DETECTION-GUIDED LUMBAR SPINE DEGENERATION CLASSIFICATION USING MULTI-SLICE ENCODED MRI

SAMIUL KARIM MAZUMDER[1], RIYADUL ISLAM[1], SIAM TAHSIN BHUIYAN[1], RASHEDUR RAHMAN[1], SEFATUL WASI[2], ASHRAFUL ISLAM[1], SAADIA BINTE ALAM[1]

[1]Center for Computational & Data Sciences, Independent University, Bangladesh, Dhaka 1229, Bangladesh
[2]Department of Computer Science and Engineering, Independent University, Bangladesh, Dhaka 1229, Bangladesh
E-MAIL: samikarim191139@gmail.com, riyadulislam9591@gmail.com, siamtbhuiyan@gmail.com, rashed.riyadh14@gmail.com, sefatulwasi@gmail.com, ashraful@iub.edu.bd, saadiabinte@iub.edu.bd

Abstract:

Lumbar Spine Degeneration Classification (LSDC) poses significant diagnostic challenges as radiologists manually assess each MRI scan which is time-consuming, subjective, and prone to inter-observer variability. This highlights the need for an automated, objective, and efficient diagnostic solution to assist radiologists. Current deep learning-based approaches rely on single-slice MRI scans, which lack contextual spatial information. To address these challenges, we propose a Detection-Guided Classification Framework that integrates automated region detection with deep learning-based severity classification. Our approach introduces a novel RGB-based pre-processing technique that encodes multi-slice Sagittal T2-weighted MRI information into a single image. We first detect the degenerative regions to ensure that the classification models focus on the relevant areas. We then train multiple Convolutional Neural Network models to classify degeneration severity. Using the RSNA dataset, the YOLOv8 detection model achieves a mAP50 score of 0.9905. In the classification models, ResNet152 scores the highest accuracy of 92.74%, and EfficientNetB5 achieves the highest accuracy of 88.18% in the holdout test. Our result suggests that using the multi-slice encoded MRI dataset, the proposed detection-guided classification framework effectively localizes degenerative regions and achieves high classification accuracy by focusing on relevant areas rather than processing entire MRI slices.

Keywords:

LSDC, Deep Learning, RGB Pre-processing, Sagittal T2-weighted MRI, Detection guided Classification, Lumbar Spine Degeneration, MRI.

1. Introduction

Lumber Spine Degeneration (LSD) refers to the progressive degeneration of the intervertebral discs which often leads to chronic lower back pain, vastly affecting day-to-day life. It is one of the leading causes of disability worldwide, affecting over 500 million individuals and significantly contributing to global healthcare expenses. Epidemiological studies show that almost 80% of adults over the age of 50 exhibit some form of lumbar disc degeneration, with symptoms ranging from mild discomfort to severe neurological issues. The socioeconomic burden of LSD is undeniable, as the United States alone spends more than $100 billion annually [1]. With the worldwide increase of spinal degenerative disorders in aging populations, improving the accuracy and efficiency of detecting and classifying LSD diagnosis is of critical importance.

In the spine, there are five intervertebral disc levels L1/L2, L2/L3, L3/L4, L4/L5, and L5/S1 where degeneration occurs. In the present day, diagnosing LSD relies on radiologists manually assessing changes in disc morphology, signal intensity, and structural integrity through Magnetic Resonance Imaging (MRI) one by one using the Pfirrmann grading system. However, this approach is highly subjective, with inter-observer agreement ranging from 40% to 70%, leading to inconsistencies in clinical decision-making [2]. Moreover, manual assessments are time-consuming and inefficient for large-scale studies, creating a demand for semi or fully automated diagnostic tools that can enhance speed while being as accurate as possible. Also, most degenerations are classified by focusing on signal intensity and structural morphology of Sagittal T2-weighted MR images, as the proteoglycan concentration and water are reflected by a decreased signal intensity [3], [4], [5]. However, the view in the sagittal plane in 2D MR images is much more limited than the axial plane, as such it may also limit the detection of early degeneration [6]. Therefore, it is problematic and time-consuming for radiologists to assess early degeneration from 2D MRI slices due to the limitations of MRI.

