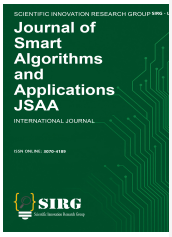


Journal of Smart Algorithms and Applications
JSAA
ISSN: 3070-4189/
Journal Homepage
https://pub.scientificirg.com/index.php/JSAA

A Comprehensive Review of Artificial Intelligence for Lumbar Spine MRI Analysis and Clinical Assessment

Marwa A. Abdelrahman [a,1] **, Ahmed Sayed Abd El Basset** [b]**, and Ahmed A. Elngar** [c]

[a] *Department of Radiology and Medical Imaging Technology, Faculty of Applied Health Sciences Technology, Beni-Suef University, Beni-Suef City, 62511, Egypt; E-mail: Marwa.AbdELNabi@fahst.bsu.edu.eg*

[b] *Department of Radiology, Faculty of Medicine, Beni-Suef University, Beni-Suef City, 62511, Egypt; E-mail: dr.ahmedbasset@gmail.com*

[c] *Department of Computer Science, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt; E-mail: elngar_7@yahoo.co.uk*

ABSTRACT

Lower back pain is one of the major diseases to disability worldwide, with lumbar intervertebral disc degeneration (IVDD), disc herniation (DH), and lumbar spinal stenosis (LSS) being the most common underlying causes. Magnetic resonance imaging (MRI) is the imaging modality of choice for assessing these conditions; however, conventional MRI interpretation continues to be time-consuming, subjective, and prone to considerable interobserver variability. Recent advances in artificial intelligence (AI), particularly deep learning and advanced AI frameworks, have shown strong potential for optimizing the accuracy, consistency, and efficiency of lumbar spine MRI analysis. This review offers an in-depth overview of AI applications for lumbar spine MRI analysis of IVDD, DH and LSS. It includes lumbar spine anatomy, pathophysiology of diseases, MRI grading systems and publicly available datasets including SPIDER, LumbarDISC and LSpineSMRI. AI methodologies are comprehensively reviewed across deep learning methods and advanced AI frameworks, covering feature–classifier hybrids, multi-stage functional cascades, cross-modal semantic fusion, and consensus ensemble systems, alongside widely adopted architectures such as CNNs, U-Net, YOLO, and transformer-based models for localization, segmentation, classification, and severity grading. Reported performance in the reviewed studies showed Dice similarity coefficients exceeding 0.90 for segmentation tasks and AUC values up to approximately 0.98 for lumbar spinal stenosis assessment. The review further discusses the key challenges to clinical translation, encompassing dataset heterogeneity, class imbalance, domain shift, limited external validation, and insufficient interpretability, while outlining future research priorities such as multimodal learning, self-supervised learning, foundation models, and explainable AI (XAI) to enhance model robustness, transparency, and clinical applicability. This review presents a clinically driven and technically structured perspective on the current landscape of AI for lumbar spine MRI analysis and identifies directions for developing more robust, interpretable, and clinically deployable diagnostic systems.

PAPER INFORMATION

HISTORY

Received: 13 May 2026
Revised: 25 June 2026
Accepted: 11 July 2026
Online: 28 August 2026

MSC

68T07; 68R10; 94A60; 68M15

KEYWORDS

Artificial Intelligence; Lumbar Spine MRI; Clinical Foundations; Deep Learning; Intervertebral Disc Degeneration.

[1]Corresponding author at Department of Radiology and Medical Imaging Technology, Faculty of Applied Health Sciences Technology, Beni-Suef University, Beni-Suef City, 62511, Egypt; E-mail: Marwa.AbdELNabi@fahst.bsu.edu.eg

