CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

Explainable Deep Learning Model for Lumbar Spinal Stenosis Diagnosis

by

Abdullah Khan

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2025

This thesis is dedicated to my Family and my Teachers. I am deeply grateful to my beloved parents and siblings for their endless support and encouragement. I owe a special debt of gratitude to my supervisor, whose constant trust in me has helped me attain this crucial milestone.

CERTIFICATE OF APPROVAL

Explainable Deep Learning Model for Lumbar Spinal Stenosis Diagnosis

by

Abdullah Khan

(MCS223011)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Ashfaq Ahmed	MY, Islamabad
(b)	Internal Examiner	Dr. Nadeem Anjum	CUST, Islamabad
(c)	Supervisor	Dr. Muhammad Furqan	CUST, Islamabad

Dr. Muhammad Furqan
Thesis Supervisor
November, 2025

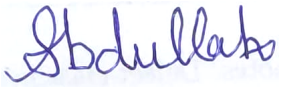
Dr. Mohammad Masroor Ahmed
Head
Dept. of Computer Science
November, 2025

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
November, 2025

Author's Declaration

I, **Abdullah Khan** hereby state that my MS thesis titled "**Explainable Deep Learning Model for Lumbar Spinal Stenosis Diagnosis**" is my work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

If my statement is found to be incorrect at any time, even after my graduation, the University has the right to withdraw my MS Degree.



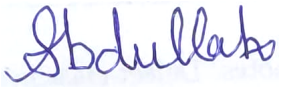
(Abdullah Khan)
Registration No: MCS223011

Plagiarism Undertaking

I solemnly declare that the research work presented in this thesis titled "**Explainable Deep Learning Model for Lumbar Spinal Stenosis Diagnosis**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.



(Abdullah Khan)
Registration No: MCS223011

Acknowledgement

First and foremost, I am deeply grateful to Almighty Allah for blessing me with knowledge, strength, courage, and patience throughout my studies. I am also very grateful to my supervisor, **Dr. Nadeem Anjum**, for his close monitoring of the progress of this thesis, providing insights at every stage, and correcting the direction whenever necessary.

I would like to express my deepest gratitude to my dearest family members: my father, my mother, and my siblings, for their unconditional support during good and bad times. They have always encouraged me to stay motivated and achieve my goals.

Lastly, I would like to take a moment to acknowledge my efforts. Reflecting on my academic journey to complete this thesis, I am grateful to myself for maintaining a positive outlook and self-belief during challenging times. Balancing social life and household responsibilities, staying dedicated to my research, and remaining steadfast through every challenge have all been pivotal in overcoming obstacles and reaching this academic milestone. These efforts have truly been essential to my success.

(Abdullah Khan)

Abstract

The Lumbar Spinal Stenosis (LSS) is a condition where the spinal canal in your lower back (lumbar spine) abnormally narrows and puts pressure on the nerves, causing numbness, weakness, and pain primarily in the legs. Diagnosis LSS via MRI remains time-consuming, expensive, and relies heavily on radiologists' expertise. For solving this problem, there are some methods available to do the classifications, i.e, deep learning (DL).

The traditional classification methods to classify LSS are hard to interpret, and the main limitations of DL models are their black-box nature. The term "black box" in deep learning refers to a model whose internal decision-making process is opaque, complex, and difficult for humans to interpret. Another limitation in DL is where the unnecessary features or patterns are mistakenly taken by the DL model during training to make the predictions. To enhance model robustness and accuracy, we require insights into decision-making processes. This necessitates reducing layer complexity while ensuring the model focuses on clinically significant classification features, achievable through Explainable AI (XAI) integration.

Our methodology is comprised of an XAI-based CNN for achieving the key features for LSS classification that are the actual representation of stenosis in MRI, while reducing the number of layers. The proposed model achieved 89.64% accuracy on our dataset. The effectiveness of the proposed model in detecting LSS within MRIs is demonstrated through XAI methods such as Grad-CAM and Occlusion Sensitivity.

Contents

List of Figures

List of Tables

Abbreviations

ANN	Artificial Neural Network
DL	Deep Learning
DNN	Deep Neural Networks
DT	Decision Tree
DSS	Dengue Shock Syndrom
FP	False Positive
GA	Genetic Algorithm
GANs	Generative Adverisal Networks
Grad-CAM	Gradient-weighted Class Activation Mapping
KNN	K-Nearest Neighbour
LSS	Lumbar Spinal Stenosis
LeakyReLU	Leaky Rectified Linear Uni
LDD	Lumbar Disc Degeneration
LR	Logistic Regression
LIME	Local Interpretable Model-Agnostic Explanations
ML	Machine Learning
OCSVM	One Class Support Vector Machine
rLDH	Recurrent Lumbar Disc Herniation
RF	Random Forest
SHAP	Shapley Additive Explanation
SVM	Support Vector Machine
TP	True Positive
XAI	Explainable AI

Symbols

$\bigoplus$	XOR Operation
X_{norm}	Normalization
X_{min}	Minimum Pixel Value
X_{max}	Maximum Pixel Value
X_{final}	Clamp to Handle Floating-Point Errors
F_θ	Convoluational Neural Network
α	Learning Rate
∂	Partial/Deviation

Chapter 1

Introduction

1.1 Spinal Stenosis Background

The spine contains the skeletal system that maintains posture. The intervertebral discs of the posterior vertebrae from the head to the pelvis and intervertebral discs. The joint of the vertebral discs allows the spine to move and connect. Another great feature of our spine is that it contains the spinal canal, a narrow fluid-filled space in the spinal column. There are many issues related to spinal which are nerve compression and mobility problems, which can significantly impact daily life. One most common causes of spinal issues is spinal stenosis, that caused by the narrowing of the spinal canal and puts pressure on the nerves of the spinal cord [1].

Spinal stenosis usually occurs when the space within the spinal canal becomes too narrow or when the spinal cord hole becomes too small, i.e., when the person bends backwards. The most common type of spinal stenosis is Lumbar Spinal Stenosis (LSS), in which patients feel pain in the lower part of their body and numbness in the legs, with increasing age due to disc degeneration [2].

Most researchers found that around 50 % of patients who underwent nonsurgical procedures usually reported that their back and leg pain reduced after 8-10 years of follow-up [3]. Furthermore, a decrease in nutrients leads to the degeneration of

the discs, and because of these degenerations, the height of the discs decreases [4].

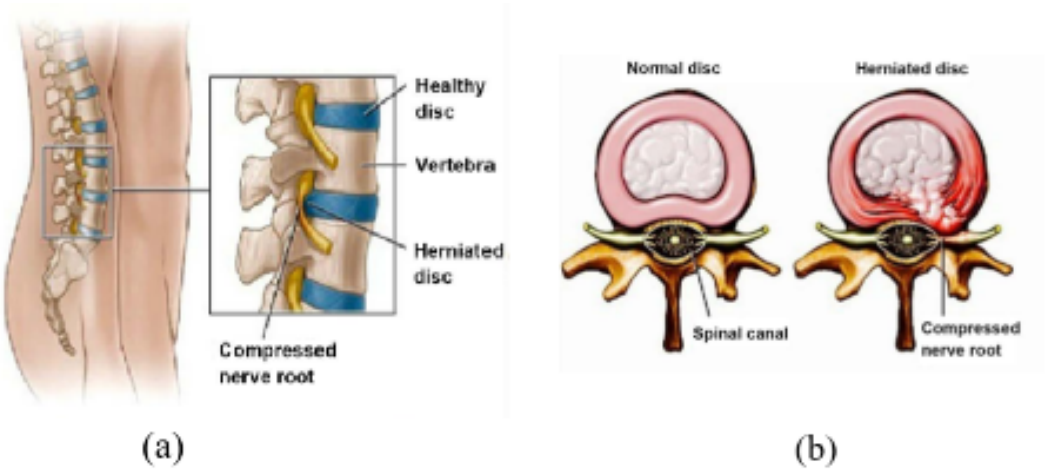


FIGURE 1.1: Lumbar Disc Herniation (a) Showing herniated disc at L4/L5 level (b) Shows normal Lumbar disc and Herniated Lumbar Disc [5]

The Lumbar Spinal Stenosis has the following main parts: Healthy and Normal Disc, Vertebra, Herniated Disc, Compressed nerve root, and Spinal Canal as shown in Figure 1.1.

1.1.1 Healthy and Normal Disc

The cushion between the bones in the intervertebral disc stays at the right height. It has a soft, jelly-like i.e., Nucleus center that has not broken down or swollen out of place, and the outer part is Annulus [6].

1.1.2 Vertebra

Little bones called vertebrae link one another to form your spine. These bones help you stand, bend, and move while maintaining the nerves inside safe conditions [7].

1.1.3 Herniated Disc

When the Annulus of the intervertebral disc becomes soft, the Nucleus slips, and the intervertebral disc eventually slips. This slipped disc causes pressure on the spinal nerves [8].

1.1.4 Compressed Nerve Root

If the space in their lower back becomes too narrow, it can squeeze the nerves. This may lead to pain, numbness pins-and-needles feeling, or weakness in your legs, buttocks, or feet [9].

1.1.5 Spinal Canal

The tunnel inside the bones of your lower back that holds your spinal cord and the nerves roots from it [10].

1.2 Types of Lumbar Spinal Stenosis

LSS is classified broadly into two main categories: foraminal stenosis, where compression of the spinal nerve, and central canal stenosis, which refers to compression of the spinal cord. Foraminal stenosis is a narrowing of the spinal openings where nerves exit the spine. This narrowing can allow nearby tissues or bones to touch or compress the nerve root due to irritation of the nerve root [11]. In contrast, the central canal stenosis compressed the intervertebral disc to put pressure on the nerve roots. It is a condition in which the degenerative central canal narrows due to degenerative disc bulging and degenerative changes in the facet joint [12].

The most common types of central canal stenosis are degenerative and herniated discs. The LSS can be classified by grading. Grades 1 and 2 indicate no stenosis, and the other one indicates moderate stenosis. In contrast, grades 3 and 4 mean severe stenosis and extreme stenosis. In this study we have discussed about three types of LSS i.e. herniated disc, thecal sac, and no stenosis [13].

The radiologist classifies these stenoses using different medical imaging. MRI and CT scans are most commonly used to identify the stenosis in the lumbar spine. The LSS can be easily identified using MRIs to obtain detailed information from the 3D angles [2].

The radiologist basically performs the LSS MRI with a focus on two primary obj-

ectives: (i) Comparison between the normal and herniated disc from the LSS images. (ii) Using the MRI images to identify the types of LSS, i.e., foraminal and central canal spine.

TABLE 1.1: Lumber Spinal Stenosis Types

Type	Image	Description
Herniated Disc		Herniated discs often occur in the lower back, especially between the L4 and L5 spinal bones [14].
Thecal Sac		Thecal sac occurs when mildly compress the outer section of the nerve roots or spinal cord [15].
No Stenosis		No stenosis means that if there is no stenosis in the patient's MRI.

Identifying and classifying LSS can be done manually, but it is the most challenging aspect due to the multiple types of LSS. Manually identifying and classifying LSS is time-consuming, laborious, and not a practical way to do it because of the large amount of MRIs [16].

Additionally, several classes of LSS categorization provide more difficulties than the two-class classification. Therefore, a consistent Introduction of an automated system is desperately needed to help radiologists overcome these obstacles related to manual LSS identification and categorization [17].

Several approaches have been developed to segment, detect, and classify LSS using

deep learning (DL) algorithms. DL models are traditional methods of image classification and object detection, but their performance is not superior to that of DL models.

1.3 Deep Learning Techniques

Medical scans, like those used to find lumbar spinal stenosis, are now totally different from how they were used before DL. These systems learn to find small details and patterns in scans using hidden layers of networks.

The advancement of identification improves understanding of complex medical conditions and facilitates better diagnostic processes in healthcare. The DL methods accelerate diagnostic efficiency to provide better and more valuable comprehension to healthcare professionals, using DL techniques to revolutionize MRIs and diagnostics

The significance of DL techniques transforms MRI images for precise diagnostics of spine stenosis, which helps healthcare professionals to better understand and effectively make decisions about diagnosis.

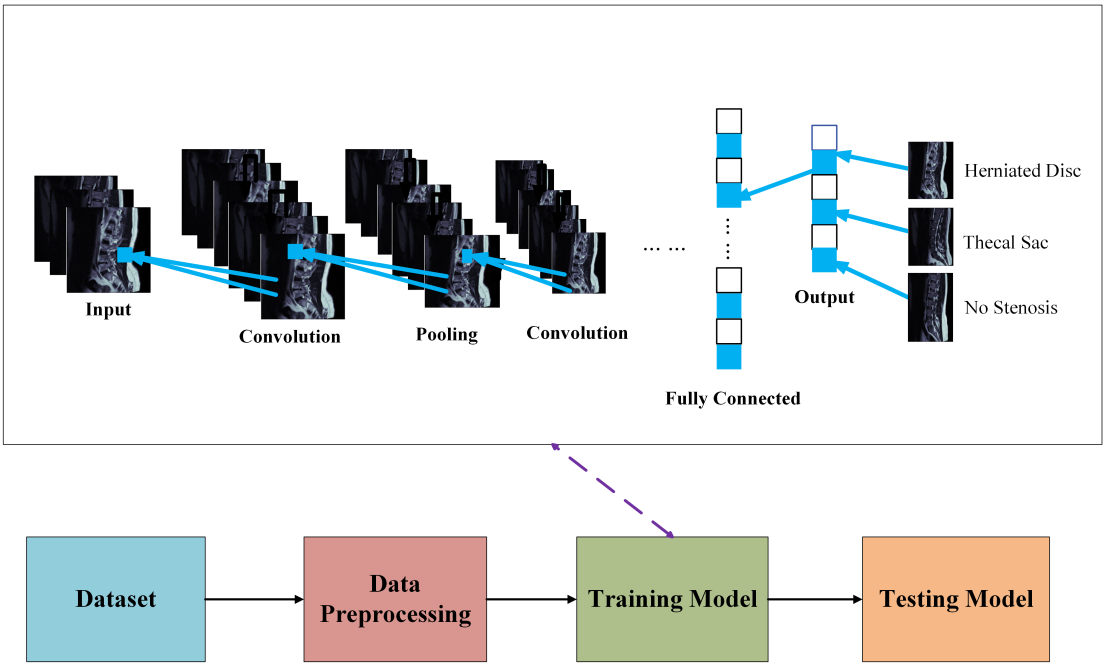


FIGURE 1.2: Traditional technique CNN to identify lumbar spinal stenosis.