Recent advancements in deep learning have significantly improved medical imaging by enabling automated detection, segmentation, and classification of pathologies with near-human accuracy. Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Transformer-based models, and hybrid architectures have demonstrated state-of-the-art performance across various imaging modalities, including radiology, pathology, and ophthalmology [7].

In musculoskeletal imaging, CNN-based models have been successfully applied to detect vertebral fractures, spinal stenosis, and intervertebral disc abnormalities, showcasing their potential to streamline LSDC diagnosis [8]. The study [9] proposed a fully automatic deep learning system for lumbar disc degeneration classification using SpineNet, a VGG11-based architecture with an attention mechanism. The model achieved an accuracy of 91.3% in classifying lumbar disk severity. Similarly, another study [10] developed a VGG16-based CNN to classify herniated lumbar discs using axial-view MRI. While the model had 94% classification accuracy, it was trained on a private dataset with limited availability, which restricted external validation. Other studies [10], [11] utilized popular CNN-based architecture, mostly trained on 2D MRI slices using transfer learning models such as VGG16, ResNet, and DenseNet, etc, for direct degeneration classification using only the degeneration-affected slices, reducing classification accuracy on real-world MRI data and overall generalizability.

In our study, we first focused on a detection-guided classification framework to counter misclassification due to the limitations of Sagittal T2-weighted slices in MR images. Detection-guided framework (DGF) in medical imaging integrates automated object detection with classification or segmentation models to improve diagnostic accuracy and efficiency. A DGF first identifies the regions of interest (ROI) from the medical images using object detection models such as You Only Look Once (YOLO), Faster R-CNN, Mask R-CNN, etc. Once an abnormality is detected, the image is cropped to retain only the relevant information, as well as discarding the unnecessary information for much better training data [11]. Finally, a classification model determines the type or severity of the abnormality. DGFs are widely used in various medical studies and applications as they focus only on relevant information, which reduces false positives and increases the efficiency and speed of a model, as there is no need to process an image entirely.

Traditional medical imaging studies often rely on single-slice MRI scans, which can lead to the loss of important contextual and spatial information that can be gathered from surrounding slices, not to mention that normal sagittal T2-weighted images have limitations in detecting early-stage degeneration [12]. Our study focuses on a detection-guided classification framework that adopts an advanced RGB pre-processing system to counter the discussed limitations in MR images. The study aims to implement this novel multi-slice RGB encoding approach and detection-guided classification to improve the accuracy and robustness of LSD assessment in MRI scans.

2. Dataset

This study uses the multi-institutional RSNA (Radiological Society of North America) 2024 Lumbar Spine MRI Dataset [13], which includes 147,320 DICOM images (.dcm) with 48,657 MRI files annotated with severity levels and single point coordinates across three degenerative conditions: Spinal Canal Stenosis (SCS), Neural Foraminal Narrowing (NFN) (Left & Right), Subarticular Stenosis (SAS) (Left & Right) at five intervertebral disc levels and each degenerative condition is classified into three severity levels: Normal/Mild, Moderate and Severe.

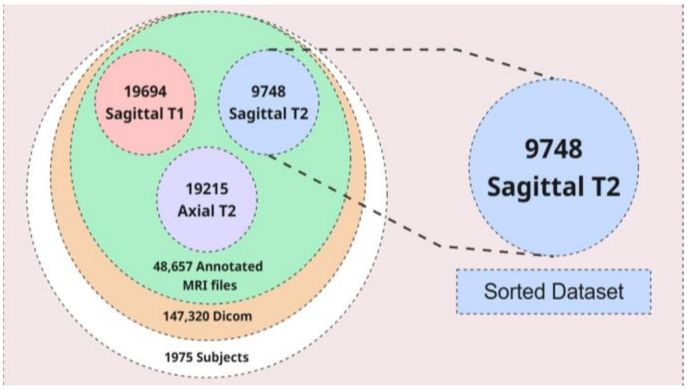

FIGURE 1. Dataset Overview.