1. INTRODUCTION

Low back pain (LBP) has become increasingly prevalent in recent decades. Lumbar intervertebral disc degeneration (IVDD) is one of its principal causes and is often associated with sciatica, which involves radiating leg pain resulting from irritation or compression of the sciatic nerve [1]. Degenerative disc disease encompasses a broad range of pathological changes, including disc bulging, disc herniation, spinal canal stenosis, neural foraminal narrowing, and facet joint arthropathy. These conditions may occur independently or in combination. As degeneration progresses, it can contribute to lumbar spinal stenosis and degenerative spondylolisthesis, producing symptoms that range from localized low back pain to radicular pain, numbness, and muscle weakness. Severe cases may result in neurological deficits [2]. Disc bulging is a circumferential extension of disc material beyond the apophyseal margins, usually involving more than 25% of the disc circumference and extending less than 3 mm [3]. Lumbar spinal stenosis (LSS), however, refers to a narrowing of the spinal canal and is often associated with pain, limited mobility, and difficulty with everyday activities [4]. Magnetic resonance imaging (MRI) is often considered the gold standard imaging modality for the assessment of intervertebral disc abnormalities and endplate degeneration due to its ability to provide high-resolution, noninvasive visualization of spinal soft tissues such as intervertebral discs, nerve roots, and the spinal cord. Conventional MRI assessment is mainly based on T1 and T2-weighted sagittal images to assess disc morphology, whereas axial and coronal views are often used for often diagnostic confirmation [5]. T2-weighted imaging (T2WI) is particularly valuable because it provides information about tissue water content and tissue relaxation properties. MRI contrast is based on T1 (longitudinal) and T2 (transverse) relaxation mechanisms that describe how nuclear spins return to equilibrium after excitation [6]. Although it has diagnostic value, the interpretation of MRI still has a considerable interobserver variability, with reported levels of agreement for common non-disc contour degenerative findings ranging from moderate to good (weighted $\kappa \approx 0.44$–0.74) [7]. This variation may affect clinical decisions and patient management. In addition, the diagnosis of low back pain remains challenging, emphasizing the need for objective and reproducible assessment methods [8]. Over the past few years, artificial intelligence (AI) has been increasingly used to improve spinal imaging analysis. AI methods can automate routine tasks, improve imaging workflows, and help radiologists make accurate diagnoses more efficiently. Currently, AI-based technologies have also been used to support diagnosis, predict disease progression, and estimate treatment outcomes, which can reduce clinicians' workload and allow faster clinical decision-making [9]. Deep learning models, particularly convolutional networks (CNNs), have shown high performance in identifying and assessing lumbar spinal stenosis and other degenerative spinal abnormalities [10]. Recent advances in artificial intelligence have expanded the use of automated methods for the analysis of medical images. Deep learning architectures can support several stages of the imaging workflow, including anatomical localization, lesion detection, image segmentation, and disease classification. Object detection models such as YOLOv8 are particularly useful for identifying regions of interest, while foundation segmentation models can refine these detections and generate more detailed anatomical boundaries. For example, Alabsi et al. combined YOLOv8 derived spatial prompts with the Segment Anything Model to perform automated brain tumor segmentation on MRI data [11]. Although that study focused on brain MRI rather than lumbar spine disorders, it illustrates how detection and segmentation models can be integrated into an automated MRI analysis pipeline. Similar strategies may be relevant to lumbar spine applications involving vertebral and disc localization, disc herniation detection, spinal canal segmentation, and the assessment of degenerative changes. However, their performance must be evaluated using pathology specific lumbar spine datasets and independently validated across different scanners, imaging protocols, and clinical populations. There has been great advancement in the use of AI for lumbar spine MRI analysis, but challenges like limited standardization of evaluation methods, dataset heterogeneity and poor generalizability across different clinical settings remain. Furthermore, the rapid development of AI techniques calls for a comprehensive and well-organized synthesis of the research already carried out. Therefore, this review provides a structured overview of AI applications in lumbar spine MRI analysis, focusing on intervertebral disc degeneration, disc herniation, and lumbar spinal stenosis. It integrates clinical knowledge of lumbar spine disorders (e.g. anatomy, pathophysiology, and commonly used grading systems (Pfirrmann, MSU) with publicly available MRI datasets, and recent advances in AI (e.g. deep learning, transformer-based architectures, explainable AI, and emerging AI frameworks). Moreover, the review also categorizes advanced AI frameworks according to their underlying computational approaches, which provides a coherent ground for comparing recent studies. A brief discussion about traditional machine learning methods is presented to provide a historical background of the development of AI in lumbar spine MRI analysis. The reviewed AI pipelines are compared based on their pre-processing, detection, segmentation, and classification techniques, with attention to their strengths, limitations, and performance in different lumbar spine disorders. In addition, this review addresses major challenges, including dataset heterogeneity, class imbalance, the lack of standardized evaluation frameworks, limited generalizability, and limited external validation, and highlights future research directions for advancing robust, clinically applicable AI systems. This review provides the following main contributions:

