Healthcare Analytics 9 (2026) 100459

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

# An explainable deep learning method for diagnosing lumbar spine disorders from medical images

Uthman Oyebanji [a,b], Ala AlKafri [a,*], Hiba Alsmadi [a], Mohammad Alkasasbeh [c], Safwan Maghaydah [d], Friska Natalia [e], Wa'el Al-Karaki [a]

[a] Teesside University, Middlesbrough, TS1 3BX, United Kingdom
[b] SSE Renewables, Glasgow, G2 6AY, United Kingdom
[c] Wye Valley Hospitals NHS Trust, Hereford, HR1 2BN, United Kingdom
[d] Abu Dhabi University, MZ39, Abu Dhabi, United Arab Emirates
[e] Universitas Multimedia Nusantara, Serpong, Indonesia

ARTICLE INFO

*Keywords:*
Artificial intelligence
Convolutional neural network
Deep learning. explainable diagnosis
Lumbar spinal stenosis
Medical imaging

ABSTRACT

Lower back pain (LBP) is a global medical condition that is suffered by up to 80% of people at least once in their lifetime. Various abnormal conditions could cause LBP, and their diagnostic importance can range from *Mild, to Severe* level of clinical importance. One of these conditions is Lumbar Spinal Stenosis (LSS), which is usually a *Serious* or *Severe* condition, hence the need for its prompt diagnosis. The manual (visual) assessment process of magnetic resonance imaging (MRI) is time-consuming, labour-intensive, and prone to delays. This makes the integration of an AI-enabled diagnostic system an important opportunity for improving clinical workflow efficiency. In this paper, two convolutional neural network (CNN) models have been built to diagnose T2-weighted (T2W) sagittal MRI scans. The first model classifies the patient's sagittal view MR images into five grades of clinical importance, while the second model detects spinal stenosis in those images. Both models achieved strong test performance, with accuracies and recalls between 97% and 98%, and the stenosis model achieved 88% accuracy and recall on an independent external validation dataset. Four explainable AI (XAI) methods - LIME, Grad-CAM, SHAP, and Integrated Gradients (IG) - were applied to both models to get explainable insights. Quantitative faithfulness testing demonstrated that SHAP consistently provided the most reliable explanations, achieving the lowest deletion mean AUCs (0.2164 and 0.4315) and highest insertion mean AUCs (0.9194 and 0.8874) across both models. These explainable diagnoses help ensure the models are not only accurate but also become reliable for potential clinical use.

## 1. Introduction

Lower back pain (LBP) is one of the most common medical conditions globally. Studies indicate that up to 80% of individuals experience this condition at least once in their lifetime [1] [2]. In the UK, for example, LBP poses serious public health concerns, with estimates suggesting it costs the government billions annually in healthcare expenses based on historical records [3]. These costs can be categorised as either direct National Health Service (NHS) primary healthcare expenses - such as GP consultations, physiotherapy sessions, and prescription medications; or indirect costs, including workplace absence, reduced productivity due to sickness, and payments through welfare benefits [4]. Many nations, including the UK, are also facing the challenges of ageing populations, which is expected to lead to a rise in LBP cases [5]. This growing demand is putting considerable pressure on the National Health Service (NHS) to provide diagnosis and treatment [6]. The situation is worsened by a shortage of radiologists, resulting in potential delays and, hence, highlights the need for AI-enabled diagnostics [7] [8].

LBP can be caused by a variety of lumbar conditions, each with different levels of clinical importance. One of the causes of serious or severe LBP is LSS, which is a condition characterised by the narrowing of the lumbar spinal canal caused by inflammation in bones or soft tissues. This narrowing exerts pressure on the spinal nerve roots, resulting in symptoms ranging from radicular pain to neurogenic claudication [9]. LSS is a major cause of disability in older adults and the leading reason for spinal surgery in individuals over 65 years old [10]. There are two

* Corresponding author.
*E-mail address:* a.alkafri@tees.ac.uk (A. AlKafri).

https://doi.org/10.1016/j.health.2026.100459
Received 23 July 2025; Received in revised form 21 November 2025; Accepted 24 March 2026
Available online 31 March 2026

main categories of LSS: central stenosis (compression of the thecal sac containing the lumbar nerve roots) and lateral recess stenosis (compression of the individual exiting nerve roots) [11].

The diagnosis of LSS and other lumbar spine abnormal conditions conventionally involves assessing MRI scans of a patient's lumbar spine to visualise the spine slice by slice in sagittal and axial views [9]. Traditionally, radiologists analyse these scans using their expertise to identify and note down critical findings within the scans. However, this manual visual analysis is time-consuming, labour-intensive, and prone to delays [12]. The integration of AI-enabled diagnostic models is expected to reduce these delays, ease the burden on radiologists and foster a less time-consuming and cost-efficient diagnostic process.

Convolutional neural networks have shown strong potential in automating clinical diagnosis from medical imaging across varying severity levels - for example, a CNN model applied to chest X-rays successfully classified normal cases, COVID-19, viral pneumonia, and bacterial pneumonia [13]. Building on this, we adopt similar deep learning principles to support diagnostic decision-making in lumbar spine MRI. This paper used the CNN algorithm to build two AI-enabled diagnostic models. The first model, *Sagittal Clinical Importance model,* aims to classify lumbar sagittal MR images into *No*, *Mild*, *Moderate, Serious and Severe* clinical importance. This first model provides the starting point for the MRI diagnosis. The second model, *Sagittal Stenosis Detection model,* aims to further detect spinal stenosis in the MR images by classifying them into *Stenosis* and *No Stenosis*. This model helps to further diagnose an MR image for spinal stenosis, particularly images with *Serious and Severe* clinical importance where spinal stenosis is categorised.

There are various LBP conditions in the dataset. The interpretation of these conditions was done by using the expert radiologist report provided along with the MRI dataset [14]. The following are some key abnormal conditions found in the report with some brief explanations as available in Refs. [15], [16] and [11].

- **Disc Bulge/Protrusion:** This occurs when the soft center (nucleus pulposus) of an intervertebral disc pushes outward against the surrounding tough outer ring (annulus fibrosus), causing the disc to bulge out of its normal shape. Usually, it doesn't cause any symptoms but can sometimes lead to pain or other abnormal conditions.
- **Disc Herniation:** This occurs when there is an annulus tear, allowing the nucleus pulposus to bulge/herniate out. This can cause pain, numbness, or weakness in the lower back or legs.
- **Sequestered Disc:** This occurs when a piece of the nucleus pulposus breaks through the annulus fibrosus and migrates into the spinal canal. This can cause nerve compression and severe pain.
- **Thecal Sac Compression:** This occurs due to the narrowing of the thecal sac, a protective membrane surrounding the spinal cord. This can be caused by disc herniation, spondylolisthesis, or other conditions.
- **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.
- **Spinal Stenosis:** This is the narrowing of the spinal canal, which can put pressure on the spinal cord and nerves. This can cause pain, numbness, or weakness in the lower back or limbs.
- **Spondylolisthesis:** This is a condition in which one lumbar vertebra slips forward over another vertebra. This occurs mostly due to fracture or degenerative changes and has similar symptoms as spinal stenosis.
- **Facet Joint Damage:** This occurs when there is damage to the facet joint because of injury, overuse, or arthritis. It can cause pain in the lower back, especially when bending, twisting, or lifting.
- **Schmorl's Nodes:** These are small herniations of the nucleus pulposus into the vertebral body. Typically, they are considered benign but can sometimes cause pain.
- **Muscle Spasm:** This is a sudden, involuntary contraction of a muscle, often occurring in response to muscle strain or injury. This can vary from mild to serious pain or stiffness in the affected area.
- **Posterior osteophytes:** These are bony growths that develop on the back of the vertebrae usually due to degenerative changes in the spine.

The data interpretation from the radiologist report was then used to classify the clinical importance into 5 classes of *No*, *Mild*, *Moderate, Serious and Severe,* and the spinal stenosis detection into 2 classes of *Stenosis* and N*o Stenosis*.

Explainable AI methods have been increasingly applied in medical imaging to align model reasoning with clinician expertise. For example, a CNN integrated with LIME was used to identify and highlight local discriminative regions in chest X-rays, effectively mimicking the diagnostic reasoning of radiologists [17]. In this paper, we also used this approach by integrating LIME's *Image Explainer* to produce intuitive pixel-level heatmaps for lumbar spine MRI predictions, thereby enhancing transparency and fostering clinician trust.

## 2. Literature review

This section explores the existing knowledge surrounding AI-enabled diagnosis of lumbar spine MR images. It provides a comprehensive understanding of the evolution of the topic by delving into previous research and highlighting significant achievements made so far.

### 2.1. AI-enabled diagnosis of lumbar spine MR images

*Salehi* et al. [1] used a deep CNN-based CAD system to diagnose three types of disc herniation conditions based on lumbar axial MR images. A dataset of 2329 scanned MR images was gathered to build the model and then extended to 9316 via *Visual Rotation* and *Mirror Reversal* techniques. An *AlexNet* architecture was set up to classify the images as *Normal*, *Bulge*, *Protrusion* and *Extrusion*. The model achieved accuracy, sensitivity, and specificity of 87.75%, 86.5% and 94.75%, respectively.

*Mbarki* et al. [18] addressed the challenges in axial lumbar disc herniation recognition and emphasised the importance of disc segmentation in herniation diagnosis. Using CNN with VGG16 architecture, an automatic system was developed to detect disc herniation. MR images of 450 patients from the Sahloul University Hospital of Souss were processed to train and test different models. The U-net model achieved the best Intersection over Union(IoU) results of 93.3%.

*Aggarwal* [19] used ten different machine learning (ML) and deep learning (DL) classification systems to predict the presence of low back pain or the absence of this condition. This study investigated the radiological risk factors for this condition and used AI modelling to interpret the relationships. The MRI data of 63 individuals were used with 27 having chronic low back pain, and 36 being asymptomatic. Results showed that disc heights and ligamentum flavum hypertrophy were statistically significant in predicting the symptoms of low back pain. Also, the logistic regression classification model (the best-performing model) used quantitative measures from lumbar imaging to predict the condition. The model achieved accuracy, precision and sensitivity of 0.806, 0.815, and 0.806, respectively.

*Guinebert* et al. [20] developed an automated tool using CNN for segmenting and detecting lumbar disc abnormal conditions in MRI scans. A Picture Archiving and Communication System (PACS) with a *DICOM viewer* was created to extract training data and implemented two CNNs for segmentation and analysis. Training the CNNs with 244 MRI scans, the research achieved high accuracy in semantic segmentations, with DICE index scores of 0.96 for intervertebral discs and 0.93 for vertebral bodies. However, the networks were less effective in detecting common pathologies, achieving an area under the precision-recall curve of 0.88 for fractures and 0.76 for degenerative disc disease. The research believed expanding the training dataset would enhance performance in

future iterations.

*Alsmirat* et al. [21] built two CNN-based CAD systems to diagnose lumbar disc herniation from MRI axial scans using 164 images of dimension $512 \times 512$. The first system detects herniation, while the second system recognises the type of herniation. The detection and recognition systems achieved accuracies of 95.65% and 91.38%, respectively.

