Teesside University

School of Computing, Engineering
& Digital Technologies

School of Computing, Engineering and Digital Technologies
Department of Computing and Games
Teesside University
Middlesbrough TS1 3BA

Enhancing Axial View of Lumbar Spine MR Image Diagnosis with AI: A Study on Clinical Significance Classification and Serious Disc Bulge Detection Using CNNs and Image Explainer

Submitted in partial requirements for the degree of *MSc Applied Artificial Intelligence*

Date: *8th of May, 2024*

Name: *Chukwurah Paul*

Student Number: *B1267718*

Supervisor: *Al kafri Ala*

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to God for giving me the strength and courage to complete this research.

I would like to say thank you to my supervisor, Al kafri Ala, whose advice, guidance, and support I had throughout the research process were of remarkable importance to my success in this project. Your expertise and insights have been instrumental in shaping this project.

My heartful thanks to my family, specially my parents, brothers, aunties etc. for their unwavering love, encouragement, and financial support provided throughout this academic journey I wouldn't have done it without you all.

I am immensely grateful to all my friends particularly Adeboye Adeoluwa, Nwanguma Chibuzor, Otensaya Oluwafeola, Pedrosa Andrea, Ezeh Jidechukwu for their steadfast support, care, love and encouragement throughout this journey. I love you guys.

Lastly, I would like to thank everyone who took part in this study or contributed helpful thoughts. Your efforts were critical in achieving relevant outcomes and conclusion.

ABSTRACT

This study investigates the potential of AI technology to improve the diagnosis of lumbar spine MR images in detecting serious disc bulge and classifying clinical significance of MRI interpretations by employing Convolutional Neural Networks (CNNs) and Image Explainer techniques. Disc bulge occurs when the inner component of the intervertebral disc protrudes from its outer wall and progresses over time which can lead to additional disc degeneration problems such as spinal stenosis. Serious bulges on the disc can put pressure on the surrounding nerve roots, causing pain to travel down the back and other parts of the body. The pressure on the nerves created by these bulges can put pressure on the sciatic nerve which can lead to sciatica. The dataset used comprises of 515 patients who reported lower back pain. It included the last 3 lumbar spine disc L3-L4, L4-L5, L5-S1 for each of the patient and a label was supplied by the clinicians about conditions experienced by each disc. This was used to train both models, the clinical significance classification ranges from no, mild and serious clinical significance. The model developed achieved an accuracy of 81%. The second model (serious disc bulge detection) was used to classify the lumbar disc into two categories; no serious disc bulge and serious disc bulge. The model achieved an accuracy of 89%. Local Interpretabile Model-Agnostic Explanations (LIME) was applied to both models to explain the model's decision and hence eliminate the black box problem of models. A GUI was developed to enhance clinicians interaction with the models. The findings shed light on how advanced computational methods can enhance medical imaging interpretation and potentially revolutionize the diagnosis and treatment of lumbar spine conditions.

Keywords: Deep Learning, Artificial Intelligence, Convolutional Nueral Networks, Disc Bulge, Disc Protrusion.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
CAD	Computer-aided diagnosis
CNN	Convolutional Neural Network
GUI	Graphic user interface
IVD	Intervertebral disc
LBP	Lower Back Pain
LIME	Local Interpretabile Model-Agnostic Explanations
LSS	Lumbar Spinal Stenosis
MAE	Mean absolute error
ML	Machine Learning
MR	Medical Resonance
MRI	Magnetic Resonance Imaging
XAI	Explainable Artificial Intelligence

1.0 INTRODUCTION

Magnetic resonance imaging (MRI) serves as a cornerstone in diagnosing lumbar disorders, offering high-quality images without ionizing radiation. Among the common lumbar intervertebral disc (IVD) injuries detected through MRI are disc bulge etc., which frequently manifest as low back pain and leg numbness (D'Antoni et al., 2022). Lower Back Pain (LBP) stands as a primary cause of global disability, imposing significant socioeconomic burdens. Its diagnosis and treatment requires a multidisciplinary, customised approach involving numerous outcome indicators, imaging data, and developing technologies. The increasing data generated throughout this process has accelerated the development of artificial intelligence (AI) methods, including computer-aided diagnosis (CAD), to augment clinical decision-making and improve patient care (D'Antoni et al., 2022).

This study focuses on detecting serious disc bulges and classifying clinical significance of MRI interpretation. Disc bulge occurs when the inner component of the intervertebral disc protrudes from its outer wall. Serious bulges can put pressure on the surrounding nerve roots, causing pain to travel down the back and other parts of the body depending on their position in the spinal column. The pressure on the nerves created by these bulges can put pressure on the sciatic nerve which can lead to sciatica, which causes leg pain and possibly tingling, numbness, and weakness that develops in the lower back and travels through the buttock and down the big sciatic nerve in the back of the leg (MC, et al., 1994). This study discusses serious disc bulge, which involves degenerated disc, protrusion, focal or asymmetric expansion of the disc beyond the interspace, and extrusion as there are no significant differences between them (Milette, et al., 1999).

This project uses AI, primarily convolutional neural networks (CNNs), to create a robust model capable of detecting serious disc bulges on axial MRI images and determining the clinical significance classification of MRI interpretation. CNNs can recognize complex patterns and characteristics inside images using big datasets and deep learning approaches, allowing them to generate accurate predictions and assist physicians with diagnosis (Elshazly & Marrero, 2020). By automating the diagnosis of serious disc bulges, AI models can help radiologists and doctors uncover subtle anomalies that could otherwise be overlooked or misconstrued. This, in turn, can result in early detection of more precise diagnoses and better patient outcomes.

In addition to developing AI models, this study underlines the significance of interpretability in AI-driven medical diagnostics. Clinicians frequently struggle to interpret AI-generated data due to the black-box nature of deep learning algorithms. As a result, this study will use image explainer approaches to provide clear insights into the AI's decision-making process in a bid to increase clinicians trust and knowledge by explaining why AI-generated predictions are made, hence facilitating AI inclusion into clinical practice.

1.1 Aims and Objectives of this study:

The aims and Objectives of this study includes:

1. Develop AI models capable of identifying serious disc bulges in the lumbar spine and categorizing MRI into different clinical significance; no, mild, and serious clinical significance.
2. Investigate the potential of convolutional neural networks (CNNs) for reliable detection of serious disc bulges and classification of lumbar spine clinical significance.
3. Design a graphical user interface (GUI) for the AI models to improve the user experience of clinicians, thereby enhancing patient care.
4. Implement image explainer techniques to provide interpretable insights into the decision-making process of the AI models, aiming to enhance trust and understanding among clinicians.
5. Contribute to the advancement of AI-driven diagnostic tools to improve patient outcomes in the diagnosis and management of lumbar spine disorders.

2.0 LITERTURE REVIEW

2.1 Exploring Pre-AI Methods for Lumbar Spine Diagnosis

Before the development of AI, the diagnosis of lumbar spine relied on a combination of traditional methods including patient history, physical examination and imaging modalities.

A two-level probabilistic model of lumbar disc localisation using clinical MRI data was put forth by Alomari et al. (2011). They identified possible changes in the spine anatomy by using both pixel-level structures and object-level characteristics. The model efficiently localizes discs by combining appearance and spatial information with generalized expectation-maximization optimization. When tested on a dataset of 105 MRI patients, the model produced encouraging results in recognizing normal and diseased lumbar disc. This technique improves efficiency while remaining robust, with the potential to improve backbone anatomical structure detection and labeling in clinical practice (Alomari & Jason J. and Chaudhary, 2011).

Hancock et al. (2011) investigated the diagnostic accuracy of neurological examinations in determining the level of disc herniation in sciatica patients. Analyzing 283 individuals with verified disc herniation, they discovered that individual neurological tests were not particularly accurate, with area under the curve (AUC) values less than 0.75. Multiple test results indicated marginally improved accuracy but lacked sensitivity and specificity. Dermatomal pain location proved to be the most informative individual test. However, a neurologist's overall assessment following a thorough examination was reasonably accurate, with AUC values of 0.79 and 0.80 for L4/5 and L5/S1 herniations, respectively (Hancock, et al., 2011). Coster, de Bruijn, and Tavy, 2009 evaluated the diagnostic value of history, physical examination, and needle EMG in predicting nerve root compression on MRI in patients with suspected lumbosacral radicular syndrome (LSRS). Analyzing 202 patients, they found that dermatomal radiation, increased pain on coughing, sneezing, or straining, positive straight leg raising, and ongoing denervation on EMG were significant predictors of radiological nerve root compression. Specifically, ongoing denervation on EMG showed the highest odds ratio (OR 4.5). Additionally, 7% of patients with ongoing denervation on EMG did not exhibit radiological nerve root compression, suggesting potential utility of EMG in such cases. (Coster & Tavy, 2009). Tomkins-Lane et al. (2016) employed a Delphi survey to reach an international consensus on the clinical diagnosis of lumbar spinal stenosis (LSS). They identified ten history items related to LSS diagnosis using a multi-phase procedure including specialist clinicians. The study found that clinicians could make an 80% accurate diagnosis with just six questions, including symptoms such as leg or buttock pain during walking and alleviation with forward flexion. This consensus-based set of "seven history items" serves as a practical criterion for defining LSS in both clinical and research contexts, potentially resulting in more cost-effective therapy and better patient outcomes.

