

SPINEFORMER: VISION TRANSFORMER-DRIVEN LUMBAR SPINE SEGMENTATION

Md Tareq Mahmud, Subhajit Chakrabarty
Louisiana State University Shreveport

Corresponding Author: Md Tareq Mahmud
Louisiana State University Shreveport
Email: mahmudm28@lsus.edu
doi: 10.34107/UDUK9890313

ABSTRACT

Lower back pain (LBP) is one of the leading causes of disability worldwide, often linked to structural changes in the lumbar spine. Accurate segmentation of vertebrae, intervertebral discs (IVDs), and the spinal canal from MRI scans is critical for diagnosis and treatment planning, yet manual annotation remains time-consuming and inconsistent. In this work, we propose *SpineFormer*, a Vision Transformer (ViT)-based hybrid architecture for automated lumbar spine MRI segmentation, developed as part of the SPIDER Challenge. The model integrates a pre-trained ViT encoder with a lightweight convolutional decoder, enabling both global contextual reasoning and fine spatial refinement. Preprocessing included resampling, intensity normalization, adaptive cropping, and data augmentation to address inter-scanner variability. Training used a Dice + cross-entropy loss to mitigate class imbalance, particularly for the spinal canal. Experiments on the SPIDER dataset demonstrated robust performance, achieving mean Dice Similarity Coefficient (DSC) scores of 0.917 for vertebrae, 0.882 for IVDs, and 0.903 for the spinal canal, with an overall DSC of 0.900. Compared to the widely used CNN baseline (nnU-Net), our model showed superior consistency in challenging cases with disc degeneration or low contrast. Qualitative analysis confirmed that ViT attention maps captured global anatomical relationships, improving segmentation reliability across diverse patient anatomies. These findings highlight the potential of transformer-based models for clinical imaging tasks. While computationally more demanding than CNNs, SpineFormer demonstrates strong generalization and paves the way for future work on 3D volumetric transformers and multimodal integration for comprehensive spinal diagnosis.

Keywords: Vision Transformers; Lumbar Spine; MRI; Segmentation; Deep Learning

INTRODUCTION

Low back pain (LBP) affects more than 600 million people worldwide and is one of the most common causes of disability and work absenteeism [1]. Clinical studies have repeatedly linked LBP to degenerative changes in the lumbar spine, including disc herniation, vertebral misalignment, and spinal canal narrowing [2]. Magnetic Resonance Imaging (MRI) plays a crucial role in diagnosing these conditions due to its high soft-tissue contrast and ability to visualize spinal structures non-invasively [3]. However, manual annotation and analysis of MRI scans are time-consuming, subject to inter-rater variability, and not scalable for large datasets or real-time clinical workflows.

Automated segmentation of spinal structures—specifically vertebrae, intervertebral discs (IVDs), and the spinal canal—can significantly enhance diagnostic precision and facilitate downstream tasks such as disease grading, surgical planning, and biomechanical modeling [4]. Over the past decade, deep learning-based approaches, particularly convolutional neural networks (CNNs) [5], have become the standard for medical image segmentation. Models like U-Net and its variants (e.g., nnU-Net) have shown strong performance across various imaging modalities and anatomical regions [6]. Yet, CNNs inherently struggle with capturing long-range dependencies due to their localized receptive fields [7], which can limit performance in anatomically variable regions like the spine.

Despite the clinical importance of lumbar spine segmentation, the volume of research in this domain remains relatively limited compared to other anatomical regions. Traditional approaches