These studies show substantial progress in applying AI to lumbar spine MRI, particularly in detecting disc herniation, segmenting spinal structures, and predicting low back pain pathology. Most approaches used CNN-based architectures such as AlexNet, VGG16, U-Net, and custom deep learning models, achieving strong performance in segmentation tasks and moderate to high performance in classification. However, a number of limitations emerge across these works. First, many datasets are institution-specific, which raises concerns regarding generalisability. Second, the predictive focus is predominantly on disc conditions rather than broader diagnostic categories. Third, performance tends to drop when models attempt to identify pathologies rather than anatomical structures, suggesting gaps in feature learning and robustness. Overall, while AI-enabled diagnosis has shown promise, the literature indicates the need for more accurate, generalisable, and clinically reliable models - particularly for conditions beyond disc herniation. This motivates further research into improved classification methods and more comprehensive diagnostic support systems for lumbar MRI.

Table 1 below compiles all the studies reviewed in this subsection.

### 2.2. Clinical importance classification of lumbar spine MR imaging using CNN

*Lu* et al. [22] developed an efficient approach for utilising expert knowledge in expansive archival data for automated lumbar spinal stenosis grading through deep learning. One of the three major contributions of this study was a multi-input, multi-task CNN for grading central canal and foraminal stenosis across axial and sagittal imaging series, guided by extracted report-derived labels. The study used a large dataset of 22,796 disc-levels from 4075 patients. Spinal stenosis was graded into clinical relevance of normal, mild, moderate and severe with achieved class accuracies ranging from 78% to 83%.

*Won* et al. [23] evaluated the feasibility of a computer-assisted spine stenosis grading system by comparing the diagnostic agreement between two experts and AI CNN classifiers. 542 L4-5 axial MR images were graded by two experts and two CNN classifiers trained with each expert's grading labels. The grading agreements between the first expert and the corresponding model were 83% (accuracy) and 75.4% (F1 scores), while that of the second were 77.9% and 74.9%, respectively. The study concluded that automatic deep learning diagnosis might be feasible for spinal stenosis grading.

*Hallinan* et al. [24] developed a two-component CNN model for automated detection and classification of lumbar central canal, lateral recess, and neural foraminal stenosis. A dataset of 446 lumbar spine MRIs was used to train the model. Experienced radiologists graded the dataset into normal, mild, moderate, and severe clinical importance. The first component of the model, a CNN, was used to extract the region of interest (ROI) and the second component was used to classify the clinically relevant gradings. With internal test set classification using the four gradings, the model achieved *agreement scores* ranging from 69% to 82%.

*Bharadwaj* et al. [25] implemented an interpretable DL model for grading central canal stenosis, neural foraminal stenosis, and facet arthropathy from lumbar spine MRI. This study used a dataset of 200 T2-weighted axial MRI lumber spine scans. A CNN classifier was built to grade the images into normal, mild, moderate and severe. This classifier achieved a Cohen's kappa score of 0.54, translating to a moderate agreement ratio.

Across existing clinical importance classification studies, a consistent pattern emerges. While these CNN-based approaches demonstrate the feasibility of automating lumbar MRI grading, their predictive performance remains moderate and insufficient for clinical use. Most models struggle with accuracies between 54 and 83%. Collectively, these findings reveal both the progress made and the limitations that remain, highlighting the need for models that achieve higher reliability and robustness - particularly before these models can support radiologists in real-world diagnostic workflows.

Table 2 compiles all the studies reviewed in this subsection.

### 2.3. Spinal stenosis detection in lumbar spine MRI using CNN

*Al-Kafri* et al. [9] proposed a semantic segmentation and delineation of MRI scans using a segmentation network (SegNet). This would help clinicians detect spinal stenosis by segmenting and delineating important boundaries in lumbar axial-view MRI images. The study used a dataset containing 515 patients with symptomatic back pains. The study developed a ground truth dataset comprising labels of four regions of interest in the images. This dataset was then used to train and test classification models for segmentation. They found out that the model's performance falls within manual labelling outcomes. Also, the ground-truth dataset demonstrated an exceptional inter-rater agreement mean score of 0.92. Visual evaluation of the outcomes affirms the sufficient accuracy of the proposed method, making the results suitable for

**Table 1**
Review of AI-enabled diagnosis of lumbar spine MR images.

| Study | Objectives | Dataset | Methods Used | Results |
|---|---|---|---|---|
| *Salehi* et al. *(2019)* [1] | Using a deep CNN-based CAD system to diagnose three types of disc herniation conditions in lumbar axial MR images. | 2329 lumbar axial MR images, which was later augmented to 9316 images. | Employed an *AlexNet* architecture to classify the MR images into four categories: Normal, Bulge, Protrusion, and Extrusion. | Accuracy: 87.75%, Sensitivity: 86.5%, Specificity: 94.75%. |
| *Mbarki* et al. *(2020)* [18] | Recognition of disc herniation in axial lumbar disc | MR images from 450 patients. | A CNN with VGG16 architecture was implemented along with the U-net model for automatic disc herniation detection and segmentation. | U-net model IoU: 93.3%. |
| *Aggarwal (2021)* [19] | Using ML & DL to predict the presence of LBP. | MRI data of 63 individuals. | Ten ML and DL classification systems were built and tested, with the logistic regression model being identified as the most effective. | Accuracy: 80.6%, Precision: 81.5%, Sensitivity: 80.6%. |
| *Guinebert* et al. *(2022)* [20] | Automated segmentation and detection of abnormal conditions in lumbar discs from MRI scans using CNN. | 244 MRI scans of lumbar discs. | Two CNNs were implemented—one for segmentation and the other for pathology analysis. | DICE index scores - for intervertebral discs: 0.96, for vertebral bodies: 0.93. Area under the precision-recall curve – for fractures: 0.88, for degenerative disc disease: 0.76. |
| *Alsmirat* et al. *(2022)* [21] | Developing CNN-based CAD systems for diagnosing lumbar disc herniation from MRI axial scans. | 164 lumbar MRI axial scans. | Two CNN-based systems were developed - first system for herniation detection and the second for recognition of herniation types. | Accuracy – for the detection system: 95.65%, for the recognition system: 91.38%. |

**Table 2**
Review of clinical importance classification of lumbar spine MRI using CNN.

| Study | Objectives | Dataset | Methods Used | Results |
|---|---|---|---|---|
| *Lu*et al. *(2018)* [22] | Automation of lumbar spinal stenosis grading using deep learning. | 22,796 disc-levels from 4075 patients. | A multi-input, multi-task CNN architecture to grade spinal stenosis into categories of normal, mild, moderate, and severe clinical relevance. | Accuracies:78% to 83%. |
| *Won*et al. *(2020)* [23] | Evaluation of the feasibility of a computer-assisted spine stenosis grading system. | 542 L4-5 axial MR images. | Comparing the diagnostic agreement between two experts and CNN classifiers. | Accuracies: First expert – 83%, second expert – 77.9%. F1 scores: First expert – 75.4%, second expert – 74.9%. |
| *Hallinan*et al. *(2021)* [24] | Automated detection and classification of lumbar central canal, lateral recess, and neural foraminal stenosis. | 446 lumbar spine MRIs. | A two-component CNN model was utilised. The first component was used to extract the region of interest (ROI) from the MRIs, and the second for classifying the extracted ROIs based on their clinically relevant gradings. | Agreement scores ranging from 69% to 82%. |
| *Bharadwaj*et al. *(2023)* [25] | Implementation of an interpretable DL model for grading central canal stenosis, neural foraminal stenosis, and facet arthropathy from lumbar spine MRI. | 200 T2-weighted axial MRI lumber spine scans. | A CNN classifier to grade the images into normal, mild, moderate and severe. | Cohen's kappa score of 0.54, translating to a moderate agreement ratio. |

CAD applications.

As a follow-up to the above two studies to automatically detect the presence of LSS, *Natalia* et al. [26] considered a location-determination algorithm. This algorithm was crucial in the LSS diagnosis procedure. The CNN boundary delineation method in the previous study was adopted, and the results were used to supply boundary points to the algorithm. The algorithm used axial-view images of the intervertebral discs from 515 patients within their Lumbar Spine MRI dataset. When tested, the algorithm reliably located all nine important points in the MR images.

While radiography is efficient at diagnosing LSS, promptly evaluating all suspected cases using MRI proves challenging when considering the modality's high cost and limited availability. To diagnose severe central LSS, *Kim* et al. [27] developed a CNN model to perform binary classification of radiographic findings into *LSS* and *normal*, and further utilise gradient-weighted class activation mapping (Grad-CAM) to evaluate radiological diagnostic features. A VGG19 model was trained with 12,442 sagittal lumbar MR images. By averaging 5-fold models, the model achieved a validation accuracy and an extra validation accuracy of 82.8% and 81.8%, respectively.

*Altun* et al. [28] proposed a deep learning-based classification model to diagnose LSS automatically. A VGG16 model was trained on a dataset, as used in Refs. [9] and [26], containing MR images of patients with and without LSS. The model achieved an accuracy of 87.70%.

Studies that focused on segmentation and boundary delineation for lumbar stenosis show strong localisation capabilities but limited exploration of end-to-end classification. Existing attempts at classification report accuracy values below 90%, particularly for the model based on sagittal MRI views - the same view used in this paper. These insights collectively indicate that while segmentation work provides a solid foundation, there is a clear opportunity to improve direct stenosis classification performance - an opportunity this study seeks to address.

Table 3 compiles all the studies reviewed in this subsection.

**Table 3**
Review of spinal stenosis detection in lumbar spine MRI using CNN.

| Study | Objectives | Dataset | Methods Used | Results |
|---|---|---|---|---|
| *Al-Kafri*et al. *(2019)* [9] | Semantic segmentation and delineation of MRI scans. | Axial-view lumber spine MRI images of 515 patients with symptomatic back pains. | Segmentation network (SegNet) classification model for segmenting and delineating important boundaries in lumbar axial-view MRI images. | Inter-rater agreement mean score of 0.92. |
| *Natalia*et al. *(2019)* [26] | Automatic detection of the presence of LSS. | Axial-view lumber spine MRI images of 515 patients with symptomatic back pains. | The CNN boundary delineation method was adopted, and the results were used to supply boundary points to a location-determination algorithm. | The algorithm reliably located all nine important points in the MRI. |
| *Kim*et al. *(2022)* [27] | Diagnosing severe central LSS | 12,442 sagittal-view lumbar MR images. | A binary classification VGG19 CNN model classifying radiographic findings into LSS and normal, and Grad-CAM to evaluate radiological diagnostic features. | Validation accuracy: 82.8%, Extra validation accuracy: 81.8%. |
| *Altun*et al. *(2023)* [28] | Diagnosing LSS using deep learning. | Axial-view lumber spine MRI images of 515 patients with symptomatic back pains. | Deep learning VGG16 model | Accuracy: 87.70%. |

### 2.4. Explainable diagnosis

As explained by *Dwivedi* et al. [29], explainable AIs (XAIs) are methods and techniques used to explain complex AI models to humans. This is to solve the "black box" problem in which AI models' developers may not be able to explain why or how an AI model is making a specific decision. Explaining the concept as it relates to the diagnosis of lumbar spine MRI, XAI can provide explanations for CNN models' diagnoses and detection of abnormalities in MR images. These explanations can help point out the specific regions in the MRI scans that led to the models' results. By understanding the regions or pixels an AI model uses to make decisions, radiologists can cross-check and potentially reduce false negatives, ensuring patients receive accurate diagnoses.