19694 files are from the Sagittal T1 plane, 9748 files are Sagittal T2 and 19215 files are Axial T2. As our study solely focuses on Sagittal T2-weighted MR images, we separated the 1974 subjects that contain the relevant 9748 slices (Fig. 1). 1 subject with no proper severity levels in the Sagittal T2 plane was found and removed. Sagittal T2-weighted MRI provides superior soft-tissue contrast and detailed visualization of spinal discs, which is ideal for assessing SCS and detecting degenerative changes in the lumbar spine [14], [15], [16], [17], [18]. These 9748 slices will then be used for RGB multi-slice encoding to prepare them for our framework.

3. Method

3.1. Pre-processing

The first step in our study, after sorting the dataset, is the implementation of a merging technique to enhance the representation of LSD in MR images. This encoding process

(Fig. 2) ensures that the model is trained with multi-slice spatial information.

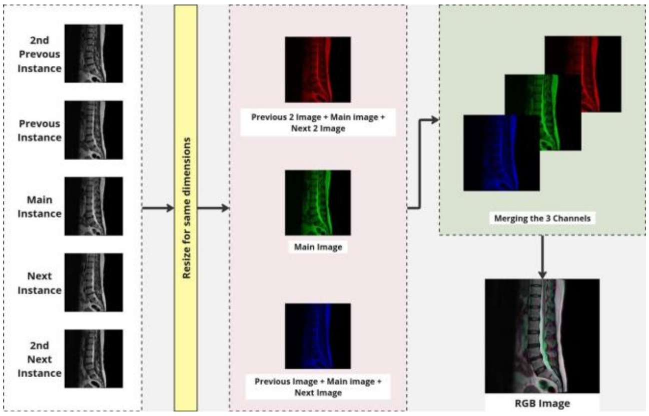

FIGURE 2. Pre-process technique for RGB image.

Firstly, we extract all the 9748 sagittal T2-weighted MRI slices from each subject. Then we identified the slice that contained the highest number of degenerative regions and labeled it as the main slice. We then selected the two previous and next adjacent slices, forming a five-slice cluster around the main slice.

Secondly, we took the following approaches to create colored channels to encode multi-slice information into a single RGB image. The Green Channel contains the main slice, which also holds the most degenerative information. Additionally, the Red Channel is the fusion of the main slice, previous, and next adjacent slices, which preserves local degeneration patterns. Lastly, the Blue Channel contains the fusion image of the entire cluster, which has the most amount of spatial LSD data.

Finally, the three channels are merged to create a single RGB image that contains degenerative variations across the intervertebral disc levels. From the total number of 9748 slices, a total of 1978 Sagittal T2-weighted RGB pre-processed images were created.

3.2. Validation strategy

We have employed a two-stage validation strategy, a holdout test set for final evaluation, and a 4-fold cross-validation during the model development.

The holdout test contains 20% of the entire Sagittal T2 dataset, which contains a total of 395 RGB pre-processed images, and remains untouched until the final validation is required. This serves as an independent evaluation dataset, which ensures an unbiased assessment of the entire framework.

The remaining 80% of the dataset is used in 4-fold cross-validation [19] to ensure robust evaluation and minimize potential biases. This technique utilizes 1583 RGB pre-processed images, which are divided into four equal folds, and each fold holds an average of 395 images. In each iteration, three folds (75% of the data) were used for training, while the remaining fold (25% of the data) served as test data. This process was repeated four times, with each fold serving as test data once. This approach prevents dataset bias and provides a more robust estimate of model performance. Additionally, this method ensures fairness in both the training and testing phases throughout the validation process, as each data sample contributes equally to both phases.

