A. Miner, and Vaishali V. Phalke. Electromyographic and magnetic resonance imaging to predict lumbar stenosis, low-back pain, and no back symptoms. *The Journal of bone and joint surgery. American volume*, 89:358–366, 2007. ISSN 0021-9355. doi: 10.2106/JBJS.E.00704.

Yu Hao, Haoyang Pei, Yixuan Lyu, Zhongzheng Yuan, John-Ross Rizzo, Yao Wang, and Yi Fang. Understanding the impact of image quality and distance of objects to object detection performance. *2023 IEEE/RSJ International Conference on Intelligent*

Robots and Systems (IROS), pages 11436–11442, 2023. doi: 10.1109/IROS55552.2023.10342139.

S. Haykin. *Redes Neurais: Princípios e Prática*. Bookman Editora, 2001. ISBN 9788577800865.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:770–778, 12 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.90.

R. M Hristev. *The ANN book*. GNU General public license, 1990. URL `http://www.pdg.cnb.uam.es/cursos/Complutense/Complutense2004/pages/12_NeuralNetworks/Hritsev_The_ANN_Book.pdf`.

Y. Ishimoto, N. Yoshimura, S. Muraki, H. Yamada, K. Nagata, H. Hashizume, N. Takiguchi, A. Minamide, H. Oka, H. Kawaguchi, K. Nakamura, T. Akune, and M. Yoshida. Prevalence of symptomatic lumbar spinal stenosis and its association with physical performance in a population-based cohort in japan: The wakayama spine study. *Osteoarthritis and Cartilage*, 20:1103–1108, 10 2012. ISSN 10634584. doi: 10.1016/j.joca.2012.06.018.

Leonid Kalichman, Robert Cole, David H. Kim, Ling Li, Pradeep Suri, Ali Guermazi, and David J. Hunter. Spinal stenosis prevalence and association with symptoms: the framingham study. *Spine Journal*, 9:545–550, 7 2009. ISSN 15299430. doi: 10.1016/j.spinee.2009.03.005.

Adiraju Karthik, Kamal Aggarwal, Aakaar Kapoor, Dharmesh Singh, Lingzhi Hu, Akash Gandhamal, and Dileep Kumar. Comprehensive assessment of imaging quality of artificial intelligence-assisted compressed sensing-based mr images in routine clinical settings. *BMC medical imaging*, 24:284, 12 2024. ISSN 14712342. doi: 10.1186/s12880-024-01463-6.

Jeffrey N Katz and Mitchel B Harris. Clinical practice. lumbar spinal stenosis. *The New England journal of medicine*, 358:818–25, 2 2008. ISSN 1533-4406. doi: 10.1056/NEJMCP0708097.

Young Uk Kim, Yu Gyeong Kong, Jonghyuk Lee, Yuseon Cheong, Se hun Kim, Hyun Kyu Kim, Jun Young Park, and Jeong Hun Suh. Clinical symptoms of lumbar spinal stenosis associated with morphological parameters on magnetic resonance images. *European Spine Journal*, 24:2236–2243, 10 2015. ISSN 14320932. doi: 10.1007/s00586-015-4197-2.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature 2015 521:7553*, 521:436–444, 5 2015. ISSN 1476-4687. doi: 10.1038/nature14539.

Juncai Lin, Honglai Zhang, and Hongcai Shangs. Convolutional neural network incorporating multiple attention mechanisms for mri classification of lumbar spinal stenosis. *Bioengineering 2024, Vol. 11, Page 1021*, 11:1021, 10 2024. ISSN 2306-5354. doi: 10.3390/BIOENGINEERING11101021.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 07 2017. doi: 10.1109/CVPR.2017.106.

Guoxu Liu, Joseph Christian Nouaze, Philippe Lyonel Touko Mbouembe, and Jae Ho Kim. Yolo-tomato: A robust algorithm for tomato detection based on yolov3. *Sensors*, 20(7), 2020. ISSN 1424-8220. doi: 10.3390/s20072145.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, October 2021. doi: 10.1109/ICCV48922.2021.00986.