*Loh* et al. [30] analysed the state of explainable AI techniques used in healthcare, uncovering trends that can guide future research towards clinical integration of XAI for transparency and reliability. This

integration is needed because the lack of explainability and interpretability of AI models can limit the trust and adoption of AI in healthcare. To promote their adoption and improvement, it's essential to compile a selection of the most outstanding works, accompanied by simple and intuitive explanations of the models and their healthcare applications. So, systematic review of 99 studies from 2011 to 2022 was carried out. These studies applied various XAI techniques like SHAP, LIME, Grad-CAM etc. It was found that SHAP is the most widely used XAI method for identifying key clinical predictors for diseases or patient results, often paired with predictive ML models like XGboost. GradCAM is the leading XAI method for visually explaining medical images by highlighting critical areas with heatmaps. The use of XAI for lumbar spine MRI diagnosis was not reported.

With the need for XAI in MRI diagnosis, *Groen* et al. [31] performed a systematic review of XAI in CAD within radiology. They showed limited use of XAI, with its use in only 37% of deep learning diagnostic test studies. For lumbar spine MRI, only two research papers were found to have considered XAI for explaining pixels as they contribute to the model's results. *Jamaludin* et al. [32] looked into automatic pinpointing of classification evidence in spinal MRI. They built a CNN model to automatically predict radiological scores of sagittal MRI scans based on identified and localised pathologies in the images. These pathological regions were termed *evidence hotspots,* and heat maps of these hotspots were predicted for each score. The map highlights pathological areas in an image, with brighter areas indicating a stronger evidence for that region to influence the classification. This study is limited to only hotspots and doesn't explain to what extent other regions contribute positively or negatively.

*Dindorf* et al. [33] proposed a pathology-independent classifier for spinal posture data to detect a wider range of pathologies than existing pathology-dependent classifiers. The classifier was trained on data from healthy patients and patients with various pathologies, including back pain, spinal fusion, and osteoarthritis. The classifier achieved good classification results for patients with spinal fusion but was less accurate for subjects with back pain. The classifier was also able to provide explanations for its predictions using LIME.

Explainability has become increasingly important in clinical AI because of its role in building trust, supporting clinical decision-making, and reducing harmful diagnostic errors. However, XAI applications in lumbar MRI remain limited. In particular, no studies were found applying explainability methods to **clinical importance classification** or **spinal stenosis detection** models, nor evaluating the *faithfulness* of these methods through quantitative testing. This reveals a significant gap: the absence of transparent and quantitatively verifiable interpretability frameworks capable of explaining model predictions in a way that clinicians can trust.

Table 4 compiles all the studies reviewed in this subsection.

**Table 4**
Review of explainable diagnosis.

| Study | Objective | Methods Used | Results |
|---|---|---|---|
| *Jamaludin*et al. *(2017)* [32] | Automatic pinpointing of classification evidence in spinal MRI scans. | A CNN model was built to automatically identify and localize *evidence hotspots* within MRI scans, which were then used to generate heat maps highlighting pathological areas corresponding to specific radiological scores. | The generated heat maps provided visual representations of *evidence hotspots* influencing classification, with brighter areas indicating stronger evidence for specific classifications. |
| *Dindorf*et al. *(2021)* [33] | Pathology-independent classifier for spinal posture data to detect a wider range of pathologies. | Built a classifier to detect spinal fusion and utilised LIME to provide explanations for the classifier's predictions. | The use of LIME allowed for interpretable explanations of the classifier's predictions. |
| *Loh*et al. *(2022)* [30] | To know the state of explainable AI techniques used in healthcare, uncovering trends that can guide future research towards clinical integration of XAI for transparency and reliability. | Systematic review of 99 studies from 2011 to 2022 that applied various XAI techniques such as SHAP, LIME, and Grad-CAM. | SHAP is the most widely used XAI for identifying key clinical predictors for diseases. GradCAM is the leading XAI method for visually explaining medical images. |
| *Groen*et al. *(2022)* [31] | To show the use of XAI for CAD in radiology. | Systematic review of XAI in CAD within radiology. | Use of XAI in only 37% of DL diagnostic test studies in radiology, indicating a relatively low incorporation of XAI techniques. For lumbar spine MRI, only two research papers were found to have considered XAI. |

## 3. Methodology

This section provides a comprehensive outline of the methodology followed to implement the models, as illustrated in Fig. 1.

### 3.1. Dataset

The dataset employed in our proposed models, which is opensource, comprises T2-weighted (T2W) sagittal MR images from 515 patients that experienced symptomatic back pain [34]. Both T1W and T2W images are useful for image classification. However, T2W images are better for visualising soft tissues, such as the intervertebral discs, nerves and spinal cord. In T2W images, intermediate signal intensity is demonstrated between the spinal cord and nerves, resulting in a maximal contrast between the cerebrospinal fluid (CSF) and the neural tissue [35]. So, T2W images have been used in this research paper because the research is about degeneration of the intervertebral discs which could cause stenosis/narrowing or irritation of the neural canal or nerves.

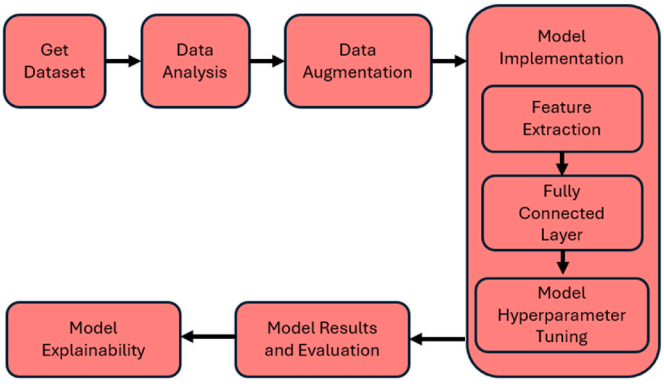


**Fig. 1.** Methodology flowchart for the AI-enabled Lumbar Spine MRI diagnosis.

These scans, as shown in Fig. 2, are sagittal images of the patients' lumbar spines obtained from the open source [34]. The lower end of the lumbar spine, where LBP occurs, includes the last three lumbar vertebrae (L3, L4, L5), the first sacral vertebrae (S1), and the discs in between

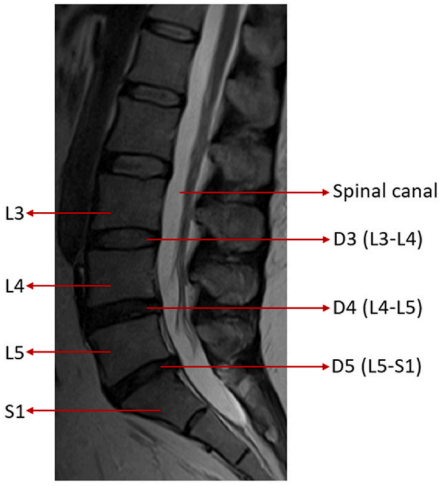


**Fig. 2.** A T2W mid-sagittal view MRI scan of lumber spine in the dataset.

them. These discs are D3, D4 and D5 between vertebrae L3-L4, L4-L5 and L5-S1, respectively.

The image data required some extraction process. Each patient has multiple sagittal slices (between 15 and 32 slices). A single mid-sagittal slice was selected for each patient. This selection process was conducted using a *MicroDicom viewer* by viewing and identifying the appropriate slice. The image was then extracted by snipping out the relevant part of the slice. This process was repeated for all patients in the dataset, thereby preparing the data for interpretation and analysis.

### 3.2. Data analysis

This section analyses the findings in the dataset for *Clinical Importance* and *Spinal Stenosis* classifications.

#### 3.2.1. Clinical importance classification

Table 5 shows various spinal conditions (in the report) classified based on their level of clinical importance (from *No* to *Severe* conditions). The imbalance in the number of images for each class was addressed during data augmentation discussed in Section 3.3.

Fig. 3 shows the counts of all the images in each class.

#### 3.2.2. Spinal stenosis classification

All conditions that lead to narrowing (stenosis) of the spinal canal were classed under *Stenosis* and those that don't were classed under *No Stenosis*. These conditions (listed below) include the two main categories of spinal stenosis as explained in Ref. [11] and pointed out in the first paragraph of Section 1. These conditions fall under serious and severe clinical importance.

- Compression of the thecal sac (central stenosis).
- Compression of the exiting nerve roots (lateral recess/foraminal stenosis).

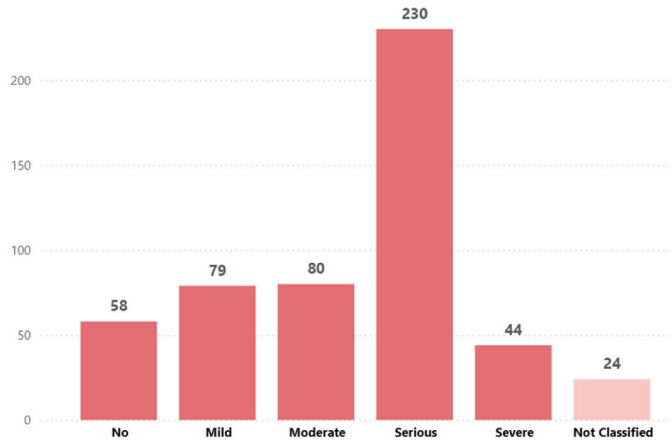


**Fig. 3.** Count of MR images by clinical importance.

- Sequestered disc (causing compression of the spinal canal – which can cause any of the two main spinal stenosis categories).

Fig. 4 shows the counts of all the images in each class.

### 3.3. Data augmentation

Image data augmentation was done to increase the data size in each class by using the *ImageDataGenerator* in *Keras*. This was to improve the models' predictive accuracies and generalisabilities [36]. So, augmentation parameters that would preserve the images' standard presentation were applied [13]. For example, the images were rotated by 5° instead of 20°, as standard MR images are not normally presented 20° tilted. The augmentation parameters considered are.

- ***rotation_range***: This parameter sets the range of angles by which the images are rotated. A value of 5 means that the images can be rotated

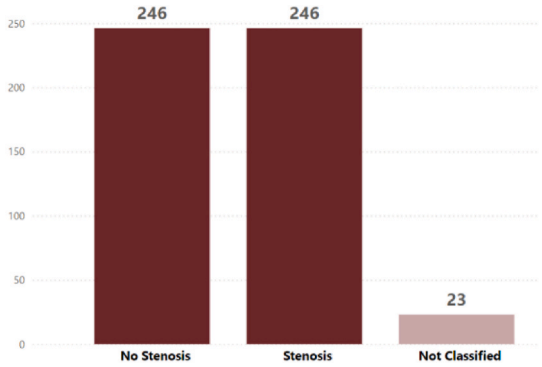


**Fig. 4.** Count of MR images by presence of spinal stenosis.

**Table 5**
Clinical importance classification of conditions in the radiologist report.

| No | Mild | Moderate | Serious | Severe |
|---|---|---|---|---|
| No evidence of any clinical condition | - Features of muscle spasm only.<br>- Small disc protrusion.<br>- disc abutting the thecal sac.<br>- Mild diffuse disc bulge.<br>- Dehydrated discal material.<br>- Posterior osteophytes | - Diffuse disc bulge only.<br>- Encroaching exit canal but no compression yet.<br>- Mild compression<br>- Abutting nerve root.<br>- Schmorl's nodes noted.<br>- Compression but adequate spinal canal. | - Narrowing of one or both neural foramina.<br>- Disc herniation.<br>- Thecal sac and exit canals compression.<br>- Annular tear.<br>- Facet joint hypertrophy.<br>- Spondylolisthesis. | - Sequestered disc.<br>- largely compressing thecal sac or/ and nerve roots.<br>- secondary neural canal stenosis. |

by any angle between −5° and 5°. This was only done to the sagittal images because the images can sometimes come slighted tilted.