Authors	Alomari *et al.* 2011	Hancock *et al.* (2011)	Coster *et al.* 2009	Tomkins-Lane *et al.* (2016)
Study	Labeling of Lumbar Discs Using Both Pixel- and Object-Level Features With a Two-Level Probabilistic Model	Diagnostic Accuracy of the Clinical Examination in Identifying the Level of Herniation in Patients with Sciatica	Diagnostic value of history, physical examination and needle electromyography in diagnosing lumbosacral radiculopathy	Consensus on the Clinical Diagnosis of Lumbar Spinal Stenosis Results of an International Delphi Study
Methodology	Proposed a two-level probabilistic model for disc localization in MRI data, capturing pixel- and object-level features.	Index tests consisting of a neurologist's overall assessment of the severity of disc herniation, individual neurological tests and multiple test results.	Bivariate and multivariate logistic regression analyses were conducted on 202 cases to identify predictors of radiological nerve root compression	Delphi method, survey, consensus meetings
Key Findings/Results	Tested model on on 105 lumbar MRI dataset.	Neurological tests had a fair accuracy (AUC < 0.75) for	47% of patients had radiological nerve root	Identified six top-ranking

	Achieved promising results in detecting normal and abnormal cases.	identifying disc herniation level; multiple test findings was slightly more accurate but with low sensitivity and specificity.	compression.While 7% had ongoing denervation on EMG without radiological nerve root compression.	history items for diagnosing lumbar spinal stenosis.
Conclusion	Two-level model efficiently localizes discs, enhancing efficiency and maintaining robustness in spine abnormality detection	A neurologist's overall assessment was relatively accurate in determining the degree of disc herniation.	More pain on sneezing, coughing, denervation on EMG can be used to predict nerve root compression on MRI.	Clinicians are 80% certain of diagnosing lumbar spinal stenosis within six questions.

Table 1: Comparism of various literatures on pre-AI method on lumbar spine diagnosis

2.2 AI diagnosis of Lumbar Spine

Lehnen et al. (2021) evaluated a convolutional neural network (CNN) trained on several MR imaging characteristics of the lumbar spine for detecting degenerative alterations. They examined 146 patients' lumbar spine MRIs with CNN to identify vertebrae, discs, and diseases such as disc herniation, bulging, spinal stenosis, nerve root compression, and spondylolisthesis. The CNN obtained complete accuracy in disc recognition and labeling, as well as moderate to high accuracy in disc herniations, extrusions, bulgings, stenoses, nerve compressions, and spondylolisthesis. The work demonstrates that employing a single comprehensive CNN to automatically diagnose numerous lumbar spine degenerative alterations is feasible, with good diagnostic accuracy for clinically important findings. (NC, et al., 2021). Al-Kafri *et al.* (2019) offer a method for doctors to detect lumbar spinal stenosis by semantic segmentation and delineation of lumbar spine MRI scans using deep learning. Their dataset consists of 515 MRI examinations of patients with symptomatic back pain that were annotated by professional radiologists. They created a ground truth dataset to train and test segmentation methods, as well as innovative measures to evaluate dataset quality. The authors tested SegNet for semantic segmentation and evaluated the outcomes using contour and region-based metrics. They discovered that their methodology produced extremely good results, comparable to manual labeling, and had great interrater agreement. Representative delineation results demonstrated accuracy appropriate for computer-aided diagnosis (Sudirman, et al., 2019). In 2019, Watanabe *et al.* created a scoliosis screening system that uses moiré topography to determine spinal alignment, Cobb angle, and vertebral rotation. A convolutional neural network (CNN) was used to predict the locations of the thoracic and lumbar vertebrae, spinous processes, and vertebral rotation angles. The technique produced a mean absolute error (MAE) of 3.6 pixels (~5.4 mm) per individual for vertebral locations, with T1 and L5 showing lower errors. The mean absolute error (MAE) between doctors' and predicted Cobb angles was 3.42°, with lesser mistakes reported in deformed spines. The MAE for spinal rotation angle was 2.9°±1.4°, which decreased with lesser abnormalities. (K, et al., 2019). Varçin et al. (2019) examined the use of artificial neural networks, specifically AlexNet and GoogleLeNet, to diagnose lumbar spondylolisthesis using X-ray images. Their dataset contained 272 photos, 136 of which depicted patients with spondylolisthesis and the remaining 136 without. GoogleLeNet has an accuracy of 93.87%, somewhat higher than AlexNet's 91.67% (Varçin, et al., 2019). Kim et al. developed convolutional neural networks (CNNs) for diagnosing severe central lumbar spinal stenosis (LSS) using radiography, and they evaluated radiological diagnostic features using gradient-weighted class activation mapping. Based on formal MRI reports, participants were divided into two groups: severe central LSS and healthy control, and radiographs were taken for both. A CNN-based transfer learning system was used to determine if radiographic findings were LSS or normal. The VGG19 model was the most accurate (82.8%), with an area under the receiver operating characteristic curve (AUROC) of 90.0%. Grad-CAM detected characteristics such as reduced disc height, limited foramina, short pedicle, and hyperdense facet joint (Tackeun Kim MD, et al., 2022).

Authors	Watanabe et al. (2019)	Varçin et al. (2019)	Lehnen et al. (2021)	Al-Kafri et al. (2019)
Study	An Application of Artificial Intelligence to Diagnostic Imaging of Spine Disease: Estimating Spinal Alignment From Moiré Images.	Diagnosis of Lumbar Spondylolisthesis via Convolutional Neural Networks.	Detection of Degenerative Changes on MR Images of the Lumbar Spine with a Convolutional Neural Network.	Boundary Delineation of MRI Images for Lumbar Spinal Stenosis Detection Through Semantic Segmentation Using Deep Neural Networks.
Methodology	Use of CNN to predict the locations of the thoracic and lumbar vertebrae, spinous processes, and	Use of artificial neural networks, (AlexNet and GoogleLeNet) to diagnose lumbar spondylolisthesis using X-ray images.	Use of CNN for detecting degenerative changes with lumbar spine.	Employed CNN for semantic segmentation and boundary delineation of lumbar spinal stenosis MR-Images.

	vertebral rotation angles			
Key Findings/Results	The technique produced a mean absolute error (MAE) of 3.6 pixels (~5.4 mm) per individual for vertebral locations, with T1 and L5 showing lower errors	GoogleLeNet achieved a higher accuracy (93.87%) when compared to AlexNet's (91.67%).	The CNN obtained complete accuracy in disc recognition and labeling, as well as moderate to high accuracy in disc herniations, extrusions, bulgings, stenoses, nerve compressions, and spondylolisthesis.	SegNet for semantic segmentation yielded high accuracy comparable to manual labeling.
Conclusion	Automated estimation of spinal alignment with a low mean absolute error.	Computer-assisted systems can be used in diagnosing of lumber spondylothesis.	The study shows the feasibility of using a single comprehensive CNN model to support the diagnosis of several lumbar degenerative changes with moderate to high diagnostic accuracy.	Results demonstrated accuracy appropriate for computer-aided diagnosis.

Table 2: Comparism of various Literatures on Lumbar Spine Diagnosis

2.3 Related Works on Disc Bulge Diagnosis

Milette et al. (1999) used MRI findings and discographic data to identify between lumbar disc protrusions, bulges, and discs with aberrant signal intensity. They discovered that loss of intervertebral height or an aberrant signal on MRI implies disc disturbances extending to or beyond the outer anulus, which is frequently linked with pain during disc injection. There were no significant differences seen between protrusions, bulges, and discs with normal contours but aberrant signals of degeneration or discogenic discomfort (Milette, et al., 1999). Ma (2015) proposed a novel pathological classification of lumbar disc protrusion, dividing it into four kinds based on intraoperative findings. The "damage-herniation" kind, which is most likely caused by injury, has soft herniation and easily detachable disc tissue, indicating that minimally invasive endoscopic surgery is likely to be successful. The "degeneration-protrusion" kind, which is distinguished by hard protrusions and degenerative alterations, may necessitate nerve decompression and posterior wall removal, whereas disc excision may not be required. The "posterior vertebral osteochondrosis with disc protrusion" type is characterized by vertebral deformities and osteochondral nodules, which require the excision of herniated disc tissue and partially projecting nodules (Ma, 2015). Pan et al. (2021) suggested an automated approach for identifying lumbar disk bulge and herniation using deep convolutional neural networks (CNNs) on magnetic resonance (MR) images. The study's goal is to simplify the interpretation of MR images, lowering radiologists' burden. The system locates vertebral bodies and disks with 100% accuracy in cross-validation. It diagnoses axial lumbar disk MR images as normal, bulging, or herniation, with accuracies ranging from 84.2% to 92.7% for various intervertebral disk levels. The created system improves diagnostic efficiency, standardizes reports, and has potential uses beyond disk problems, such as recognizing lumbar anomalies and cervical spondylosis (Q, et al., 2021). Kieffer et al. (1982) employed amipaque myelography to differentiate between diffusely bulging lumbar intervertebral disks and herniated disks that cause nerve root compression. They distinguish between two types of disks: bulging disks, which have rounded, symmetrical deformities that do not extend beyond the disk space, and herniated disks, which have angular deformities that extend over or below the disk level and expanded impacted nerve roots. In a study of 33 patients undergoing laminectomy, these criteria accurately identified all six bulging disks and 96% of disk herniations (SA, et al., 1982).

Authors	**Milette et al. (1999)**	**Ma (2015)**	**Pan et al. (2021)**	**Kieffer et al. (1982)**
Study	Differentiating Lumbar Disc Protrusions, Disc Bulges, and Discs with Normal Contour but Abnormal Signal Intensity.	A New Pathological Classification of Lumbar Disc Protrusion and Its Clinical Significance.	Automatically Diagnosing Disk Bulge and Disk Herniation with Lumbar Magnetic Resonance Images by Using Deep Convolutional Neural Networks: Method Development Study.	Bulging lumbar intervertebral disk: myelographic differentiation from herniated disk with nerve root compression.
Methodology	Discography correlations.	Pathological analysis, intraoperative findings.	Deep convolutional neural networks, image analysis, classification.	Amipaque myelography, differentiating criteria based on extradural deformity and nerve root properties.
Key Findings/Results	All 23 protrusions (100%) and 12 of 15-disc bulges (80%) were linked to Stage 2 or 3 annular abnormalities.	Proposed classification system for lumbar disc protrusion.	Development of an automated technique for identifying disk bulge and herniation with excellent accuracy. Achieved an average accuray of 88%.	Accurate differentiation between bulging and herniated disks using myelography.
Conclusions	Loss in disc height or unusual signal intensity is particularly predictive of	Improved understanding of lumbar disc protrusion,	The automatic diagnosis method developed could classify images of normal disks, disk	Help patients with low back pain and suspected disk pathology make

	symptomatic rips extending into or beyond the outer anulus.	with potential for better diagnosis and therapy.	bulges, and disk herniation.	surgical decisions and arrange their treatments.