- It connects lumbar spine anatomy, disease pathophysiology, MRI grading systems, and recent AI developments in one integrated clinical–technical framework.
- It offers a structured overview of publicly available lumbar spine MRI datasets, grouped into foundational, benchmark, and clinical/morphometric categories, to support reproducible AI research.

- It introduces a new categorization of advanced AI frameworks based on their computational approach, distinguishing conventional deep learning from four advanced AI framework categories: feature–classifier hybrids, multi-stage cascades, cross-modal fusion, and consensus ensembles.
- It compares AI pipelines across localization, segmentation, classification, and grading tasks, with particular attention to their strengths, limitations, and potential clinical use.
- It highlights the main barriers to clinical translation, including dataset heterogeneity, domain shift, limited external validation, and poor interpretability, while also discussing future directions such as multimodal learning and explainable AI.

The rest of this paper is organized as follows: Section 2 describes the literature search methodology and study selection process. Section 3 covers the clinical background of lumbar disc diseases along with publicly available MRI datasets. Section 4 reviews recent related surveys in the field. Section 5 is devoted to MRI based grading systems and quantitative assessment methods. Section 6 presents deep learning techniques and advanced AI frameworks for lumbar spine MRI analysis. Section 7 discusses the main challenges, explainable AI considerations and future research directions. Finally, Section 8 presents the conclusions.

2. LITERATURE SEARCH METHODOLOGY

A structured literature search was performed in four major electronic databases: PubMed/MEDLINE, Scopus, Web of Science, and IEEE Xplore. The original search was performed in September 2025 and updated through May 2026 to include recently published studies. Articles published in peer-reviewed journals, conference proceedings and early-access journal papers between January 2018 and May 2026 were searched. The search strategies included Medical Subject Heading (MeSH) and keywords in the title and abstract combined with Boolean operators (AND, OR) to expand the search. The standardized search query used across the databases was as follows:

> *(“lumbar spine” OR “intervertebral disc” OR “disc herniation” OR “lumbar spinal stenosis” OR “disc degeneration” OR “Modic changes” OR “foraminal stenosis”) AND (“magnetic resonance imaging” OR “MRI” OR “T2-weighted” OR “T1-weighted”) AND (“artificial intelligence” OR “deep learning” OR “convolutional neural network” OR “CNN” OR “transformer” OR “foundation models” OR “explainable AI” OR “transfer learning” OR “automated grading”)*

The selection process involved several screening stages. A total of 318 records were retrieved from the four electronic databases. After removing 115 duplicate records, 203 unique records were screened by title and abstract. During this screening stage, 125 records were excluded because they were not relevant to the lumbar spine (n = 58), involved non-MRI modalities (n = 42), or represented other irrelevant AI or medical studies (n = 25). The remaining 78 reports were assessed for eligibility according to the predefined inclusion and exclusion criteria. Studies were eligible if they met the following criteria: (1) original peer-reviewed research published as journal articles or conference proceedings; (2) proposed or evaluated AI-based methods for automated analysis of lumbar spine MRI; (3) used lumbar spine MRI data; and (4) addressed one or more tasks, including detection, localization, segmentation, classification, ordinal severity grading, or quantitative/morphometric analysis, with quantitative results relevant to the evaluated task. Based on these criteria, 48 full-text articles were excluded: 20 did not meet the predefined eligibility criteria, 20 lacked extractable quantitative results, and 8 were non-English publications. Ultimately, 30 original studies met the eligibility criteria and formed the basis of this review. Of these, 20 studies that reported sufficiently detailed architectural and performance data underwent structured comparative tabulation, while the remaining 10 studies contributed supporting evidence and context throughout the narrative discussion. The screening and study selection process is illustrated in the PRISMA 2020 flow diagram presented in **Figure 1**.

