A Hybrid Dual-Backbone Deep Learning Framework for Lumbar Spinal Stenosis Classification from MRI with External Validation

1st Sadia Hamid
Department of Information and Communication Technology
Bangladesh University of Professionals
Dhaka-1216, Bangladesh
obonyhamid02@gmail.com

2nd Mugdha Farheen Kibria
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
mugdhafarheenkibria@gmail.com

3rd Fatema Tuz Zohora Panna
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
zohoratuzfatema786@gmail.com

4th Marjana Khanam Roza
Department of Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh
mk.roza1999@gmail.com

5th Mansura Afsana Rima
Department of Computer Science and Engineering
World University of Bangladesh
Dhaka-1230, Bangladesh
afsanarima85@gmail.com

6th MD Rezaul Karim
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
reza15361382@gmail.com

7th Tauhidur Rahman Sakib
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
tauhidurrahmansakib@gmail.com

***Abstract*—The automated diagnosis of lumbar spinal stenosis from magnetic resonance imaging (MRI) remains challenging due to anatomical complexity and variability in interpretation. This study introduces SpineNet-Fusion, a hybrid deep learning framework that combines EfficientNet-B3 and DenseNet-121 to capture both global and fine-grained features. The model is trained on a publicly available Kaggle dataset and evaluated using 5-fold cross-validation and external clinical data from Bangabandhu Sheikh Mujib Medical University (BSMMU). The experimental results demonstrate robust performance, achieving a mean cross-validation accuracy of 97.4% and an external validation accuracy of 92.9%. Grad-CAM visualization further enhances interpretability by highlighting clinically relevant regions. The framework demonstrates strong generalization capability and provides support for clinical decision-making in real-world healthcare settings.**

Keywords—Lumbar Spinal Stenosis, SpineNet-Fusion, EfficientNet, DenseNet, MRI Classification, Feature Fusion, Computer-Aided Diagnosis, Grad-CAM, Explainable AI.

I. INTRODUCTION

The diagnosis of lumbar spinal stenosis (LSS) requires accurate assessment because this degenerative spinal disorder leads to chronic pain, reduced mobility, and neurological disability. The condition develops when the spinal canal narrows, resulting in nerve root compression that can lead to severe disability. Radiologists face difficulties when manually examining lumbar MRI scans because the process is time-consuming and complex and may produce varying results depending on the radiologist's level of expertise and subtle differences in spinal anatomy. Traditional diagnostic approaches depend on expert assessment, which creates challenges in scaling their application and maintaining consistent results across different medical facilities. Deep learning (DL) methods, which operate through automated feature extraction, enable systems to perform objective analysis while learning complex patterns directly from imaging data. The use of DL in spinal disorder classification improves diagnostic efficiency, reduces human bias, and enhances system consistency in computer-aided diagnosis.

Recent studies demonstrate the effectiveness of deep learning (DL) in spinal imaging for automated diagnosis and classification tasks. DeepSPINE achieved Dice scores near 0.93 and AUC values up to 0.983 for segmentation, localization, and grading tasks [1]. VGG16-based models have been used for lumbar spinal stenosis classification, achieving accuracy of up to 87.70% [2]. Multi-stage frameworks integrating segmentation and classification further improved diagnostic performance, reporting AUROC values between 0.95 and 0.963 with strong agreement with radiologists [3], [4]. Clinical and biomechanical studies have also emphasized the need for accurate, consistent, and standardized diagnostic tools for spinal disorders [5]. Two-stage DL systems demonstrated high reliability ($\kappa \approx 0.96$) for multi-region stenosis detection by effectively handling anatomical variations [6]. Alternative imaging approaches, including radiograph-based models, achieved AUROC values around 0.90 [7], while a recent meta-analysis reported an overall AUC of approximately 0.92 for AI-based lumbar stenosis diagnosis, highlighting persistent challenges related to dataset heterogeneity and model generalization [8]. CT-based deep learning methods achieved classification accuracies of up to 88% [9], although MRI remains the preferred imaging modality due to its superior soft-tissue