Table 3: Comparism of literatures on related works on serious disc bulge

The table above provides a comparative overview of the key aspects of the literatures reviewed in this *section*, including their titles, publication years, main focuses, methodologies, key findings/results, and implications/conclusions.

2.4 Interpretability of Deep Learning Models

AI models have been able to achieve human-like performance in healthcare, but their application still remains low because they are viewed as a black box. As a result, explainable artificial intelligence (XAI) was presented as a technique to provide confidence in model's forecast by describing how the prediction is generated. Hui et al. (2022) conducted a systematic review to investigate the potential use of explainable artificial intelligence (XAI) in healthcare. They found 99 high-quality articles on various XAI approaches, including SHAP, LIME, GradCAM, and rule-based systems. The review underlined the need for more attention to detecting irregularities in 1D biosignals and finding critical clinical text. By resolving these shortcomings, the XAI research community may boost trust in AI models and encourage their incorporation into healthcare systems (Loh, et al., 2022). Dindorf et al. (2021) used spinal posture data to create a pathology-independent classifier that predicted probabilities and explained classification decisions. They used a one-class support vector machine and Platt's approach to convert outputs to probability distributions. Local Interpretable Model-Agnostic Explanations (LIME) made interpretation easier. The approach successfully categorized participants with spinal fusion but struggled with those with back pain. The performance was comparable to that of binary classifiers (Dindorf, et al., 2021). Jeong-Woon et al. 2023 created a deep learning model to estimate bone mineral density (BMD) from CT scans, which had a correlation coefficient of.90 with DXA-measured BMD. With a maximum F1 score of.875 in abnormal/normal categorization, the model shows potential as an auxiliary tool in clinical practice. Explainable AI approaches indicated that the network concentrated on tissues surrounding the vertebral foramen (Kang Jeong-Woon, 2023). In 2022, Otaki et al. created a deep learning model called CAD-DL that outperformed conventional techniques in the detection of obstructive coronary artery disease (CAD) from SPECT myocardial perfusion imaging (MPI). When CAD-DL is integrated with clinical software, doctors can receive quick results and explanations of predictions. When tested, CAD-DL outperformed quantitative analysis and visual reading in terms of diagnostic accuracy (AUC: 0.83). This work demonstrates how explainable AI may be used in clinical settings to improve diagnostic confidence in the diagnosis of CAD after MPI (Yuka Otaki, et al., 2022).

Authors	Kang et al. (2022)	Dindorf et al. (2021)	Otaki et al. (2022)	Hui Wen Loh et al. (2022)
Study	Prediction of bone mineral density in CT using deep learning with explainability	Classification and Automated Interpretation of Spinal Posture Data Using a Pathology-Independent Classifier and Explainable Artificial Intelligence (XAI)	Clinical Deployment of Explainable Artificial Intelligence of SPECT for Diagnosis of Coronary Artery Disease	Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)
Methodology	Created a deep learning model to estimate bone mineral density, used XAI for model interpretation	Used pathology independent classifier and XAI to interpret spinal posture data	Created deep learning model for obstructive CAD detection from SPECT MPI	Compared journal databases for XAI articles, focus was on Q1 journals
Key Findings/Results	Correlation coefficient between bone mineral density estimates and DXA BMD: 0.90; achieved F1 score of 0.875 for diagnostic test; XAI revealed	Model is useful for interpretation of spinal posture data	CAD-DL model outperformed quantitative analysis and clinical reading. AUC	Found areas om healthcare needing more XAI research. Surveyed 99 XAI articles, focusing on SHAP, LIME, GradCAM and others

	network's focus on local areas around vertebral foramen		of 0.83 in testing.	
Conclusion	Model was found suitable as auxiliary tool in clinical practice	Beneficial for interpretation of predictions	CAD-DL improved CAD diagnosis, XAI facilitates AI acceptance in clinical practice	Detected abnormalities in 1D biosignals and key clinical text areas needing more XAI research

Table 4: Comparism of literatures on Interpretability of deep learning models

2.5 Research Gaps Identified

1. Milette *et al.,* (1999) used discography correlations on MRI to identify and differentiate between lumbar disc protrusions, bulges, and discs with aberrant signal intensity. His model achived an accuracy of 80% for detecting disc bulge which gives room for improvement. Pan *et al.,* (2021) also created a deep learning model using CNN to automatically diagnose disk bulge in his study. His model achieved an average accuracy of 88%. Furthermore, these study only concentrated their effort on identify disc bulge which in some cases might not be a cause for alarm as every human walks around with a form of bulge in their spine and not every bulge means that a patient could be experiencing lower back pains. However serious disc bulge resulting in the compression of the thecal sac and nerve root are more important indicators of serious LBP (Michael & Peter, 2006) which is what this study looked into.
2. Despite the promising results achieved by models developed in similar studies, there is still a lack of studies that have applied explainable AI techniques to disc bulge detection. Most studies to date have focused on the development of deep learning models and have not looked into the possibilities of explainable AI to improve their interpretability and reliability. This research gap highlights the need for further studies that explore the application of explainable AI in serious disc bulge detection and its possible impact on clinical decision-making.

3.0 METHODOLOGY

This section outlines the detailed steps employed in achieving the objectives of this study, which includes the development of AI models with convolutional neural networks (CNNs) for serious disc bulge detection and clinical significance classification of lumbar spine MRI, the design of a graphical user interface (GUI) to improve clinician interaction, and the use of image explainer techniques to provide interpretable insights into the AI model's decision-making process. Fig. 1 below describes the flowchart of the methodology employed in this study.

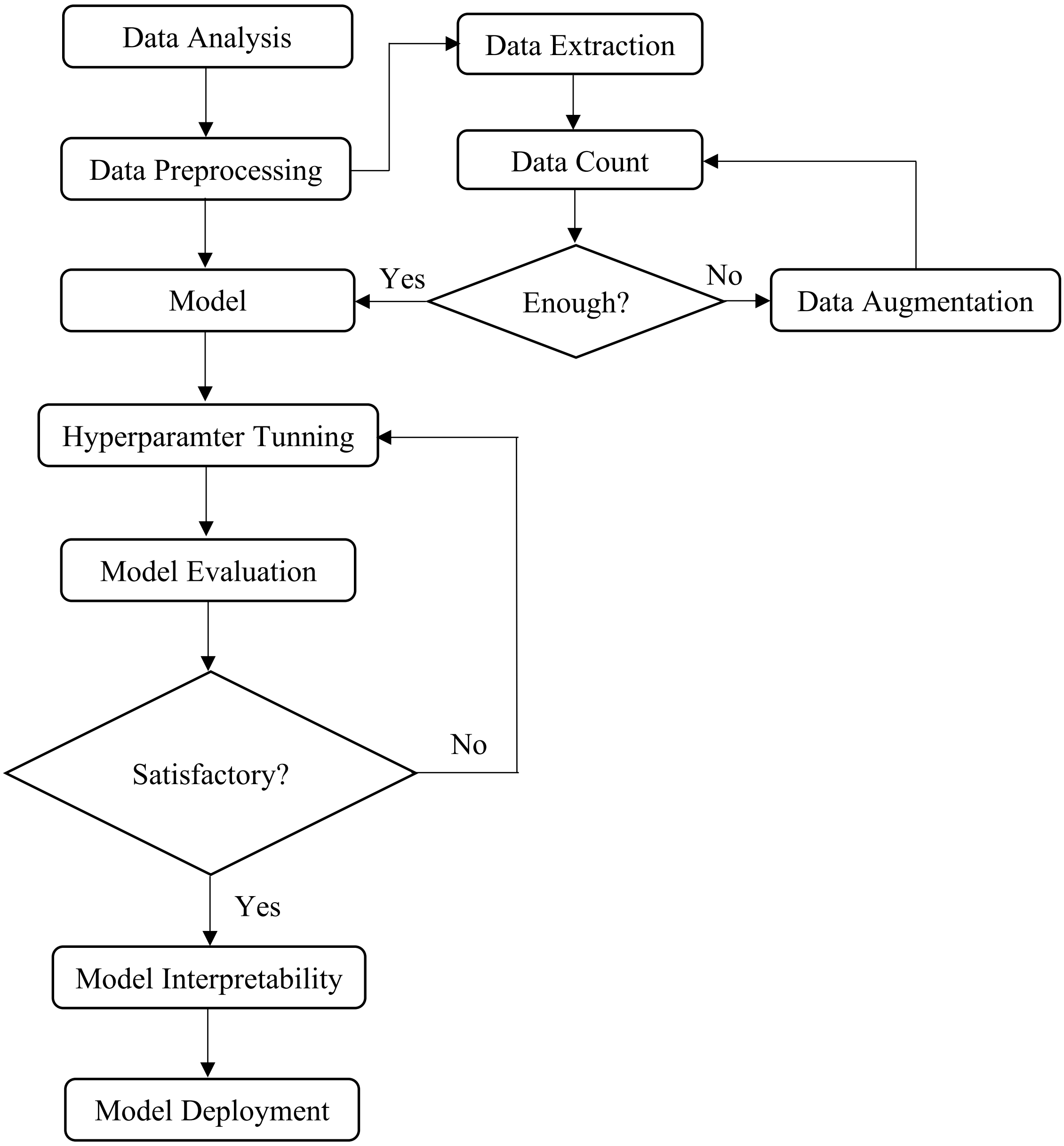


Figure 1: Flowchart describing the methodology of this study

3.1 Description of Dataset

The dataset used for this study comprises of MR composite images of 515 patients that reported symptomatic back pain. The images were sourced from the ground truth data provided by (Sudirman, et al., 2019). This study uses the axial view of the MRI, which is primarily obtained from the final three IVDs, including the one between the last vertebrae and the sacrum. The figure below shows the traverse of an axial view of a lumbar MRI. The labeling was guided by (Meidia, et al., 2018). The majority of the slices have a 320x320 pixel image resolution, and all of the pixels have 12-bit per pixel precision, which is greater than that of ordinary 8-bit greyscale images. For all axial-view slices, the slice thickness is uniformly 4 mm, with a centre-to-centre distance of 4.4 mm. The axial-view slices have a constant horizontal and vertical pixel spacing of 0.6875 mm.