- ***width_shift_range***: This parameter sets the fraction of the width of the image that is shifted. A value of 0.05 means that the images can be shifted by up to 5% of their width.
- ***height_shift_range***: This parameter sets the fraction of the height of the image that is shifted. A value of 0.05 means that the images can be shifted by up to 5% of their height.
- ***brightness_range:*** This parameter sets the brightness range applied to the images. A value of (0.8, 1.2) means that the brightness of the images can be multiplied by any value between 0.8 and 1.2.
- ***horizontal_flip***: This parameter sets the images to be randomly flipped horizontally. A value of *True* means that the images can be flipped horizontally with a probability of 50%. This was only done to the axial images because the anatomy of discs is generally quite symmetrical.
- ***fill_mode:*** This parameter controls the method used to fill in the gaps created by the transformations. Setting "nearest" means filling the gaps with the nearest pixel value.

The applied augmentation parameters settings are summarised as:

$$[rotation\_range = 5, width\_shift\_range = 0.05, height\_shift\_range = 0.05, brightness\_range = (0.8, 1.2), fill\_mode='nearest']$$

Tables 6 and 7 are augmentation tables for each of the classifications in Section 3.2.

### 3.4. Model implementation using CNN

This section explains the implementation of the two models. These models are the *Clinical Importance model* and the *Stenosis Detection model*. These models were developed in the *Keras* environment.

#### 3.4.1. Data preprocessing

Each model's respective image data and labels were processed and resized to 300 × 150. These images were converted to NumPy arrays, and the image data was then split into an 80% training set and a 20% validation set.

#### 3.4.2. Models architectures

1. Clinical Importance model

The architecture in Fig. 5 was built for this model. The *Sequential* model is used, which allows the addition of one layer at a time. The first layer is a *Cropping2D* layer using the parameters *cropping= ((top_crop, bottom_crop), (left_crop, right_crop))*. This cropping targets the region of interest (ROI) which is the region covering the 3 lumbar discs and the spinal canal. The cropping parameter is set to *cropping= ((100, 30), (5, 50))*, which crops the input images at the top and bottom by 100 and 30 pixels, and on the left and right by 5 and 50 pixels, respectively. This layer removes unnecessary borders and potentially irrelevant regions, allowing the model to focus on the ROI.

**Table 6**
Augmentation table for clinical importance classes.

| Clinical Importance | Number of MR images | Augmentation factor[a] | Generated images | Total images |
|---|---|---|---|---|
| No | 58 | 55 | 3132 | 3190 |
| Mild | 79 | 40 | 3081 | 3160 |
| Moderate | 80 | 40 | 3120 | 3200 |
| Serious | 230 | 14 | 2990 | 3220 |
| Severe | 44 | 73 | 3168 | 3212 |

[a] *These factor numbers were selected to keep the total images balanced and above 3000 per class.*

**Table 7**
Augmentation table for spinal stenosis classes.

| Stenosis presence | Number of MR images | Augmentation factor[a] | Generated images | Total images |
|---|---|---|---|---|
| Stenosis | 246 | 20 | 4674 | 4920 |
| No stenosis | 246 | 20 | 4674 | 4920 |

[a] *These factor numbers were selected to keep the total images balanced and close to 5000 per class.*

Next is the convolutional layers with twelve layers comprising *Conv2D*, *MaxPooling2D*, and *Dropout* layers. The *Conv2D* layers have 64, 128, 256, and 512 filters, each with a kernel size of 3 × 3, a *'relu'* activation function, and 'same' padding. These layers are applied to the input MR image to detect various features like edges, textures, and patterns. The filters slide over the image with a 3 × 3 kernel, and the 'same' padding ensures the output retains the same spatial dimensions as the input. The 'relu' activation function introduces non-linearity, allowing the network to learn complex patterns. The *MaxPooling2D* layers have a pool size of 2 × 2 to reduce the spatial dimensions of the image by keeping the most important information (the features detected by the Conv2D layers) while reducing computational complexity. This helps the model focus on essential features and prevents overfitting. The *Dropout* layers have a rate of 0.2, meaning that 20% of the neurons will be dropped out during training.

The increasing depth of the convolutional layers (from 64 to 512) allows the model to learn and extract increasingly complex features from the image. Initial layers might capture simple patterns like edges, while deeper layers can recognise more complex structures inherent in the MR images. Also, *Dropout* can be crucial to MR images by providing a regularisation effect and preventing overfitting.

The *Flatten* layer is added next, which flattens the input so that it can be fed into the fully connected layers. Then three sequential *Dense* layers with 256, 125, and 64 neurons were added, each with a *'relu'* activation function. After extracting features using convolutional layers, the fully connected layers can integrate these features for classification decisions of the 5 classes at the output. Given the complexity and the nature of medical images, having a series of dense layers can help in making a more accurate decision.

Finally, the *Output* layer is added with five neurons and a *softmax* activation function shown is Equation (1). Given that this is a multi-class classification task with five potential classes, the *softmax* activation is used which ensures that the output from the network is a probability distribution across the classes. This is useful in medical imaging tasks, where understanding the confidence of a model's decision can be as vital as the decision itself.

$$\boldsymbol{Softmax}(\boldsymbol{x_i}) = \frac{\boldsymbol{e^x i}}{\sum_{j=1}^{5} \boldsymbol{e^x j}} \text{ for all i in j} = \{1, \ldots, 5\} \quad (1)$$

where:

$x_i$ = the input value *x* for class *i* before applying the *softmax* function - from the output layer, after the fully connected dense layers.

2. Stenosis Detection model

The architecture for this model in Fig. 6 is similar to that of the *Clinical Importance* model in Fig. 5 except for the following.

- **Cropping2D:** The parameters are set to *cropping= ((100, 30), (5, 50))* since the new ROI is the spinal canal.
- **Second Dense layer:** This was set to *Dense(units=128, activation='relu')*.
- **Output layer:** This was set to *Dense(unit=1, activation='sigmoid')*. The *Sigmoid* activation function, as shown in Equation (2), is used

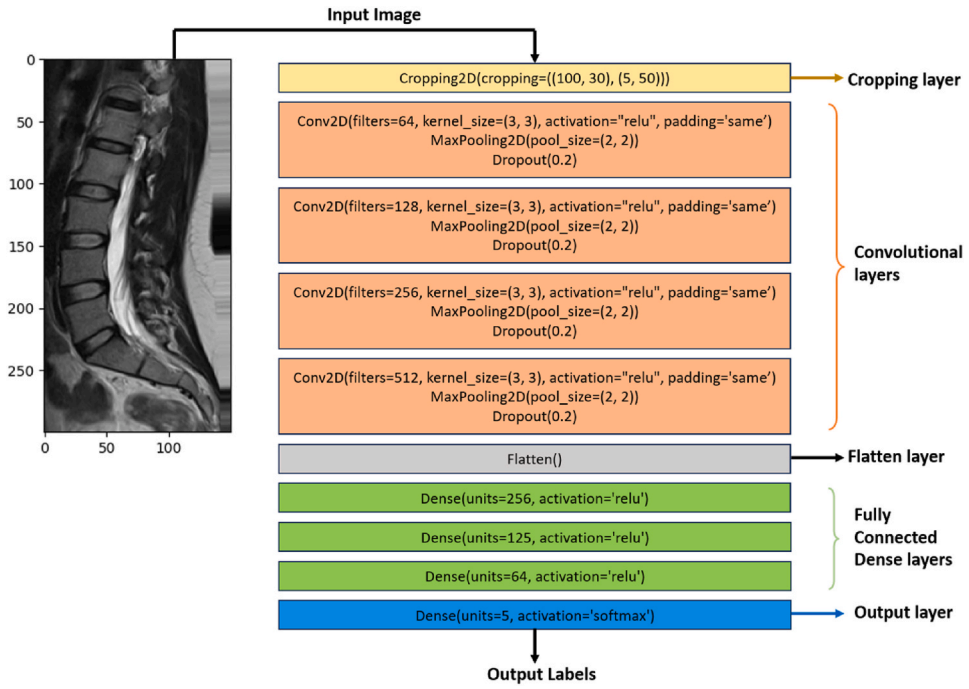


**Fig. 5.** Architecture for the Clinical Importance model.

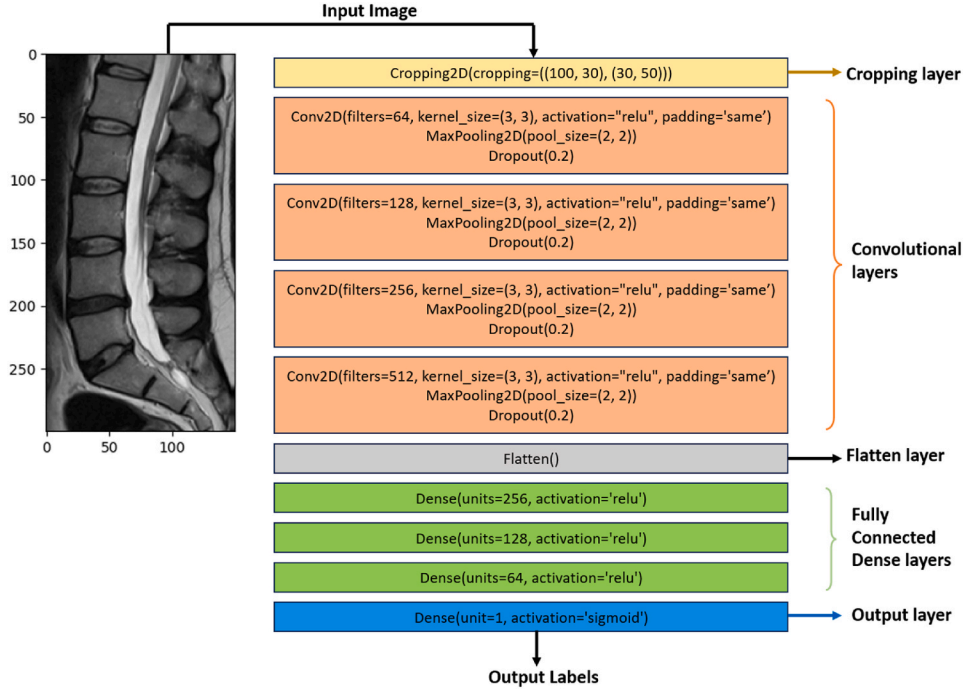


**Fig. 6.** Architecture for the Stenosis Detection model.

because this is a binary classification model. The function maps the output to a value between 0 and 1. The classification threshold is set by default at 0.5 with values above this threshold classed as *Stenosis* and values below classed as *No Stenosis*.

$$\boldsymbol{Sigmoid}(\boldsymbol{x}) = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}^{-\boldsymbol{x}}} \tag{2}$$

where:

$x$ = the input value before applying the *sigmoid* function - from the output layer, after the fully connected dense layers.