The researchers applied several LSS identification and classification methods that

are based on DL techniques. Figure 1.2 shows the traditional way to classify lumbar spinal stenosis using the CNN architecture. Some other traditional techniques can exist [2, 18]. DL have hidden layers, which are black boxes in nature.

It challenges researchers to understand the behavior of deep learning-based models when making decisions. This lack of understanding concerns many areas where decision-making is required.

1.4 Explainable Artificial Intelligence

XAI is the application of methods to improve human comprehension of artificial intelligence decision-making systems. XAI stresses openness by showing how input data is turned into usable outputs [19].

1.4.1 XAI Methods

The methods of XAI are as follows

1.4.1.1 Model-agnostic

It is a technique which are particularly valuable in the context of providing insights into the AI model without the need to access or alter the underlying algorithms [20]. The model-agnostic approach uses the tools to help explain predictive models for patient outcomes, regardless of whether they are based on LIME, SHAP, or some complex neural networks [21].

1.4.1.2 Model-specific

It is another method that is used to get the insights of the internal decision-making process by using some specific algorithms [22]. The main demerits of this technique is that there is a limitation in determining a model when there is a need of a particular type of explanation, and requires some extensive knowledge

to modify the train a model [23].

1.4.1.3 Local Explainer

It is technique which targets the particularly a relevant in a spatial context because of data is individual feature attributions can be visualised spatially, and any spatial patterns can be observed [24]. The main adcantages are to identify the necessary components which leads to the prediction, find biases and errors, and establish trust in the models are all advantages [25].

1.4.1.4 Global Explainer

It is an alternative method of local explainer where it provides a high degree of understanding of how the model is making decisions on the entire dataset, but rather on explaining the predictions on the basis of the individual features. It helps the researchers to find out the overall patterns and trends on which the model is being trained [26].

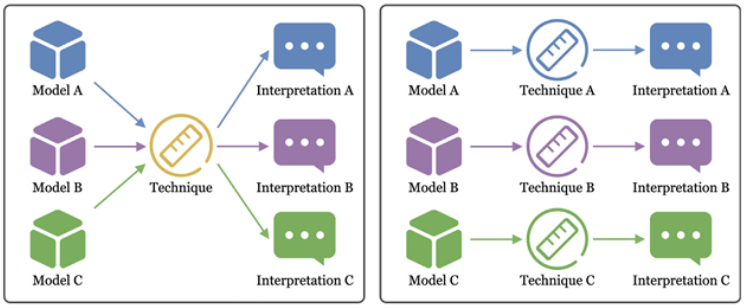


FIGURE 1.3: Model-agnostic and Model-specific ML [27]

1.4.2 Types of XAI

1.4.2.1 Gradient-weighted Class Activation Mapping

Grad-CAM is an explainable artificial intelligence (XAI) method presenting visual explanations for CNN output decisions. It produces a coarse localization heatmap by using gradients about a desired class score that backpropagate into the final c-

onvolutional layer [28].

These gradients are fed through global average pooling in order to derive importance weights for each feature map, subsequently combined in order to indicate salient regions in the input image where most influence was exerted for producing a prediction [29].

The main disadvantage of this technique is that it provides a broad, generalized level of visual explanations, but it lacks the pixel-level explanation required to identify the anomalies in medical images [30].

1.4.2.2 Local Interpretable Model-agnostic Explanations

LIME is an XAI technique that is used to explain the predictions of any classifier or regressor reliably by approximating it locally with an interpretable model, and it highlights the specific feature using saliency maps or heatmap [31].

It aims to provide local fidelity. The local fidelity means that the explanation for individual predictions should at least be locally faithful [32].

1.4.2.3 Shapley Additive Explanations

SHAP is a model-agnostic methodology that determines the significance of each characteristic in a specific prediction. Feature importance is calculated using Shapley values, a concept derived from game theory. The concept involves equitably allocating a reward across participants [33].

The prize should be contingent upon the varying contributions of players to the victorious squad. Shapley values are utilized to ascertain the contribution of characteristics to the total decision output by indicating the impact of their presence or absence on the decision-making process [34].

1.4.2.4 Occlusion Sensitivity Explainer

It is a model-agnostic, visual-based explainability method. It uses a local explainer to generate a saliency or heatmap independently of the model by covering the inp-

ut image's pixels with an occlusion mask for further class prediction [35].

The occlusion occurs due to the given variation in the model's prediction based on the input image, as shown using a heatmap and numerically using metrics of feature relevance. The impact of masking certain parts of the given image based on the predicted class will then be considered for the explanation [36].

1.4.3 Comparison between XAI Types

The following table Comparison between Grad-CAM, LIME, and SHAP is as follows:

TABLE 1.2: Comparison between Grad-CAM, LIME, and SHAP

Feature	**Grad-CAM**	**LIME**	**SHAP**
Explanation Type	It is a model-specific design for CNN.	It is a model-agnostic approach that can be used with any ML model.	It is a model-agnostic approach that can be used with any DL and ML model.
Mechanism	It uses gradients to create a localization map (heatmap) that highlights the important regions in an image that contribute to the model's prediction.	It explains a single prediction by highlighting the features (e.g., words in text, pixels in an image) that are most important for that specific prediction.	It uses Shapley values from cooperative game theory to calculate the contribution of each feature to a model prediction.
Computational Complexity	It is comparatively computationally inexpensive since it takes advantage of the gradients already available from the forward and backward passes of the CNN.	LIME can be computationally expensive as it needs to create a large number of perturbed samples and obtain predictions for all of them to train the localized, condensed model.	The computation of precise SHAP values is highly computationally expensive since the requirement of checking all conceivable permutations of features has to be met.

1.5 Deep Learning Techniques with XAI

Deep Neural Networks (DNNs) sometimes operate as opaque 'black boxes,' hindering trust in their conclusions. In healthcare, where transparency is paramount,

professionals and patients require explicit explanations for AI-generated diagnoses. Explainable AI (XAI) elucidates these models by disclosing their decision-making processes, so providing reliable and accountable medical AI.

XAI methods include layer visualization, feature analysis, and understandable explanations that help to elucidate how deep learning models decide. Users confidence is enhanced by XAI, which enhances the changes the model transparency, enabling efficient debugging, and enhancing system stability. The use of neural networks in practical applications where trust and transparency are critical necessitates XAI [37].

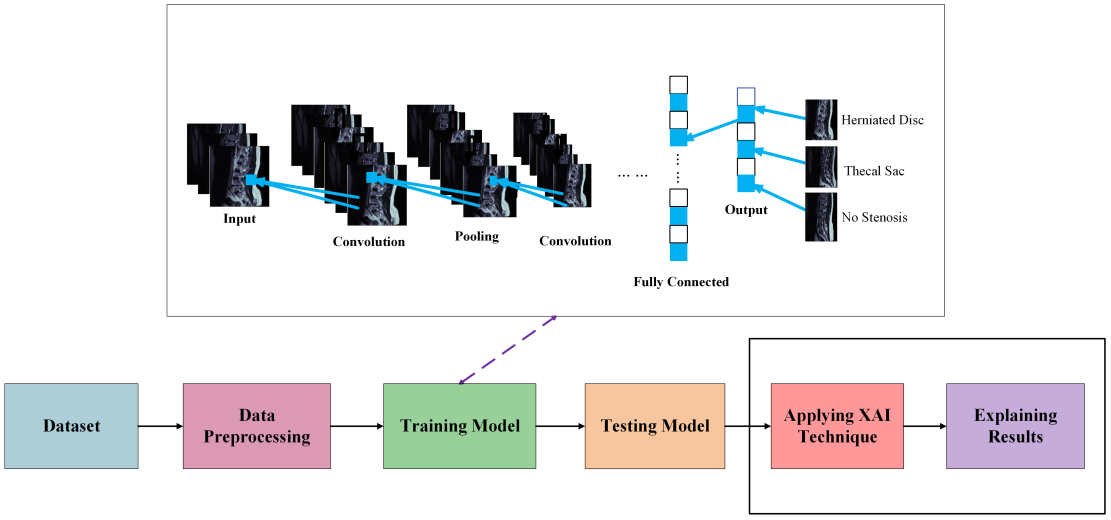


FIGURE 1.4: Modern CNN with XAI for lumbar spinal stenosis classification.

In recent years, several DL approaches with XAI have been developed. Figure 1.4 shows a typical XAI-based DL methodology for lumbar spinal stenosis classification.

Nevertheless, the lack of appropriate visualization in all of these previous studies that have applied XAI results in a lack of explainability. This absence of interpretation results in the model's complexity, as it concentrates on superfluous features, rendering the entire system questionable. Additionally, the model's efficiency is not being improved; rather, XAI is being employed solely to enhance its explainability.

1.6 Motivation

The DL models always improve performance in classifying LSS, surpassing tradi-

tional methods like machine learning (ML) with better and remarkable accuracies. The models proposed by other authors for LSS had complex structures that were difficult to interpret [38]. Furthermore, the absence of transparency is especially concerning in healthcare environments, where trust and understanding of artificial intelligence (AI) systems are essential. DL models may also have layers or parameters that do not change the output and may give inaccurate results. For that reason, knowledge regarding how a system makes decisions must be shared to clarify things. Additionally, XAI can simplify this method while increasing its accuracy and dependability.

1.7 Problem Statement

The LSS classification models can have additional layers and parameters which may produce imprecise results. A sophisticated model needs to make the system more robust and accurate with a reduced number of layers and make the model learn to focus on important features (stenosis) that can be possible by explainability.

1.8 Research Questions

1. How efficiently diagnosis LSS be diagnosed using a DNN architecture?
2. What techniques and methodologies can be used to provide insights into the model decision process to reduce the number of layers of LSS?
3. How to implement XAI to achieve explainability from the proposed LSS diagnosis prediction model?

1.9 Objectives

- Proposed a novel LSS diagnosis prediction model.
- Highlighting the necessary features on each layer and then making changesing the model architecture according to them.

- Implementation of XAI on the proposed model to simplify the decision-making process.

1.10 Significance of the Problem

Accurate diagnosis of the LSS is time-consuming, resource-intensive, and requires high technical expertise. The DL models can be used to empower the traditional methods, but the DL models have some shortcomings; i.e., DL models select some unnecessary features to train a prediction model, and also their internal decision-making process is hard to understand.

1.11 Contribution of Proposed Work

The novelty of our proposed work is by integrating XAI methods directly into the model to understand the internal process of LSS diagnosis. Unlike previous approaches that use XAI only post-hoc for interpretation, our method uses Layer Grad-CAM iteratively to identify and remove those unnecessary layers in the CNN, to reducing the model complexity while improving interpretability and maintaining high accuracy 89.64 %. These results are more efficient, transparent, and clinically trustworthy which focuses on anatomically necessary regions which is significant to step beyond the conventional black-box deep learning models in medical imaging.

1.12 Thesis Organization

This chapter in this thesis is organized as follows:

- Chapter 1 introduces lumbar spinal stenosis and its subtypes, and includes motivation, problem statement of the proposed work, and the importance of the solution.
- Chapter 2 conducts a comprehensive literature review of LSS classification using deep learning approaches with and without XAI.

- Chapter 3 details the methodology used, focusing on the use of XAI to make the LSS classification model less complex and interpretable.
- Chapter 4 presents the results and a detailed discussion of the findings.
- Chapter 5 summarizes the conclusions drawn from the study results.

Chapter 2

Literature Review

Many researchers have proposed methods to classify LSS. Our main study involves classifying LSS using MR images based on deep learning techniques. This literature focuses on diagnosing LSS using a CNN model and explainable AI.

2.1 Survey of Existing Techniques

2.1.1 Deep Learning Methods with Explainable AI

Suzuki et al, [39] proposed a model to diagnose the presence or absence of LSS using the CNN. They used the dataset related to 150 patients from January 2022 to August 2022. To strengthen external validation, the study included 25 additional patients who went for surgery at two separate hospitals. For training a model, they used a CNN, which has specialized layers, including convolutional filters, pooling, and fully connected layers. Grad CAM was used to highlight the highly featured areas extracted by CNN. In their paper, they could use the Layer Grad CAM on each CNN layer to see what basis their model is predicting.

In Kim T et al. [17], the study proposed using a pretrained model based on CNN for diagnosing LSS using MR images. They used the CNN-based transfer learning algorithms to train the pretrain model and they used May 1, 2005, to December

31, 2017 on LSS-related datasets. In this study, they used four algorithms to conduct the experiments: VGG16, VGG19, ResNet50, and EfficientNet1.

They have selected VGG19 and achieved around 81.9 % accuracy with 5-fold. To evaluate the features proposed network by inspecting the spinal radiographs were inspected using the Grad-CAM algorithm by heatmap of LSS could be categorized into reduced disc height, narrow foramina, short pedicle, and facet degeneration. It helps to identify the error in feature extraction, including the back muscle and organs. The major gap in this paper is that they could not modify their model's architecture to improve their model's accuracy because they used the pretrained model, i.e., VGG19.