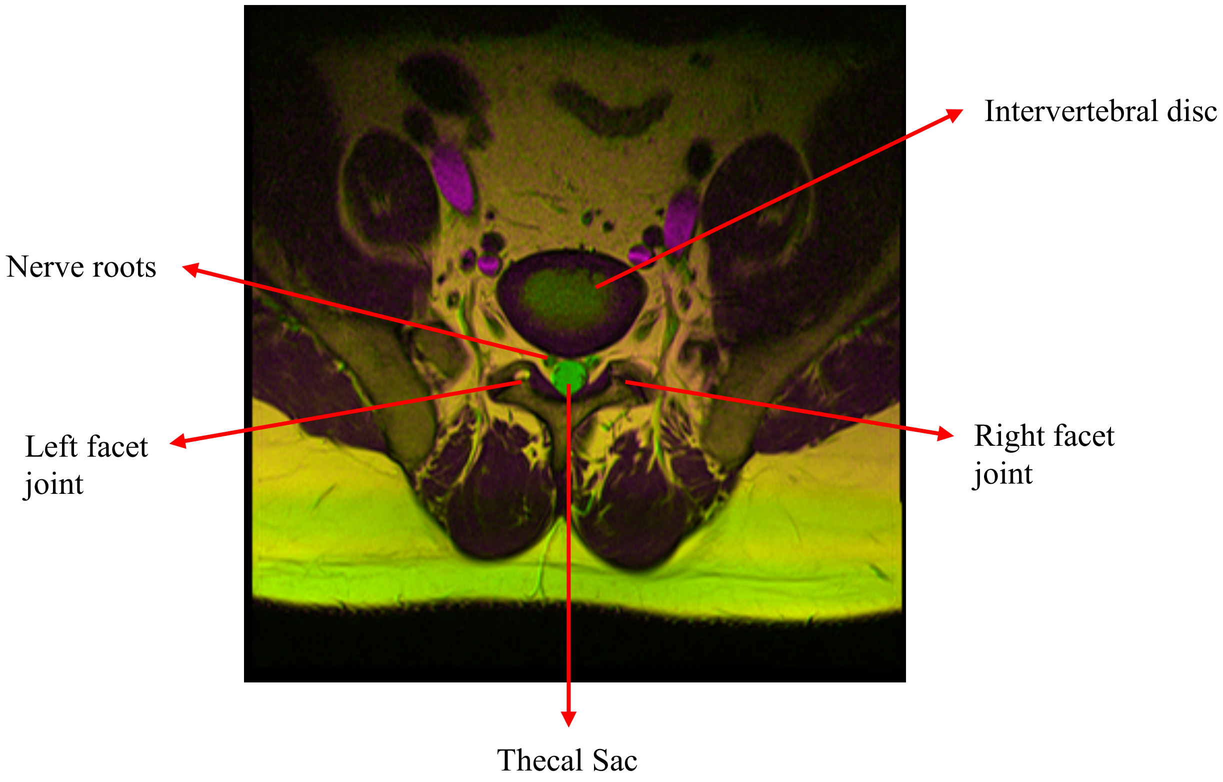


Figure 2: A composite traverse MRI image of L5-S1 intervertebral disc

3.1.1 Data Analysis

Various abnormal conditions were reported on numerous MRIs in the report provided by the radiologist. The following conditions, each with a brief description, were reported:

Disc Bulge/Protrusion: A bulging disc is a condition in which the inner portion of the intervertebral disc begins to protrude from the outer wall of the disc. This condition usually develops over time and can cause other disc degeneration conditions, such as spinal stenosis (Eguino, 2023).

Nerve Root Compression: This occurs due to the narrowing of the neural foramen, where the spinal nerves exit the spinal cord. This narrowing compresses the nerve roots, causing pain, numbness, and tingling in the lower back and legs (Hopkins, 2023).

Spinal Stenosis: Lumbar spinal stenosis (LSS) can be defined as any narrowing of the lumbar spinal canal, nerve root canals, and/or intervertebral foramina that may encroach on the nerve roots of the lumbar spine (Whitman & Fritz, 2017).

Spondylolisthesis: It is a disorder in which one vertebral body slips with regard to the adjacent vertebral body, causing radicular or mechanical symptoms or discomfort. This could occur due to fracture or degenerative changes (Steven & Christopher, 2023).

Facet Joint Damage: This occurs when there is damage to the facet joint because of injury, overuse, or arthritis. It is a common cause of lower back pain, especially when bending, twisting, or lifting (Perolat, et al., 2018).

Disc Herniation: Lumbar disc herniation is a mechanical condition caused by deterioration in disc structures that allow for movement between spinal discs and protect against injury. This may result in discomfort, numbness, or weakness in the lower back or legs (Kılıç, 2015).

Table 5 below shows the summary of the different conditions present in the report.

	D3	D4	D5
Bulge	89	255	126
Compression	74	256	149
Spinal Stenosis	6	10	2
Spondylolisthesis	1	6	6
Herniation	7	54	78
Facet Joint Damage	0	3	1

Table 5: Breakdown of different conditions present in the Radiologist's report

3.2 Data Preprocessing

Sudirman et al. (2019) supplied an axial view of the composite lumbar MRI dataset in '.png' format, and an excelsheet containing labeling notes from the radiologist. However, not all notes were provided for every disc present in the file. The sheet contained information for 515 patients, the ones with missing or unclear information were removed, bringing the total number of records down to 496. Each patient's data included images of three discs: L3-L4 (D3), L4-L5 (D4), and L5-S1 (D5). This study consist of two sections: Clinical Significance Classification and Serious Disc Bulge Detection. The first section was classifying each patient's discs (D3, D4, D5) into three level of clinical significance: No Clinical Significance, Mild Clinical Significance, Serious Clinical Significance. Table 5 below shows a summary of each significance level alongside their different conditions which was categorised using previously published studies (Michael & Peter, 2006), (Milette, et al., 1999), (Ma, 2015). Three separate columns were added to the excelsheet and 3 corresponding folders were created respectively with titles No_CS, Mild_CS, Serious_CS indicating (No, Mild and Serious Clinical Significance). Each image was cross-referenced against the clinician's notes to confirm that the labeled conditions were appropriately represented. The spreadsheet was populated by comparing the clinician's notes to Table 5 below, assigning a value of 1 for true and 0 for false for each case. Each MRI was then moved to it's designated folder.

No Clinical Significance	Mild Clinical Significance	Serious Clinical Significance
	Features of muscle spasm	Narrowing both neural foramina
	Small disc protrusion	Narrowing right neural foramen
	Abutting the thecal sac.	All abnormal condition
	Mild diffuse disc bulge	Compressing the thecal sac and exit canals.
	Features of muscle spasm	Mild disc bulge with annular tear
	Dehydrated discal material	Sequested disc
	Posterior osteophytes	Facet joint hypertrophy
	Diffuse disc bulge	Largely compressing
	Encroaching exit canal	Compressing thecal sac and encroaching exiting neural canals

	Encroaching nerve root	Secondary neural canal stenosis
	Abutting nerve root	Annular tear
	Schmorl's nodes noted	
	Compression but adequate spinal canal	

Table 6: Clinical Significance Classification of conditions in the radiologist report

The count of all images for each class is as follows:

- No Clinical Significance: 801
- Mild Clinical Significance: 236
- Serious Clinical Significance: 387

Similar steps were employed for the second phase of this study, focusing on serious disc bulge detection. This section consists of two classes, No serious bulge and Serious bulge. In the radiologist report, conditions like diffuse disc bulge compressing the thecal sac and exiting the canals, disc protrusion compressing the thecal sac and narrowing the left neural foramen etc. were considered serious conditions of serious disc bulge, while cases of mild diffuse disc bulge, and diffuse disc bulge mildly compressing the thecal sac etc. were considered not to be serious cases hence no serious disc bulge similar to what was has been reported by (Jun, et al., 2009). Two columns were added to the excelsheet and 2 corresponding folders were created respectively with titles no_serious_db, serious_db indicating (No and Serious disc bulge). The spreadsheet was populated by assigning a value of 1 for true and 0 for false for each case. Each MRI was then moved to it's designated folder.

The count of all images for each class in this section is as follows:

- No Serious Bulge: 1074
- Serious Bulge: 366

The total number of images generated for each case was not enough to train the model as this resulted in poor performance. Deep convolutional neural networks rely heavily on big data to avoid underfitting. Underfitting occurs when a machine learning model is too simple to capture underlying pattern of the data. This can be identified as low accuracy on both training and testing dataset. In some cases, it could overfit where a network learns a function with extremely high variation in order to perfectly model the training data (Shorten & Khoshgoftaar, 2019).

