ARTIFICIAL INTELLIGENCE SEGMENTATION FRAMEWORK FOR IDENTIFYING SIGNIFICANT PATHOLOGICAL AREAS CAUSING LUMBAR SPINAL STENOSIS

[1]SAIED SALEM, [2]AHMED S. ELBADAWY, [3]RAZA MUKHLIS, [4]ERTAN BÜTÜN, [5]MUGAHED A. AL-ANTARI

[1,2,3,5]AISSLab, Department of Artificial Intelligence, College of Software & Convergence Technology, Daeyang AI Center, Sejong University, Seoul, 05006, Korea
[4]Department of Computer Engineering, Firat University, Elazığ, Turkey
E-mail: [1]en.mualshz@sejong.ac.kr

Abstract - Lumbar Spinal Stenosis (LSS) is a condition characterized by the narrowing of the spinal canal in the lumbar region, resulting in pressure on the spinal cord and nerves. This can cause symptoms like pain, numbness, weakness, or tingling sensations in the lower back, buttocks, and legs. In this study, we introduce a new AI-based multi-class segmentation framework designed to identify significant pathological regions, including the Intervertebral Disc (IVD), Posterior Element (PE), Thecal Sac (TS), and the Area between Anterior and Posterior (AAP) vertebrae. The proposed computer-aided diagnosis (CAD) framework comprises four main steps: medical MRI data acquisition and collection, preprocessing, optimization, selection of a backbone XAI model, and quantitative and qualitative performance evaluation. We utilized the Mendeley public benchmark LSS dataset to build and train the proposed AI segmentation framework. Meanwhile, the entire pipeline is verified and validated using our private MRI dataset (i.e., AISSLab LSS MRI dataset) where it is collected under the international research collaboration protocol between Korea and Turky. Among the three segmentation AI models tested (UNet, SwinUNETR, and UNet++), the UNet model demonstrated slightly better segmentation performance, achieving a class-wise dice scores of97.51% ($\pm$0.014), 92.78% ($\pm$0.033), 90.75% ($\pm$0.0718), and 78.88% ($\pm$0.0734) in terms of IVD, PE, TS and APP, respectively. Our proposed framework exhibits promising reliability and trustworthiness for medical industrial applications in LSS segmentation.

Keywords - Lumber Spinal Stenosis (LSS), Back Pain, Herniated Spinal Disk, Nerve Compressing, AI-based Segmentation, Spinal MRI Axial Images, Explainable AI (XAI).

I. INTRODUCTION

Lumbar Spinal Stenosis (LSS) continues to pose significant challenges in the realm of spinal pathology, contributing prominently to chronic lower back pain and neurological symptoms[1, 2]

This condition, characterized by a narrowing of the lumbar spinal canal, often arises from various degenerative processes affecting the vertebral column's components, including intervertebral discs, facet joints, and ligaments[3, 4].

Fig. 1 illustrates the fundamental concept of back pain associated with LSS and its manifestation in the human back region.

The resultant compression of spinal nerve roots manifests clinically as radicular back pain, atypical leg pain, sciatica, numbness and tingling in buttocks and legs, weakness in the legs, and pain that gets better with sitting[3, 5].

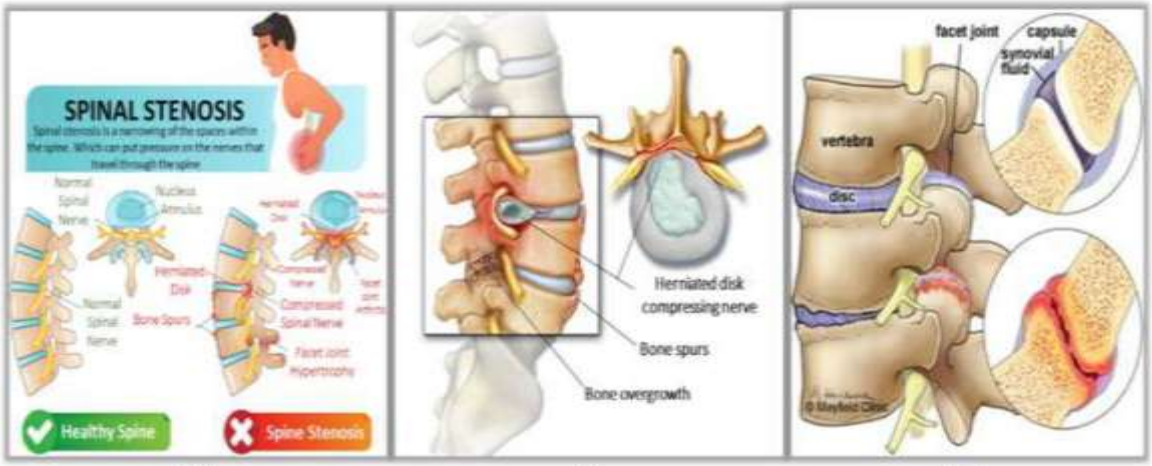
Such harmful pain profoundly impactsthelife quality of patients and their daily functional capacity[5]. Indeed, magnetic resonance imaging (MRI) stands as the foremost medical modality for diagnosing LSS challenges[6], offering a precise depiction of the dynamic spinal anatomy.Although MRI axial and sagittal scans are viable for inspecting LSS, the axial scan provides more detailed insights into the vertebra-specific pathological structures[7].