visualization capabilities [10]. Segmentation-based approaches such as U-Net reported Dice scores of approximately 92.9% [11], while EfficientNet-based models achieved classification accuracies of up to 91.56% for lumbar spinal stenosis analysis [12]. Additional studies explored machine learning-based predictive modeling [13], YOLO-based real-time detection systems [14], advanced semantic segmentation techniques [15], CT-based cervical stenosis classification [16], and multilevel grading frameworks with expert-level agreement reaching 96.8% [17]. More recently, EfficientNet introduced compound model scaling for improved feature extraction efficiency [18], whereas DenseNet employed dense connectivity to enhance feature reuse and gradient propagation [19]. These architectures have demonstrated strong performance across medical imaging applications and provide complementary strengths for feature representation. Despite these advances, most existing approaches rely on single-backbone architectures, creating a need for hybrid frameworks that effectively combine global contextual information, fine-grained anatomical features, and model interpretability.

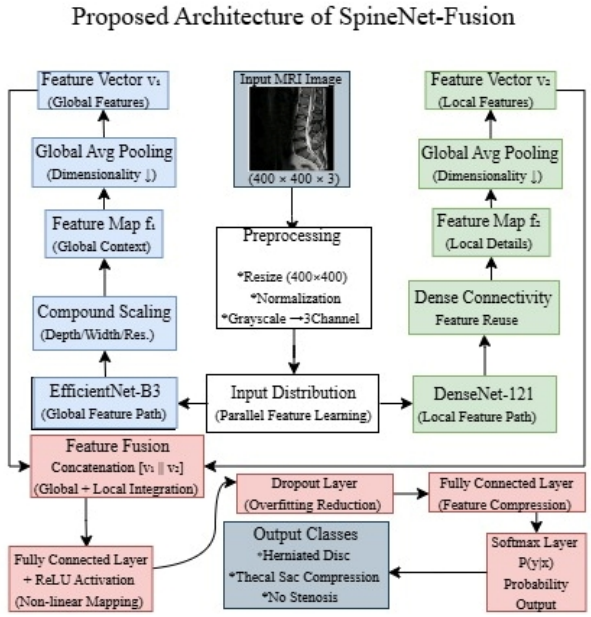
Despite recent advances, most existing methods rely on single-backbone architectures, limiting their ability to capture both global contextual information and fine-grained anatomical details. Furthermore, limited feature fusion and poor interpretability remain challenges for clinical adoption. To address these issues, this paper proposes SpineNet-Fusion, a dual-backbone framework that combines EfficientNet-B3 and DenseNet-121 for lumbar MRI classification, while Grad-CAM provides visual interpretability of clinically relevant regions. The main contributions of this work are as follows:

- **Hybrid Dual-Backbone Architecture** – Integration of EfficientNet-B3 and DenseNet-121 for multi-representational learning.
- **Improved Classification Performance** – Robust and accurate lumbar MRI classification.
- **Explainable AI Integration** – Grad-CAM-based interpretability for clinical relevance.
- **External and Cross-Validation** – Demonstration of strong generalization using cross-validation and external clinical data.
- **Efficient Diagnostic Framework** – Scalable solution for computer-aided spine analysis.

The remainder of this paper is organized as follows. Section II presents the methodology. Section III presents the results and analysis. Section IV concludes the paper.

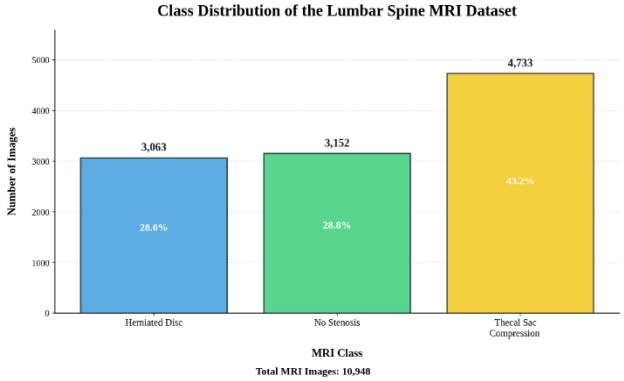
II. METHODOLOGY

This study proposes SpineNet-Fusion, a hybrid deep learning framework for classifying lumbar spinal disorders from MRI scans. It combines EfficientNet-B3 and DenseNet-121 to capture global and fine-grained features, followed by feature fusion and classification. Grad-CAM provides visual explanations by highlighting clinically relevant regions that influence model predictions.