3.2.1 Data Augmentation

Data augmentation was used to enrich model training data and create fake random variations to improve the model's robustness. The data augmentation technique involves adding modest random stochastic modifications to the training image dataset and training the model with both the source and supplemented data. The use of data augmentation has been shown to improve model resilience and test accuracy (Krizhevsky et al., 2012; Zoph et al., 2019). Seven typical data augmentation procedures are used, random cropping, horizontal flipping, rotation, zooming in, symmetric wrapping, and modifying the brightness and contrast of the images. For this study, the augmentation parameters considered are:

- ***width_shift_range***: It horizontally shifts the image to the left or right. If the value is float and <=1, it will use the percentage of total width as the range. For example, if the image width is 100px, and width_shift_range = 0.2 it will take -20% to +20% means -20px to +20px. It will shift image randomly between this range. For both models developed during the data augmentation a value of 0.1 was used.
- ***height_shift_range***: It functions similarly to width_shift_range but shifts vertically (up or down). A value of 0.1 was utilized in both cases which takes out 10px from the top and 10px from the bottom.
- ***horizontal_flip***: This causes the images to be randomly flipped horizontally. True indicates that the images can be flipped horizontally with a probability of 50% which was used for both models during development.
- ***fill_mode***: It defines rules for newly moved pixels in the input area. Setting "nearest" means filling gaps with the nearest pixel value.

The following code snippet demonstrates the use of the *ImageDataGenerator* class from the Keras library to perform data augmentation:

```
ImageDataGenerator (
        width_shift_range=0.1
        height_shift_range=0.1
        horizontal_flip=True
        fill_mode='nearest')
```

Tables 5 and 6 below show a summary of the data count before and after data augmentation performed in each section of this study.

	Before Augmentation	Augmentation factor	No. of Augmented Images	Total
No Clinical Significance	801	2.15	1721	2522
Mild Clinical Significance	236	7.79	1839	2075
Serious Clinical Significance	387	5.48	2121	2509

Table 7: Summary of data augmentaion done for clinical significance classification

	Before Augmentation	Augmentation factor	No. of Augmented Images	Total
No Serious Bulge	1074	1.33	1426	2500
Serious Bulge	366	5.83	2134	2500

Table 8: Summary of data augmentation done for serious disc bulge detection

3.3 Model Implemmentation

Convolutinal Neural Networks was used for the development of this model. CNN is a classification system that classifies images into labeled classes. They are made up of neurons with learnable weights and many layers to extract information from images and then learn to classify them (Shaw, et al., 2021).

3.3.1 CNN's Architecture

CNN employs several layers to extract features from images. The architecture used in this study is made up of 4 layers which includes:

- Cropping layer – This layer was added to get the center of focus by removing 45 from the top, 25 from the bottom, left and right of the image with an input shape of 150px by 150px.

- Convolutional and Pooling layers - This layer is used to extract different features from the input images. This layer performs the mathematical operation of convolution between the input image and a filter of a specific size. Each models had 4 layers, Each layer had filters with the first layer having 64, second layer 128, third layer 256 and fourth layer 512 and an activation function of relu, a padding “same”, each layer had a max pooling 2D with a pool_size 2, 2 and a dropout function of 0.2.

- Flatten and Fully Connected layers - Consists of weights and biases as well as neurons and is used to connect neurons from different layers. It is often the final few layers of a CNN architecture. The fully connected layer had 3 dense layers with filters of 512, 256, 128 and an activation function of relu for each layer.

- Output Layer – This is the final layer of the CNN where predictions are obtained. 2 neurons and a Softmax activation function was added which ensures that the output from this network is splitted into three classes for the first section (Clinical Significance) and two classes for the second section (Serious disc bulge).

For both models developed in each section of this study, relu and softmax activation functions were utilised. Activation functions is one of CNN most important parameters. They are used to learn and approximate any continuous and complex relationship that exists between network variables. RMSprop was used as the optimizer and a learning_rate of 0.001was utilised this helps change the wieghts and learning rate to reduce losses.

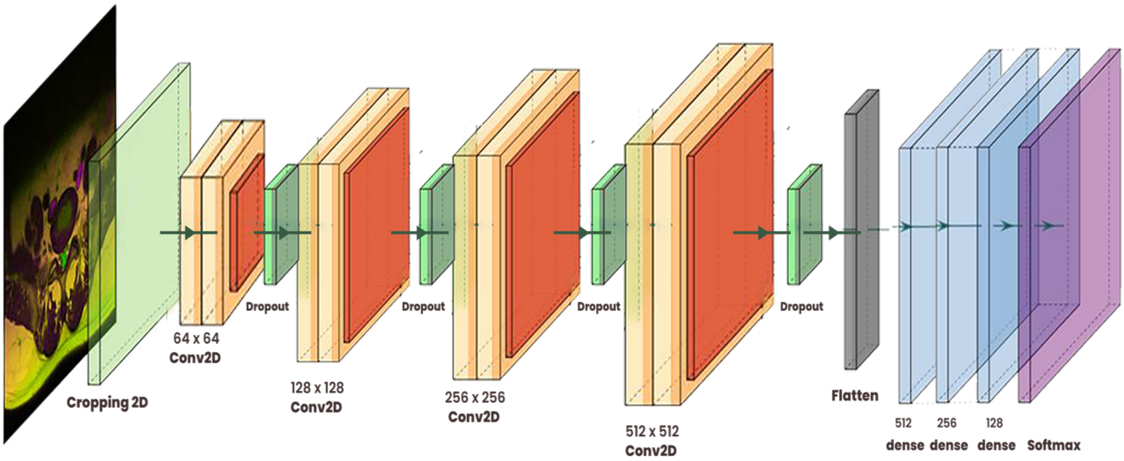


Figure 3: CNN's architecture

Layer (type)	Output Shape	Param #
cropping2d_2 (Cropping2D)	(None, 80, 100, 3)	0
conv2d_8 (Conv2D)	(None, 80, 100, 64)	1792
max_pooling2d_8 (MaxPooling2D)	(None, 40, 50, 64)	0
dropout_10 (Dropout)	(None, 40, 50, 64)	0
conv2d_9 (Conv2D)	(None, 40, 50, 128)	73856
max_pooling2d_9 (MaxPooling2D)	(None, 20, 25, 128)	0
dropout_11 (Dropout)	(None, 20, 25, 128)	0
conv2d_10 (Conv2D)	(None, 20, 25, 256)	295168
max_pooling2d_10 (MaxPooling2D)	(None, 10, 12, 256)	0
dropout_12 (Dropout)	(None, 10, 12, 256)	0
conv2d_11 (Conv2D)	(None, 10, 12, 512)	1180160
max_pooling2d_11 (MaxPooling2D)	(None, 5, 6, 512)	0
dropout_13 (Dropout)	(None, 5, 6, 512)	0
flatten_2 (Flatten)	(None, 15360)	0
dense_8 (Dense)	(None, 512)	7864832
dense_9 (Dense)	(None, 256)	131328
dense_10 (Dense)	(None, 128)	32896
dense_11 (Dense)	(None, 64)	8256
dense_12 (Dense)	(None, 3)	195
Total params: 9588483 (36.58 MB) Trainable params: 9588483 (36.58 MB) Non-trainable params: 0 (0.00 Byte)		

Table 9: Serious Disc Bulge Detection Model Summary

3.3.2 Model Training and Evaluation Metrics

The dataset was split into 70% for training, 20% for testing and 10% for validation as suggested by (Blackman, et al., 2016). The model was trained on the training set, tested on the testing set and validated using the validation data set. External validation is an important step in the validation of a deep-learning-based, predictive model to avoid overfitting to the training data set, potentially resulting in an overestimation of the model's diagnostic performance (S.H & K, 2018), (J.R. & Cheng, 2019). A batch size of 128 and 50 epochs were used for both models. The epoch is the complete number of pass of the training dataset through the model (Afaq & Rao, 2020). Callback parameters were also included by utilizing ModelCheckpoints to monitor validation accuracies and save the best model. The models were evaluated by plotting the training and validation accuracies, a confusion matrix and classification reports.
The following classification metrics were used to evaluate and compare the performance of the models developed in this study:

• Accuracy: is simply the percentage of correct label predictions against the sum of all the predictions as shown below, it is a commonly used metric for evaluating the performance of classification models.

Accuracy=$\frac{Correct\ label\ prediction}{Sum\ of\ all\ prediction}$

• Precision: is the ratio of true positive (TP) to the sum of false positives (FP) and TP as shown below, it measures the capability of a model to not predict positive for a negative input (Sklearn, 2024).

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

• Recall: is the ratio of true positive (TP) to the sum of false negative (FN) and TP, it measures the ability of a model to find all the positive samples (Sklearn, 2024)

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

• F1-score: is the harmonic mean of both precision and recall (Sklearn, 2024),

F1=$\frac{2*(precision * recall)}{precision+recall}$

3.4 Model Interpretability using LIME

Model Interpretability or Explainable AI as the name implies, is a type of Artificial Intelligence that allows for the explanation of complex models such as deep learning models, CNN etc. It focuses on why the AI arrived at a particular decision analyzing its logical paradigms, as opposed to the inherent black box nature of AI (Vishwarupe, et al., 2024). They are referred to as black box because of their difficulty in analyzing and visualising the role of their inputs in producing the predicted output (Lundberg & Lee, 2017). Many libraries and packages aim to simplify and explain complex models, reducing the "black box" problem. Local Interpretabile Model-Agnostic Explanations (LIME) was used in this study to explain how different areas of the MRI contributes to the model's decision. LIME generates an interpretable model by training a local linear model around the prediction point.

3.5 Web App Development For Model

A graphical user interface GUI was developed to interact with the model. This GUI was created using streamlit which is an open-source python library used to create dynamic apps (Streamlit-Docs, 2024). The best models for each category was saved and passed to the streamlit application. User will be able to upload new MRI that hasn't been seen by the model using the GUI and the model will give predictions and XAI will be applied to the model diagnosis and both results will be available to the user on the GUI.