Zelong Liu, Andrew Tieu, Nikhil Patel, Georgios Soultanidis, Louisa Deyer, Ying Wang, Sean Huver, Alexander Zhou, Yunhao Mei, Zahi A. Fayad, Timothy Deyer, and Xueyan Mei. Vis-mae: An efficient self-supervised learning approach on medical image segmentation and classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 15242 LNCS: 95–107, 2 2024. doi: 10.1007/978-3-031-73290-4_10.

I. Livshin. *Artificial Neural Networks with Java: Tools for Building Neural Network Applications*. Apress, 2019. ISBN 9781484244210.

Greger Lønne, Bent Ødegård, Lars Gunnar Johnsen, Tore K. Solberg, Kjell Arne Kvistad, and Øystein P. Nygaard. Mri evaluation of lumbar spinal stenosis: Is a rapid visual assessment as good as area measurement? *European Spine Journal*, 23:1320–1324, 2 2014. ISSN 14320932. doi: 10.1007/s00586-014-3248-4.

Wei-Bang Ma, Yang Yang, and Wai-Chi Fang. An effective tuberculosis detection system based on improved faster r-cnn with roi align method. *BioCAS 2023 - 2023 IEEE Biomedical Circuits and Systems Conference, Conference Proceedings*, pages 1–5, 2023. doi: 10.1109/BioCAS58349.2023.10388704.

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, Dec 1943. ISSN 1522-9602. doi: 10.1007/BF02478259.

Akshit Mehra. Understanding yolov8 architecture, applications & features, 2023. URL `https://www.labellerr.com/blog/understanding-yolov8-architecture-applications-features/`.

M. A. Mikhael, I. Ciric, J. A. Tarkington, and N. A. Vick. Neuroradiological evaluation of lateral recess syndrome. *Radiology*, 140:97–107, 7 1981. ISSN 00338419. doi: 10.1148/RADIOLOGY.140.1.7244248.

Rachel A. Moses, Wenyan Zhao, Lukas P. Staub, Markus Melloh, Thomas Barz, and Jon D. Lurie. Is the sedimentation sign associated with spinal stenosis surgical treatment effect in sport? *Spine*, 40:129–136, 2015. ISSN 15281159. doi: 10.1097/BRS.0000000000000672.

Amir Mousavi. Lumbar spine degenerative classification using yolo v8 and deepscorenet. *medRxiv*, 2024. doi: 10.1101/2024.12.06.24318595.

Van Hiep Phung and Eun Joo Rhee. A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, 9:4500, 10 2019. doi: 10.3390/app9214500.

RangeKing. Brief summary of yolov8 model structure, 2023. URL `https://www.atmosera.com/blog/deep-learning/`.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. doi: 10.48550/arXiv.1506.02640.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 6 2015. ISSN 01628828. doi: 10.1109/TPAMI.2016.2577031.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *IEEE Access*, 9:16591–16603, 5 2015. ISSN 21693536. doi: 10.1109/ACCESS.2021.3053408.

RSNA. Rsna 2024 lumbar spine degenerative classification — kaggle, 2024. URL `https://www.kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification`.

Branimir Rusanov, Ghulam Mubashar Hassan, Mark Reynolds, Mahsheed Sabet, Jake Kendrick, Pejman Rowshanfarzad, and Martin Ebert. Deep learning methods for enhancing cone-beam ct image quality toward adaptive radiation therapy: A systematic review. *Medical Physics*, 49:6019–6054, 9 2022. ISSN 2473-4209. doi: 10.1002/MP.15840.