A. Dataset Overview

The study used a publicly available Kaggle lumbar spine MRI dataset containing 10,948 images from three classes: Herniated Disc, Thecal Sac Compression, and No Stenosis.



As shown in Fig. 1, the dataset has moderate class imbalance, with more Thecal Sac Compression samples. Stratified splitting, class-weighted categorical cross-entropy, and uniform data augmentation were applied to reduce bias and improve generalization. For external validation, 580 anonymized lumbar MRI images were collected from Bangabandhu Sheikh Mujib Medical University (BSMMU), Bangladesh, including 180 Herniated Disc, 190 No Stenosis, and 210 Thecal Sac Compression cases. This independent dataset was used to assess the model's generalization in real-world clinical settings.

Fig. 1. Class distribution of the lumbar spine MRI dataset.

B. Data Preprocessing and Augmentation

The research team conducted data preprocessing steps to standardize MRI image size and intensity values, which enhanced model performance. The team performed image resizing to convert all images to a common resolution of 400 × 400 pixels, matching the backbone network input specifications while maintaining consistent spatial dimensions. Because the MRI images were grayscale, they were replicated into three channels ensuring compatibility

with pretrained convolutional networks that use RGB inputs. Pixel intensity normalization scaled image values to the range [0,1] using the following definition:

(1)

where denotes the original pixel intensity at spatial location x, y, and channel c. This normalization helps stabilize gradient updates and accelerate model convergence during training.

During training, data augmentation methods were applied, including random rotation, horizontal flipping, zooming, and slight translation, to achieve better generalization and reduce overfitting. These transformations simulate variations in MRI acquisition, enabling the system to perform better on unseen data. The researchers applied only resizing and normalization to the validation data to maintain evaluation fairness.

C. *Proposed SpineNet-Fusion Architecture*

The proposed SpineNet-Fusion architecture is a hybrid dual-backbone deep learning framework designed to capture complementary feature representations from lumbar MRI images. By integrating EfficientNet-B3 and DenseNet-121, the model simultaneously learns global anatomical context and fine-grained structural characteristics associated with lumbar spinal abnormalities. This complementary learning strategy enhances feature discrimination and improves classification performance. The overall architecture of the proposed SpineNet-Fusion framework is illustrated in Fig. 2.

(2)

where and represent feature maps extracted from EfficientNet-B3 and DenseNet-121, respectively. EfficientNet-B3 employs a compound scaling strategy to efficiently capture high-level semantic and global contextual features [18], while DenseNet-121 utilizes dense connectivity to promote feature reuse and improve gradient propagation for learning fine-grained local representations [19]. Global Average Pooling (GAP) processes the output feature maps from both branches by reducing spatial dimensions while preserving essential information. The resulting feature vectors are then concatenated to form a unified representation for the final classification process. This feature fusion strategy enables the model to leverage complementary information from both backbones, resulting in more discriminative feature representations for lumbar spinal abnormality classification.

(3)

This fusion mechanism combines complementary global and local feature representations, enhancing the discriminative capability for lumbar spinal abnormality classification. The fused feature vector is processed through fully connected layers with ReLU activation and dropout regularization, followed by a Softmax classifier that generates normalized class probabilities for final prediction. Overall, this design further strengthens the model's ability to produce reliable and consistent predictions across different data samples.

(4)

where represents the output logits corresponding to each class. The model uses this probabilistic framework to generate confidence scores for every class, enabling it to accurately distinguish between different lumbar spinal disorders while supporting computer-aided diagnosis systems.

Fig. 2. Proposed architecture of SpineNet-Fusion.

D. *Training Configuration and Evaluation Metrics*

SpineNet-Fusion was trained using a transfer learning strategy with ImageNet-pretrained EfficientNet-B3 and DenseNet-121 backbones. Training was performed using the Adam optimizer with a batch size of 32, an initial learning rate of 1×10^{-4}, and 25 epochs. Validation data were used to monitor convergence and reduce overfitting. The multi-class classification task employed categorical cross-entropy loss,



while model performance was evaluated using accuracy and macro-averaged F1-score, defined as:

(5)

(6)