Timely identification of the location and evaluation of the extent of LSS are crucial for diagnosis and treatment[5, 8].

Recognizing the significance of robust data acquisition in axial MRI [9], our study emphasizes utilizingT2-weighted slice images to automatically segment the most important spin vertebral regions: (1) Intervertebral Disc (IVD), (2) Posterior Element (PE), (3) Thecal Sac (TS), and (4) the Area between Anterior and Posterior (AAP) vertebrae.

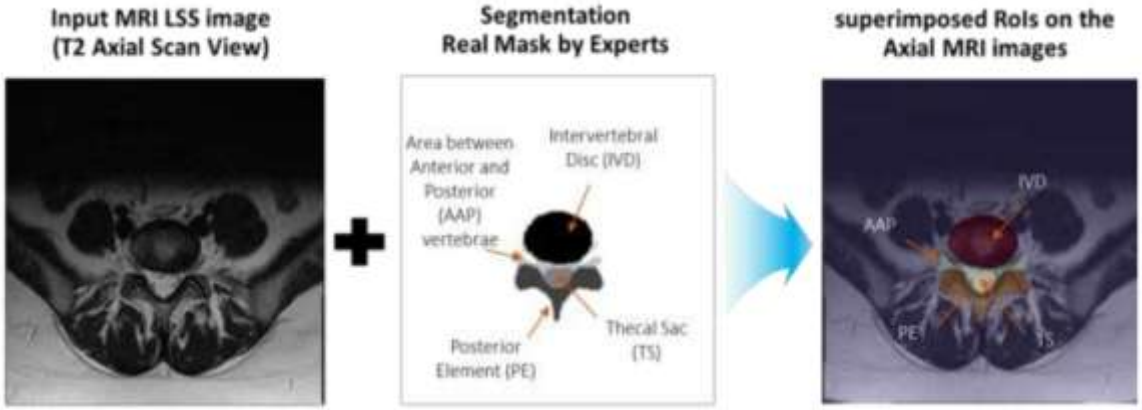
Fig. 2(a) depicts an Axial MRI scan exhibiting the pathological structure of the spinal vertebra, accompanied by expert annotations of these regions.

Recognizing the utility and reliability of supine MRI for lumbar spine segmentation, particularly in the context of aging, we endeavor in this paper to introduce a comprehensive approach to segmenting lumbar spinal structures.



(a) (b) (c)

Fig. 1: The fundamental concept of the Lumber Spinal Stenosis (LSS).



(a) (b)

Fig. 2: Axial MRI slice (T2_00138_D5) to show the pathological structure of the spinal vertebra's regions of interest (RoIs): IVD, PE, TS, and AAP.

Our goal is to establish a foundation for enhanced diagnostic precision and treatment strategizing in LSS patients through large-scale analyses.Similarly, Al-kafri et al. [9] employed a SegNetAI model to segment theaxial spine image anatomyinto four classes achievingIoUs for IVD (92%), PE (78%), TS (85%), and APP (53%).

Meanwhile, Li et al. [10]proposed a multi-scale attention U-shaped network (MANet) for semanticsegmentation of axial MR imageswhere the model it was adapted to different scale targets during the retaining process.In 2022, Pang et al. introduced the SGRNet model, incorporating a segmentation-guided attention module (SGAM) to generate attention-aware features[11].

These features guided the regression path, enhancing the performance of spine indices measurement. The automatic estimation of these indices remains under investigation, with future research poised to focus on this as the next step in our work.The main contributions in this study are briefly summarized here:

- A new multi-class computer-aided diagnosis (CAD) framework is proposed to segment the most important regions in the spinal vertebrae to cause the LSS: IVD, PE, TS, and AAP.
- A comprehensive experimental study has been conducted to automatically optimize the hyperparameters of the three state-of-the-art AI segmentation models using the iterative trials of random search approach[12].
- Real private MRI dataset is collected to validate the robustness of the proposed CAD pipeline where it was trained on another public dataset.

The rest of this article is managed as follows. Section 2 presents the details of the proposed methodology, while section 3 shows theexperimental results. Section 4 concludes the most findings in this study.

II. MATERIALS AND METHOD

2.1. Mendeley LSS MRI Dataset

To achieve the goal of the proposed CAD segmentation framework, we use the open-access

LSS MRI dataset called Mendeley public benchmark LSS dataset [9]. This data was collected from 515 patients with 1,545 axial T1 and T2 slices. Due to the labelling challenges, only three axial MR image slices were selected and annotated by the experts. These slices were carefully selected and extracted to be close to the half-height of L3–L4, L4–L5, and L5–S1 IVDs. In total,1,545 MRI T1 and T2-weighted images to perform the proposed research experiments. In this study, we use the Axial T2-weighted because they have been labeled by experts for four class labels: IVD, PE, TS, and AAP. Such segmentation labels support future work in how to automate the process ofindices measurements of that could indicate the existence of the stenosis[13].