3.6 Software Tools Used

1. **Google Colab Pro**: This Jupyter notebook environment allows python codes to be written and executed in the cloud. All the models in this research were built in this environment, with the Python version being 3.10.12. This is a paid version of *Google Colab*, allowing access to GPUs for faster training of the models.
2. **Keras**: This is an open-source API built on top TensorFlow for building and training neural network models in Python.
3. **LIME**: This library is used to explain the predictions of machine learning models. It is very useful for black-box models like deep neural networks where internal workings are not easily interpretable.
4. **Streamlit**: This is an open-source Python library used to create interactive web applications. Streamlit provides a simple API for designing user interace for showcasing projects.

4.0 RESULT AND DISCUSSION

This section is divided into 5 parts which discusses the model evaluation results, the effect of LIME when applied on the model, web application developed for interfacing with the model, ethical and professional issues with the development of an AI model, and research contribution.

4.1 Model Evaluation Results

Five evaluation metrics were used to evaluate the performance of the models. Accuracy, recall, precision, F1-score and confusion matrix. The formulas for each metric has been provided in the methodology of this study. The model for Clinical Significance Classification achieved an average precison score of 81%, recall score of 81% using the macro average and an overall accuracy of 81%. The "macro" average was used because it is an averaging method that treats all class labels as equals and calculates the metric for each label and outputs their unweighted mean (Sklearn, 2024). While that of serious disc bulge detection achieved a precision score of 89%, recall score of 89% and an higher accuracy of 89%. A summary of each model's evaluation metrics is presented in the table below.

	Precision	Recall	F1-Score
No Clinical Significance	0.81	0.83	0.82
Mild Clinical Significance	0.75	0.76	0.75
Serious Clinical Significance	0.87	0.83	0.85
Accuracy			0.81
Macro avg	0.81	0.81	0.81
Weighted avg	0.81	0.81	0.81

Table 10: Model Evaluation Summary for Clinical Significance Model

	Precision	Recall	F1-Score
No Bulge	0.90	0.88	0.89
Serious Bulge	0.88	0.90	0.89
Accuracy			0.89
Macro avg	0.89	0.89	0.89
Weighted avg	0.89	0.89	0.89

Table 11: Model Evaluation Summary for Serious Disc Bulge Detection Model

Figure 3 shows a plot of the validation accuracy over the training accuracy of the clinical significance model over 50 epochs. During the initial 13 epochs, both training and validation accuracy remained below 75%. The training accuracy experienced growth reaching a peak of 99.42%, while the highest validation accuracy 81% was observed around the 27th epoch.

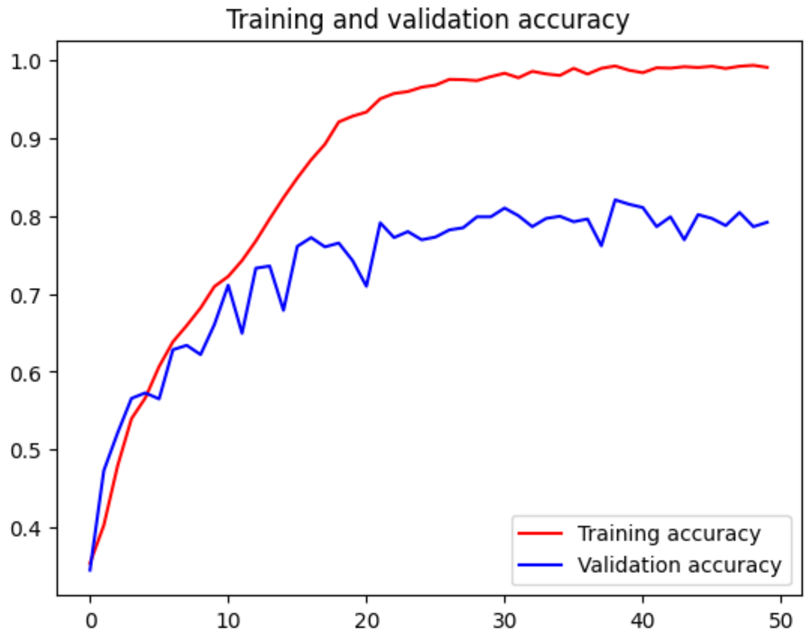


Figure 4: Plot of training over validation accuracy for clinical significance model

Figure 4 below shows the validation accuracy plot over the training accuracy of the disc bulge detection model over 50 epochs. The first 9 epochs recorded a training and validation accuracy under 80%. The training accuracy experienced a steady growth and climaxed at 99.48% while the highest validation accuracy of 89% was observed around the 27th epoch.

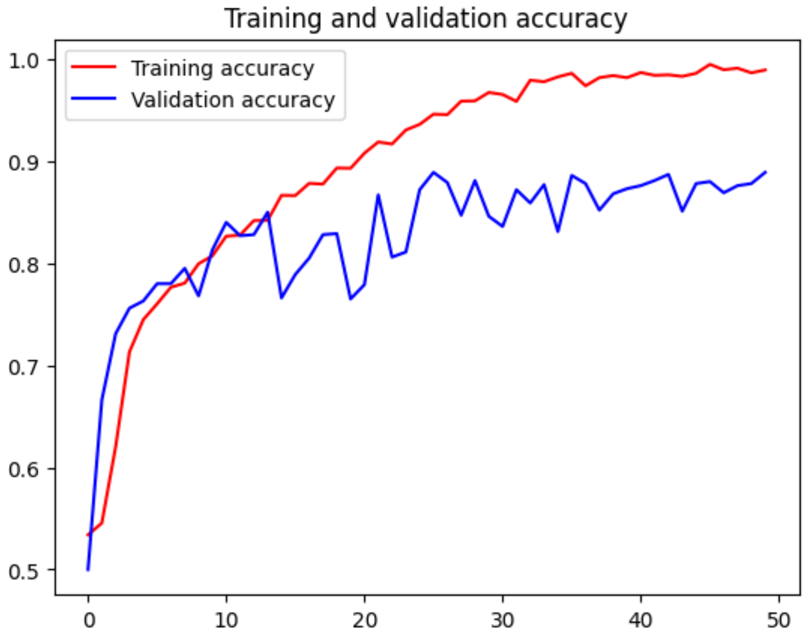


Figure 5: Plot of training over validation accuracy for serious disc bulge detection model

A confusion matrix plot was generated for both models to access where errors were made in the models. Figures 5 and 6 below show the confusion matrix plot of actual against predicted for the serious disc bulge detection model and the clinical significance classification model. The confusion matrix for the clinical significance classification model reveals that it correctly identified 395 cases with no clinical significance, 310 cases with mild clinical significance, and 428 cases with serious clinical significance. Meanwhile, for the serious disc bulge detection model, it accurately classified 444 cases with no disc bulge, and 443 cases with serious disc bulge.