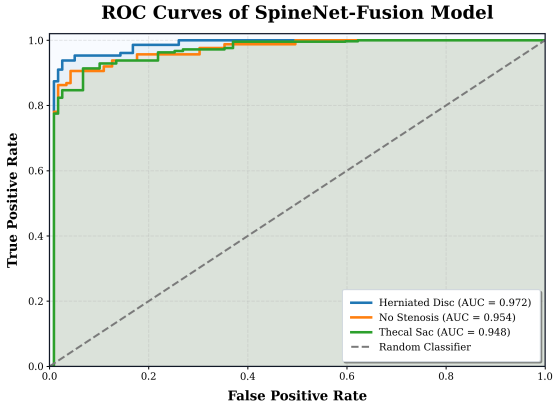
where and represent precision and recall for the class, respectively, and is the number of classes. These metrics provide a balanced evaluation because they effectively assess datasets with uneven class distributions. The research experiments were executed using Python-based deep learning frameworks in a GPU-enabled environment while maintaining fixed random seed values to ensure both reproducibility and consistent evaluation results.

III. RESULTS AND DISCUSSION

This section evaluates the proposed SpineNet-Fusion framework for lumbar spinal condition classification from MRI images. The model evaluation process involves two separate assessment methods, including quantitative and qualitative testing. The quantitative assessment includes three components: training convergence, class-wise metrics, and confusion matrix analysis, while the qualitative assessment uses Grad-CAM to interpret model predictions. The model performance evaluation uses baseline and state-of-the-art methods to demonstrate its accuracy and generalization capabilities. The proposed approach demonstrates robust performance through cross-validation and external validation results, confirming its effectiveness.

A. *Training Convergence Analysis*

Fig. 3 presents the training and validation accuracy and loss curves of the proposed SpineNet-Fusion framework. The close alignment between the training and validation performance, coupled with the smooth reduction in loss



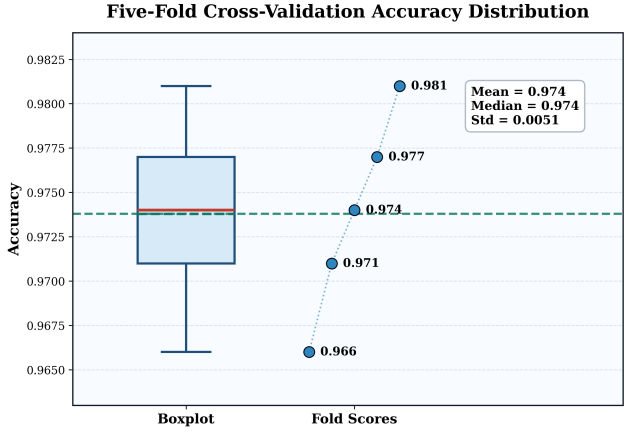
throughout the training process, demonstrates stable optimization and effective convergence. Furthermore, the absence of noticeable divergence between the training and validation curves indicates strong generalization capability and confirms that the model did not exhibit overfitting during training.

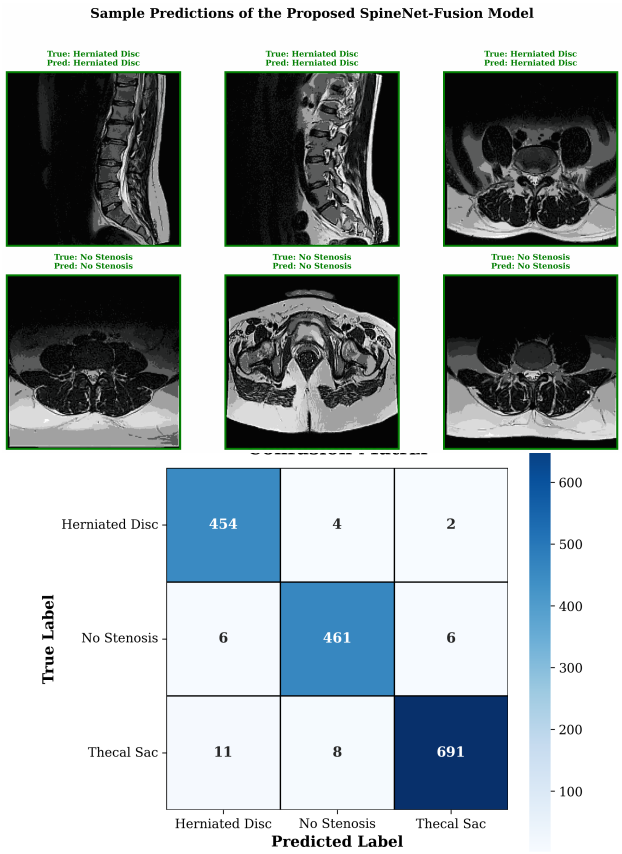
Fig. 3. Training and validation accuracy and loss curves for SpineNet-Fusion.