2.2. AISSLab LSS MRI Dataset

To validate the capabilities of the AI models to deal with unseen and different dataset that the model didn't train on, a real LSS dataset is collected from Firat university hospital, Turkey from August 2013 till September 2023. This dataset is collected based on the official international research project supported by National Research Foundation (NRF) in Korea and TUBITAK in Turkey. This dataset contains Axial and Sagittal scans for 508 patients (174 male and 334 female) associated with the text description by experts for each case. For each patient, the 13 Sagittal slices are collected with thickness of 4 mm and slice range of 0.4 mm.Meanwhile, 5 Axial slices between each disc are collected with thickness of 4 mm and 0.4 mm slice range.The MRI Sagittal images are acquired using Philips ingenia 1.5 Tesla, while the Axial scans are acquired using Philips ingenia 3 tesla MRIdevices. The data annotation is still in progress, and we use it just for validationtasks.

2.3. Preprocessing

To achieve standardization, all MRI scans are resized to a pre-defined resolution. The optimal resolution (either 128×128 or 256×256 pixels) is determined through an automated hyperparameter optimization process. For intensity normalization,the Z-score normalization is applied to all images. This step ensures a consistent intensity range across all pixels, potentially improving segmentation performance[14]. Then, we have applied the data splitting in the patient level to be 70% for training, 10% for validation, and 20% for testing. For training data augmentation to enlarge its size, we use the strategy of augmentation on the fly[15], where the system randomly selects the most effective augmentation techniques from two predefined packages during automatic hyperparameter optimization. (1) Random rotation (-35° to +35°), Horizontal flip, Shift-scale rotation, Random Gamma correction. (2) Vertical flip, random adjustment of brightness and contrast, elastic transformation.This approach allows the training process to dynamically choose the most suitable augmentation techniques for optimal performance.

AI Model	Optimizer	Batch size	LR	Backbone	Loss	Class weighting	Aug.
Unet	AdamW	16	23e-4	se_resnet 152	Log cosh dice	✖	✓
SwinUNETR	Adam	16	26e-4	-	Dice	✓	✖
Unet++	Adam	64	39e-4	-	Focal	✓	✓

Table 1. The configuration of the automatic hyperparameter optimization using Mendeley LSS MRI dataset

2.4. Automatic Hyperparameter Optimization

To optimize the hyperparameters of all models, the automated hyperparameter tuning with random search method is used [12].The configuration parametersthat have been selected to be tuned during the automatic optimization areoptimizer, loss function, batch size, image size, learning rate, include the background or not, weighted loss, network architecture with different backbones, augmentation process, and the measurement of mean dice score. Fig. 4 shows a screenshot of one sweep trail as an example of automated hyperparameter execution. The best optimization result is recorded in Table 1 for each AI model (i.e., UNet, SwinUNETR and UNet++).

2.5. The proposed CAD Framework

The proposed CAD framework for back pain LSS segmentation involves five key steps (see Fig. 3). Initially, a patient experiencing back pain consults with radiology experts at a hospital. The radiologists order MRI scans: Axial or Sagittal, T1 or T2 weighted. Axial T2-weighted scans preferred for reliable manual analysis because the cerebrospinal fluid is better distinguished by radiologists. Experts then annotate selected Axial scans, outlining relevant structures (IVD, PE, TS, AAP) to create training and evaluating data for supervised AI models. Then, images undergo preprocessing for compatibility with the AI models.

Hyperparameter optimization automates model selection and fine-tuning for optimal segmentation results. Finally, three robust segmentation models are recommended and deployed during the automatic hyperparameter optimization process: UNet, UNet++, and SwinUNETR. The first two of these models utilize convolutional neural networks, while the third employs an attention-based vision transformer architecture. To support the healthcare staff, we propose this CAD system to automatically and rapidly segment and provide the fine counter of various spinal regions to be easily inspection and assure the stenosis case or not. By this way, the diagnosing process take less time and minimize the labor cost of doctors as well as the patient.