### 3.4.3. Models training, testing, validation and evaluations

Both models were compiled using *RMSprop* optimiser (with a learning rate set to 0.001). The *RMSprop* optimiser can be particularly effective for datasets with non-stationary distributions, like MR images, where the data characteristics might vary across different slices or patients [37]. For loss function, *categorical cross-entropy* and *binary cross-entropy* were used for the *Clinical Importance* and the *Stenosis Detection* models, respectively.

The models were trained (i.e., *fit*) on the training set and validated on the validation test with *batch_size=128* and *epochs=30*. *Callback*

arguments were also incorporated by using *ModelCheckpoints* to monitor the validation accuracies and save only the best (i.e., save_*best_only*=*True* and *monitor*='*val_accuracy*'). The models are evaluated using classification reports and confusion matrices.

For additional external validation, an independent dataset of lumbar spine MRI scans was obtained from the RSNA 2024 Lumbar Spine Degenerative Classification repository [38]. The dataset consisted of DICOM (.dcm) images with radiologist labels provided in CSV format. From this source, 40 mid-sagittal images were extracted - 20 labelled as *stenosis* and 20 as *no/mild stenosis*. These were used to evaluate the generalisability of our *Stenosis Detection* model on unseen clinical data.

### 3.4.4. Models explainability

*Algorithm 1* below was implemented to show how image regions contribute to the models predictions using four XAI methods (LIME, Grad-CAM, SHAP and Integrated Gradients (IG).

ALGORITHM 1: Models Explainability using XAI (LIME, Grad-CAM, SHAP, IG)

INPUT:
- *img_array*: Array containing MR images
- *model*: Trained and saved model

OUTPUT:
- Visual explanations showing image regions contributing to model predictions.

ALGORITHM:
1. Import all required XAI modules (LIME, SHAP, TensorFlow/Keras, Grad-CAM utilities, Integrated Gradients).
2. Select an MRI image
   a. Choose *image_to_explain* from *img_array*.
   b. Convert to appropriate shape and data type.
3. Define a prediction function:
   a. Reshape the input image array.
   b. Return predictions using the *model*.
4. Generate four attribution maps:
   (I) LIME
   a. Initialise a LimeImageExplainer.
   b. Run explain_instance(.) with 1000 samples.
   c. Convert superpixel weights into a 2D heatmap.
   (II) Grad-CAM
   a. Identify the last convolution layer of the model.
   b. Compute gradients of the target class with respect to this layer.
   c. Weight and combine feature maps to form a 2D heatmap.
   (III) SHAP (DeepExplainer)
   a. Select a background subset of images.
   b. Compute SHAP values for image_to_explain.
   c. Aggregate across channels to produce a 2D heatmap.
   (IV) Integrated Gradients (IG)
   a. Define a baseline image (e.g., zeros).
   b. Compute gradients along interpolated steps from baseline to the input image.
   c. Average gradients and multiply by (image – baseline) to obtain a 2D map.
5. Normalise each heatmap
   a. Scale values to a standard range (e.g., 0–1 or −1–1 depending on the method).
6. Identify most positively contributing regions
   a. Threshold each attribution map is set at top 20%.
   b. Create a binary mask highlighting top-importance pixels.
7. Visualise XAI outputs for each method:
   a. The original MRI image.
   b. Highlighted important regions using contour or boundary overlays.
   c. A full heatmap showing pixel-wise or region-wise contributions.
8. Visualise and display:
   a. Display the original MR Image.
   b. Highlight the most positively contributing regions of the image.
   c. Show a heatmap representing each region's contribution.

*Algorithm 2* below was implemented to evaluate the faithfulness of the explanations in *Algorithm 1* using deletion test, insertion test, and Area Under the Curve (AUC) metrics.

1. **Deletion Test**: This test involves progressively deleting pixels in the image, starting with those identified as the most important by the XAI method. If the XAI method is faithful, removing its "most important" pixels should cause the model's confidence in the prediction to drop quickly [39]. Steeper decreases therefore indicate higher faithfulness [40].
2. **Insertion Test**: This test involves progressively inserting pixels back into a baseline image (e.g., a black or blurred image), starting with those identified as the most important. If the XAI method is faithful, adding its "most important" pixels should cause the model's confidence in the prediction to increase quickly [39]. Faster increases indicate a more faithful attribution [40].
3. **Area Under the Curve (AUC) for Deletion and Insertion Curves**: For both deletion and insertion curves, the AUC provides a summary metric of faithfulness. For *deletion curves*, smaller areas indicate that removing important pixels rapidly reduces the model's confidence, indicating that the explanation has accurately identified critical regions. For *insertion curve*, larger areas indicate that restoring important pixels quickly boosts the model's confidence, meaning the XAI method correctly highlights influential regions [39] [40]. Together, deletion and insertion AUCs offer a quantitative and comparable measure of how reliably each XAI method captures the features the model uses to make its decisions.

ALGORITHM 2: Faithfulness Evaluation of XAI Methods (Deletion, Insertion, AUC)

INPUT:
- *model:* trained classifier
- *XAI methods:* {LIME, Grad-CAM, SHAP, IG}
- *images:* 10 images from test set (for clinical importance classification model) and 10 images from validation set (for stenosis detection model)
- *true labels:* ground-truth classes

OUTPUT:
- Deletion and insertion curves.
- Mean AUC scores for each XAI method.

ALGORITHM:
1. For each image:
   a. Predict class probabilities with the model.
   b. Select the true class for confidence measurement.
2. For each XAI method:
   a. Generate its attribution map.
   b. Normalise the map to [0,1].
   c. Flatten and rank pixels by importance.
3. Deletion test:
   a. Start with the original image.
   b. Iteratively remove the most important pixels (replace with baseline).
   c. Record the model's confidence in the true class at each step.
4. Insertion test:
   a. Start with a baseline (blurred/constant) image.
   b. Iteratively restore the most important pixels.
   c. Record confidence at each step.
5. Aggregate curves across all images
   a. Average deletion and insertion curves for each XAI method.
6. Compute mean AUC metrics:
   a. Insertion AUC (the higher, the better)
   b. Deletion AUC (the lower, the better)
7. Compare XAI methods using AUCs and curve trends to determine faithfulness.

## 4. Results analysis and discussion

This section presents the analysis, and the discussion of the results got from running the models.

**Table 8**
Evaluation report for clinical importance model.

| *approx. Values* | Accuracy | Weighted average Recall | Weighted average Precision | Weighted average F1-score |
|---|---|---|---|---|
| Clinical Importance Classification | 0.97 | 0.97 | 0.97 | 0.97 |

### 4.1. Clinical Importance model

#### 4.1.1. Model evaluation

This model's evaluation report and confusion matrix are presented in Table 8 and Fig. 7, respectively.

#### 4.1.2. Model explainability

Fig. 8 presents qualitative explanations for an image in the test set, illustrating how different XAI methods highlight the regions that drive the model's decision-making. Across LIME, SHAP, and IG, the attributions consistently emphasise clinically relevant anatomical structures – including the spinal canal, L4, L5, S1 and D5 vertebral levels; and the regions in between them where compressions, disc herniations and other conditions could occur. In contrast, Grad-CAM highlights only a subset of these critical areas, capturing the decision-relevant zones less completely than the other methods.

In Fig. 9, the XAI *faithfulness* is evaluated using deletion and insertion tests performed on 10 test images, with each test conducted over 20 incremental steps. The resulting confidence curves are then summarised using the AUC metric for each method. Table 9 presents the mean AUC values across the steps. SHAP obtained the lowest deletion mean AUC (0.2164) and the highest insertion mean AUC (0.9194), indicating the strongest correspondence between its highlighted regions and the model's decision behaviour. LIME and IG showed intermediate performance, with similar insertion mean AUCs (0.8026 and 0.8045) and moderately low deletion AUCs. Grad-CAM produced the highest deletion mean AUC (0.4576) and the lowest insertion mean AUC (0.6704).

### 4.2. Stenosis detection model

#### 4.2.1. Model evaluation

This model's evaluation reports and confusion matrices are presented in Tables 10 and 11 and Figs. 10 and 11, respectively.

#### 4.2.2. Model explainability

Fig. 12 presents qualitative explanations for an image in the validation set, illustrating how different XAI methods highlight the regions that drive the model's decision-making. Across LIME, SHAP, and IG, the attributions consistently emphasise clinically relevant anatomical structures – including the spinal canal, L3, L4, L5, S1 and D5 vertebral levels; and the regions in between them where compressions leading to stenosis could occur. In contrast, Grad-CAM highlights only a subset of these critical areas, capturing the decision-relevant zones less completely than the other methods.

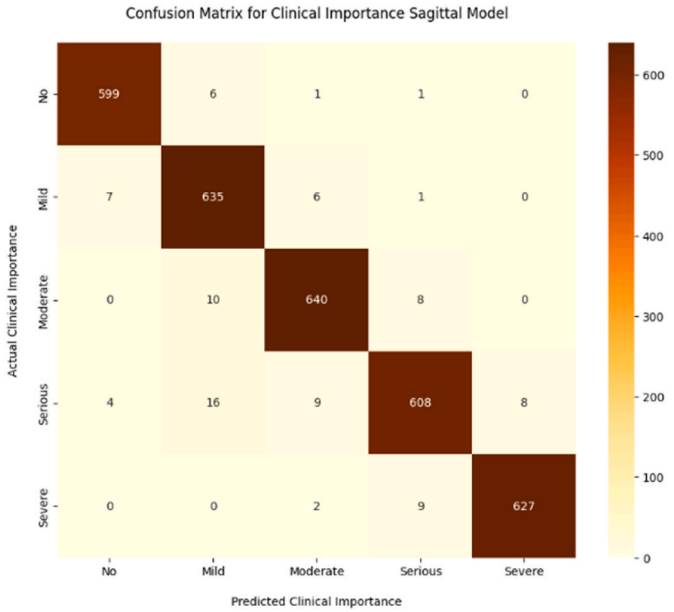


**Fig. 7.** Confusion matrix for clinical importance model.

In Fig. 13, the XAI faithfulness is evaluated using deletion and insertion tests performed on 10 test images, with each test conducted over 20 incremental steps. The resulting confidence curves are then summarised using the AUC metric for each method. Table 12 presents the mean AUC values across the steps. SHAP achieved the lowest deletion mean AUC (0.4315) and the highest insertion mean AUC (0.8874), indicating that removing or restoring its most influential regions produced the strongest and most consistent changes in model confidence. IG showed similar behaviour, with competitive deletion (0.4638) and insertion (0.8826) mean AUC values. LIME performed moderately well, yielding a slightly higher deletion mean AUC (0.6673) and a comparable insertion mean AUC (0.8752). Grad-CAM produced the weakest faithfulness scores, with the highest deletion mean AUC (0.7835) and the lowest insertion mean AUC (0.8573), suggesting that its highlighted regions were less aligned with the model's true decision-making process.