B. *Quantitative Performance Evaluation*

The proposed SpineNet-Fusion model was assessed by evaluating its classification capabilities using class-based precision, recall, and F1-score metrics. Results on the Kaggle dataset (Table I) show strong and balanced performance across all classes, with consistently high precision and recall values. The Thecal Sac Compression class achieves the highest precision, while the other classes maintain comparable F1-scores, indicating that performance remains consistent despite class imbalance. On the other hand, the model was tested on an external dataset from Bangabandhu Sheikh Mujib Medical University (BSMMU) (Table II) to evaluate its generalization capability. The model demonstrates consistent performance across all classes because it maintains its effectiveness despite changes in imaging conditions and data distribution across domains. Overall, the results show reliable performance, as the accuracy on the BSMMU dataset decreased only slightly compared to the Kaggle dataset while maintaining strong generalization capability. The proposed framework shows potential for real-world clinical applications due to its strong generalization performance.

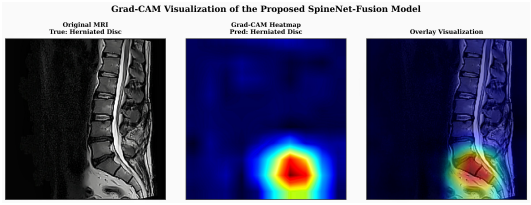
I. CLASS-WISE CLASSIFICATION PERFORMANCE





Class	Precision	Recall	F1-Score	Support
Herniated Disc	0.963	0.987	0.975	460
No Stenosis	0.974	0.974	0.974	473
Thecal Sac	0.988	0.973	0.980	710
Macro Avg	0.975	0.978	0.976	1643
Weighted Avg	0.977	0.977	0.977	1643

II. CLASS-WISE CLASSIFICATION PERFORMANCE ON THE EXTERNAL DATASET



Class	Precision	Recall	F1-Score	Support

Herniated Disc	0.938	0.927	0.932	180
No Stenosis	0.921	0.934	0.927	190
Thecal Sac	0.949	0.927	0.938	210
Macro Avg	0.936	0.929	0.932	580
Weighted Avg	0.937	0.929	0.933	580

C. Confusion Matrix Analysis

The proposed SpineNet-Fusion model shows high classification accuracy according to the confusion matrix presented in Fig. 4, as most predictions fall along the diagonal. The model effectively distinguishes Thecal Sac Compression from the other classes, resulting in only minor confusion between Herniated Disc and No Stenosis due to their similar structural features. The model demonstrates robust performance, as its low misclassification rate confirms its reliability and effectiveness.

Fig. 4. Confusion matrix for SpineNet-Fusion classification results.

D. ROC Curve Analysis

The proposed SpineNet-Fusion model displays the ROC curves for all three classes in Fig. 5. The model achieves strong discriminative performance, with AUC values of 0.972 for Herniated Disc, 0.954 for No Stenosis, and 0.948 for Thecal Sac Compression. The curves remain near the top-left region, indicating that the system achieves high true positive rates while maintaining low false positive rates. These results demonstrate the model's ability to effectively distinguish between different lumbar spinal conditions and confirm its robust classification performance.

Fig. 5. ROC curves of the proposed model.

E. Statistical Analysis

Five-fold cross-validation was conducted on the Kaggle dataset to evaluate model reliability and stability (Fig. 6). The proposed framework achieved a mean accuracy of 97.4% with a standard deviation of 0.0051, indicating consistent performance across different data partitions. Although the external BSMMU dataset exhibited slightly lower accuracy due to variations in data distribution and imaging conditions, the model maintained strong generalization capability in real-world clinical settings.

Fig. 6. Five-fold cross-validation accuracy distribution of the proposed model.