3.3. Detection

We employed YOLOv8 [20] due to its high detection speed, superior localization accuracy, and ability to handle complex medical images. The model detects SCS degeneration in intervertebral disc levels using sagittal T2 MR images. The detection model is trained using bounding boxes that were manually derived from single-point severity coordinates provided in the dataset. We systematically tested multiple bounding box sizes and selected an optimal configuration based on experimental performance. With the most relevant parameters for our study (Optimizer: AdamW, Learning Rate: 1e-4 with cosine annealing, Batch Size: 16, Epochs: 100, IoU Threshold: 0.5), we train the model to detect the region. The following evaluation metrics are used: mAP50 and mAP50-95. mAP50 evaluates the detection performance at an Intersection over Union (IoU) threshold of 0.5, while mAP50-95 calculates the mean average precision across IoU thresholds from 0.5 to 0.95, with a step size of 0.05. These metrics provide a comprehensive assessment of detection performance as shown in the following equations (1), (2).

$$mAP@50 = \frac{1}{N}\sum_{i=1}^{N} AP(IoU \geq 0.5) \quad (1)$$

$$mAP@50-95 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{10}\sum_{t=0.5,0.55,0.6\ldots0.95} AP(IoU \geq t) \quad (2)$$

After degenerative regions are detected, they are cropped and extracted, which ensures that unnecessary background information other than the required regions is not used to train the classification model, aiding the classification model in focusing only on the degenerated areas.

3.4. Classification

CNN classification models DenseNet121 [21], EfficientNetB5 [22], MobileNetV2 [23], and ResNet152 [24] were used for their range of model sizes, as well as on numerous tasks and benchmarks. After detecting the region of

SCS degeneration, the cropped regions are categorized based on the severity, which is Normal/Mild, Moderate, and Severe. Also, a random shuffle was performed among subjects to prevent overlapping of the same subjects across folds and ensure the result's validity by reducing potential bias. These classification models are trained using the RSNA dataset with relevant parameters for our study (Optimizer: Adam, Learning Rate: 1e-4, Batch Size: 8, Epochs: 50) with a classifier GlobalAveragePooling 2D, Dense layers 1024 and 512 with ReLU activation, Dense layer 3 with Softmax activation. For classification, the following evaluation metrics are used, Accuracy, Precision, Recall, AUC (Area Under the Curve) micro average, and F1 score, as shown in the following equations (3), (4), (5), (6) and Grad-CAM (Gradient- weighted Class Activation Mapping) [25].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1\ score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \tag{6}$$

To represent the areas of interpretability and better emphasize clinically significant regions in the pictures, Grad-CAM was applied to each model. The method makes it easier to fully understand the model's explainability and how the model assigns various severity ratings to each subject.

4. Results & Discussion

4.1. Detection Results

YOLOv8 model detection result is generated with ground truth labels provided by the RSNA dataset.

TABLE 1. Detection Results on Ground Truth

mAP50	mAP50-95	Box precision	Box recall	Box f1
0.9905	0.84375	0.97925	0.99075	0.98496

The detection model achieved an mAP50 of 0.9905 and a box recall of 0.99075 in the final epoch, demonstrating high detection accuracy.

4.2. Classification Results

The classification results demonstrate a noticeable performance, as highlighted in Table 2.

TABLE 2. Classification Results on Ground Truth

Models	Accuracy	Precision	Recall	F1 score
DenseNet121	0.9111	0.9106	0.9106	0.9106
EfficientNetB5	0.9167	0.9118	0.9118	0.9118
ResNet152	**0.9274**	**0.9164**	**0.9164**	**0.9164**
MobileNetV2	0.9156	0.9095	0.9095	0.9095

From the classification models, the ResNet152 has the best result with the highest Precision and Recall value of **0.9164**. In the case of Micro Average ROC, MobileNetV2 achieved a score of **0.9450**, outperforming other models across all severity levels (Normal/Mild, Moderate, and Severe) (Fig. 3).