The authors in [40] used the Support Vector Machine (SVM) based algorithm to do the binary classification tasks. To capture he density of the class majority, they used OCSVM, which was trained on training samples of healthy patients from the given training set only. The random hyperparameters are set on the parameters called kernel, nu, and gamma using the healthy samples related to the validation set for the evaluation performance. For the training of an OCSVM-based model, they have used only the healthy patients, around 900 samples, and 150 unhealthy patients were removed.

They have used the XAI technique to get the explanation of the model, e.g., OCSVM or binary classifier, using the tool Local Interpretable Model-agnostic Explanations (LIME). To understand the black box of the model, they used a single prediction using LIME, which uses the approximation of a local prediction of a black box. For this paper, the authors used a closed-source dataset to train a model. Furthermore, they could implement the different data preprocessing methods, i.e., masking, rotation, and resizing, to clean, make consistent, and make suitable for further processing for training a binary classification-based SVM model.

The authors in [41], proposed a model to predict Lumbar Disc Degeneration (LDD) based on four machine learning models, e.g., XGBoost, Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT). The accuracies of the models were

XGBoost 82.1 %, LR 80.8 %, RF 79.5 %, and DT 76.3 %. For training the model they have 518 patients, then they have simply applied an algorithm called Synthetic Minority Oversampling Technique (SMOTE) to mitigate class imbalance in the training set, and then they randomly split 70 % training set, and 30 %.

They have also done the interpretability analysis on their proposed model to get insights from these trained models. They have used the Shapley Additive Explanations (SHAP) algorithm to determine the importance of various predictor variables optimal model's prediction and then allow to make decision-making. The values from SHAP allow for the analysis of potential risk factors for LDD, and the SHAP values are the key features associated with an individual patient's prediction. In this research, the authors used a private dataset, and they missed the implementation of the data preprocessing methods to clean the inconsistencies to make the model's predictions more accurate.

The authors in [42], proposed models, e.g., XGBoost, DT, RF, SVM, and LR to predict the stenosis on the patient's posterior lumbar spine, and do the hyperparameter fine tuning was then performed based on grid search combined with the 5-fold cross-validation to optimize the model performance evaluation. The dataset that they have used to train these models, using January 2018 to June 2020, from the Zhongda Hospital of Southeast University, and they used 1,016 MRIs. During the data preprocessing, they learned that there were 10 % variables with missing values.

To handle these missing values, they used the imputation technique called the K-Nearest Neighbor (KNN) strategy based on k = 5. To get the model insights on which basis the model is predicting, the author used SHAP. SHAP is used as an additive explanatory model where all the features are considered, and the model then generates a prediction value for each prediction sample where the individual feature's score is based on the SHAP value.

Out of all, the trained they have selected XGBoost because it is giving the highest accuracy among the other models. For this study, they used the private dataset, and there are missing data preprocessing methods, i.e., zooming, resizing, and no-

rmalization.

The authors in [43], they used the pretrained model EfficientNet-B5, which is based on the ImageNet database, with a learning rate 3×10^{-5} with cosine and Adam optimization. They used the January 2012 to February 2021, which was approved by the Institutional Review Board of Chuncheon Sacred Hospital (CHUNCHEON) and for training a model, and their training dataset contained around 162,257 lumbar X-ray images from the given patients 31,149 For testing their model they used 18,014 lumbar X-ray images which contained 3,512 patients.

During training, a model their model is controlled by four-fold cross-validation. Lastly, they used Grad-CAM and applied this technique on the last layer of their CNN to verify their trained model to get the proper insights and try to find out whether their model is predicting lumbar stenosis based on the specific regions or not. For this research, they could perform some of the data preprocessing methods, i.e., resizing, normalization, and used the private dataset.

The authors in [44] explained the Recurrent Lumbar Disc Herniation (rLDH), which describes the herniation of the disc from the prior surgical procedure in the lower part of the spine. The basic role of rLDH is to support body weight and the continuous strain on the lower part of the body, then the lumbar spine is vulnerable to disc herniation.

Their study involved predicting the Dengue Shock Syndrome (DSS) and which is the most common cause of lumbar back pain. They used many methods to develop the ML algorithms, e.g., XGBoost, Random Forest (RF), Artificial Neural Network (ANN), Support Vector Machine (SVM) too prediction of dengue shock syndrome. Training a model, they used the 230 MRI dataset, where there were 45 patients related to rLDH surgery, and 180 patients without rLDH surgery.

During the data preprocessing, the missing values were handled with average or mode imputation techniques, and then it was further use the normalization method to normalize the MR images. They used SHAP to check whether their models are predicting correctly or not. It allows the features to predict the average when the SHAP value is positive; the opposite would be true when the SHAP value is nega-

tive. For this research, the authors could use the data preprocessing methods like deblurring (bacause mostly MRIs are blurred), resizing, and zooming.

The authors in [45], an RF model to predict the LSS based on surgical recommendations which were compared to individual doctor recommendations. Their model used the 500-based MRI of the patients, which also included the patients who had the stenosis, and all of these were determined by the radiologist. The given dataset was randomly split into 80 % for fine-tuning and training a model, whereas for testing, they used the remaining for making predictions. They used

SHAP to check whether the model is predicting on the best possible given values were computed from the random forest model to predict the surgical recommendations. These values are varied due to the predicted likelihood of being denied surgery, which is around 0.195.

The denial is due to absence due to there was no stenosis present or not. Whereas the SHAP probability, which is around 0.851, means that the vignette agrees to do the surgery.

The authors in [46], using Explainable AI (XAI), the authors explored the multi-determined drivers of opioid and NSAID prescriptions for non-specific chronic low back pain (ns-cLBP). The authors used a large multi-modal dataset of electronic medical records from 4,077 MRI images to train predictors of prescription patterns in tree models. In addition, SHAP analysis was used to control for the explanations of predictive choices. To the surprise of the authors, the results showed psychosocial determinants (e.g., anxiety/depression, race/ethnicity, social support) and the year of treatment—representing changing opioid prescribing guidelines—were often better predictors than MRI-derived assessments of tissue pathology. This research adds a clear, data-driven model that measures the considerable contribution of the biopsychosocial model to clinical decision-making, thus challenging the common belief that prescription patterns are mostly based on radiological information.

The authors in [47], they selected a retrospective cohort consisting of 580 patients with tuberculous spondylitis and employed LASSO regression in the feature sele-

ction process. The authors subsequently employed techniques like multivariable logistic regression to create a prognostic nomogram. The derived model successfully recognized the seven key surgical and clinical predictors. To improve the interpretability of the trained model, the authors use the XAI technique, such as SHAP, which evaluates and visualizes the impact of each variable on the risk estimation, depending on patients experiencing an extended postoperative hospital stay.

The authors in [48], the authors have mentioned multiple applications, including image enhancement, anatomical segmentation, and diagnosis. The authors used Deep learning methods such as CNNs, ResNet, GANs to train a model to predict the LSS. However, the interpretability and clinical remain significant challenges. To overcome these hurdles, the researchers of this paper used the XAI method, such as SHAP, to highlight those regions on which the model is predicting.

TABLE 2.1: Analysis of Deep Learning Methods with XAI

Ref.	**Dataset**	**Technique**	**XAI**	**Limitations**
[17]	–	VGG16/19, ResNet50, EfficientNet1	Grad-CAM	No Occlusion Sensitivity, Layer Grad-CAM, and no data preprocessing.
[39]	–	CNN	Grad-CAM	Private dataset and no proper data preprocessing
[40]	900 MRI	OCSVM	LIME	Private dataset and No proper data preprocessing.
[41]	–	XGBoost, LR, RF, DT	SHAP	Private dataset and No proper data preprocessing, i.e., masking.
[42]	1K MRI	XGBoost, DT, RF, SVM, LR	SHAP	Private dataset and proper data preprocessing.
[43]	ImageNet	EfficientNet-B5	Grad-CAM	No proper data preprocessing like masking.

TABLE 2.2: Continue from previous page

Ref.	Dataset	Technique	XAI	Limitations
[44]	230 MRI	XGBoost, ANN, RF, SVM	SHAP	Occlusion Sensitivity, and private dataset.
[45]	500 MRI	RF	SHAP	Data preprocessing methods were not used.
[46]	4,077 MRI	RF, AdaBoost, XGBoost, Bagging	SHAP	Lacks detailed pain descriptors, treatment rationales.
[47]	–	LASSO, Logistic	SHAP	Depending on patients experiencing an extended postoperative.
[48]	–	CNN, ResNet, GANs	SHAP	Data preprocessing methods were not used.

2.2 Research Gap

From the above literature reviews, the authors mentioned the papers about the LSS classification using MRIs to diagnose stenosis. The classification models have a complex structure and often make them difficult to interpret. The existing studies that applied XAI do not properly provide visualization and lack of explainability. Furthermore, the decision-making processes of the model are difficult to understand. Moreover, LSS classification models have additional layers that make no difference in output, but make it complex and may produce inappropriate results. For enhancing transparency and better understandings, there is a need to present insights of the internal model process of dicision making.

Chapter 3

Methodology

3.1 Proposed Model

The proposed approach is to dianosing the LSS and to interpret the internal decision making processes which is illustrated in Figure 3.1. In this thesis, the internal decision making process is being visualized using the most commonly XAI techniques like Layer Grad CAM and Occuslion Sensitivity.

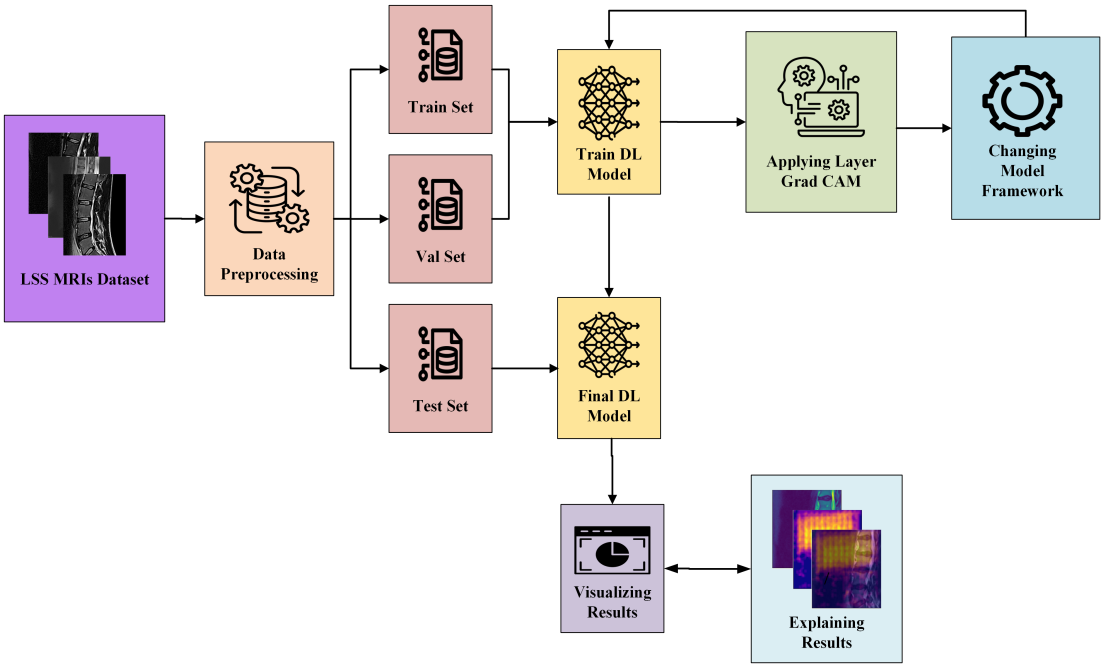


FIGURE 3.1: Proposed Model

The proposed approach has two major steps: first of all training a model and repeat this process using Layer Grad CAM until improving the results, second using Occulsion Sensitivity technique to interpret and visualize the internal decision making processes.

3.2 Dataset

The Sud Sudirman [49] dataset related to lumbar spinal stenosis is the lumbar spinal MRI dataset. This dataset includes 13,686 MRIs related to lumbar spinal, which are available in IMA formats. There is a radiologist review available as well, which helps us to distribute them into labels and then classify these images based on the presence or absence of two types of stenosis, resulting in three distinct classes in the dataset:

1. Herniated Disc

   According to 3.2, the intervertebral discs are squeezed, and therefore the nerve roots are compressed.

2. Thecal Sac

   According to 3.3, the intervertebral discs are mildly squeezed and compress the thecal sac.

3. No Stenosis

   According to 3.4, there is no stenosis in the patientś MRIs.