F. Qualitative Prediction Analysis

The test dataset contains example predictions, which are shown in Fig. 7 through its display of MRI images together with their actual and predicted labels. The results showed that predicted labels matched ground-truth labels across multiple cases. The model demonstrates its capability to identify various spinal disorders through different anatomical patterns and imaging conditions, which enables it to perform well across all patient data.

Fig. 7. Example predictions showing true and predicted lumbar spinal classes.

G. Interpretability using Grad-CAM

Grad-CAM was used to visualize the regions influencing model predictions. As shown in Fig. 8, the heatmaps highlight clinically relevant areas such as the spinal canal, intervertebral discs, and adjacent soft tissue. The high-activation regions align with key diagnostic indicators, including disc protrusion and compression, and correspond to the actual anatomical structures. This confirms the model's ability to identify essential features, supporting its trustworthiness and effectiveness in clinical decision support.

Fig. 8. Grad-CAM visualization highlights important spinal regions.

H. Performance and Complexity Analysis

As shown in Table III, SpineNet-Fusion was compared with CNN- and transformer-based models, including ViT and Swin Transformer. The proposed model achieved the highest accuracy and Macro-F1 score while maintaining competitive computational complexity and inference efficiency, owing to the complementary fusion of EfficientNet-B3 and DenseNet-121 for effective global and local feature learning.

III. Performance and Computational Complexity Comparison

Model	**Para-meters (M)**	**FLOPs (G)**	**Inference Time (ms)**	**Acc-(%)**	**Macro-F1**
CNN (Baseline)	8.4	1.9	11.8	81.32	0.8024
EfficientNet-B3	12.0	1.8	15.6	83.26	0.8284
Vision Transformer	86.0	17.6	42.3	93.80	0.9272
Swin Transformer	28.3	4.5	24.8	94.50	0.9141
SpineNet-Fusion	22.7	3.9	19.4	97.75	0.9769

IV. Conclusion

This research established SpineNet-Fusion as a deep learning framework that automatically classifies lumbar spinal disorders through MRI image analysis. The proposed architecture uses a dual-backbone system that combines EfficientNet-B3 and DenseNet-121, enabling the model to capture both global contextual information and fine-grained anatomical features present in spinal MRI scans. Through feature fusion, the model combines complementary representations from both backbones to achieve improved discriminative performance. The experimental results demonstrate that SpineNet-Fusion outperforms baseline architectures, including EfficientNet-B3 and conventional CNN models, achieving a classification accuracy of 97.75% and a Macro-F1 score of 0.9769. The hybrid architecture performs effectively across all classes and remains robust in the presence of class imbalance. Using Grad-CAM visualization, the model demonstrates its focus on clinically

significant areas such as the spinal canal and intervertebral discs, making its behavior more understandable and enhancing clinical trustworthiness. Overall, the framework provides an accurate and fully interpretable solution with strong generalization capability for real-world medical decision-support systems. Future research will evaluate the model on multi-institutional datasets, including various clinical reports and CT imaging data, while exploring architectural optimizations for real-time clinical deployment.

Author Contributions

Sadia Hamid contributed to conceptualization, data curation, and drafting. Mugdha Farheen Kibria handled preprocessing, augmentation, and implementation, while Fatema Tuz Zohora Panna conducted model training and experiments. Marjana Khanam Roza performed analysis, validation, and cross-validation, and Mansura Afsana Rima prepared the visualizations. MD Rezaul Karim supported software development, external validation, and result evaluation. Tauhidur Rahman Sakib conceived and supervised the study, designed the SpineNet-Fusion architecture, led the methodology and project administration, and reviewed and edited the manuscript as corresponding author. All authors approved the final manuscript.