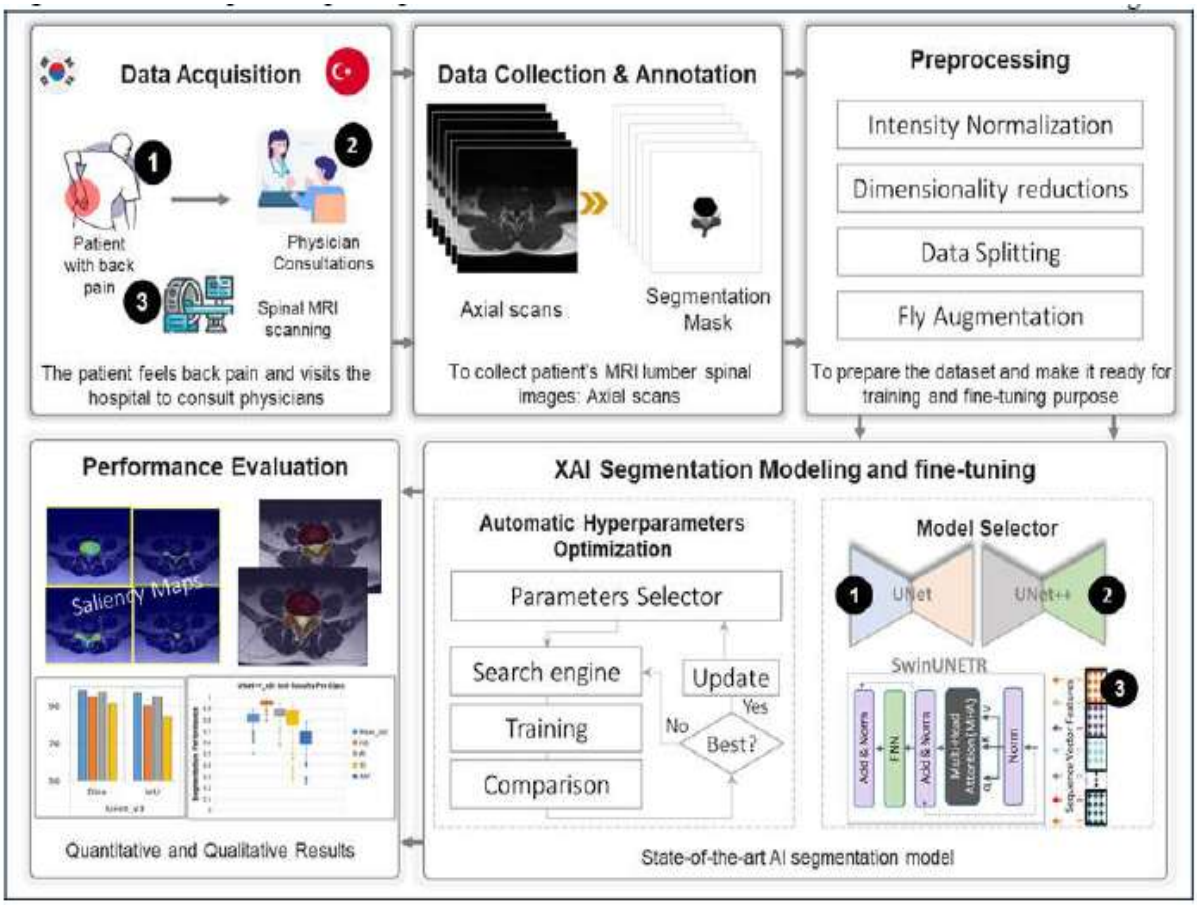


Fig. 3: The proposed end-to-end CAD segmentation framework

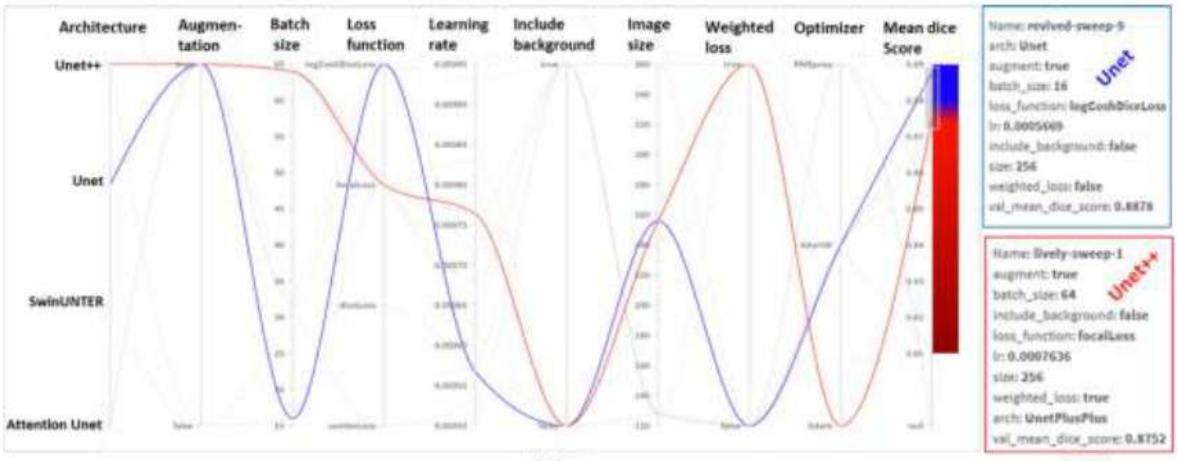


Fig. 4: (a) The automated hyperparameters optimization sweep path, and (b) the best selected hyperparameters.

