

End-to-End Deep Learning–Based Grading of Lumbar Facet Joint Degeneration and Analysis of Segmentation-Induced Performance Degradation

Baisen Chen (Nantong, China), Zhiming Cui (Nantong, China), Guanhua Xu (Nantong, China), Yuyu Sun (Nantong, China)

Introduction

Facet joint degeneration is an important imaging biomarker associated with chronic low back pain. Clinical grading is performed at the joint level and requires both accurate localization and severity assessment. However, most existing studies focus on segmentation or image-level classification, with limited evaluation of end-to-end grading performance. As multiple joints with different severities may coexist in a single scan, facet joint grading is inherently an instance-level task. This study aims to develop an end-to-end automated grading framework and analyze the impact of segmentation errors on severe degeneration detection.

Materials and Methods

A total of 145 lumbar CT scans were included, each annotated with five facet joints labeled as background, non-severe, or severe degeneration.

A 3D nnU-Net was trained for binary joint segmentation. Connected component analysis was used to extract joint instances, from which volumetric, morphological, and intensity features were derived to build a joint-level severe degeneration classifier.

Two evaluation settings were designed: (1) an oracle setting using ground-truth joint regions; and (2) an end-to-end setting using automatically segmented joints.

A direct three-class nnU-Net model was trained as a baseline. Performance was assessed at joint and case levels using AUC, F1-score, accuracy, and confusion matrices.

Results

The joint segmentation model achieved a mean Dice coefficient of 0.77 on the validation set.

Under the oracle setting, joint-level severe degeneration classification achieved an AUC of 0.76, indicating moderate discriminative ability when joint localization was accurate.

In the end-to-end setting, joint-level AUC decreased to 0.73, and severe degeneration recall dropped substantially to 0.16. Case-level severe detection accuracy was 0.45 under an any-severe aggregation rule. Error analysis demonstrated that missed detections, fragmented joint regions, and instance mismatches during segmentation were the primary contributors to the reduced severe recall.

The direct three-class segmentation approach showed unstable performance for joint-level severe degeneration assessment and did not provide reliable clinical grading results.

Conclusion

Lumbar facet joint degeneration grading is fundamentally an instance-level clinical task. Although semantic segmentation models may achieve satisfactory voxel-wise performance, segmentation errors significantly impair downstream severity assessment in end-to-end pipelines. These findings suggest that accurate joint localization and instance-aware modeling are critical for developing clinically reliable automated grading systems for facet joint degeneration.

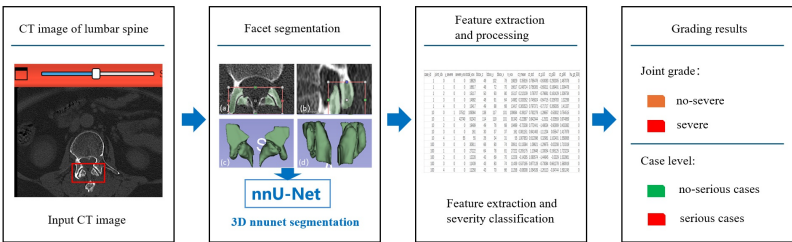


Figure 1. Overview of the proposed end-to-end framework for automated lumbar facet joint degeneration grading. Lumbar CT images were first processed using a 3D nnU-Net model for facet joint segmentation. Connected component analysis was applied to extract individual joint instances. Volumetric, morphological, and intensity-based features were then derived for joint-level severe degeneration classification. Case-level severity was determined using an any-severe aggregation strategy.

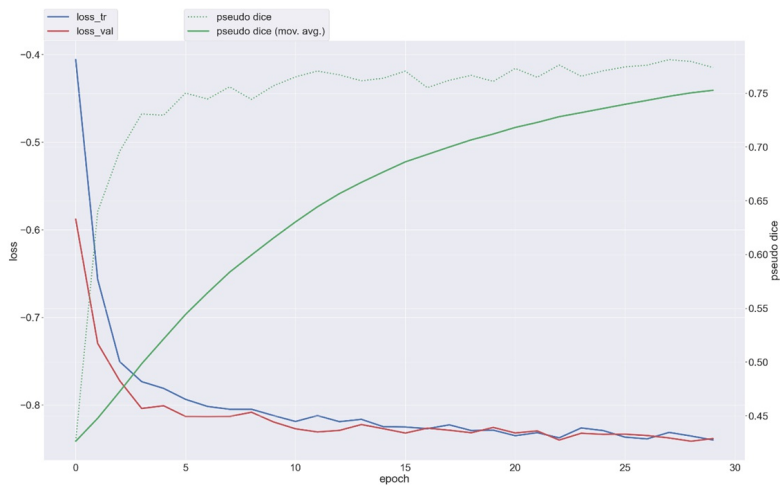


Figure 2. Training and validation performance of the 3D nnU-Net segmentation model. The training loss (blue) and validation loss (red) decreased steadily across epochs, while pseudo Dice scores (green) improved progressively, reaching approximately 0.77 at convergence. The results indicate stable optimization and satisfactory segmentation performance.