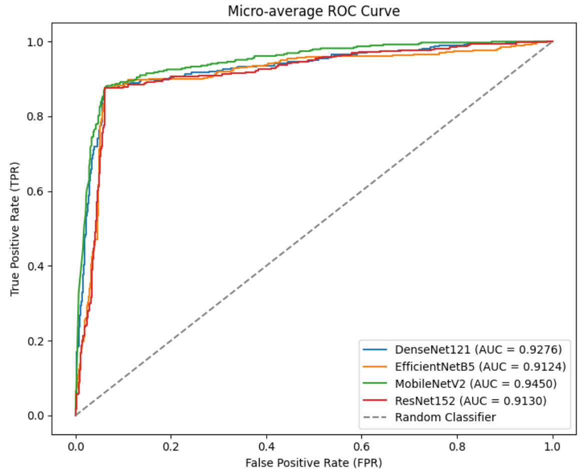

FIGURE 3. Micro Average ROC Curve.

4.3. Full Framework Results

The holdout results of the proposed framework demonstrated robust performance in classifying SCS across the three severity classes. 395 subjects were selected from the dataset for the holdout test, ensuring an unbiased evaluation result for the model's generalizability. In Table 3 and Table 4,

TABLE 3. Detection Results on Holdout Test

mAP50	mAP50-95	Box precision	Box recall	Box f1
0.9932	0.851	0.9830	0.9910	0.9869

TABLE 4. Classification Results on Holdout Test

Models	Accuracy	Precision	Recall	F1 score
DenseNet121	0.8644	0.8724	0.8724	0.8724
EfficientNetB5	**0.8818**	**0.8780**	**0.8780**	**0.8780**
ResNet152	0.8772	0.8720	0.8720	0.8720
MobileNetV2	0.8808	0.8762	0.8762	0.8762

Detection and Classification Results on the Holdout Test are shown for each model.

For this Holdout Test, the EfficientNetB5 model performs better than any other model (Table 4), showcasing the model's ability to accurately classify SCS severity. These results indicate that the model maintains high diagnostic accuracy even when applied to unseen data.

In Figure 4, we showcase every class for each model performance. The ROI correctly detected from the images for the EfficientNetB5 was better than any other heatmap. The visualizations demonstrate a balanced focus across the SCS, with the best regional heatmap. For the DenseNet121 and MobileNetV2, the regional interest was correct, but the region was across the maximum portion of the image.

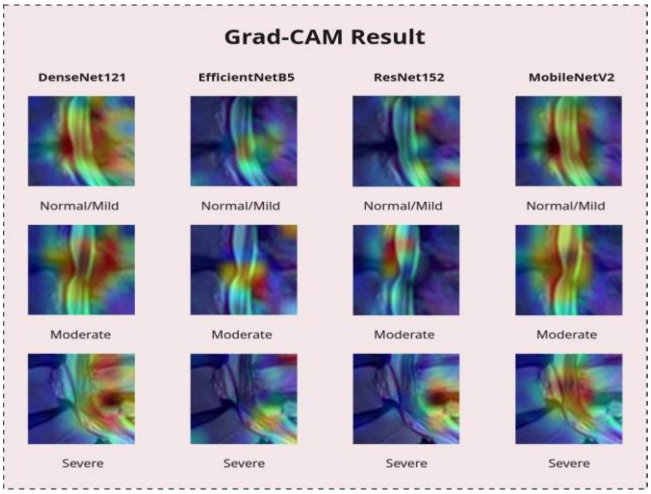

FIGURE 4. Grad-CAM Results.

Our proposed framework has several key design choices. Firstly, the RGB preprocessing technique enabled the model to leverage the multi-slice spatial information, which provides a more comprehensive view of degenerative patterns across SCS disc degeneration. This can address the limitations of single-slice MRI scans by utilizing sufficient contextual information on disc degeneration for accurate diagnosis.