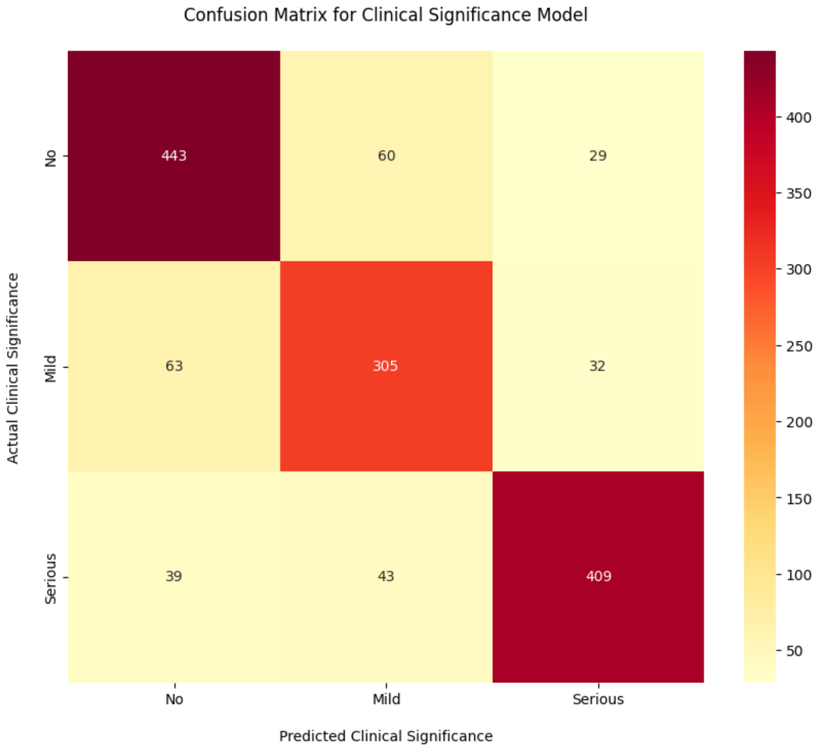


Figure 6: Confusion matrix for clinical significance classification model

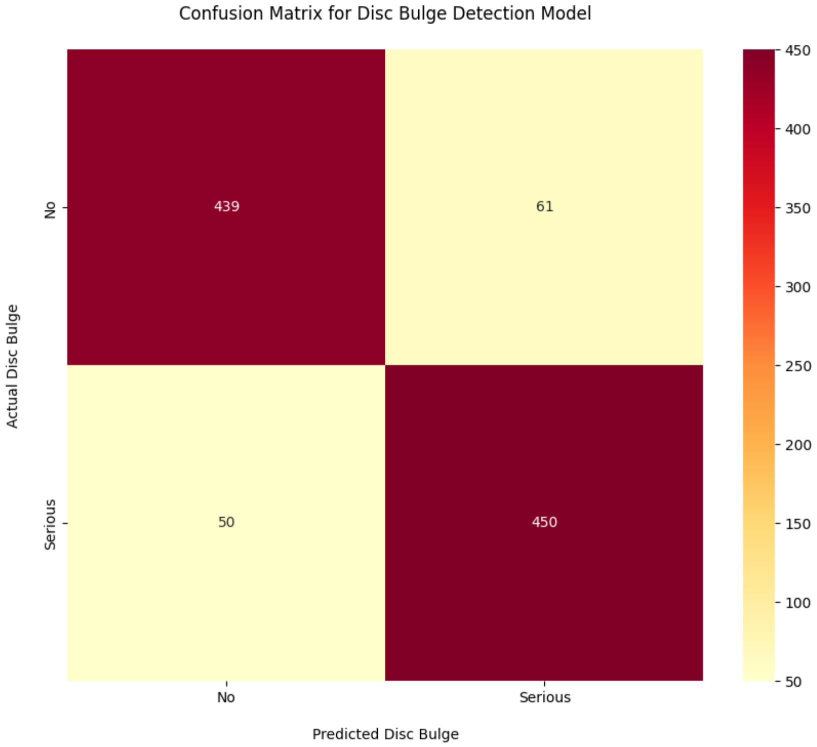


Figure 7: Confusion matrix for serious disc bulge detection model

4.2 Model Interpretability Results

In this section, we provide the interpretability findings of both models: serious disc bulge detection and clinical significance classification using lumbar spine magnetic resonance imaging (MRI). Model interpretability is critical for assuring the confidence, transparency, and comprehension of artificial intelligence (AI) models in medical imaging. By investigating how the models produce predictions, we want to get insights into their decision-making process, thereby increasing their clinical utility and promoting trust among healthcare practitioners. LIME was selected for this study as it is model agnostic and can be applied to any machine learning model.

Figures 7 and 8 show the output of LIME reflecting the contribution of each region to the models prediction. The heatmap on the right breaks down the level of contribution for each region.

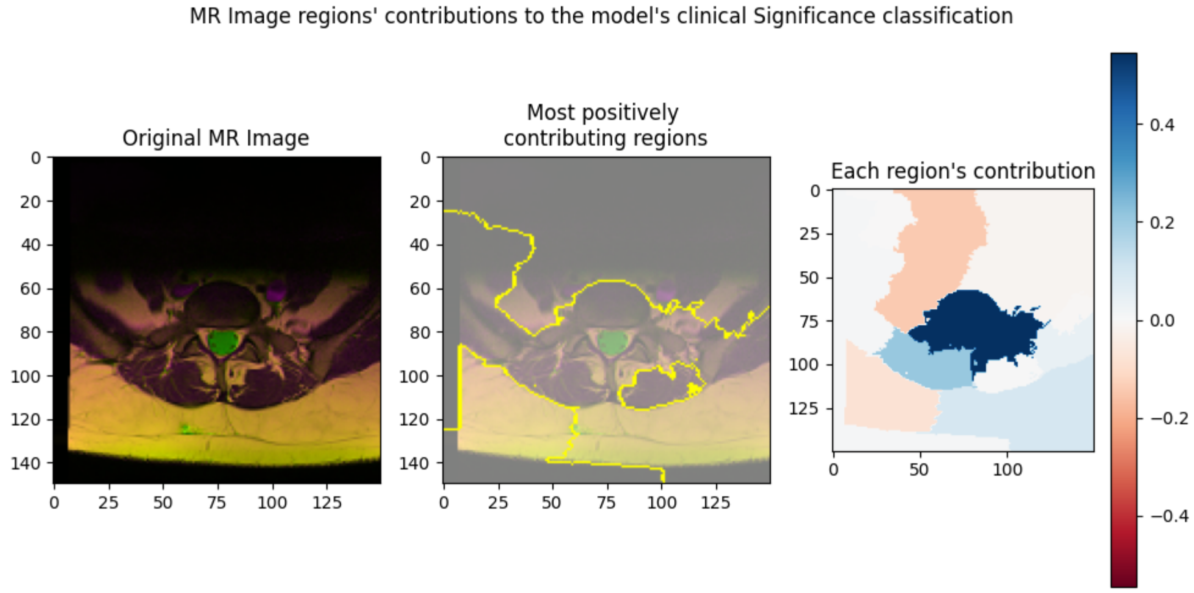

MR Image regions' contributions to the model's clinical Significance classification
Original MR Image
Most positively contributing regions
Each region's contribution
0
20
40
60
80
100
120
140
0
25
50
75
100
125
0
25
50
75
100
125
0
50
100
0.4
0.2
0.0
−0.2
−0.4

Figure 8: Contribution of each region to clinical significance classification

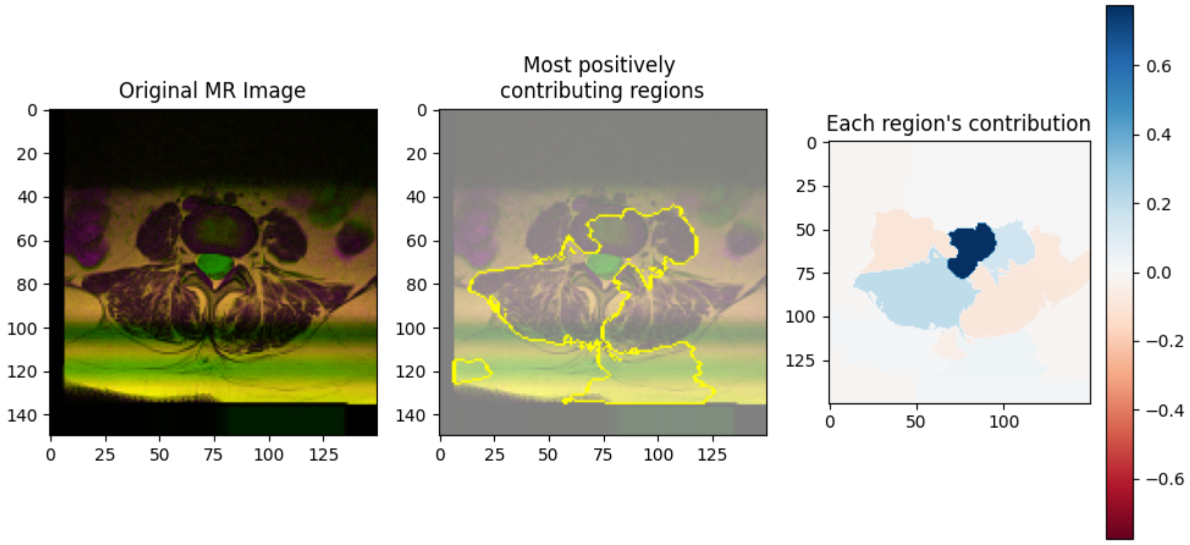
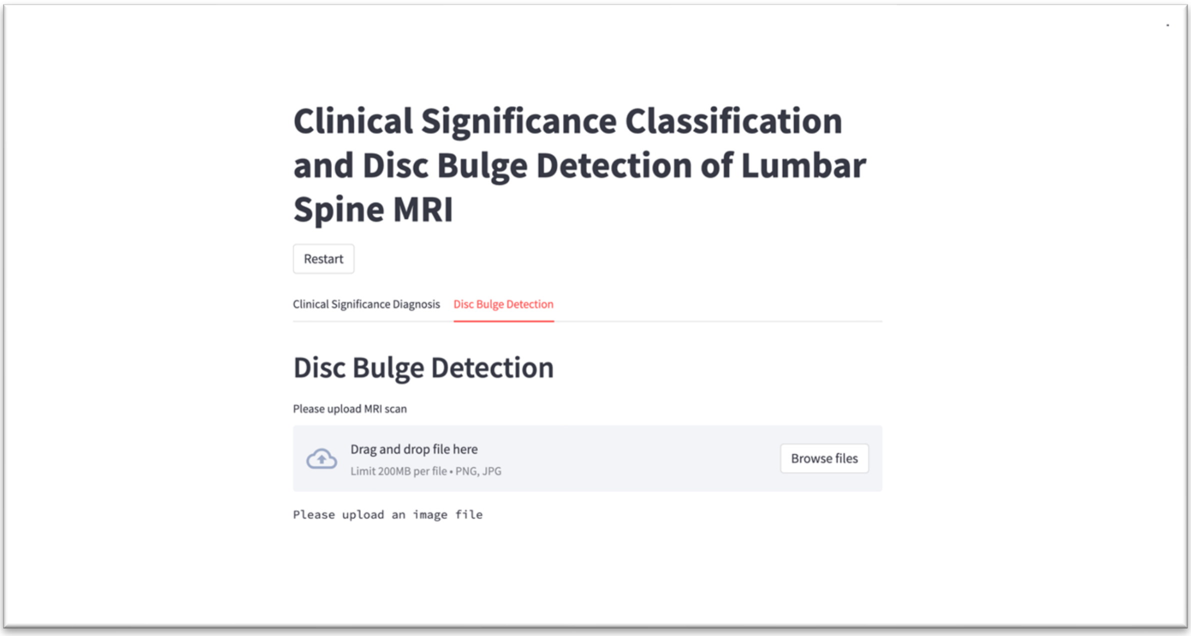


Figure 9: Contribution of each region to serious disc bulge detection

4.3 Web Application for Model

A Graphical User Interface (GUI) was created using streamlit. It is a one page web app with two tabs for the Clinical Significane Classification and disc bulge detection. The application allows you upload an MRI and runs a prediction using the saved best models. During the developement of both models, the best models were saved for each category as a '.h5' file which is a common pratice in keras/tensorflow as it saves the entire model architecture including the weights, training configuration, optimizer and its state. An app.py file was created with the codes as the entry file. When an MRI is uploaded through the GUI the MRI is stored in an image directory which the model picks it up from and makes prediction on. When the predictions are done LIME is applied to the image. The GUI presents you back with the prediction score and a LIME explanation applied to the image. Fig. 9 below shows a demonstration of the GUI.