References

1. U. U. Bharadwaj *et al.*, "Deep learning for automated, interpretable classification of lumbar spinal stenosis and facet arthropathy from axial MRI," *European Radiology*, vol. 33, no. 5, pp. 3435–3443, Mar. 2023, doi: 10.1007/s00330-023-09483-6.
2. S. Altun, A. Alkan, and İ. Altun, "LSS-VGG16," *Clinical Spine Surgery a Spine Publication*, vol. 36, no. 5, pp. E180–E190, Jan. 2023, doi: 10.1097/bsd.0000000000001418.
3. T. R. Sakib *et al*., "CDFNet Based Fusion Network for Interpretable Alzheimer's Disease Prediction Using Hybrid Imaging and Clinical Meta-Ensemble Learning," in *Proc. 2026 IEEE 18th Int. Conf. Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, Lviv, Ukraine, 2026, pp. 1–6, doi: 10.1109/TCSET65181.2026.11461211.
4. V. Tumko *et al.*, "A neural network model for detection and classification of lumbar spinal stenosis on MRI," *European Spine Journal*, vol. 33, no. 3, pp. 941–948, Dec. 2023, doi: 10.1007/s00586-023-08089-2.
5. H. Suzuki *et al.*, "Deep learning-based detection of lumbar spinal canal stenosis using convolutional neural networks," *The Spine Journal*, vol. 24, no. 11, pp. 2086–2101, Jun. 2024, doi: 10.1016/j.spinee.2024.06.009.
6. J. T. P. D. Hallinan *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, no. 1, pp. 130–138, May 2021, doi: 10.1148/radiol.2021204289.
7. T. Kim *et al.*, "Diagnostic triage in patients with central lumbar spinal stenosis using a deep learning system of radiographs," *Journal of Neurosurgery Spine*, vol. 37, no. 1, pp. 104–111, Jan. 2022, doi: 10.3171/2021.11.spine211136.
8. T. Wang *et al.*, "Machine Learning and Deep Learning for Diagnosis of Lumbar Spinal Stenosis: Systematic Review and Meta-Analysis," *Journal of Medical Internet Research*, vol. 26, p. e54676, Dec. 2024, doi: 10.2196/54676.
9. K.-Y. Li, J.-J. Weng, H.-L. Li, H.-B. Ye, J.-W. Xiang, and N.-F. Tian, "Development of a Deep-Learning model for diagnosing Lumbar Spinal Stenosis based on CT images," *Spine*, vol. 49, no. 12, pp. 884–891, Dec. 2023, doi: 10.1097/brs.0000000000004903.
10. D. Won, H.-J. Lee, S.-J. Lee, and S. H. Park, "Spinal stenosis grading in magnetic resonance imaging using deep convolutional neural networks," *Spine*, vol. 45, no. 12, pp. 804–812, Jan. 2020, doi: 10.1097/brs.0000000000003377.
11. İ. Altun, S. Altun, and A. Alkan, "LSS-UNET: Lumbar spinal stenosis semantic segmentation using deep learning," *Multimedia Tools and Applications*, vol. 82, no. 26, pp. 41287–41305, Apr. 2023, doi: 10.1007/s11042-023-15205-3.
12. H. Genç, E. Seyyarer, and F. Ayata, "Deep Learning-Driven MRI analysis for accurate diagnosis and grading of lumbar spinal stenosis," *Measurement*, vol. 251, p. 117294, Mar. 2025, doi: 10.1016/j.measurement.2025.117294.
13. J. Abbas, M. Yousef, N. Peled, I. Hershkovitz, and K. Hamoud, "Predictive factors for degenerative lumbar spinal stenosis: a model obtained from a machine learning algorithm technique," *BMC Musculoskeletal Disorders*, vol. 24, no. 1, p. 218, Mar. 2023, doi: 10.1186/s12891-023-06330-z.
14. T. Wang *et al.*, "A novel deep learning system for automated diagnosis and grading of lumbar spinal stenosis based on spine MRI: model development and validation," *Neurosurgical FOCUS*, vol. 59, no. 1, p. E6, Jul. 2025, doi: 10.3171/2025.4.focus24670.
15. A. S. Al-Kafri *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, Jan. 2019, doi: 10.1109/access.2019.2908002.
16. Y.-L. Zhang *et al.*, "Automated Classification of Cervical Spinal Stenosis using Deep Learning on CT Scans," *Spine*, Jun. 2025, doi: 10.1097/brs.0000000000005414.
17. D. Won, H.-J. Lee, S.-J. Lee, and S. H. Park, "Lumbar spinal stenosis grading in multiple level magnetic resonance imaging using deep convolutional neural networks," *Global Spine Journal*, vol. 15, no. 4, pp. 2309–2317, Nov. 2024, doi: 10.1177/21925682241299332.
18. M. Tan and Q. Le V., "EfficientNet: Rethinking model scaling for convolutional neural networks," *arXiv (Cornell University)*, May 2019, doi: 10.48550/arxiv.1905.11946.
19. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.