### 4.3. Discussion

#### 4.3.1. Performance and explainability of the models

Across both studies - clinical importance classification and stenosis detection, the proposed CNN models demonstrated substantially higher performance than previously published approaches. Earlier studies on clinical importance grading reported accuracies and agreement scores ranging from 54% to 83%, whereas our model achieved 97% accuracy and 97% recall, indicating a significant improvement in identifying clinically relevant categories. Similarly, prior research using sagittal MR images for lumbar spinal stenosis classification (such as [27]) achieved accuracies below 90%, with the best sagittal-view model reporting 82.8%. Our stenosis detection model reached 97% accuracy and 98% recall, representing a meaningful performance gain in a clinically important task.

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

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

where:

$TP = True\ Positive$, $TN = True\ Negative$, $FP = False\ Positive$, $FN = False\ Negative$.

The accuracies (using Equation (3)) of both models show that the overall correctness of the decision made by the model is 97%. The high recalls (using Equation (4)) of both models indicate very low *False Negatives*. In the clinical context of this research, it means that the *Clinical Importance* model rarely misclassifies the severity level of lumbar abnormalities, indicating the model's ability to mostly classify the 5 classes correctly. It also means that the *Stenosis Detection* model rarely misses actual LSS cases, indicating the model's ability to correctly identify almost all patients with spinal stenosis. This is important because, for example, the consequences of the model failing to identify/flag spinal stenosis patients are more serious compared to the model flagging a patient without stenosis as having stenosis.

However, performance alone is insufficient for real-world deployment. Clinical AI systems must also provide transparent and trustworthy explanations. To address this, four XAI methods - LIME, Grad-CAM, SHAP, and Integrated Gradients - were applied to both models. The qualitative explanations showed that, for most images, LIME, SHAP, and IG consistently highlighted anatomically relevant structures such as the spinal canal and adjacent vertebral levels (L3–S1), aligning with radiological expectations. Grad-CAM occasionally focused on broader or less precise regions, showing its limitation in anatomically precise localisation.

To quantitatively assess the reliability of these explanations, deletion

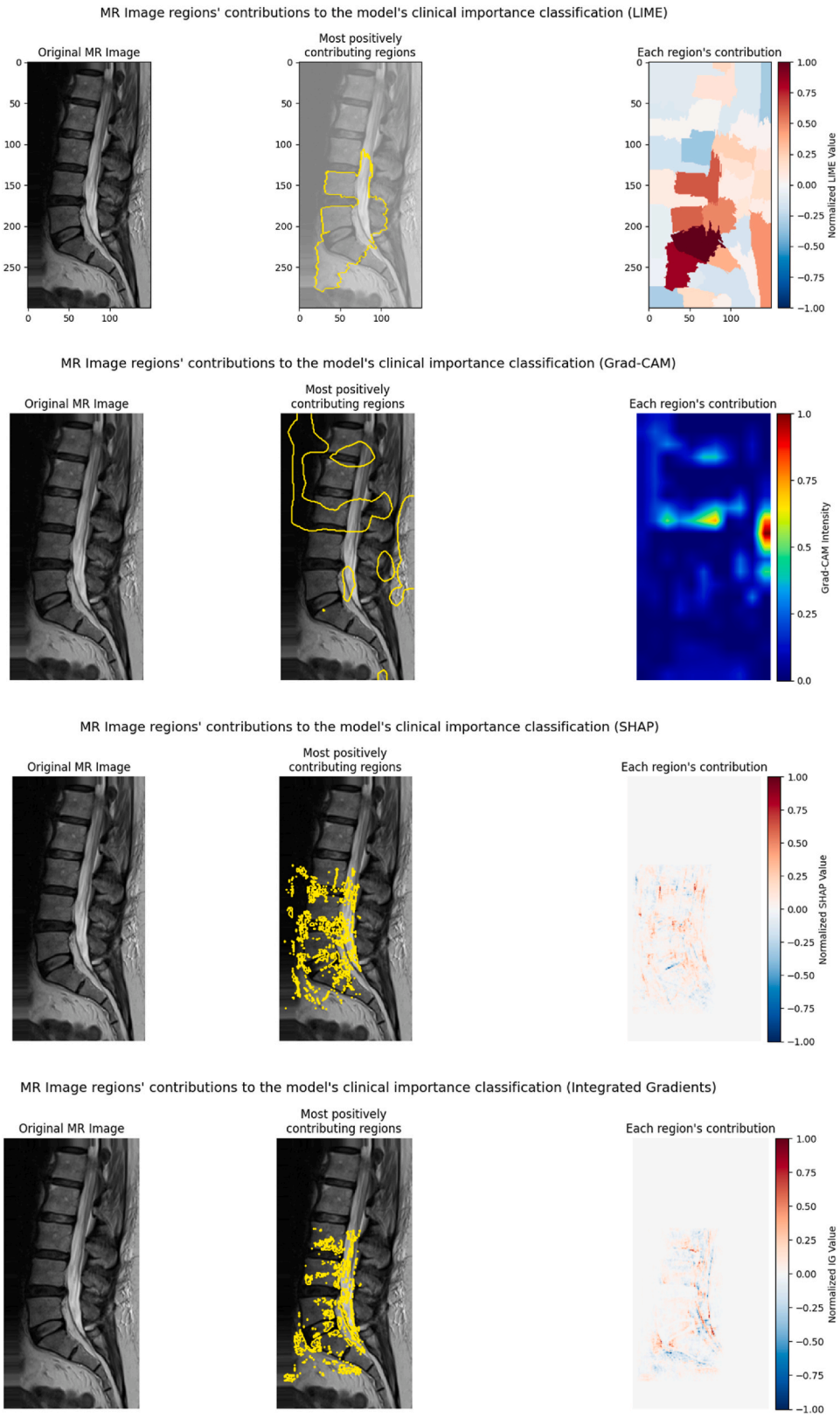


**Fig. 8.** Explainability for Clinical Importance model.

and insertion faithfulness tests were carried out to evaluate the fidelity of each explanation to the model's decision-making. For both models, SHAP consistently demonstrated the most faithful behaviour, producing the lowest deletion AUCs and highest insertion AUCs. IG and LIME showed moderate-to-high fidelity, with IG generally outperforming LIME in the *Stenosis Detection* model, whereas LIME showed slightly better faithfulness in the *Clinical Importance* model. In contrast, Grad-CAM was consistently least faithful. These quantitative results reinforce the value of multi-method XAI evaluation rather than relying on a single interpretability tool.

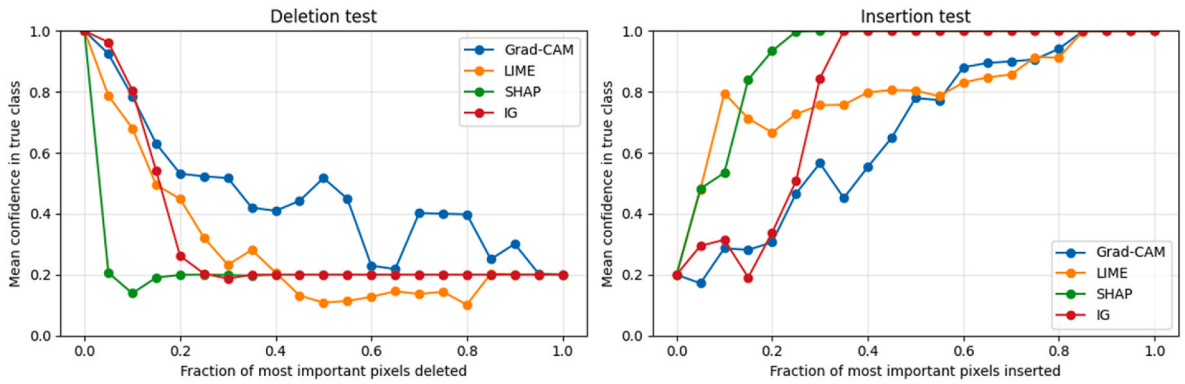


**Fig. 9.** Deletion and Insertion curves for the 4 XAI methods (Clinical Importance Model).

**Table 9**
Faithfulness evaluation of the XAI methods measured in terms of mean AUC (*Clinical Importance Model*).

| *XAI Method* | Deletion test (mean AUC) | Insertion test (mean AUC) |
|---|---|---|
| SHAP | 0.2164 | 0.9194 |
| LIME | 0.2832 | 0.8026 |
| IG | 0.3079 | 0.8045 |
| Grad-CAM | 0.4576 | 0.6704 |

**Table 10**
Evaluation report for stenosis detection model (test set).

| *approx. Values (test set)* | Accuracy | Weighted average Recall | Weighted average Precision | Weighted average F1-score |
|---|---|---|---|---|
| Spinal Stenosis Detection | 0.97 | 0.98 | 0.98 | 0.98 |

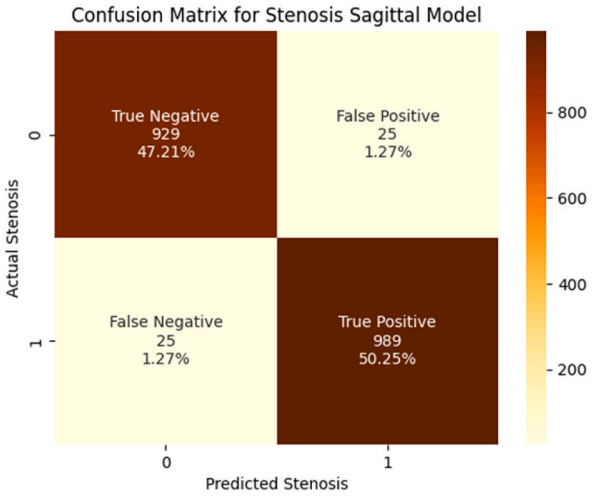


**Fig. 10.** Confusion Matrix for Stenosis Detection Model (test set).

*4.3.2. Clinical translation, limitations, ethical implications, and regulatory considerations*

Overall, the integration of high-performing models with robust explanation and faithfulness evaluation strengthens the reliability of the models towards clinical adoption. Despite the models showing impressive performances, limitations remain. The dataset, while adequate for experimentation, may not represent the full demographic and scanner variability seen in clinical practice, introducing risks of reduced generalisability. External validation on an independent dataset for the *Stenosis Detection* model showed promising results (88% in accuracy and recall), but still preliminary. Additionally, regulatory and clinical adoption require rigorous multi-centre testing, bias assessment, and model monitoring workflows, which were beyond the scope of this study. Addressing these limitations is essential before the models are safe for clinical adoption.

From an ethical perspective, deploying AI models in clinical settings raise concerns regarding algorithmic bias, over-reliance on automated outputs, and the potential for incorrect predictions to influence clinical decisions. There is also the risk of "right prediction with the wrong reasoning" where the model's prediction appears correct but is based on irrelevant image features. Faithfulness testing helps mitigate this risk, but continuous monitoring and radiologist oversight remain essential. Ensuring that AI tools assist rather than replace human judgement is critical for safe adoption.