2.6. Training Execution Environment

The technical experiments for all models using both public and private datasets are conducted in a unique learning environment. This is to assure the correct and direct comparison performance among all AI models.Once the auto-parameter optimization is finished, the best configuration of each AI model is summarized in Table 2.Trainingexecution is launched over 50 epochs.

2.7. Evaluation Strategy

We evaluate the effectiveness of the proposed CAD framework quantitively and qualitatively. For quantitative evaluation in terms of class-wise and overall average performance, we evaluate the models using the mean intersection over union (mIoU), dice score, F1-score, precision (Pre), recall (Re), specificity (Spe), and accuracy (Acc)[16-18]. For qualitative evaluation, we visually compare the real

and predicted masks by superimposing them over the original input images.

III. RESULTS AND DISCUSSION

In this section, we will summarize the evaluation segmentation results of the best top three models. Table 2 shows the performance achievement of UNet with Squeeze-and-Excitation based ResNet backbone (se_resnet152) per class. It is clearly shown an encouraged segmentation results for all classes with an average dice and IoUscores of 89.98% and 82.84%. The average pre, Re, and Spe record promising outcomes by 94.11%, 94.65%, and 94.65%, respectively.

Class	UNet						
	Dice	IoU	F1-score	Pre	Re	Spe	Acc
IVD	97.51	95.18	97.51	97.24	97.83	99.89	99.82
PE	92.78	86.70	92.78	93.03	92.77	99.89	99.77
TS	90.75	83.78	90.75	90.12	92.29	99.97	99.95
APP	78.88	83.78	78.88	78.09	81.01	99.88	99.77
Avg.	**89.98**	**82.84**	**94.35**	**94.11**	**94.65**	**94.65**	**99.83**

Table 2.Segmentation evaluation performance (%) per class using the testing set MRI images

Class	SwinUNETR						
	Dice	IoU	F1-score	Pre	Re	Spe	Acc
IVD	97.33	94.84	97.33	97.13	97.59	99.89	99.80
PE	91.84	85.17	91.84	92.53	91.47	99.88	99.75
TS	90.18	82.98	90.18	90.26	91.23	99.97	99.95
APP	77.74	94.21	77.74	77.04	97.92	99.87	99.76
Avg.	**89.27**	**81.80**	**93.87**	**93.82**	**93.99**	**93.99**	**99.81**

Table 3.Segmentation evaluation performance (%) per class using the testing set MRI images

Similarly, Table 3 depicts the segmentation outcomes by the second top SwinUNETR. The evaluation performance is similar in the UNet. The averages achieved dice and IoU score are 89.27% and 81.80%, respectively.

It shows that the highest performance isa achieved for the IVD class, while the lowest performance is recorded for APP class. This could be due to the minority of the APP pixels among all other classes. This occurseven though the weighted loss function is considered during the training phase. The F1-score can measure the imbalance pixels among various classes to record the average performance of 93.87%. While the UNet++model demonstrates the lowest segmentation performance in Table 4, the difference compared to other models is minimal.

Class	Unet++						
	Dice	IoU	F1-score	Pre	Re	Spe	Acc
IVD	97.45	95.05	97.45	97.07	97.88	99.89	99.81
PE	92.15	85.63	92.15	91.61	92.95	99.86	99.75
TS	88.41	80.27	88.41	88.88	89.34	99.97	99.94
APP	77.09	63.26	77.09	76.03	79.47	99.86	99.75
Avg.	**88.77**	**81.06**	**93.88**	**93.37**	**94.46**	**94.46**	**99.81**

Table 4.Segmentation evaluation performance (%) per class using the testing set MRI images

The comparison among the top AI segmentation models is presented in Fig. 5 in terms of dice score for APP and IoU for TS classes. It is clearly shown that the segmentation performance of these models is slightly similar, while the top rank is recorded for UNet model.

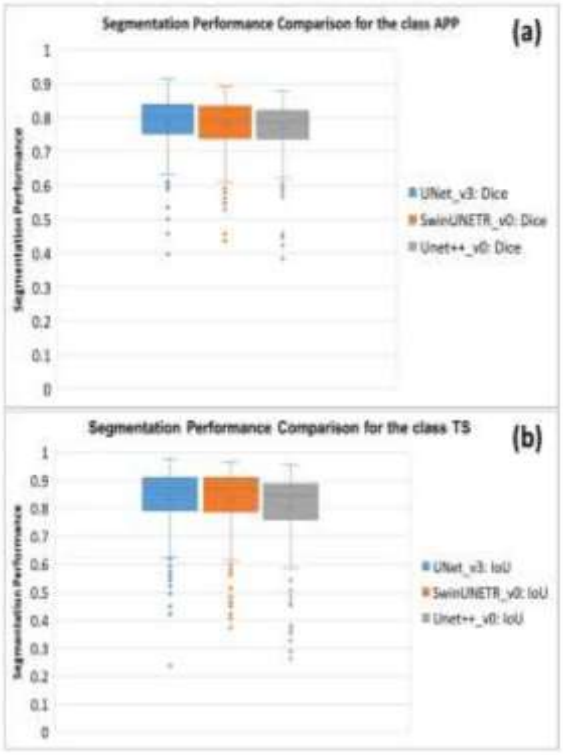


Fig. 5: Segmentation performance comparison among the top three models (UNet, SwinUNETR, and UNet++) in terms of (a) Dies score for APP class and (b) IoU for TS class.