Evelien I.T. De Schepper, Gijsbert M. Overdevest, Pradeep Suri, Wilco C. Peul, Edwin H.G. Oei, Bart W. Koes, Sita M.A. Bierma-Zeinstra, and Pim A.J. Luijsterburg. Diagnosis of lumbar spinal stenosis: An updated systematic review of the accuracy of diagnostic tests. *Spine*, 38, 4 2015. ISSN 03622436. doi: 10.1097/BRS.0B013E31828935AC.

N. Schönström and J. Willén. Imaging lumbar spinal stenosis. *Radiologic Clinics of North America*, 39:31–53, 1 2001. ISSN 0033-8389. doi: 10.1016/S0033-8389(05)70262-1.

Youssef Skandarani, Pierre Marc Jodoin, and Alain Lalande. Gans for medical image synthesis: An empirical study. *Journal of Imaging*, 9:69, 3 2023. ISSN 2313433X. doi: 10.3390/jimaging9030069.

Johann Steurer, Simon Roner, Ralph Gnannt, and Juerg Hodler. Quantitative radiologic criteria for the diagnosis of lumbar spinal stenosis: A systematic literature review. *BMC Musculoskeletal Disorders*, 12:1–9, 7 2011. ISSN 14712474. doi: 10.1186/1471-2474-12-175.

Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35:1299–1312, 5 2016. ISSN 1558254X. doi: 10.1109/TMI.2016.2535302.

T.T. Teoh. *Convolutional Neural Networks for Medical Applications*. SpringerBriefs in Computer Science. Springer Nature Singapore, 2023. ISBN 9789811988141.

Juan Terven and Diana Cordova-Esparza. A comprehensive review of yolo: From yolov1 and beyond. *arXiv*, 2023. doi: 10.48550/arXiv.2304.00501.

Christy Tomkins-Lane, Markus Melloh, Jon Lurie, Matt Smuck, Michele C. Battié, Brian Freeman, Dino Samartzis, Richard Hu, Thomas Barz, Kent Stuber, Michael Schneider, Andrew Haig, Constantin Schizas, Jason Pui Yin Cheung, Anne F. Mannion, Lukas Staub, Christine Comer, Luciana Macedo, Sang Ho Ahn, Kazuhisa Takahashi, and Danielle Sandella. Issls prize winner: Consensus on the clinical diagnosis of lumbar spinal stenosis. *Spine*, 41:1239–1246, 8 2016. ISSN 15281159. doi: 10.1097/BRS.0000000000001476.

Christy Tomkins-Lane, Markus Melloh, and Arnold Wong. Diagnostic tests in the clinical diagnosis of lumbar spinal stenosis: Consensus and results of an international delphi study. *European Spine Journal*, 29:2188–2197, 9 2020. ISSN 14320932. doi: 10.1007/s00586-020-06481-w.

Vladislav Tumko, Jack Kim, Natalia Uspenskaia, Shaun Honig, Frederik Abel, Darren R. Lebl, Irene Hotalen, Serhii Kolisnyk, Mikhail Kochnev, Andrej Rusakov, and Raphaël Mourad. A neural network model for detection and classification of lumbar spinal stenosis on mri. *European Spine Journal*, 33:941–948, 3 2024. ISSN 14320932. doi: 10.1007/s00586-023-08089-2.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv*, 7 2018. doi: 10.48550/arXiv.1807.03748.

Jasper W. van der Graaf, Liron Brundel, Miranda L. van Hooff, Marinus de Kleuver, Nikolas Lessmann, Bas J. Maresch, Myrthe M. Vestering, Jacco Spermon, Bram van Ginneken, and Matthieu J.C.M. Rutten. Ai-based lumbar central canal stenosis classification on sagittal mr images is comparable to experienced radiologists using axial images. *European Radiology*, pages 1–9, 9 2024a. ISSN 14321084. doi: 10.1007/s00330-024-11080-0.