Finally, regulatory compliance is essential for any clinical AI system. Under the new European Union (EU) AI Act (*Regulation (EU) 2024/1689*) [41], medical diagnostic AI is classified as a high-risk application

**Table 11**
Evaluation report for sagittal stenosis detection model (validation set).

| *approx. Values (validation set)* | Accuracy | Weighted average Recall | Weighted average Precision | Weighted average F1-score |
|---|---|---|---|---|
| Spinal Stenosis Detection | 0.88 | 0.88 | 0.88 | 0.87 |

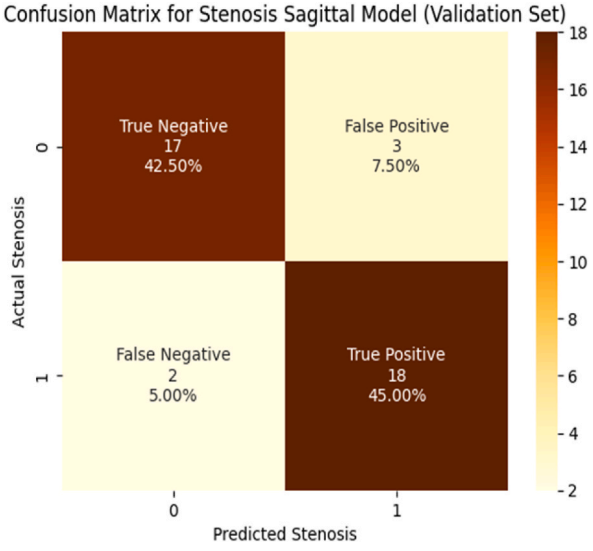


**Fig. 11.** Confusion Matrix for Stenosis Detection Model (validation set).

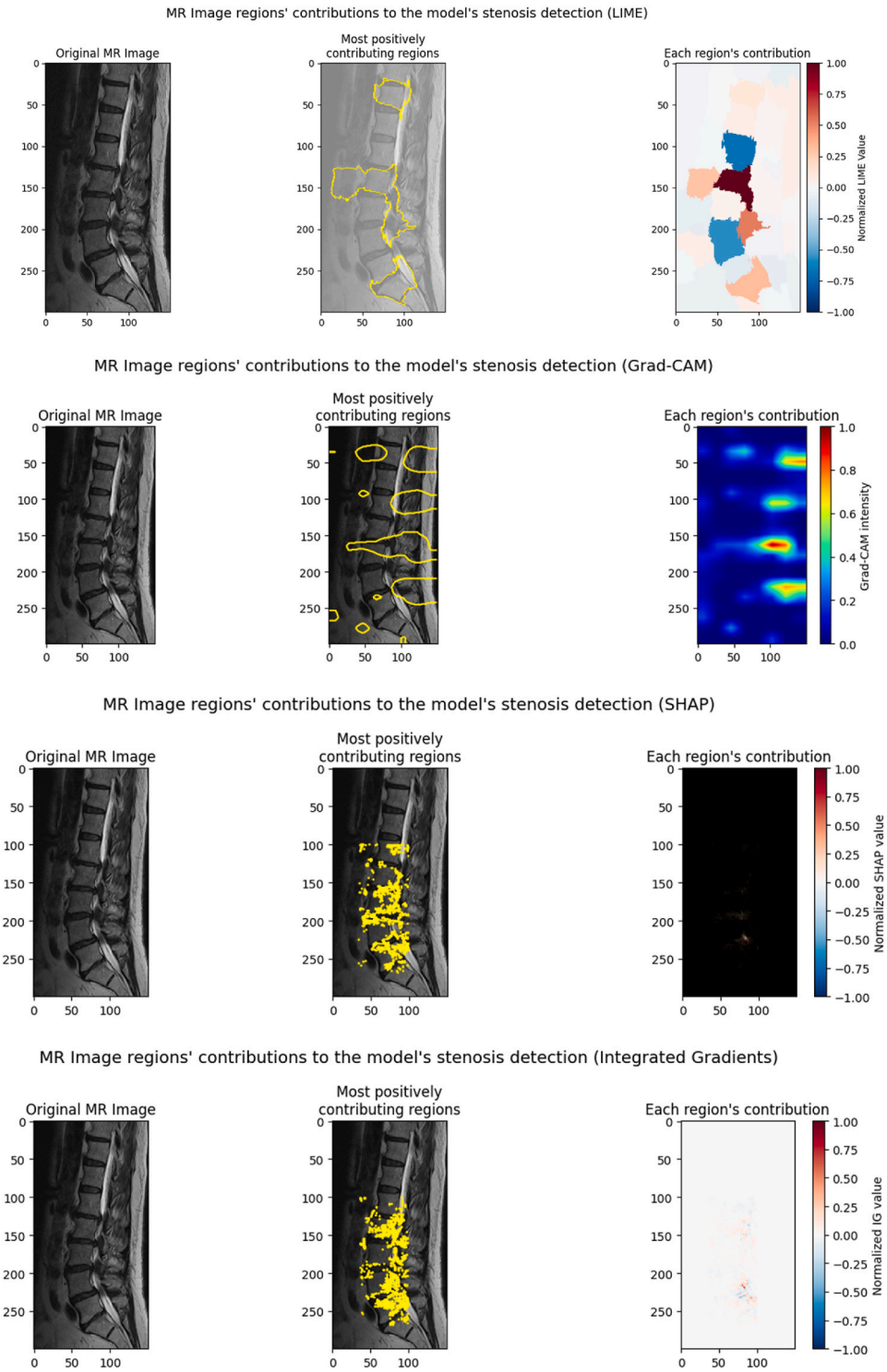


**Fig. 12.** Explainability for Stenosis Detection model.

and is therefore subject to stringent obligations, including comprehensive technical documentation, risk-management procedures, demonstrable transparency, post-market monitoring, bias assessment, and mandatory human oversight. These requirements emphasise that AI systems used in healthcare must not only be accurate but also interpretable, robust, and accountable. Similarly, in the United Kingdom, the Medicines & Healthcare products Regulatory Agency (MHRA) classifies AI-driven diagnostic software as a medical device, requiring manufacturers to follow a risk-based regulatory pathway involving clinical evaluation, lifecycle management, post-market surveillance, change-control plans, and clear evidence that the system is safe, and supports - rather than replaces - clinical judgement [42]. The MHRA's strategic approach to AI also highlights fairness, transparency, and human-factors considerations as essential components for regulatory clearance.

The models presented in this study incorporate multiple explainability mechanisms and quantitative faithfulness evaluation, aligning with these emerging regulatory expectations across the EU and UK. Nonetheless, achieving full regulatory readiness will require larger multi-centre validation, prospective clinical testing, robustness

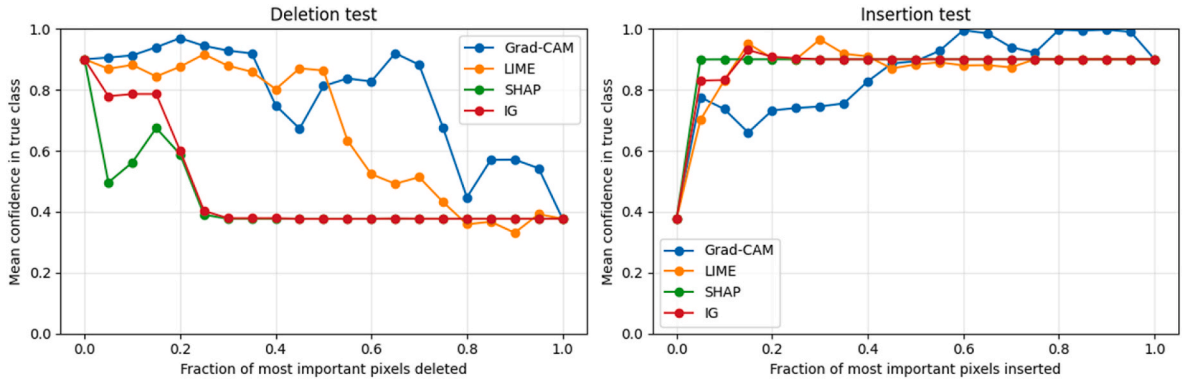


**Fig. 13.** Deletion and Insertion curves for the 4 XAI methods (Stenosis Detection Model).

**Table 12**
Faithfulness evaluation of the XAI methods measured in terms of mean AUC (*Stenosis Detection Model*).

| *XAI Method* | Deletion test (mean AUC) | Insertion test (mean AUC) |
|---|---|---|
| SHAP | 0.4315 | 0.8874 |
| IG | 0.4638 | 0.8826 |
| LIME | 0.6673 | 0.8752 |
| Grad-CAM | 0.7835 | 0.8573 |

verification across scanners and acquisition protocols, and formal conformity assessments before safe deployment in real-world clinical environments.

## 5. Conclusion

This paper explored the use of CNN to enhance the clinical importance classification of LBP conditions and LSS detection in MRI scans - two important clinical tasks. Both models recorded impressive performance, achieving 97–98% accuracy and recall on their respective test sets. External evaluation using an independent dataset further confirmed the robustness of the stenosis detection model, which achieved 88% accuracy and recall - indicating good generalisability beyond the training environment.

A major contribution of this work is the integration of explainable AI techniques into the diagnostic pipeline. While prior studies have focused predominantly on classification accuracy, this research incorporates four XAI methods—LIME, Grad-CAM, SHAP, and Integrated Gradients—to provide transparency into model decision-making. Qualitative visualisations showed that most methods consistently highlighted clinically relevant anatomical regions, such as the spinal canal and adjacent vertebral levels. For the *Clinical Importance* model, SHAP demonstrated the strongest faithfulness, achieving the lowest deletion mean AUC (0.2164) and the highest insertion mean AUC (0.9194) among all XAI methods. For *Stenosis Detection* model, SHAP also produced the most faithful explanations, achieving the lowest deletion AUC (0.4315) and highest insertion AUC (0.8874). This indicates that SHAP has the strongest alignment between its highlighted regions and the model's true decision behaviour.