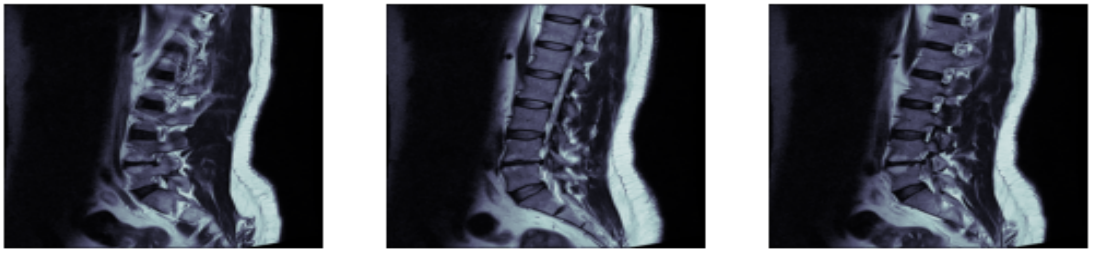


FIGURE 3.2: Herniated Disc

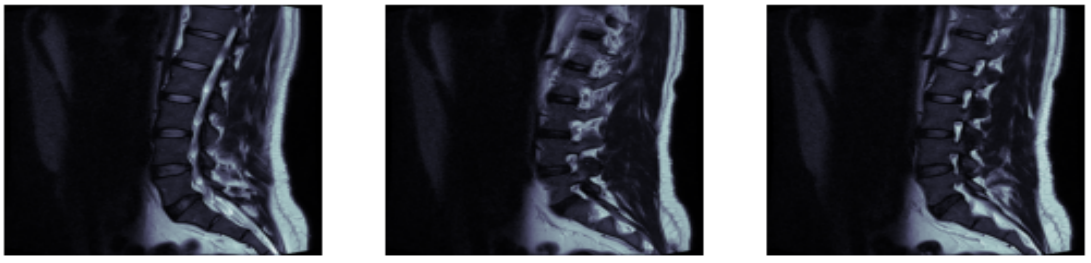


FIGURE 3.3: Thecal Sac

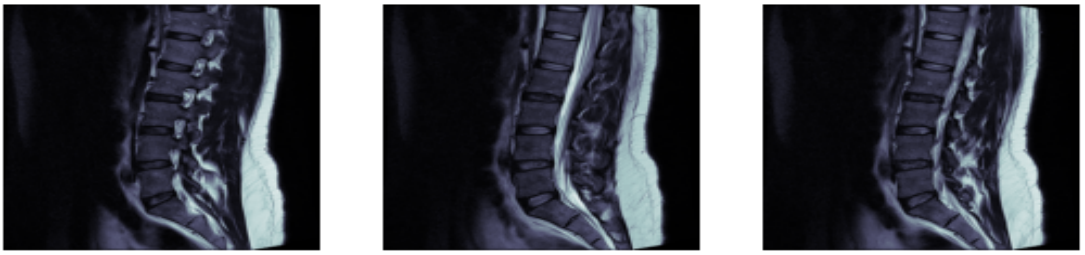


FIGURE 3.4: No Stenosis

3.3 Data Preprocessing

Effective data preparation is a foundational step important for meaningful dataset exploration. The data preparation critically governs the success of the dataset investigation and the training of the model using DL. The data preprocessing is essential for training a CNN for our proposed work. Although CNNs are powerful feature extractors, they depend on consistent, high-quality input.

3.3.1 Masking

Masking is a data preparation technique employed before training machine learning models to address missing or unnecessary data by designating individual values or entire features as uninformative.

In this thesis, we have used the grayscale images, and the main reason behind using the masking is to remove the white and black spots, noise, and outliers which might be present in these images [50, 51].

This operation expands the bright areas in the image. It first takes an image as input X, and then takes kernel B of the same size as the original to compute the maximal pixel value overlapped by the kernel. Then it replaces all the pixels in the image at the anchor point z with the maximal value according to (eq. 3.1).

$$X \oplus B = \{x + b : x \in X, b \in B\} \tag{3.1}$$

After applying the (eq. 3.1) we can obtain the results as shown in figure 3.5.

FIGURE 3.5: Masking

3.3.2 Resizing Images

Resizing helps to reduce computational timing because different or large-sized images take a lot of resources for computation. There were different sizes of the images in the given dataset, i.e., 320 x 320 x 1, 640 x 640 x 1, and 384 x 384 x 1, respectively. Therefore, all the images were resized to 400 x 400 x 1, which impact the results. The Figure 3.6, shows the results after using the resizing images.

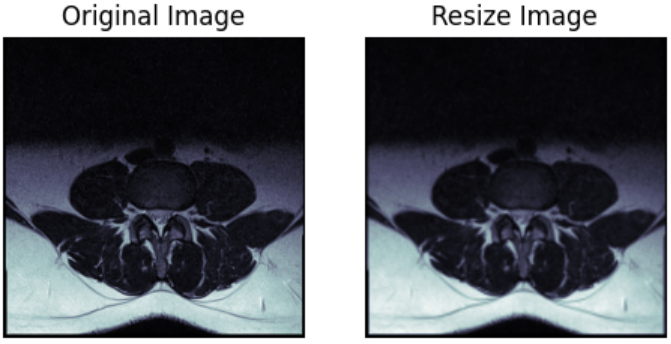


FIGURE 3.6: Resizing Image

3.3.3 Zoom

Zooming enhances model performance by strategically focusing on regions of interest (e.g., stenosis in MRIs) while balancing computational efficiency and feature relevance. It zoom in, crops, and resized images [52, 53].

The main benefit is to reduces memory overhead by isolating critical details (e.g., fine textures or small anomalies) and standardizing input dimensions, aligning with neural network architectures requiring fixed-size inputs [54]. The Figure 3.7, shows the results after applying this method.

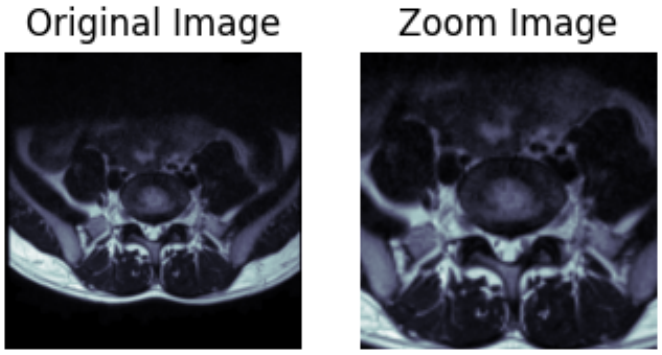


FIGURE 3.7: Zooming Image

3.3.4 Unblurred Images

Using Zooming make all the images blur, then for unblurring images is the method which makes the images unblur or increasing the intensity of the images. The following mathematical equation helps us to understand how it is unblurred images.

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - F_\theta(X_i))^2 \tag{3.2}$$

The results shown in Figure 3.8.

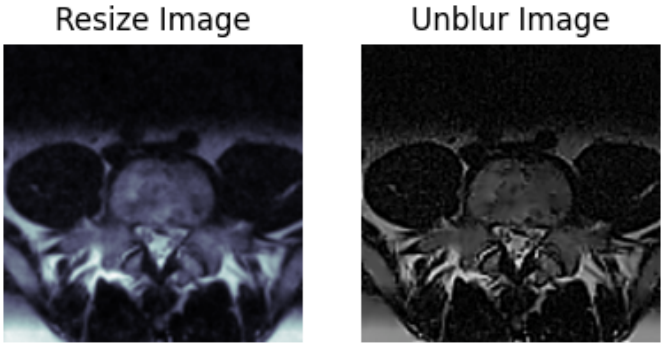


FIGURE 3.8: Unblurred Image

In equation 3.2, F_θ be the CNN allows to unblur the images, $Y_1, Y_2, \ldots, Y_n$ be the input images, and their counterpart of the blur images, $X_1, X_2, \ldots, X_n$. In order to find out the θ parameter then for the CNN, which would minimize the mean squared error on every pixel of the given image.

3.4 CNN and XAI Based LSS Prediction

In this phase, we have introduced an XAI that integrates a CNN designed for multi-class spinal image classification, focusing on LSS. The model architecturally embeds XAI to prioritize clinically critical features, such as herniated disc, thecal

sac, and no stenosis, during classification, enhancing diagnostic accuracy while ensuring transparency. Unlike conventional approaches, it directly aligns learned features with LSS [55, 56], improving interpretability without compromising performance.

Our research primarily goal is to focus on the XAI architecture to get the interpretation of the trained model, and also improving the medical image analysis, like multi-class image classification of spinal stenosis [57, 58].

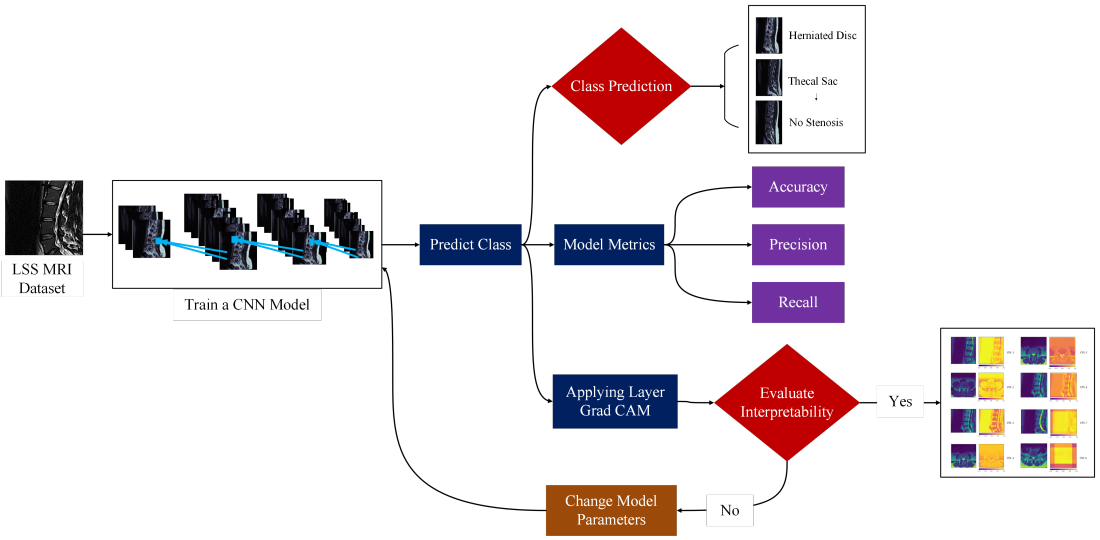


FIGURE 3.9: XAI with the CNN - Proposed Approach

Our approach main focus is to diagnosis the stenosis by reducing the number of layers. The main advantage of our proposed model is to help out in the field of medical automated systems, offering reliability, trust-enhanced performance, and improving the accuracy of the multi-class classification scenarios.

3.4.1 CNN

The CNN architecture helps to solve the problem to do the classification relaed to lumbar spine. Essential elements of CNNs comprise convolutional layers [59, 60], activation functions, pooling layers, adaptive average pooling, and fully connected layers. Convolutional layers employ convolutional filters on input images, generating feature maps highlighting distinct characteristics, such as edges, textures, or patterns.

For this research, we used the filters to traverse the picture and execute element-wise multiplications aggregated to generate the feature maps. As for the activation

functions we used the Rectified Linear Unit (ReLU), which allows the non-linearity into the model, enabling it to acquire more complex representations [61, 62].

For reducing the spatial dimensions of feature maps, the pooling layers technique. It reduces the computational demands and alleviating the risk of overfitting [63, 64]. For reducing the spatial dimensions of feature maps, we used the technique called Adaptive Average Pooling. It allows to calculate the average and adjust the units by considering the mean of activation characteristics [65]. It also reducing the computational demands and alleviating the risk of overfitting [63, 64].

The fully connected (FC) layers were used to make all the layers for following up to many iterations of convolutional and pooling layers. These layers integrate all the characteristics to generate final predictions regarding the class of the input picture. In the context of multi-class classification, as in our work, the output layer typically employs a softmax activation function to produce a probability distribution over the various classes, indicating the likelihood of each class being correct.

3.4.2 Model Training

Our proposed technique is basically based on CNN, which is trained for doing image classification tasks and consists the multiple layers that are responsible for handling the process of the input data. Our model first takes the $400 \times 400 \times 1$ grayscale images as the input to convolutional layers, gives 8 as the output channels with the kernel size of 3×3, then applies the **LeakyReLU** as the activation function, and we also set the padding as **"same"** which allows to add zeros on all outer regions of an image so that no pixels should not be ignores. After this, we add a layer of max pooling, which allows for down-sampling the feature maps of, the input image by a factor of 2 in both width and height.

This process repeats multiple times with the increasing number of output channels in exponential (16, 32, 64, 128, and 256), and takes the same kernel size in all the CNN layers. Next, in our proposed model we have added the batch normalization layer which allows to make the mean and standard deviation (0 to 1), and it makes the model to converge even faster, then we use the adaptive average pooling which makes the down-sampling the feature maps, and it also responsible to reduce the

dimension makes it to 1×1.

Making all the preceding CNN layers fully connected, we used the linear layer. Finally, our model has the final layer, which is a softmax activation function, which gives the probability distribution of the three classes to select the highest probability class among the other.

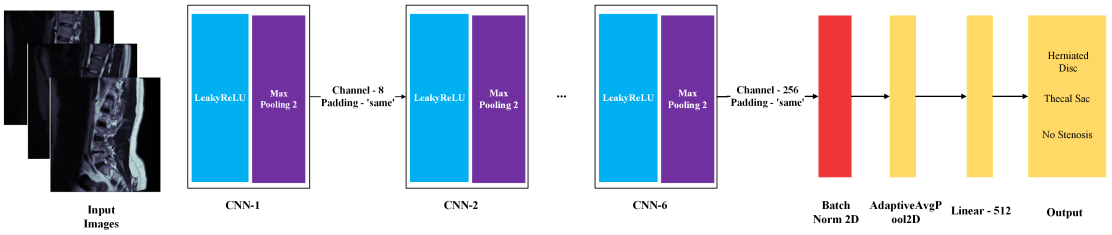


FIGURE 3.10: Internal Model Training - Layers

The intricate structure of LSS classification models frequently complicates their interpretability, resulting in difficulties in comprehending the decision-making processes of classification.

3.4.3 Layer Grad-Cam

Grad-CAM is an XAI-based technique used to understand the importance of different layers in the trained CNN model. Grad-CAM helps to visualize differents, parts on the basis of an input image to contribute to the model's predictions [66]. It is model-specific and is used in CNN-based models. It works on the last layer. However, in our study, we applied Grad-CAM to the preceding layers in the CNN model, highlighting those regions that the model has selected for prediction.