Secondly, the detection-guided classification framework ensures the model is focused on relevant ROI, reducing the impact of Sagittal T2 MRI limitations and unnecessary background information. Our use of YOLOv8 was highly effective with a mean average precision (mAP50) of **0.99322**, showing it accurately identifies degenerative regions.

Finally, a 4-fold cross-validation strategy during model development ensures a robust and generalizable framework. On the other hand, the independent holdout test set simulates real-world deployment scenarios to further validate its clinical applicability for use.

Our proposed framework has significant potential for improving the diagnosis and classification of lumbar spine degenerative conditions. This framework can reduce the workload of radiologists by reducing the time to diagnosis. Furthermore, because the dataset is contributed from different global regions of the world, it contains a wide range of subject variability.

Despite the results of the framework, it has certain limitations. First, the study is solely focused on T2-weighted MR images, which can be extended in the future to Axial T2-weighted and Sagittal T1-weighted MR images to detect and classify Neural Foraminal Narrowing (Left and Right), and Axial Subarticular Stenosis (Left and Right). Also, Further validation of external datasets is necessary to confirm the framework's robustness.

5. Conclusion

This study presents a detection-guided classification framework for LSD by integrating an RGB pre-processing technique. By taking advantage of multi-slice spatial encoding (RGB pre-processing), the study overcomes the limitations of single-slice MRI analysis and enables the model to learn complex degenerative patterns. Our results demonstrate the effectiveness of this approach, with YOLOv8 achieving a mAP50 score of **0.9905,** while ResNet152 achieves the highest accuracy of **92.74%** among the other classification models in 4-fold-cross-validation and EfficientNetB5 achieves the highest accuracy of **88.18%** in the holdout test, which proves the study's ability to generalize to unseen data. The RGB multi-slice encoding technique was a key factor in improving model performance. Thus, the proposed framework offers a reliable, AI-assisted diagnostic tool to detect and classify LSD from Sagittal T2-weighted MRI scans, which can reduce radiologists' workload while improving accuracy.

References

[1] Vos, T., Lim, S.S., Abbafati, C., Abbas, K.M., and others., "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019", The Lancet, Vol. 396, No. 10258, pp. 1204-1222, 2020.

[2] Pfirrmann, C.W., Metzdorf, A., Zanetti, M., Hodler, J. and Boos, N., "Magnetic Resonance Classification of Lumbar Intervertebral Disc Degeneration", Spine, Vol. 26, No. 17, pp. 1873-1878, 2001.

[3] SCHNEIDERMAN, G., FLANNIGAN, B., KINGSTON, S., THOMAS, J., DILLIN, W.H. and WATKINS, R.G., "Magnetic Resonance Imaging in the

Diagnosis of Disc Degeneration: Correlation with Discography", Spine, Vol. 12, No. 3, pp. 276-281, 1987.

[4] Tertti, M., Paajanen, H., Laato, M., Aho, and others.., "Disc Degeneration in Magnetic Resonance Imaging: A Comparative Biochemical, Histologic, and Radiologic Study in Cadaver Spines", Spine, Vol. 16, No. 6, pp. 629-634, 1991.

[5] Gunzburg, R., Parkinson, R., and others.. "A Cadaveric Study Comparing Discography, Magnetic Resonance Imaging, Histology, and Mechanical Behavior of the Human Lumbar Disc", Spine, Vol. 17, No. 4, pp. 417-426, 1992.

[6] Watanabe, A., Benneker, L.M., Boesch, C., Watanabe, T., Obata, T. and Anderson, S.E. "Classification of Intervertebral Disk Degeneration with Axial T2 Mapping", PubMed, 2007.

[7] Litjens, G., Kooi, T., Bejnordi, B.E., and others. "A Survey on Deep Learning in Medical Image Analysis", Medical Image Analysis, Vol. 42, pp. 60-88, 2017.