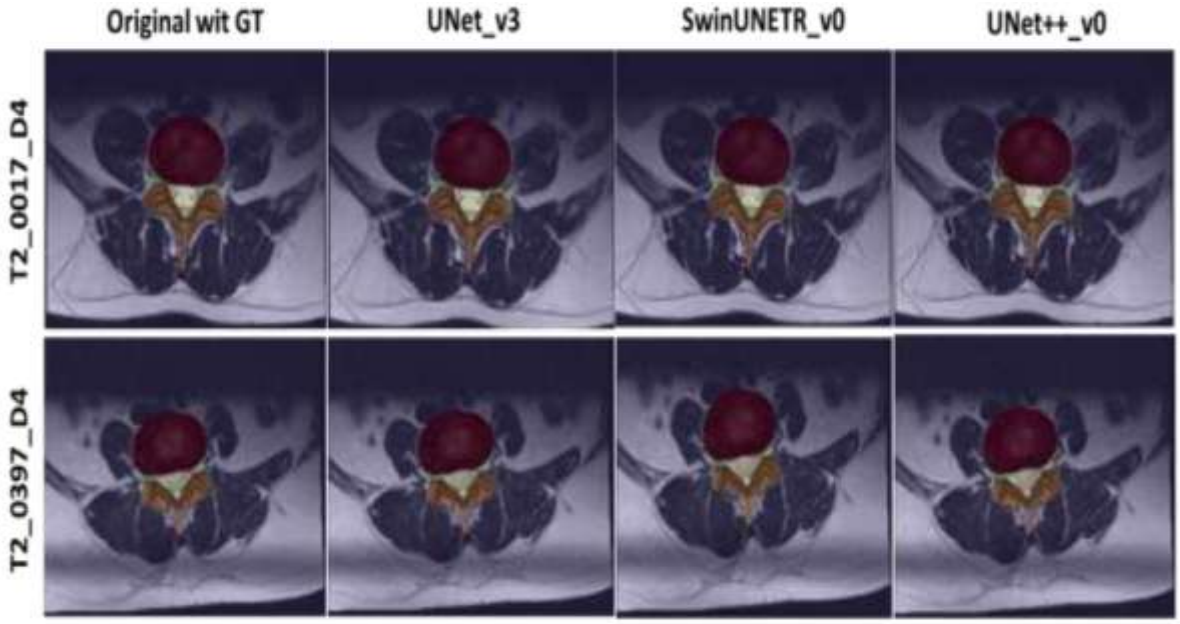


Fig. 6. Examples of multi-class segmentation results using UNet_v3, SwinUNETR_v0, and UNet++_v0 to show the outcome comparison among four classes IVD, PE, TS, and APP.

Fig. 6 visually compares the performance of all AI models against the ground truth, revealing qualitatively similar segmentation results across models, with all four classes being accurately segmented. Such impressive results provide a promising AI segmentation in practical industrial applications.