Layer Grad-CAM uses the gradients to any target concept on the basis of the localization map, highlighting the important or necessary regions in the image for predicting the target.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.3}$$

According to equation 3.3, the localization map layer Grad-CAM $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$ where the u and v are the width and height for the class c, and where $\mathbb{R}$ is some real input image. We first do to compute the gradient of the score of the class c,y^c (before getting the results from the softmax) with the feature map activation

(the output of a convolutional layer after an activation function i.e. ReLU is applied to it) A^k of the convolutional layer, i.e. $\frac{\partial y^c}{\partial A^k_{ij}}$. These gradients flow back is a global average, pooled over the Z will be the $height \times width$ dimensions of the image. The i, j also denote the width and height, and thus we will get the localization mapping for Layer Grad-CAM. During the computation of the α^c_k while backpropagating gradients for activation functions.

The computation of the successive matrix products of the weight matrices and the gradient concerning the activation function till the final CNN layer, where the gradient is being propagated. This weight α^c_k, which represents a partial linearization of the DL based on A, and thus captures necessary feature maps k for a target class c.

$$L^c_{Grad-CAM} = ReLU(\sum \alpha^c_k A^k) \tag{3.4}$$

According to equation 3.4, it is applied to the ReLU activation function, which involves a linear combination of the feature maps. We focus only on the positive classes, which have high-intensity pixels and are the desired class, and ignore all the negative pixels, which have low intensity. Without the ReLU activation function and Grad-CAM, the localization mapping may sometimes highlight those regions that are not the desired class.

3.4.3.1 Algorithm for Layer Grad CAM

The algorithmic steps of layer Grad-CAM as follows:

1. Output: Calculate weights for each feature map of α^c_k.
2. Input: Corresponding activation map of a spcific layer of A with k, and where k $\in$ 1, 2, ..., K represents of each feature map.
3. For n = 1 to k:
   (a) Let H and W be the height and width of an image.

(b) Initialize gradSum:= 0, Z:=0

(c) For i = 1 to H:

i. For j = 1 to W:

A. gradSum:= gradSum + gradient[i][j]

(d) Compute normalization factor: Z = H × W.

(e) Compute importance weight: alphaCK:= gradSum / Z

4. Compute Layer Grad CAM using Relu $L^c_{Grad-CAM} = ReLU(\sum \alpha^c_k A^k)$

3.4.4 Modifying Model Framework

During our research, we achieved the results by integrating the Grad-CAM layer with the layer attribution in our proposed model. After incorporating, we identified the layers that can improve the model's performance. To improve the model's performance, we changed the model's architecture. We removed all layers that decreased the model's performance during our research.

3.5 Evaluation of the Proposed Model

During our research, we have employed XAI techniques to interpret the model's decision-making processes and make the model more comprehensible.

3.5.1 Occlusion Sensitivity

Understanding the operation of the CNN-based model requires the interpretation of the feature activity in intermediate layers. A primary difficulty in image classification is determining whether models accurately locate and identify target items in images or mostly depend on incidental contextual patterns (e.g., background textures or neighboring objects) for predictions [67].

Occlusion Sensitivity is the simplest technique used to get an explanation of the model and develop an understanding of what features the model uses to predict it. This technique identifies critical regions in an image by systematically masking

portions of the input (typically with a gray square) and observing how the model's prediction probability for a target class changes as the mask moves. This probability, with the mask's position, the method quantifies the importance of specific image regions to the model's decision [68].

Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}^C$ be a trained model that maps an input $x \in \mathbb{R}^d$ (i.e., an image) to a probability distribution over C classes. For a target class c, the model's confidence is $f_c(x)$. Occlusion sensitivity identifies regions of x critical to the prediction c by systematically masking parts of x and measuring the resulting change in $f_c(x)$.

Algorithm steps to compute the occlusion sensitivity:

1. Slide a window w of size $k \times k$ over x with stride s.
2. For each position (i, j), compute $S(R_i, j)$, where R_i, j is the region covered by w at (i, j).
3. Aggregate $S(R_i, j)$ into a heatmap.

The heatmap H is a discrete approximation of the importance function:

$$H_i, j = \frac{1}{N} \sum_{n=1}^{N} S(R_{i,j}^{(n)}) \tag{3.5}$$

Where N is the number of occlusions overlapping at (i, j). As $k \longrightarrow 1$ and $s \longrightarrow 1$, H converges to a continuous sensitivity map.

3.6 Evaluation Metric

The classification of lumbar spinal stenosis subtypes follows a systematic method with models trained on an annotated dataset to learn discriminative features, fine-tuned via a validation set to optimize hyperparameters, and rigorously evaluated on an independent test dataset.

3.6.1 Accuracy

Accuracy evaluates a model's overall classification performance by computing the proportion of correct predictions relative to the total predictions made across all

classes. This metric is important in multi-class classification models, as it uniformly evaluates predictive consistency across diverse categories, providing a balanced performance measure even when classes are unevenly distributed.

$$Accuracy = \frac{1}{N}\sum_{i=1}^{n} 1(\hat{y}_i = y_i) \tag{3.6}$$

According to equation 3.6, N is the total number of samples. The numeration from $i = 1$ to n where it checks the $\hat{y}_i$ (predicted class) with y_i (actual class). If the predicted and actual classes are equal, then it gives us the value of 1; otherwise, 0. After all of this, it will simply divide all of these by N.

3.6.2 F1-Score

The F1-score is a crucial statistic in classification tasks, providing a comprehensive integration of accuracy (the ability to reduce false positives) and recall (the ability to identify genuine positives) into a single measure. The F1-score, derived from the harmonic mean of these measures, offers a balanced assessment of a model's diagnostic reliability, which is especially vital in unbalanced datasets frequently seen in medical applications such as LSS classification.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.7}$$

3.6.3 Precision

Precision is a vital assessment parameter in machine learning classification tasks, defined as the ratio of properly predicted positive examples (true positives, TP) to the total number of instances projected as positive (true positives plus false positives, FP). It is mathematically represented as:

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

This precision evaluates the ability of the model to reduce false positives, therefore ensuring that, should it project a positive class, the prediction is rather accurate.

3.6.4 Recall

Recall, sometimes referred to as sensitivity or true positive rate, is an important assessment parameter in classification tasks that assesses a model's accurately identify all pertinent occurrences of a positive class. The metric is defined as the ratio of true positives (TP) to the aggregate of true positives and false negatives (FN):

$$Recall = \frac{TP}{TP + FN} \tag{3.9}$$

Chapter 4

Implementation and Evaluation

In this chapter, we will briefly explain the experimentation we conducted to assess the proposed methodology discussed in Chapter 3.

4.1 Tools and Techniques

During our experimentation, we utilized the following tools and techniques described below.

4.1.1 Google Colab Notebook

Google Colab notebook is an online platform for researchers to conduct experiments. Researchers can use Google Colab to write Python code for experimentation. It offers high performance with GPUs and TPUs, allowing researchers to train their ML and DL models.

4.1.2 Python Programming Language

Python is the most popular programming language, enabling researchers to conduct their experiments by training models using ML or DL techniques. We can utilize Python's most popular libraries, such as PyTorch, to develop the DL model.

Scikit-learn is another popular library that allows the use of ML techniques.

4.1.3 PyTorch Library

PyTorch is the most popular open-source DL framework. It enables researchers to apply DL techniques to train models using Python. This library accelerates the research, prototyping, and deployment process. PyTorch allows researchers to seamlessly transition between modes to optimize functionality, speed, ease of use, and flexibility.

This library enables researchers to train models in image classification, reinforcement learning, and Natural Language Processing. It offers multiple algorithms related to image classification.

4.1.4 Data Visualization

Data visualization is a tool that allows researchers to present complex information in a clear, concise, and visually appealing manner. This would enable researchers to identify patterns and trends and share these findings with others.

For data visualization, a popular library such as matplotlib allows creating static images. It is used to create different types of graphs, line plots, histograms, and scatter plots. It is helpful to show the MRIs of the patients before and after data preprocessing.

4.1.5 Google Colab Paid

For this study, we used the paid version of Google Colab, as the free version has limited resources. In the paid package, there are some extra features like GPU, TPU, and RAM, which allow us to take few time to train a model. Researchers who want quick, high-performance computing resources depend on this tool.

4.1.6 XAI Libraries

LIME is an approach for explaining the individual predictions made by complic-

ated predictive machine learning models that doesn't depend on the model. LIME makes explanations that people can understand by using less complicated models (like linear classifiers) to approximate the behaviour of a "black box" model locally around a specific case. Shapley values are a part of cooperative game theory that SHAP uses to figure out how much each trait affects a model's prediction. It gives both local and global answers and makes sure that theoretical claims are true, such as that things are consistent and add up.

The AI researchers of Meta's team, developers of Captum is an open-source PyTorch library is used to interpredication that employs gradient- and perturbation-based attribution methods to interpret the decision-making mechanisms of deep learning models. This library addresses the imperative for transparency by enabling researchers to systematically analyze and validate neural network behaviors. But in contrast to Captum is natively integrated with PyTorch, leveraging its autograd system to compute exact gradients for deep learning models. Unlike LIME and SHAP, which are model-agnostic and approximate attributions via perturbations, locally and globally,

4.2 Dataset for Research

In this study, we have used a public dataset related to MRI lumbar spine for image classification tasks. In our dataset, there are around 13,686 patients' data, and a radiologist is labeling each, and they have mentioned what has happened with their patients in their review report, i.e., **radiologist_report.xlxs**. This dataset has been divided by the patient's ID, i.e., 1,2,3, and so on, and in each patient's ID folders, there is a folder called **L-SPINE_LSS_20160309_091629_240000** respectively. There were three distinct views, i.e., sagittal, axial, and coronal, and for this study, we selected the sagittal and axial views. This dataset has two MRI-weighted, i.e., T1 and T2, and for this study, we have selected these two T2-weighted MRIs, i.e., **T2_TSE_SAG** and **T2_TSE_TRA**.

According to Figure 4.1, the dataset has three distinct types, i.e., Herniated Disc, Thecal Sac, and No Stenosis. This dataset is perfect for medical imaging, lumbar

spinal identification, and classification ML, and DL model development and testing due to its large size and labeling. A wide variety of types are included:

1. Herniated Disc
2. Thecal Sac
3. No Stenosis

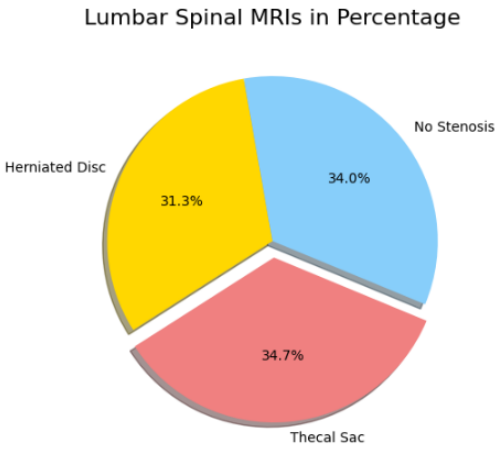


FIGURE 4.1: Lumbar Spinal Dataset

For this study, we have divided into training, validation, and testing sets consisting of the MRIs of 13,686 patients with the ratio of 90 %, 5 %, and 5 %.

The images in this dataset were initially different in size and did not consistently focus on the lumbar spine section, which presented challenges for effective analysis. Conducted an exhaustive preprocessing procedure to resolve these concerns. To ensure that only the lumbar spine section was retained, the images were resized, and do the zooming to find the region of interest. According to the radiologist's report, our preprocessed dataset with properly manually labeled is available on kaggle [1]

To enhance the performance of ML algorithms by guaranteeing uniformity throughout the dataset, the images were normalized after cropping to standardize the pixel intensity values. And finally, the images were resized to 400 × 400 × 1 pixels. Not

[1]"https://www.kaggle.com/datasets/abdullahkhan70/lumbar-spinal-mri-dataset"

only do these resizing methods help to make the images consistent, but they also help to train the models because most of the algorithms like CNN also require all the images in the same size.

4.3 Implementation of Model

In this study, we have implemented an XAI-based CNN model using the PyTorch library. This implemented model consists of the CNN along with the Layer Grad CAM. The CNN has a series of multiple convolutional layers, which are responsible for extracting the features from the given image. The Layer Grad CAM is an XAI-based technique that allows us to get to know what features and layers are important from the given CNN-trained model. By combining the CNN and the layer Grad-CAM methods improve the trained model's internal workings and performance in image classification.

We have started training a model with the first layer of a convolutional network that was accepted 400×400×1 Grayscale images (Lumbar Spine MRIs), which used 8 filters with the kernel size of 3×3, padding with a value of 'same', and then it passes through a ReLU activation function. Next, we have max pooling layers to reduce the feature map size by half. In the next layer, we have used the convolutional layer along with the LeakyReLU instead of ReLU activation because the LeakyReLU resolves the problems like dead neurons and stops learning. After along with the max pooling layers are repeated, with the filter count increasing exponentially (8, 16, 32, 64, 128, 256) respectively. After this, we have added a batch normalization layer (Scaling and shifting layer activations to have zero mean and standard deviation within each mini-batch); we majority add it after the fully connected or convolutional operation in a convolutional layer, and before the non-linear activation function, i.e., ReLU.

Followed by the Adaptive average pooling performs average pooling (reduces the spatial dimensions of the feature map by calculating the average value within small windows or regions of the feature map) operations, which reduces the spatial dimensions (adjusts its pooling window size dynamically to achieve a target output

dimension). And after this, we have flattened and reshaped the tensor by combining all dimensions except the batch size into one, and it has followed one linear layer with 512 neurons along with the three distinct classes using softmax activation. Then we have used the layer Grad CAM along with the layer attribution that allows us to show the importance of individual layers of feature mapping, which were calculated based on individual gradients.

The average of these gradients to obtain weights was performed, which reflects the importance of each feature map for the target class score. Lastly, the Grad Cam map is obtained by performing a weighted sum of the feature maps.