These findings reinforce that clinical AI systems must not only perform well but must also demonstrate trustworthy and clinically grounded reasoning. A correct prediction derived from irrelevant or misleading features poses significant risk in medical settings. Thus, the combination of high predictive performance and faithful interpretability is essential for safe and responsible clinical deployment of these models. Although the results are promising, further work remains. Broader validation across multi-centre datasets, radiologist-in-the-loop evaluation, and stress-testing under varied imaging conditions will be necessary to ensure robustness, regulatory readiness, and real-world clinical safety. With continued refinement, the models developed in this study could contribute meaningfully to AI-assisted diagnosis of lumbar spine disorders.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

I have shared the DOI to the data

## References

[1] E. Salehi, S. khanbare, H. Yousefi, H. Sharpasand, O.S. Sheyjani, Deep convolutional neural networks for automated diagnosis of disc herniation on axial MRI, 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT) (2019) 1–6, https://doi.org/10.1109/EBBT.2019.8741895.
[2] P. Shokri, M. Zahmatyar, M.F. Tafti, M. Fathy, M.R. Tolzali, A.G. Jolfayi, S. A. Nejadghaderi, M.J.M. Sullman, A. Kolahi, S. Safiri, Non-spinal low back pain: global epidemiology, trends, and risk factors, Health Sci. Rep. 6 (9) (2023) 1533, https://doi.org/10.1002/hsr2.1533.
[3] G. Coates, P. Clewes, C. Lohan, H. Stevenson, R. Wood, T. Tritton, R.D. Knaggs, A. J. Dickson, D.A. Walsh, Chronic low back pain with and without concomitant osteoarthritis: a retrospective, longitudinal cohort study of patients in England, Int. J. Clin. Pract. (2023), https://doi.org/10.1155/2023/5105810.
[4] J. Kigozi, K. Konstantinou, R. Ogollah, K. Dunn, L. Martyn, S. Jowett, Factors associated with costs and health outcomes in patients with back and leg pain in primary care: a prospective cohort analysis, BMC Health Serv. Res. 406 (2019) 19, https://doi.org/10.1186/s12913-019-4257-0.
[5] A.Y. Wong, J. Karppinen, D. Samartzis, Low back pain in older adults: risk factors, management options and future directions, Scoliosis. Spinal Disord. 12 (14) (2017), https://doi.org/10.1186/s13013-017-0121-3.
[6] A. Cheshire, M. Polley, D. Peters, D. Ridge, Patient outcomes and experiences of an acupuncture and self-care service for persistent low back pain in the NHS: a mixed methods approach, BMC Complement Altern Med 13 (300) (2013), https://doi.org/10.1186/1472-6882-13-300.
[7] S. Ghosh, V. Chaudhary, Supervised methods for detection and segmentation of tissues in clinical lumbar MRI, Comput. Med. Imag. Graph. 38 (7) (2014) 639–649, https://doi.org/10.1016/j.compmedimag.2014.03.005.
[8] S. Kalidindi, S. Gandhi, Workforce crisis in radiology in the UK and the strategies to deal with it: is artificial intelligence the saviour? Cureus 15 (8) (2023) 43866 https://doi.org/10.7759/cureus.43866.
[9] A.S. Al-Kafri, S. Sudirman, A. Hussain, D. Al-Jumeily, F. Natalia, H. Meidia, N. Afriliana, W. Al-Rashdan, M. Bashtawi, M. Al-Jumaily, Boundary delineation of MRI images for lumbar spinal stenosis detection through semantic segmentation using deep neural networks, Deep Learn. Comput. aided Medi. Diagn. 7 (2019) 43487–43501, https://doi.org/10.1109/ACCESS.2019.2908002.
[10] L. Wu, S. Munakomi, R. Cruz, Lumbar spinal stenosis, in: StatPearls, National Library of Medicine, 2024 [Internet], https://www.ncbi.nlm.nih.gov/books/NBK531493.
[11] N.E. Epstein, R.D. Hollingsworth, Nursing review section of surgical neurology international part 2: lumbar spinal stenosis, Surg. Neurol. Int. 8 (139) (July 2017), https://doi.org/10.4103/sni.sni_150_17.
[12] A.S. Al-Kafri, S. Sudirman, A.J. Hussain, P. Fergus, D. Al-Jumeily, H.A. Smadi, M. Khalaf, M. Al-Jumaily, W. Al-Rashdan, M. Bashtawi, Detecting the disc herniation in segmented lumbar spine MR image using centroid distance function,

in: 2017 10th International Conference on Developments in Esystems Engineering (Dese), Paris, France, 2017, pp. 9–13, https://doi.org/10.1109/DeSE.2017.20.
[13] R. Thangaraj, P. P, J. Ramakrishnan, N. R, S. Eswaran, A deep convolution neural network for automated COVID-19 disease detection using chest X-ray images, Healthcare. Anal. 4 (2023) 100278, https://doi.org/10.1016/j.health.2023.100278.
[14] S. Sudirman, A. Al-Kafri, F. Natalia, H. Meidia, N. Afriliana, W. Al-Rashdan, M. Bashtawi, M. Al-Jumaily, Radiologists notes for lumbar spine MRI dataset. https://doi.org/10.17632/s6bgczr8s2.2, 3 April 2019.
[15] C.J. Donnally, A. Hanna, M. Varacallo, Lumbar Degenerative Disk Disease, StatPearls [Internet] | National Library of Medicine, 2023. https://www.ncbi.nlm.nih.gov/books/NBK448134/.
[16] Cleveland Clinic, Lumbar spine, Cleveland Clinic [Online]. Available, https://my.clevelandclinic.org/health/articles/22396-lumbar-spine, 2023. (Accessed 20 August 2023).
[17] S.P. Koyyada, T.P. Singh, An explainable artificial intelligence model for identifying local indicators and detecting lung disease from chest X-ray images, Healthcare. Anal. 4 (2023) 100206, https://doi.org/10.1016/j.health.2023.100206.
[18] W. Mbarki, M. Bouchouicha, S. Frizzi, F. Tshibasu, L.B. Farhat, M. Sayadi, Lumbar Spine Discs Classification Based on Deep Convolutional Neural Networks Using Axial View MRI, Elsevier, 2020, https://doi.org/10.1016/j.inat.2020.100837.
[19] N. Aggarwal, Prediction of low back pain using artificial intelligence modeling, J. Med. Artif. Intell. 4 (2021), https://doi.org/10.21037/jmai-20-55.
[20] S. Guinebert, E. Petit, V. Bousson, S. Bodard, N. Amoretti, B. Kastler, Automatic semantic segmentation and detection of vertebras and intervertebral discs by neural networks, Comput. Meth. Program. Biomed. Update. 2 (2022), https://doi.org/10.1016/j.cmpbup.2022.100055.
[21] M. Alsmirat, N. Al-Mnayyis, M. Al-Ayyoub, A. Al-Mnayyis, Deep learning-based disk herniation computer aided diagnosis system from MRI axial scans, IEEE Access 10 (2022) 32315–32323, https://doi.org/10.1109/ACCESS.2022.3158682.
[22] J.-T. Lu, S. Pedemonte, B. Bizzo, S. Doyle, K.P. Andriole, M.H. Michalski, R. G. Gonzalez, S.R. Pomerantz, Deep spine: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning, Proceedings of Machine Learning Research 85 (2018) 1–16, in: http://proceedings.mlr.press/v85/lu18a.html.
[23] D. Won, H.-J. Lee, S.-J. Lee, S.H. Park, Spinal stenosis grading in magnetic resonance imaging using deep convolutional neural networks, SPINE 45 (12) (2020) 804–812, https://doi.org/10.1097/BRS.0000000000003377.
[24] J. Hallinan, L. Zhu, K. Yang, A. Makmur, D. Algazwi, Y. Thian, S. Lau, Y. Choo, S. Eide, Q. Yap, Y. Chan, J. Tan, N. Kumar, B. Ooi, H. Yoshioka, S. Quek, Deep learning model for automated detection and classification of central canal, lateral recess, and neural foraminal stenosis at Lumbar spine MRI, Radiology 300 (1) (2021) 130–138, https://doi.org/10.1148/radiol.2021204289.
[25] U.U. Bharadwaj, M. Christine, S. Li, D. Chou, V. Pedoia, T.M. Link, C.T. Chin, S. Majumdar, Deep learning for automated, interpretable classification of lumbar spinal stenosis and facet arthropathy from axial MRI, Eur. Radiol. 33 (2023) 3435–3443, https://doi.org/10.1007/s00330-023-09483-6.
[26] F. Natalia, H. Meidia, N. Afriliana, A. Al-Kafri, S. Sudirman, Methodology to determine important-points location for automated lumbar spine stenosis diagnosis procedure, in: ICIMH 2019: Proceedings of the 2019 International Conference on Intelligent Medicine and Health, July 2019, pp. 53–57, https://doi.org/10.1145/3348416.3348426.
[27] T. Kim, Y.-G. Kim, S. Park, J.-K. Lee, C.-H. Lee, S.-J. Hyun, C.H. Kim, K.-J. Kim, C. K. Chung, Diagnostic triage in patients with central lumbar spinal stenosis using a deep learning system of radiographs, J. Neurosurg. 37 (1) (2022) 104–111, https://doi.org/10.3171/2021.11.SPINE211136.
[28] S. Altun, A. Alkan, İ. Altun, LSS-VGG16 : diagnosis of lumbar spinal stenosis with deep learning, Clinic. Spine. Surg. 36 (5) (2023) 180–190, https://doi.org/10.1097/BSD.0000000000001418.
[29] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, R. Ranjan, Explainable AI (XAI): core ideas, techniques, and solutions, ACM Comput. Surv. 55 (9) (2023) 1–33, https://doi.org/10.1145/3561048.
[30] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022), Comput. Methods Progr. Biomed. 226 (2022) 1–21, https://doi.org/10.1016/j.cmpb.2022.107161.
[31] A.M. Groen, R. Kraan, S.F. Amirkhan, J.G. Daams, M. Maas, A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: limited use of explainable AI? Eur. J. Radiol. 157 (2022) https://doi.org/10.1016/j.ejrad.2022.110592.
[32] A. Jamaludin, T. Kadir, A. Zisserman, SpineNet: automatically pinpointing classification evidence in spinal MRIs, Med. Image Anal. 41 (2017) 63–73, https://doi.org/10.1016/j.media.2017.07.002.
[33] C. Dindorf, J. Konradi, C. Wolf, B. Taetz, G. Bleser, J. Huthwelker, F. Werthmann, E. Bartaguiz, J. Kniepert, P. Drees, U. Betz, M. Fröhlich, Classification and automated interpretation of spinal posture data using a pathology-independent classifier and explainable artificial intelligence (XAI), Sensors 21 (6323) (2021), https://doi.org/10.3390/s21186323.
[34] S. Sudirman, A.A. Kafri, F. Natalia, H. Meidia, N. Afriliana, W. Al-Rashdan, M. Bashtawi, M. Al-Jumaily, Lumbar spine MRI dataset, Mendeley Data (2019), https://doi.org/10.17632/k57fr854j2.2.
[35] G.-U. Kim, M.C. Chang, T.U. Kim, G.W. Lee, Diagnostic modality in spine disease: a review, Asian Spine J. 14 (6) (22 December 2020) 910–920, https://doi.org/10.31616/asj.2020.0593.
[36] S.G. Paul, A. Saha, M. Assaduzzaman, A real-time deep learning approach for classifying cervical spine fractures, Healthcare Analytics 4 (2023) 100265, https://doi.org/10.1016/j.health.2023.100265.
[37] Y. Zhang, Y. Wang, G. Luo, A new optimization algorithm for non-stationary time series prediction based on recurrent neural networks, Future Gener. Comput. Syst. 102 (2020) 738–745, https://doi.org/10.1016/j.future.2019.09.018.
[38] Radiological Society of North America, Kaggle - RSNA 2024 lumbar Spine degenerative classification. https://www.kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification/data, 2025. Accessed November 2025.
[39] Z. Cheng, Y. Wu, Y. Li, L. Cai, B. Ihnaini, A comprehensive review of explainable artificial intelligence (XAI) in computer vision, Sensor 25 (13) (2025) 4166, https://doi.org/10.3390/s25134166.
[40] M.N. Jahromi, S.M. Muddamsetty, A.S.S. Jarlner, A.M. Høgenhaug, T. Gammeltoft-Hansen, T.B. Moeslund, SIDU-TXT: an XAI algorithm for NLP with a holistic assessment approach, Natural Language Processing Journal 7 (2024) 100078, https://doi.org/10.1016/j.nlp.2024.100078.
[41] European Paliament, EU AI Act: First Regulation on Artificial Intelligence, European Union, 2025 [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.
[42] Medicines & Healthcare products, "Software and Artificial Intelligence (AI) as a Medical Device," UK Government , [Online]. Available: https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device?utm_source.