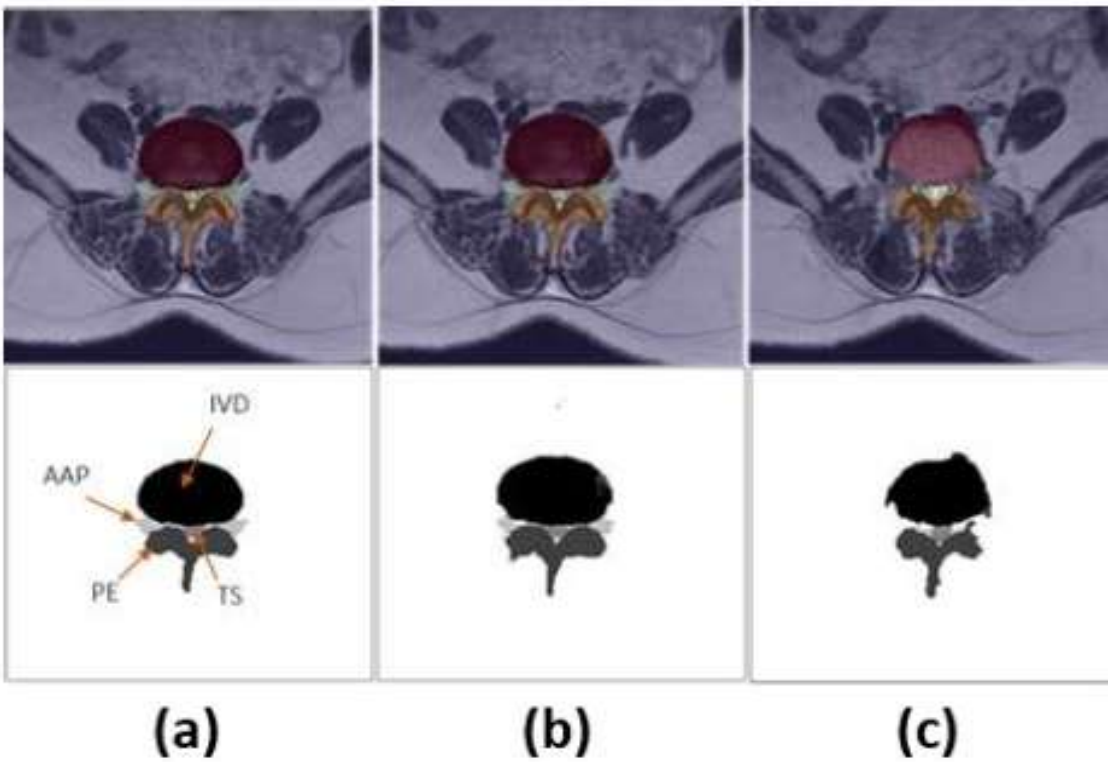


Fig. 7. Validation performance using the private LSS dataset for the same slice of PATIENT346270: (a) UNet, (b) SwinUNETR, and (c) UNet++.

Fig. 7 displays the visual predictions from all AI models on the private LSS dataset. The predicted masks demonstrate the proposed CAD segmentation framework's ability to generalize effectively and offer a smart solution even when applied to datasets not encountered during training. It is clearly shown that UNet achieves superior results, while UNet++ faces some limitations. SwinUNETR has competitive segmentation results compared to UNet.

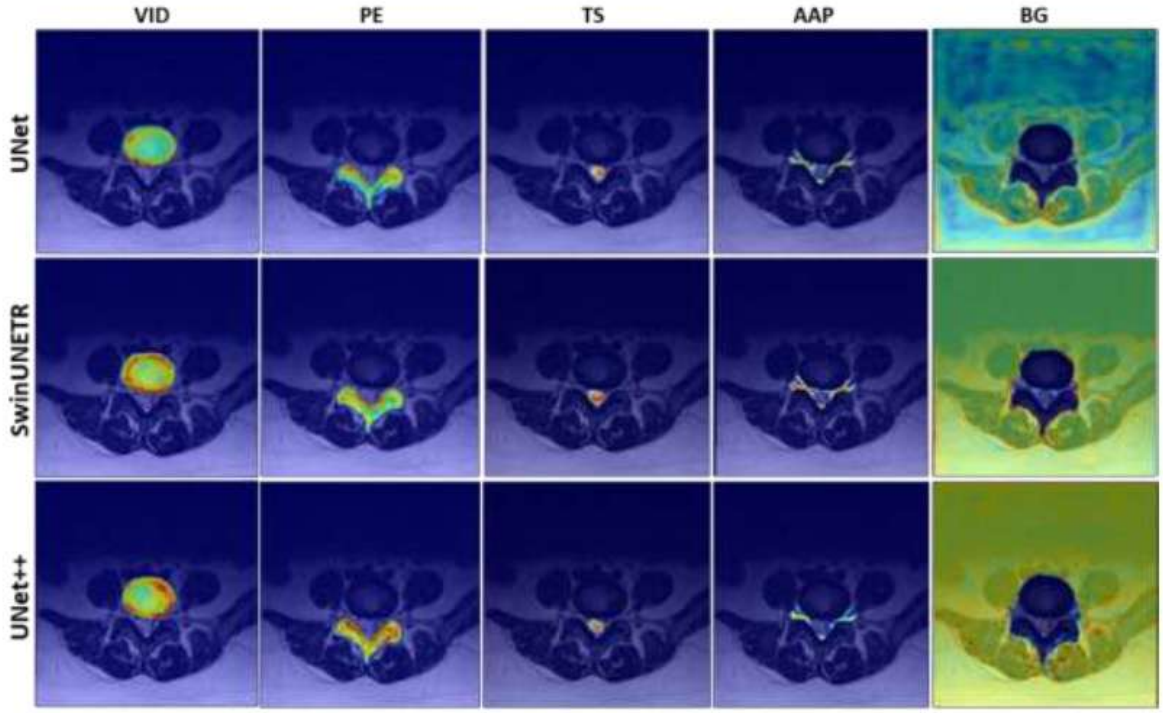


Fig. 8. Visual saliency XAI maps for each class based on the attention Grad-CAM mechanism.

Finally, we use the XAI Grad-CAM to generate the explanation heatmaps for each classas shown in Fig. 8. Indeed, the XAI heatmaps are essential in the segmentation task of LSS, offering valuable insights into model confidence, subtle features, clinical decision-making, model performance evaluation, and communication. By incorporatingheatmaps into the diagnostic workflow, clinicians can improve the accuracy and efficiency of LSS diagnosis and treatment planning.