TABLE 4.1: Model Architecture Specifications

Parameter	**Configuration**
Convolution Layers	6
Filter Sequence	8, 16, 32, 64, 128, 256
Kernel Dimensions	3×3
Padding Strategy	Same
Activation Function	Leaky ReLU ($\alpha = 0.1$)
Pooling Layers	6
Pooling Operations	Max Pooling, Adaptive Average
Pool Window Size	2×2
Pool Stride	2
Optimization Algorithm	Adam
Loss Metric	Cross-Entropy Loss
Mini-batch Size	32
Training Epochs	100

4.4 Iterated Model Layer Grad CAM

Layer Grad-CAM (Gradient-weighted Class Activation Mapping) is an enhancement of the conventional Grad-CAM method that produces class-discriminative visual elucidations for particular layers inside a deep neural network. It indicates the specific areas of an input image that a given layer in the model concentrates on during prediction. We have distributed our analysis based on multiple model training iterations.

4.4.1 Conventional Model 1st Iteration

In this iteration, we have used the CNN layers with the **out_channels** of (8, 16,

32, 64, 128, 256, 512). It also has the same activation functions and max pooling with the same hyperparameter values. But this iteration makes it different than the other previous iterations in that it uses Batch Normalization to make all the values be normalized/standardized on the basis of the mini-batch. After this, we used the adaptive average pooling. Finally, to make all the layers fully connected then we used linear with **out_features**, i.e., (512, 1024), to make all the layers together.

The Figure 4.2 shows the layer Grad CAM applied on all the individual CNN layers.

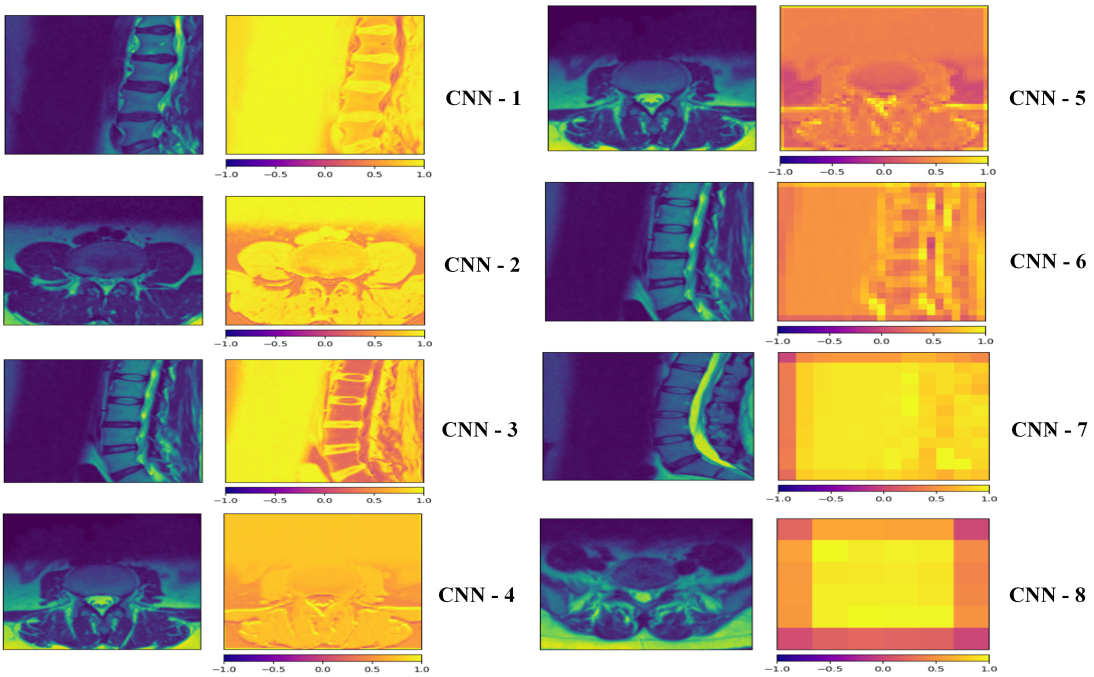


FIGURE 4.2: Conventional Model - (1st Iteration)

4.4.2 2nd Iteration

Figure 4.3 shows the layer Grad Cam after training a model in the 2nd iteration.

This iteration also trained based on the same CNN layers **out_features**, i.e., (8, 16, 32, 64, 128, 256, 512), except we have used **ReLU** in the first layer, but in all the remaining layers we have used **LeakyReLU** as the activation functions. Same Batch normalization, Adaptive Average Pooling, and make all the layers fully connected with linear layers.

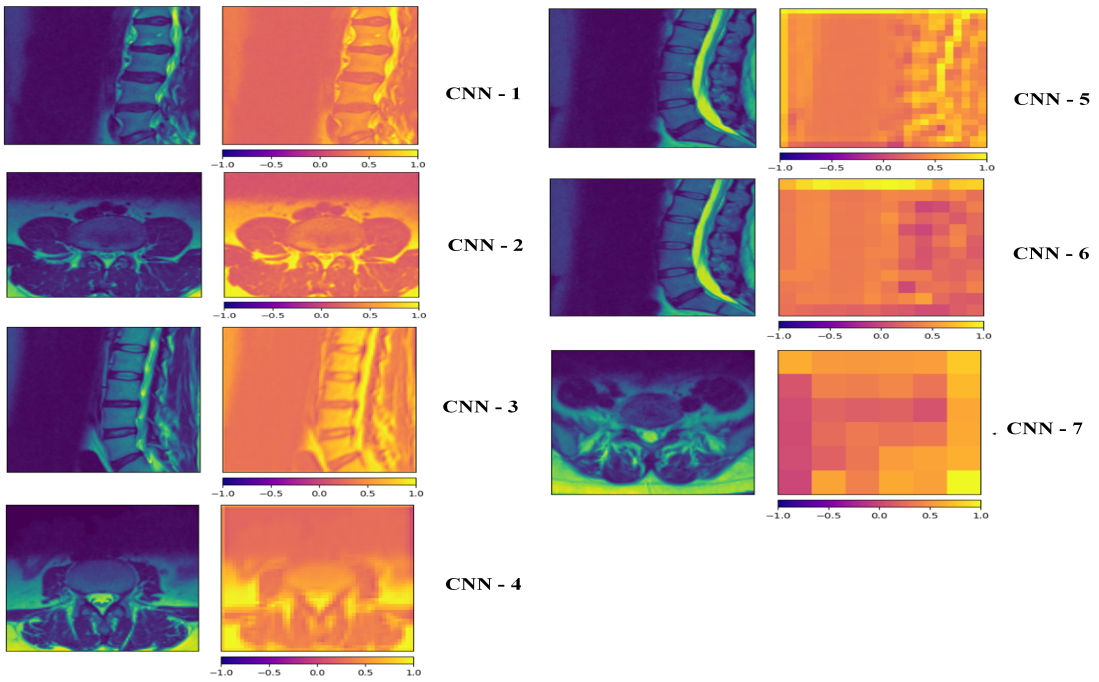


FIGURE 4.3: 2nd Iteration - Layer Grad CAM

4.4.3 3rd Iteration

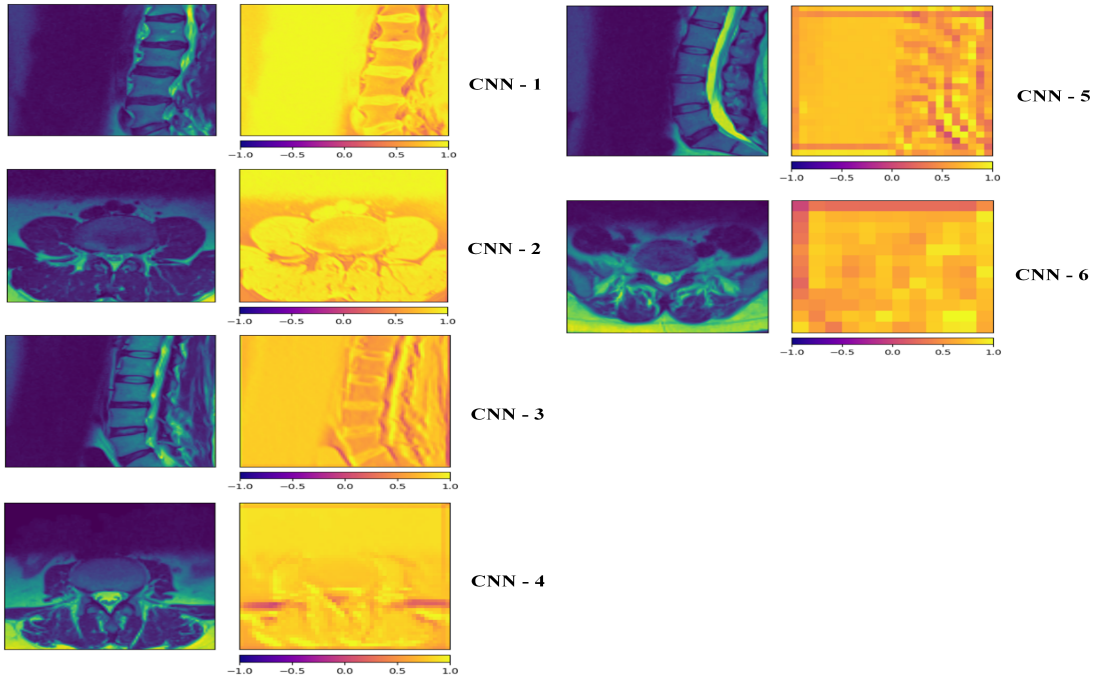


FIGURE 4.4: 3rd Iteration - Layer Grad CAM

In the 3rd iteration, we trained the model using CNN layers **out_features**, i.e., (8, 16, 32, 64, 128, 256), except we have used **ReLU** in the first layer, but in

all the remaining layers we have used **LeakyReLU** as the activation functions. Same Batch normalization, Adaptive Average Pooling, and make all the layers fully connected with linear layers.

Figure 4.4 shows the layer Grad Cam after training a model in the 3rd iteration.

4.4.4 Proposed Model 4th Iteration

The 4th iteration is the last iteration where we have selected it as our proposed model. We have already mentioned how we have trained our proposed model.

Figure 4.5 shows the necessary features that this model has selected so far.

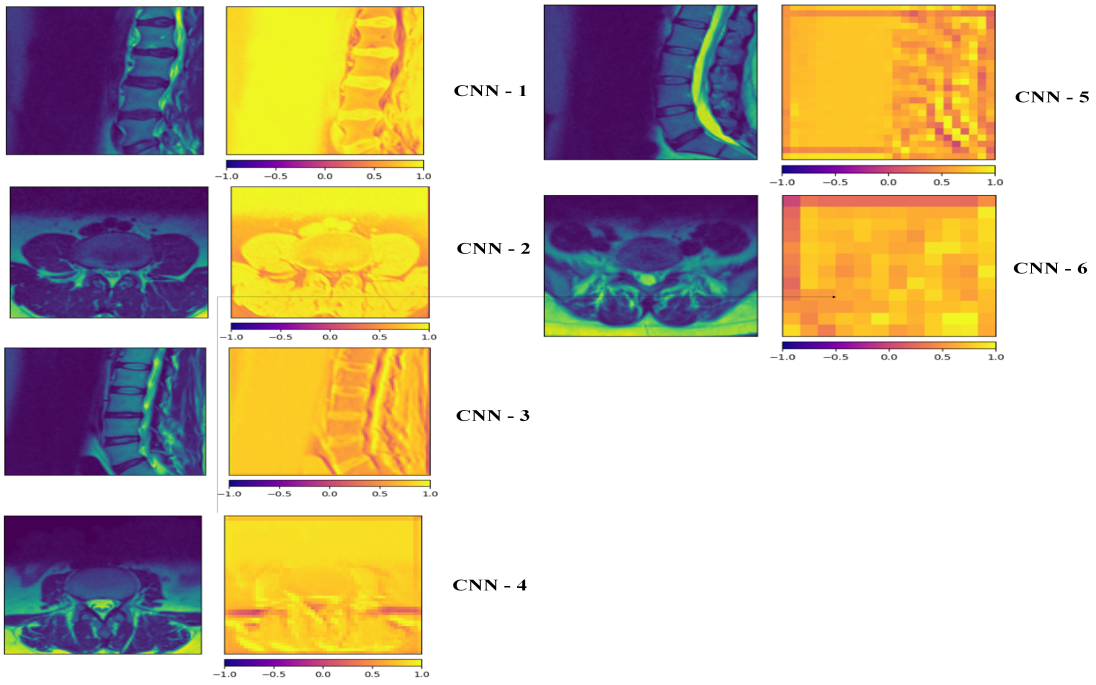


FIGURE 4.5: 4th Iteration - Layer Grad CAM

4.5 Explaining Model

The proposed model is improved and clarified by the interpretation. Occlusion Sensitivity helps us understand our model using XAI. This carefully assesses the model's decision-making abilities by selecting key lumbar spinal MRI regions for stenosis diagnosis. This approach uses input masking and estimations for identif-

ying compression of the nerves. The feature maps highlighted specific characteristics (such as a herniated disc, a thecal sac, and the absence of stenosis) which effect the model's classifying process, ensuring radiologists' significance and scientific accountability in automated spinal assessment.

4.5.1 Results Achieved by Occlusion Sensitivity

Occlusion Sensitivity is an explainable AI (XAI) method that identifies which regions of an input (e.g., an image) a model relies on for predictions. The features are highlighted by the yellow color. According to the figures below, all the figures are divided by original, Positive Attribution, Negative Attribution, Masked, and Blended Masked Image. So, in the Positive Attribution, there are some yellow color spots which indicate the necessary features in the given image. And in Negative Attribution, all the yellow spots indicate the unnecessary features. In the Mask, it takes the most highlighted parts from the/ original image by ignoring the background. Finally, in Blended Masked Image, it combines the Masked image with on the top of Positive Attribution.