Jasper W. van der Graaf, Miranda L. van Hooff, Constantinus F.M. Buckens, Matthieu Rutten, Job L.C. van Susante, Robert Jan Kroeze, Marinus de Kleuver, Bram van Ginneken, and Nikolas Lessmann. Lumbar spine segmentation in mr images: a dataset and a public benchmark. *Scientific Data 2024 11:1*, 11:1–9, 3 2024b. ISSN 2052-4463. doi: 10.1038/s41597-024-03090-w.

I. Vasilev, D. Slater, G. Spacagna, P. Roelants, and V. Zocca. *Python Deep Learning: Exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow, 2nd Edition*. Packt Publishing, 2019. ISBN 9781789349702.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-December:5999–6009, 6 2017. ISSN 10495258. doi: 10.48550/arXiv.1706.03762.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv*, 2022. doi: 10.48550/arXiv.2207.02696.

Gang Wang, Yanfei Chen, Pei An, Hannyu Hong, Jinghu Hu, and Tiange Huang. Uav-yolov8: A small-object-detection model based on improved yolov8 for uav aerial photography scenarios. *Sensors*, 2023a. doi: 10.3390/s23167190.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature 2023 620:7972*, 620:47–60, 8 2023b. ISSN 1476-4687. doi: 10.1038/s41586-023-06221-2.

William C. Watters, Jamie Baisden, Thomas J. Gilbert, Scott Kreiner, Daniel K. Resnick, Christopher M. Bono, Gary Ghiselli, Michael H. Heggeness, Daniel J. Mazanec, Conor O'Neill, Charles A. Reitman, William O. Shaffer, Jeffrey T. Summers, and John F. Toton. Degenerative lumbar spinal stenosis: an evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spinal stenosis. *Spine Journal*, 8: 305–310, 3 2008. ISSN 15299430. doi: 10.1016/j.spinee.2007.10.033.

Dongho Won, Hye Jin Lee, Sang Joon Lee, and Sung Hoon Park. Lumbar spinal stenosis grading in multiple level magnetic resonance imaging using deep convolutional neural networks. *Global Spine Journal*, 15(4):2309–2317, 2025. doi: 10.1177/21925682241299332.

Fengze Wu, Marion Chiariglione, and Xiaoyi Raymond Gao. Automated optic disc and cup segmentation for glaucoma detection from fundus images using the detectron2's mask r-cnn. *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 567–570, 2022. doi: 10.1109/ISMSIT56059.2022.9932660.

Lite Wu, Sunil Munakomi, and Ricardo Cruz. Lumbar spinal stenosis. *StatPearls*, 1 2024. URL `https://www.ncbi.nlm.nih.gov/books/NBK531493/`.

Shoji Yabuki, Norio Fukumori, Misa Takegami, Yoshihiro Onishi, Koji Otani, Miho Sekiguchi, Takafumi Wakita, Shin Ichi Kikuchi, Shunichi Fukuhara, and Shin Ichi Konno. Prevalence of lumbar spinal stenosis, using the diagnostic support tool, and

correlated factors in japan: a population-based study. *Journal of Orthopaedic Science*, 18:893–900, 11 2013. ISSN 0949-2658. doi: 10.1007/S00776-013-0455-5.

Elzat Elham-Yilizati Yilihamu, Fan-Shuo Zeng, Jun Shang, Jin-Tao Yang, Hai Zhong, and Shi-Qing Feng. Gpt4lfs (generative pretrained transformer 4 omni for lumbar foramina stenosis): enhancing lumbar foraminal stenosis image classification through large multimodal models. *The Spine Journal*, 25(9):2071–2080, 2025. ISSN 1529-9430. doi: 10.1016/j.spinee.2025.03.011.

Fabio Zanchi, Raphaël Richard, Mahmoud Hussami, Arnaud Monier, Jean-François Knebel, and Patrick Omoumi. MRI of non-specific low back pain and/or lumbar radiculopathy: do we need T1 when using a sagittal t2-weighted dixon sequence? *Eur Radiol*, 30(5):2583–2593, February 2020.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. doi: 10.48550/arXiv.1710.09412.