[8] Jamaludin, A., Lootus, M., and others. "Automation of Reading of Radiological Features from Magnetic Resonance Images (MRIs) of the Lumbar Spine Without Human Intervention Is Comparable with an Expert Radiologist." European Spine Journal, 2017.

[9] DING, Z., LI, H., CHEN, L., CHEN, B., SUN, H., HOU, W. and XIA, C "Fully Automatic Classification of Lumbar Disc Degeneration Based on Deep Learning." Chinese Journal of Clinical Research, 2024.

[10] Mbarki, W., Bouchouicha, M., Frizzi, S., Tshibasu, F., Farhat, L.B. and Sayadi, M. "Lumbar Spine Discs Classification Based on Deep Convolutional Neural Networks Using Axial View MRI." Interdisciplinary Neurosurgery, vol. 22, 2020, pp. 100905.

[11] Liawrungrueang W, Kim P, and others. Automatic Detection, Classification, and Grading of Lumbar Intervertebral Disc Degeneration Using an Artificial Neural Network Model. Diagnostics (Basel). 2023.

[12] Lee, G. and Fujita, H. eds. "Deep Learning for Medical Image Analysis: Challenges and Advancements." Journal of Biomedical AI, 2023.

[13] Tyler Richards, Jason Talbott, Robyn Ball, Errol Colak, Adam Flanders, Felipe Kitamura, John Mongan, Luciano Prevedello, and Maryam Vazirabad. RSNA 2024 Lumbar Spine Degenerative Classification. https://kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification, 2024. Kaggle

[14] Sheehan, N.J. "Magnetic Resonance Imaging for Low Back Pain: Indications and Limitations." Postgraduate Medical Journal, vol. 86, no. 1016, 2010, pp. 374-378.

[15] Chou, D., Samartzis, D., Bellabarba, C., Patel, A., Luk, K.D., Kisser, J.M.S. and Skelly, A.C. "Degenerative Magnetic Resonance Imaging Changes in Patients with Chronic Low Back Pain: A Systematic Review." Spine, vol. 36, no. 21 Suppl, 2011, pp. S43-S53.

[16] Wassenaar, M., van Rijn, R.M., and others. "Magnetic Resonance Imaging for Diagnosing Lumbar Spinal Pathology in Adult Patients with Low Back Pain or Sciatica: A Diagnostic Systematic Review." European Spine Journal, vol. 21, no. 2, 2012, pp. 220-227.

[17] Roudsari, B. and Jarvik, J.G. "Lumbar Spine MRI for Low Back Pain: Indications and Yield." American Journal of Roentgenology, vol. 195, no. 3, 2010, pp. 550-559.

[18] Sollmann, N., Mönch, S., and others "Imaging of the Degenerative Spine Using a Sagittal T2-Weighted DIXON Turbo Spin-Echo Sequence." European Journal of Radiology, vol. 131, 2020, pp. 109204.

[19] Bradshaw, Tyler J., and others "A guide to cross-validation for artificial intelligence in medical imaging." Radiology: Artificial Intelligence 5, no. 4 (2023): e220232.

[20] Varghese, Rejin, and M. Sambath. "Yolov8: A novel object detection algorithm with enhanced performance and robustness." In 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), pp. 1-6. IEEE, 2024.

[21] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. "Densely Connected Convolutional Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700-4708.

[22] Tan, Mingxing, and Quoc Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." Proceedings of the International Conference on Machine Learning (ICML), PMLR, vol. 97, 2019, pp. 6105-6114.

[23] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C. "MobileNetV2: Inverted Residuals and Linear Bottlenecks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520.

[24] He, K., Zhang, X., Ren, S. and Sun, J. "Deep Residual Learning for Image Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[25] Selvaraju, R.R., Cogswell, M.,and others "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." International Journal of Computer Vision, vol. 128, no. 2, 2020, pp. 336-359.