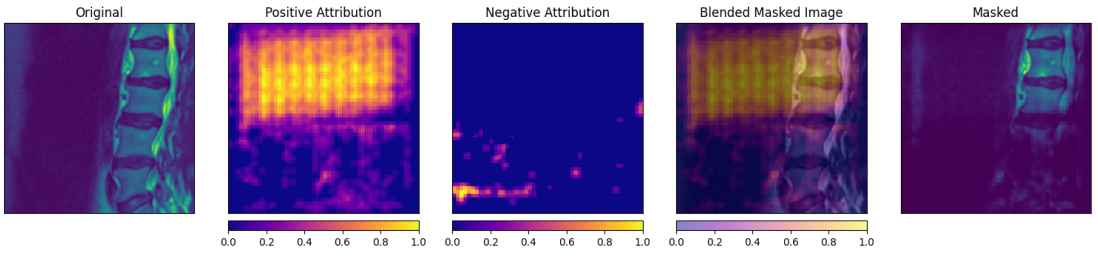


FIGURE 4.6: Herniated Disc 1 - Occlusion Sensitivity

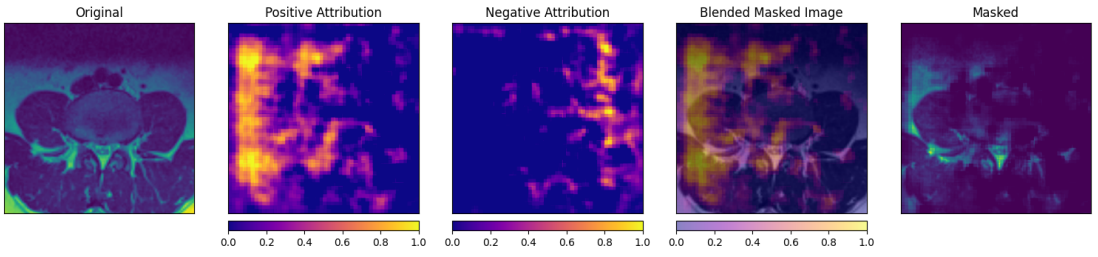


FIGURE 4.7: Herniated Disc 2 - Occlusion Sensitivity

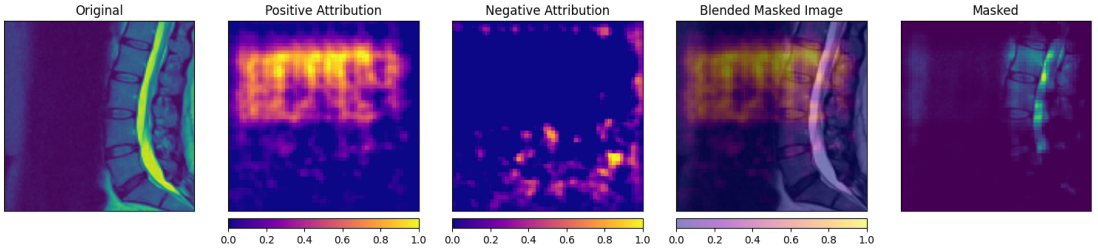


FIGURE 4.8: Thecal Sac 1 - Occlusion Sensitivity

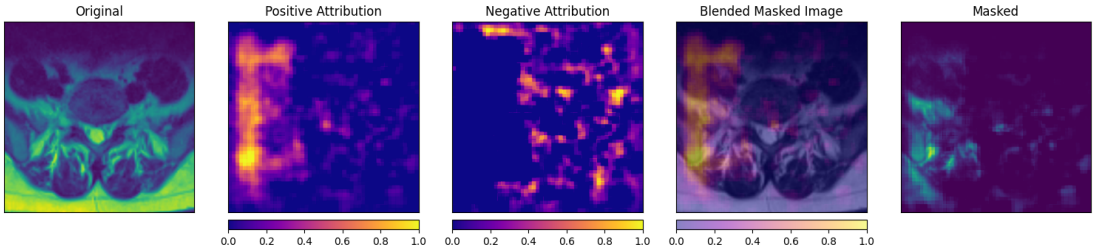


FIGURE 4.9: Thecal Sac 2 - Occlusion Sensitivity

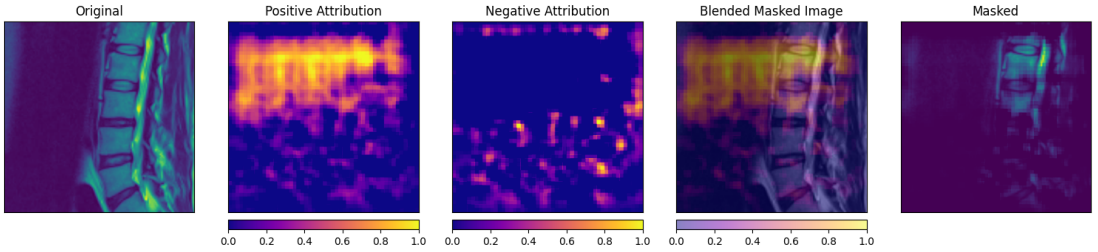


FIGURE 4.10: No Stenosis 1 - Occlusion Sensitivity

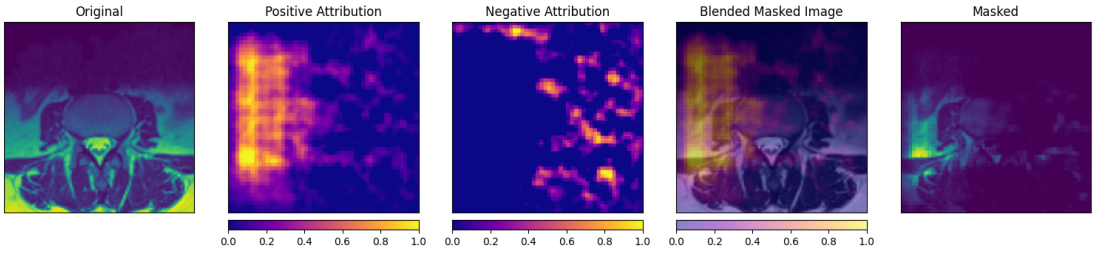


FIGURE 4.11: No Stenosis 2 - Occlusion Sensitivity

4.6 Classification Results Analysis

Table 4.2 compares the performance of our proposed model with the conventional model. The optimal results for our model were achieved in the 4th iteration, with

both accuracy and F1-score reaching 82.75 %. However, performance declined in subsequent iterations. To shows the analysis of our result in graphical form, already mentioned in table 4.2.

TABLE 4.2: Comparison of Multiple Iterations

	Accuracy	**F1-Score**	**Precision**	**Recall**
Conventional Model 1st Iteration	80.21%	80.19%	80.50%	80.20%
2nd Iteration	77.66%	77.70%	77.94%	77.66%
3th Iteration	85.69%	85.69%	85.83%	85.69%
Proposed Model 4th Iteration	89.69%	89.60%	89.63%	89.64%

The graph of model performance of accuracy, F1-Score, Precision, and Recall is given below.

4.7 Results Discussion

The essential features that influence the predictions of the proposed model were explained by the explainability results obtained through Layer Grad-CAM and Occlusion Sensitivity assessments.The XAI-enriched architecture substantially improved the precision of identifying LSS, as evidenced by a comparison assessment of Layer Grad-CAM visuals between the conventional and proposed models. It was also supported through the Occlusion Sensitivity mappings, which showed that the most important predictive variables were mostly linked to troublesome LSS patterns. Regarding classification, the suggested XAI-based model performed significantly better than the first model, achieving an accuracy of 82.75 % on the dataset. However, we were unable to perform further iterations to train a model due to overfitting.

XAI approaches, particularly Layer Grad-CAM, enhanced the model's ability to identify and emphasize LSS by training it to concentrate on the most significant indicators of a problem. The model was able to improve by focusing on radiologically significant locations in the input MRI as a result of XAI's focused approach. The all over implementation code, along with the mentioned results, can be found at Github [2].

[2]"https://github.com/abdullahkhan70/lss-classification-xai"

Model performance improved by removing layers likely due to preventing overfitting, forcing the model to learn more generalized and robust features rather than memorizing noise in the training data. The peak accuracy achieved at the 4th iteration signifies the optimal point of early stopping, where the model has learned the most critical features without beginning to over-specialize. Continuing training beyond this point would likely lead to a decrease in validation accuracy as the model begins to overfit to the training set.

The main limitation of our proposed model is that it is based on three main common classes of stenosis (i.e., herniated disc, thecal sac, and no stenosis). The others have stenosis classes which are not very common (i.e. facet joints, pedicle), still need to be addressed. Furthermore, in the proposed work, LSS patients were not properly categorized in terms of age and gender. Age and gender could be important factors influencing LSS risk in different age groups. For accurate diagnosis of LSS in medical studies, age should be considered along with MRI images, and a CNN model can produce an accurate diagnosis with the consideration of age.

4.8 Comparison with Existing Methods

To further confirm our findings, we compared the performance of our proposed model with many other similar related studies of LSS classes. The comparison is presented in the table 4.3. The proposed model has it shows better performance in all the measures than the other methods. With an accuracy of 89.64 %, precision of 89.63 %, recall of 89.64 %, and F1-score of 89.60 %, it provides more accurate and reliable classification of LSS, highlighting the effectiveness of integrating XAI techniques in CNN architectures.

Table 4.3: Comparison of Existing Techniques with Proposed Method

Ref.	**Technique**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
[41]	XGBoost	81.9%	81.4%	80.4%	79.65%
[48]	CNN	83%	75%	79%	82%
Our Work	**Proposed Model**	**89.64%**	**89.63%**	**89.64%**	**89.60%**

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Our approach uses XAI to categorize lumbar spinal MRI images (e.g., thecal sac, herniated disc, absence of stenosis) by concentrating on essential anatomical characteristics (e.g., spinal canal constriction).

Important results and contributions of our model consist of:

1. Distinguishing genuine contributory stenosis from spinal cord MRIs to enhance the overall efficiency of LSS classification.

2. Attaining an interpretable model that offers transparency in the decision-making process, hence rendering it superior to current non-interpretable methodologies.

3. Surpassing benchmark methodologies by offering transparency and classification grounded around essential stenosis characteristics with a diminished layer count.

Including XAI in CNN architecture marks a significant development since it helps the model make wiser decisions regarding feature emphasis. This improvement

generates quite visible and dependable models and increases classification accuracy. Therefore, it is appropriate to include models in practical systems where openness and trust are vital, since it also increases user confidence, helps debug, and conforms with legal requirements. Together, our dataset preprocessing techniques, Layer Grad-CAM-based model, and application of XAI techniques, including Occlusion Sensitivity for model interpretability, produce a model whose classification decisions on actual LSS make sense.

Applicable to a wider spectrum of medical imaging datasets and circumstances, the approach of the efficiency of the suggested model is utilized to classify LSS based on MRIs without considering the complicated feature extraction methods, thus relieving one of the concerns. Further improving the XAI's fit for a variety of medical image analysis uses is its ability to effortlessly do classifications and provide explainability.

5.2 Future Work

There are several interesting upcoming directions to look into:

1. LSS MRIs dataset, which can distinguish thecal sac from the herniated disc more clearly, improving the model's accuracy. Predict even more in a better way on unseen data across other LSS subtypes.
2. Develop and test out the trained model for the capabilities of radiologists to ensure that it is predicting correctly.
3. Research on different advanced XAI-based methods for further feature analysis and improving the model interpretability.

Bibliography

[1] C. Jin, L. Zhao, J. Wu, L. Jia, L. Cheng, and N. Xie, "Traumatic cervical spinal cord injury: relationship of mri findings to initial neurological impairment," *European Spine Journal*, vol. 30, pp. 3666–3675, 2021.

[2] S. Altun and A. Alkan, "Lss-net: 3-dimensional segmentation of the spinal canal for the diagnosis of lumbar spinal stenosis," *International Journal of Imaging Systems and Technology*, vol. 33, no. 1, pp. 378–388, 2023.

[3] J.-w. Kwon, S.-H. Moon, S.-Y. Park, S.-J. Park, S.-R. Park, K.-S. Suk, H.-S. Kim, and B. H. Lee, "Lumbar spinal stenosis: review update 2022," *Asian Spine Journal*, vol. 16, no. 5, p. 789, 2022.

[4] T. Deer, D. Sayed, J. Michels, Y. Josephson, S. Li, and A. K. Calodney, "A review of lumbar spinal stenosis with intermittent neurogenic claudication: disease and diagnosis," *Pain medicine*, vol. 20, no. Supplement_2, pp. S32–S44, 2019.

[5] S. Ratish, Z.-X. Gao, H. M. Prasad, Z. Pei, D. Bijendra *et al.*, "Percutaneous endoscopic lumbar spine surgery for lumbar disc herniation and lumbar spine stenosis: Emphasizing on clinical outcomes of transforaminal technique," *Surgical Science*, vol. 9, no. 02, p. 63, 2018.

[6] K. Lundon and K. Bolton, "Structure and function of the lumbar intervertebral disk in health, aging, and pathologic conditions," *Journal of orthopaedic & sports physical therapy*, vol. 31, no. 6, pp. 291–306, 2001.

[7] E. Been, A. Barash, H. Pessah, and S. Peleg, "A new look at the geometry of the lumbar spine," *Spine*, vol. 35, no. 20, pp. E1014–E1017, 2010.

[8] A. S. Zhang, A. Xu, K. Ansari, K. Hardacker, G. Anderson, D. Alsoof, and A. H. Daniels, "Lumbar disc herniation: diagnosis and management," *The American journal of medicine*, vol. 136, no. 7, pp. 645–651, 2023.

[9] K. Wang, Q.-Z. Yang, H.-N. Wen, Y.-X. Hai, G.-D. Gao, and M. Song, "Nerve root compression due to lumbar spinal canal tophi: A case report and review of the literature," *Medicine*, vol. 101, no. 45, p. e31562, 2022.