Clinical Significance Classification and Disc Bulge Detection of Lumbar Spine MRI

Disc Bulge Detection

Please upload MRI scan

Please upload an image file

Restart
Clinical Significance Diagnosis
Disc Bulge Detection
Drag and drop file here
Limit 200MB per file • PNG, JPG
Browse files

Figure 10: GUI for model deployment

4.4 Ethical Issues

One major ethical concern is patient privacy and data security. As this study involves the use of medical imaging data, ensuring patient confidentiality and adhering to data protection regulations are critical. Transparency regarding the use of patient data for research purposes are essential ethical duties. It is critical to ensure that AI models are unbiased and do not exacerbate existing healthcare disparities. It is also imperative to ensure that the AI models are rigorously evaluated and validated to guarantee their accuracy and reliability.

4.5 Research Contribution

This research significantly advances AI-enabled MRI diagnosis, specifically in lumbar spine clinical significance classification and serious disc bulge detection. This research has successfully closed the research gaps identified in *Section 2.5*. of this study which are:

1. Implementation of AI models for identifying serious disc bulges in the lumbar spine, and clinical classification significance into no, mild, and serious categories.
2. Developement of a graphical user interface (GUI) for the AI models to enhance the user experience of clinicians and hence patient care.

3. Evaluation of the clinical importance of AI-based diagnosis in the context of lumbar spine MRI interpretation.

4. An investigation on the potential of convolutional neural networks (CNNs) for reliable serious disc bulge detection and lumbar spine clinical significance classification.

5. Application of image explainer techniques to provide interpretable insights into the AI model's decision-making process, enhancing trust and understanding among clinicians.

6. Contribution to the advancement of AI-driven diagnostic tools for improved patient outcomes in the diagnosis and management of lumbar spine disorders.

5.0 CONCLUSION AND FUTURE WORKS

5.1 Conclusion

This study explored the development of two models using deep learning techniques CNN to detect serious disc bulge and the clinical significance classification of lumbar spine MRI. Both models achieved great accuracy, precision and recall metrics ranging from 81% to 89%. This study demonstrates the potential use of AI in detecting and classifying various spinal abnormalities which can provide support to radiologists and clinicians. It further emphasizes the importance of AI interpretability in healthcare decision-making. The transparency and trustworthiness of the models were improved by incorporating explainable AI techniques, allowing clinicians to understand and validate AI-generated diagnoses. The use of LIME helped eliminate the black box present in model diagnosis.

5.2 Limitations and Future Works

While the models that were developed demonstrated remarkable metrics, further improvements can be made to the AI models' performance and clinical application. Firstly, expanding the dataset to include a larger and more diverse range of lumbar spine MRI images, encompassing various pathologies and patient demographics. This will enable for the creation of more robust and generalizable models capable of properly detecting serious disc bulges and classifying a wider range of spinal abnormalities. Incorporating multi-modal data, such as clinical notes and patient history, into AI models improves diagnosis accuracy and provides more thorough clinical insights. Also, improving the interpretability of AI models in order to increase their usability and reliability in healthcare contexts by using advanced explainable AI approaches, such as attention mechanisms and saliency mapping, to provide clinicians with clear and interpretable explanations of the AI-generated diagnoses

6.0 REFERENCES

Ma, X.-l., 2015. *A new pathological classification of lumbar disc protrusion and its clinical significance..* s.l.:Orthopaedic Surgery.

Q, P. et al., 2021. *Automatically Diagnosing Disk Bulge and Disk Herniation With Lumbar Magnetic Resonance Images by Using Deep Convolutional Neural Networks.* s.l.:JMIR Med Inform.

Milette, P. et al., 1999. *Differentiating Lumbar Disc Protrusions, Disc Bulges, and Discs With Normal Contour but Abnormal Signal Intensity: Magnetic Resonance Imaging With Discographic Correlations..* s.l.:s.n.

NC, L. et al., 2021. *Detection of Degenerative Changes on MR Images of the Lumbar Spine with a Convolutional Neural Network: A Feasibility Study.* s.l.:s.n.

Sudirman, A.-K.A. S. a. et al., 2019. *Boundary Delineation of MRI Images for Lumbar Spinal Stenosis Detection Through Semantic Segmentation Using Deep Neural Networks.* s.l.:IEEE Access.

D'Antoni, F. et al., 2022. *rtificial Intelligence and Computer Aided Diagnosis in Chronic Low Back Pain: A Systematic Review.* s.l.:s.n.

K, W., Y, A. & M., M., 2019. *An Application of Artificial Intelligence to Diagnostic Imaging of Spine Disease: Estimating Spinal Alignment From Moiré Images.* s.l.:Epub.

Varçin, F. a. E. et al., 2019. *Diagnosis of Lumbar Spondylolisthesis via Convolutional Neural Networks.* s.l.:s.n.

Tackeun Kim MD, M. et al., 2022. *Diagnostic triage in patients with central lumbar spinal stenosis using a deep learning system of radiographs.* s.l.:s.n.

Bardin, L. D., King, P. & Maher, C. G., 2017. *Diagnostic triage for low back pain: a practical approach for primary care.* s.l.:s.n.

Hancock, M. J. P., Koes, B. P., Ostelo, R. P. & Peul, W. P., 2011. *Diagnostic Accuracy of the Clinical Examination in Identifying the Level of Herniation in Patients with Sciatica.* s.l.:s.n.

Coster, S. & Tavy, S. F. T. M. d. B. &. D. L. J., 2009. *Diagnostic value of history, physical examination and needle electromyography in diagnosing lumbosacral radiculopathy.* s.l.:s.n.

Alomari, R. S. a. C. & Jason J. and Chaudhary, V., 2011. *Labeling of Lumbar Discs Using Both Pixel- and Object-Level Features With a Two-Level Probabilistic Model.* s.l.:IEEE Transactions on Medical Imaging.

Loh, H. W. et al., 2022. *Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022).* s.l.:s.n.

Dindorf, C. et al., 2021. *Classification and Automated Interpretation of Spinal Posture Data Using a Pathology-Independent Classifier and Explainable Artificial Intelligence (XAI)..* s.l.:s.n.

SA, K., RG, S., DE, W. & RB., K., 1982. *Bulging lumbar intervertebral disk: myelographic differentiation from herniated disk with nerve root compression.* s.l.:PubMed.

Kang Jeong-Woon, P. C. ,. L. D.-E. ,. Y. J.-H. ,. K. M., 2023. *Prediction of bone mineral density in CT using deep learning with explainability.* s.l.:Frontiers in Physiology.

Yuka Otaki, M. P. et al., 2022. *Clinical Deployment of Explainable Artificial Intelligence of SPECT for Diagnosis of Coronary Artery Disease..* s.l.:s.n.

Sudirman, S. et al., 2019. *Lumbar Spine MRI Dataset.* s.l.:Mendeley Data.

Shorten, C. & Khoshgoftaar, T., 2019. *A survey on Image Data Augmentation for Deep Learning.* s.l.:s.n.

Whitman, J. & Fritz, J., 2017. *Lumbar Spinal Stenosis.* s.l.:Elsevier.

MC, J. et al., 1994. *Magnetic resonance imaging of the lumbar spine in people without back pain.* s.l.:J Med.

Kılıç, B., 2015. *Lumbar Disc Herniation.* s.l.:Adv. Environ. Bio.

Steven, T. & Christopher, G. C., 2023. *Spondylolisthesis.* s.l.:StatPearls Publishing.

Shaw, R. N., Ghosh, A., Balas, V. E. & Bianchini, M., 2021. *Artificial Intelligence for Future Generation Robotics.* s.l.:Elsevier.

S.H, P. & K, H., 2018. *Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction.* s.l.:s.n.

J.R., E. & Cheng, P., 2019. *Artificial intelligence for medical image analysis: A guide for authors and reviewers..* s.l.:s.n.

Blackman, D. A. et al., 2016. *The 70:20:10 model for learning and development: an effective model for capability development.* s.l.:ResearchGate.

Afaq, S. & Rao, D. S., 2020. *Significance Of Epochs On Training A Neural Network.* s.l.:INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH V.

Vishwarupe, V. et al., 2024. *Explainable AI and Interpretable Machine Learning: A Case Study in Perspective.* s.l.:Procedia Computer Science.

Lundberg, S. & Lee, S.-I., 2017. *A unified approach to interpreting model predictions. Advances in neural information processing systems.* s.l.:s.n.

Streamlit-Docs, 2024. s.l.:https://docs.streamlit.io/.

Eguino, M., 2023. *Bulging disc - Causes, symptoms & treatment.* s.l.:Bonati Spine Institute.

Michael, A. & Peter, R., 2006. *What is Intervertebral Disc Degeneration, and What Causes It?.* s.l.:Spine.

Meidia, N. F. a. et al., 2018. *Development of Ground Truth Data for Automatic Lumbar Spine MRI Image Segmentation.* s.l.:s.n.

Hopkins, J., 2023. *Radiculopathy.* [Online]
Available at: https://www.hopkinsmedicine.org/health/conditions-and-diseases/radiculopathy#:~:text=When%20a%20nerve%20root%20is,in%20the%20arms%20or%20legs

Perolat, R. et al., 2018. *Facet joint syndrome: from diagnosis to interventional management.* s.l.:National Library of Medicine.

Jun, Z. et al., 2009. *Dynamic Bulging of Intervertebral Discs in the Degenerative Lumbar Spine.* s.l.:Spine.

Ma, X.-l., 2015. *A New Pathological Classification of Lumbar Disc Protrusion and Its Clinical Significance.* s.l.:Orthopaedic Surgery.