IV. CONCLUSION

This study introduces a novel multi-class segmentation framework for identifying key pathological regions causing back pain in lumbar spinal stenosis (LSS). The framework utilizes three AI models: two CNN-based and one transformer-based, demonstrating impressive performance in simultaneous multi-class segmentation. The accuracy of this approach holds promise for developing robust automatic measurement indices across all lumbar vertebrae in future research.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-00166402 and RS-2023-00256517).

REFERENCES

[1] B. H. Lee, S.-H. Moon, K.-S. Suk, H.-S. Kim, J.-H. Yang, and H.-M. Lee, "Lumbar Spinal Stenosis: Pathophysiology and Treatment Principle: A Narrative Review," Asian Spine Journal, vol. 14, pp. 682-693, 2020.

[2] Kwon, Ji-won, Seong-Hwan Moon, Si-Young Park, Sang-Jun Park, Sub-Ri Park, Kyung-Soo Suk, Hak-Sun Kim, and Byung Ho Lee. "Lumbar spinal stenosis: review update 2022." Asian spine journal 16, no. 5 (2022): 789.

[3] J. N. Katz and M. B. Harris, "Lumbar Spinal Stenosis," New England Journal of Medicine, vol. 358, pp. 818-825, 2008.

[4] S. J. Atlas and A. Delitto, "Spinal Stenosis: Surgical versus Nonsurgical Treatment," Clinical Orthopaedics and Related Research®, vol. 443, pp. 198-207, 2006.

[5] J.-W. Kwon, S.-H. Moon, S.-Y. Park, S.-J. Park, S.-R. Park, K.-S. Suk, et al., "Lumbar Spinal Stenosis: Review Update 2022," Asian Spine Journal, vol. 16, pp. 789-798, 2022.

[6] F. Cheng, J. You, and Y. R. Rampersaud, "Relationship between spinal magnetic resonance imaging findings and candidacy for spinal surgery," Can Fam Physician, vol. 56, pp. e323-30, Sep 2010.

[7] J. Hartman, M. Granville, and R. E. Jacobson, "Radiologic Evaluation Of Lumbar Spinal Stenosis: The Integration Of Sagittal And Axial Views In Decision Making For Minimally Invasive Surgical Procedures," Cureus, 2019.

[8] J. T. P. D. Hallinan, L. Zhu, K. Yang, A. Makmur, D. A. R. Algazwi, Y. L. Thian, et al., "Deep Learning Model for Automated Detection and Classification of Central Canal, Lateral Recess, and Neural Foraminal Stenosis at Lumbar Spine MRI," Radiology, vol. 300, pp. 130-138, 2021.

[9] A. S. Al-Kafri, S. Sudirman, A. Hussain, D. Al-Jumeily, F. Natalia, H. Meidia, et al., "Boundary Delineation of MRI Images for Lumbar Spinal Stenosis Detection Through Semantic Segmentation Using Deep Neural Networks," IEEE Access, vol. 7, pp. 43487-43501, 2019.

[10] H. Li, H. Luo, W. Huan, Z. Shi, C. Yan, L. Wang, et al., "Automatic lumbar spinal MRI image segmentation with a multi-scale attention network," Neural Computing and Applications, vol. 33, pp. 11589-11602, 2021.

[11] C. Pang, Z. Su, L. Lin, G. Lin, J. He, H. Lu, et al., "Automated measurement of spine indices on axial MR images for lumbar spinal stenosis diagnosis using segmentation-guided regression network," Medical Physics, vol. 50, pp. 104-116, 2023.

[12] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, et al., "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," WIREs Data Mining and Knowledge Discovery, vol. 13, 2023.

[13] F. Natalia, H. Meidia, N. Afriliana, J. C. Young, R. E. Yunus, M. Al-Jumaily, et al., "Automated measurement of anteroposterior diameter and foraminal widths in MRI images for lumbar spinal stenosis diagnosis," PLOS ONE, vol. 15, p. e0241309, 2020.

[14] R. Kumar, A. Gupta, H. S. Arora, and B. Raman, "CBSN: Comparative measures of normalization techniques for brain tumor segmentation using SRCNet," Multimedia Tools and Applications, vol. 81, pp. 13203-13235, 2022.

[15] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," Array, vol. 16, p. 100258, 2022.

[16] Z. Al-Huda, B. Peng, R. N. A. Algburi, M. A. Al-antari, A.-J. Rabea, O. Al-maqtari, et al., "Asymmetric dual-decoder-U-Net for pavement crack semantic segmentation," Automation in Construction, vol. 156, p. 105138, 2023.

[17] Z. Al-Huda, B. Peng, R. N. A. Algburi, M. A. Al-antari, A.-J. Rabea, and D. Zhai, "A hybrid deep learning pavement crack semantic segmentation," Engineering Applications of Artificial Intelligence, vol. 122, p. 106142, 2023.

[18] M. A. Al-antari, Z. F. Shaaf, M. M. A. Jamil, N. A. Samee, R. Alkanhel, M. Talo, et al., "Deep learning myocardial infarction segmentation framework from cardiac magnetic resonance images," Biomedical Signal Processing and Control, vol. 89, p. 105710, 2024.

★★★