[10] B. Sassack and J. D. Carrier, "Anatomy, back, lumbar spine," in *StatPearls [Internet]*. StatPearls Publishing, 2023.

[11] E. Sartoretti, M. Wyss, A. Alfieri, C. A. Binkert, C. Erne, S. Sartoretti-Schefer, and T. Sartoretti, "Introduction and reproducibility of an updated practical grading system for lumbar foraminal stenosis based on high-resolution mr imaging," *Scientific Reports*, vol. 11, no. 1, p. 12000, 2021.

[12] Y. L. Guen, W. L. Joon, S. C. Hee, O. Kyoung-Jin, and S. K. Heung, "A new grading system of lumbar central canal stenosis on mri: an easy and reliable method," *Skeletal radiology*, vol. 40, pp. 1033–1039, 2011.

[13] M. Minetama, M. Kawakami, M. Teraguchi, S. Matsuo, Y. Enyo, M. Nakagawa, Y. Yamamoto, T. Nakatani, N. Sakon, W. Nagata *et al.*, "Mri grading of spinal stenosis is not associated with the severity of low back pain in patients with lumbar spinal stenosis," *BMC Musculoskeletal Disorders*, vol. 23, no. 1, p. 857, 2022.

[14] A. M. Awadalla, A. S. Aljulayfi, A. R. Alrowaili, H. Souror, F. Alowid, A. M. M. Mahdi, R. Hussain, M. M. Alzahrani, A. N. Alsamarh, E. A. Alkhaldi *et al.*, "Management of lumbar disc herniation: a systematic review," *Cureus*, vol. 15, no. 10, 2023.

[15] F. Watanabe, T. Kojima, M. Miyazu, and H. Kitoh, "Efficacy of spinal ultrasonography just before caudal epidural block for identifying tethered cord syndrome in urological cases with sacral dimples: a retrospective descriptive study," *Journal of Anesthesia*, pp. 1–8, 2025.

[16] V. Tumko, J. Kim, N. Uspenskaia, S. Honig, F. Abel, D. R. Lebl, I. Hotalen, S. Kolisnyk, M. Kochnev, A. Rusakov *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, 2024.

[17] T. Kim, Y.-G. Kim, S. Park, J.-K. Lee, C.-H. Lee, S.-J. Hyun, C. H. Kim, K.-J. Kim, and C. K. Chung, "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, 2022.

[18] B. Gaonkar, D. Villaroman, J. Beckett, C. Ahn, M. Attiah, D. Babayan, J. Villablanca, N. Salamon, A. Bui, and L. Macyszyn, "Quantitative analysis of spinal canal areas in the lumbar spine: an imaging informatics and machine learning study," *American Journal of Neuroradiology*, vol. 40, no. 9, pp. 1586–1591, 2019.

[19] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, 2022.

[20] Y. Wang, "A comparative analysis of model agnostic techniques for explainable artificial intelligence," *Research Reports on Computer Science*, pp. 25–33, 2024.

[21] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 9, no. 4, p. e1312, 2019.

[22] Z. F. Hu, T. Kuflik, I. G. Mocanu, S. Najafian, and A. Shulner Tal, "Recent studies of xai-review," in *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 2021, pp. 421–431.

[23] I. Kök, F. Y. Okay, Ö. Muyanlı, and S. Özdemir, "Explainable artificial intelligence (xai) for internet of things: a survey," *IEEE Internet of Things Journal*, vol. 10, no. 16, pp. 14 764–14 779, 2023.

[24] Z. Li, "Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost," *Computers, Environment and Urban Systems*, vol. 96, p. 101845, 2022.

[25] T.-T.-H. Le, A. T. Prihatno, Y. E. Oktian, H. Kang, and H. Kim, "Exploring local explanation of practical industrial ai applications: A systematic literature review," *Applied Sciences*, vol. 13, no. 9, p. 5809, 2023.

[26] A. L. Alfeo, A. G. Zippo, V. Catrambone, M. G. Cimino, N. Toschi, and G. Valenza, "From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks," *Computer Methods and Programs in Biomedicine*, vol. 236, p. 107550, 2023.

[27] Z. Chen, L. Xiao, F. Guo, and J. Yan, "Interpretable machine learning for building energy management: A state-of-the-art review," *Advances in Applied Energy*, vol. 9, p. 100123, 01 2023.

[28] Y. H. Chel and L. L. Poh, "Brain tumor classification in mri: Insights from lime and grad-cam explainable ai techniques," *IEEE Access*, 2025.

[29] D. Song, J. Yao, Y. Jiang, S. Shi, C. Cui, L. Wang, L. Wang, H. Wu, H. Tian, X. Ye *et al.*, "A new xai framework with feature explainability for tumors decision-making in ultrasound data: comparing with grad-cam," *Computer Methods and Programs in Biomedicine*, vol. 235, p. 107527, 2023.

[30] M. Ennab and H. Mcheick, "Advancing ai interpretability in medical imaging: a comparative analysis of pixel-level interpretability and grad-cam models," *Machine Learning and Knowledge Extraction*, vol. 7, no. 1, p. 12, 2025.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[32] H. T. T. Nguyen, H. Q. Cao, K. V. T. Nguyen, and N. D. K. Pham, "Evaluation of explainable artificial intelligence: Shap, lime, and cam," in *Proceedings of the FPT AI Conference*, 2021, pp. 1–6.

[33] K. Roshan and A. Zafar, "Utilizing xai technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation (shap)," *arXiv preprint arXiv:2112.08442*, 2021.

[34] C. Dewi, R.-C. Chen, H. Yu, and X. JIANG, "Xai for image captioning using shap." *Journal of Information Science & Engineering*, vol. 39, no. 4, 2023.

[35] R. Dakhli and W. Barhoumi, "Toward a quantitative trustworthy evaluation of post-hoc xai feature importance maps using saliency-based occlusion," in *2024 IEEE/ACS 21st International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 2024, pp. 1–8.

[36] S. A. Ali, K. R. Arain, N. A. Mushtaq, and O. ul Rehman, "Interpretable deep learning for brain tumor diagnosis: Occlusion sensitivity-driven explainability in mri classification," *VFAST Transactions on Software Engineering*, vol. 13, no. 2, pp. 135–146, 2025.

[37] G. Bleser[1], E. Bartaguiz[1], J. Kniepert, and P. Drees, "Towards a better understanding of spinal differences between healthy subjects and subjects with back pain using explainable artificial intelligence (xai)," in *Proceedings of the 9th International Performance Analysis Workshop and Conference & 5th IACSS Conference*, vol. 1426. Springer Nature, 2022, p. 97.

[38] M. A. Al-antari, S. Salem, M. Raza, A. S. Elbadawy, E. Bütün, A. A. Aydin, M. Aydoğan, B. Ertuğrul, M. Talo, and Y. H. Gu, "Evaluating ai-powered predictive solutions for mri in lumbar spinal stenosis: a systematic review," *Artificial Intelligence Review*, vol. 58, no. 8, p. 221, 2025.

[39] H. Suzuki, T. Kokabu, K. Yamada, Y. Ishikawa, A. Yabu, Y. Yanagihashi, T. Hyakumachi, H. Tachi, T. Shimizu, T. Endo *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, 2024.

[40] C. Dindorf, J. Konradi, C. Wolf, B. Taetz, G. Bleser, J. Huthwelker, F. Werthmann, E. Bartaguiz, J. Kniepert, P. Drees *et al.*, "Classification and automated interpretation of spinal posture data using a pathology-independent classifier and explainable artificial intelligence (xai)," *Sensors*, vol. 21, no. 18, p. 6323, 2021.

[41] T. Li, W. Qi, X. Mao, G. Jia, W. Zhang, X. Li, H. Pan, and D. Wang, "Prediction of lumbar disc degeneration based on interpretable machine learning models: Retrospective cohort study," *The Spine Journal*, 2025.

[42] P. Wang, L. Liu, Z. Xie, G. Ren, Y. Hu, M. Shen, H. Wang, J. Wang, Y. Wang, and X.-T. Wu, "Explainable machine learning models for prediction of surgical site infection after posterior lumbar fusion surgery based on shap," *World Neurosurgery*, p. 123942, 2025.

[43] J.-H. Kim, S.-E. Lee, H.-S. Jung, B.-S. Shim, J.-U. Hou, and Y.-S. Kwon, "Development and validation of deep learning-based algorithms for predicting lumbar herniated nucleus pulposus using lumbar x-rays," *Journal of Personalized Medicine*, vol. 12, no. 5, p. 767, 2022.

[44] S. Mani, S. Thakar, and R. S. Rachakonda, "An explainable machine learning framework for prediction of recurrent lumbar disc herniation," in *International Conference on Computing and Machine Learning*. Springer, 2024, pp. 311–322.

[45] A. De Barros, F. Abel, S. Kolisnyk, G. C. Geraci, F. Hill, M. Engrav, S. Samavedi, O. Suldina, J. Kim, A. Rusakov *et al.*, "Determining prior authorization approval for lumbar stenosis surgery with machine learning," *Global Spine Journal*, vol. 14, no. 6, pp. 1753–1759, 2024.

[46] M. W. Tong, K. Ziegeler, V. Kreutzinger, and S. Majumdar, "Explainable ai reveals tissue pathology and psychosocial drivers of opioid prescription for non-specific chronic low back pain," *Scientific Reports*, vol. 15, no. 1, p. 30690, 2025.

[47] P. Yasin, H. Luan, C. Peng, and X. Song, "Development and validation of an interpretable nomogram for predicting the risk of the prolonged postoperative length of stay for tuberculous spondylitis: a novel approach for risk stratification," *BMC Musculoskeletal Disorders*, vol. 26, no. 1, p. 539, 2025.

[48] S. Lee, J.-Y. Jung, A. Mahatthanatrakul, and J.-S. Kim, "Artificial intelligence in spinal imaging and patient care: a review of recent advances," *Neurospine*, vol. 21, no. 2, p. 474, 2024.

[49] "Lumbar Spine MRI Dataset — data.mendeley.com," https://data.mendeley.com/datasets/k57fr854j2/2, [Accessed 14-06-2025].

[50] A. R. Beeravolu, S. Azam, M. Jonkman, B. Shanmugam, K. Kannoorpatti, and A. Anwar, "Preprocessing of breast cancer images to create datasets for deep-cnn," *IEEE Access*, vol. 9, pp. 33 438–33 463, 2021.

[51] R. Muthu, C. Rani, S. Saritha, and S. Pearl Mary, "Morphological operations in medical image pre-processing," in *International conference on advanced computing and communication systems*, 2017, pp. 2065–2070.

[52] X. Zhang, Q. Chen, R. Ng, and V. Koltun, "Zoom to learn, learn to zoom," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3762–3770.

[53] C. Thavamani, M. Li, F. Ferroni, and D. Ramanan, "Learning to zoom and unzoom," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5086–5095.

[54] J. Rama, C. Nalini, and A. Kumaravel, "Image pre-processing: enhance the performance of medical image classification using various data augmentation technique," *ACCENTS Transactions on Image Processing and Computer Vision*, vol. 5, no. 14, p. 7, 2019.

[55] S. Degadwala, V. N. D. Krishnamurthy, and D. Vyas, "Deepspine: Multi-class spine x-ray conditions classification using deep learning," in *2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL)*. IEEE, 2024, pp. 8–13.

[56] B. Badada, G. Delina, R. Krishnaraj, and M. M. Thiruthuvanathan, "Application of xai in integrating democratic and servant leadership to enhance the performance of manufacturing industries in ethiopia," in *International Conference on Innovations in Computational Intelligence and Computer Vision*. Springer, 2024, pp. 123–138.

[57] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of imaging*, vol. 6, no. 6, p. 52, 2020.

[58] R. Pugalenthi, M. Rajakumar, J. Ramya, and V. Rajinikanth, "Evaluation and classification of the brain tumor mri using machine learning technique," *Journal of Control Engineering and Applied Informatics*, vol. 21, no. 4, pp. 12–21, 2019.

[59] A. W. Salehi, S. Khan, G. Gupta, B. I. Alabduallah, A. Almjally, H. Alsolai, T. Siddiqui, and A. Mellit, "A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope," *Sustainability*, vol. 15, no. 7, p. 5930, 2023.

[60] X. Dong, C. J. Taylor, and T. F. Cootes, "Defect classification and detection using a multitask deep one-class cnn," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 3, pp. 1719–1730, 2021.

[61] Y. Bai, "Relu-function and derived function review," in *SHS web of conferences*, vol. 144. EDP Sciences, 2022, p. 02006.

[62] M. S. Job, P. H. Bhateja, M. Gupta, K. Bingi, and B. R. Prusty, "Fractional rectified linear unit activation function and its variants," *Mathematical Problems in Engineering*, vol. 2022, no. 1, p. 1860779, 2022.

[63] J. Brownlee, "A gentle introduction to pooling layers for convolutional neural networks," 2019. [Online]. Available: https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/

[64] R. Nirthika, S. Manivannan, A. Ramanan, and R. Wang, "Pooling in convolutional neural networks for medical image analysis: a survey and an empirical

study," *Neural Computing and Applications*, vol. 34, no. 7, pp. 5321–5347, 2022.

[65] A. Stergiou and R. Poppe, "Adapool: Exponential adaptive pooling for information-retaining downsampling," *IEEE Transactions on Image Processing*, vol. 32, pp. 251–266, 2022.

[66] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," *International journal of computer vision*, vol. 128, pp. 336–359, 2020.

[67] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," 2013. [Online]. Available: https://arxiv.org/abs/1311.2901

[68] M. Aminu, N. A. Ahmad, and M. H. M. Noor, "Covid-19 detection via deep neural network and occlusion sensitivity maps," *Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4829–4855, 2021.