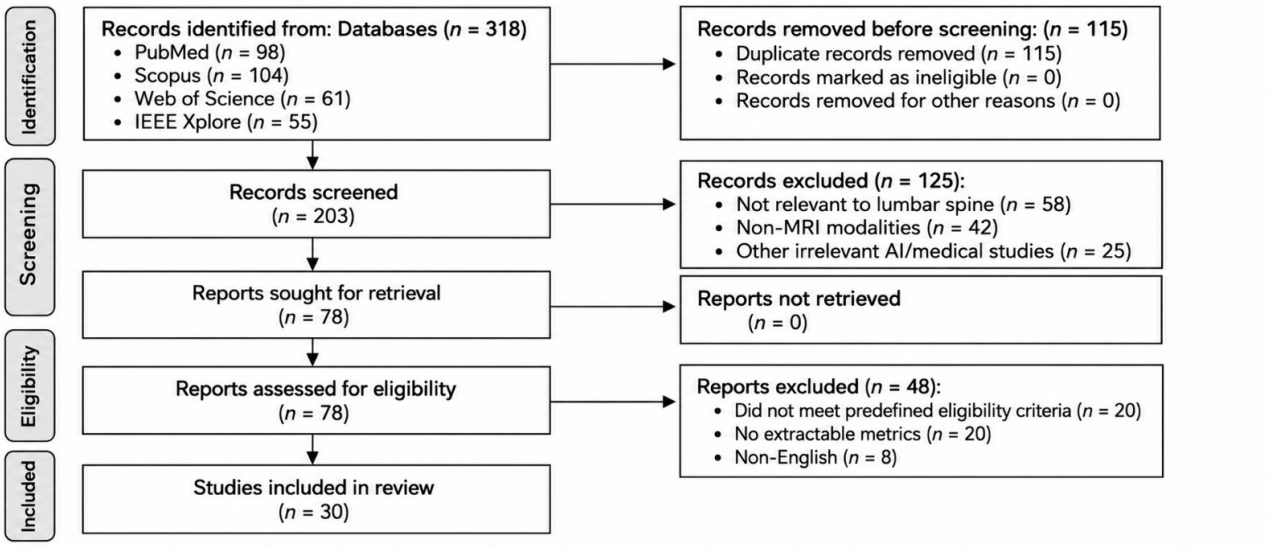


Figure 1: PRISMA Flow Diagram

2.1 Methodological Quality and Evidence Assessment

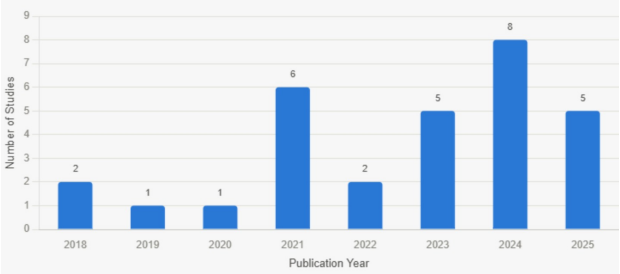
To support the interpretation of the reported performance metrics, the included studies were evaluated according to their methodological quality, dataset scale, and validation depth. The included studies showed considerable methodological heterogeneity across the evaluated AI frameworks. First, dataset sizes varied considerably across studies. While a few benchmark studies used large-scale or multi-institutional repositories, many studies relied on small, single-center patient cohorts (fewer than 150 MRIs). This can increase the risk of selection bias and reduce the clinical representativeness of complex spinal pathologies. Second, the quality of validation varied considerably. Most frameworks used only internal k-fold cross-validation or a simple train–test split based on data collected from the same scanner. Prospective external validation on geographically independent cohorts was rarely reported. As a result, giving equal weight to all findings may overestimate diagnostic reliability. It is important to distinguish between early proof-of-concept models developed using homogeneous single-center data and multicenter frameworks that generalize across different clinical settings [12, 13]. The principal elements of the literature search, including the databases, search period, search query, and inclusion and exclusion criteria, are summarized in **Table 1**.

Table 1: Summary of the Literature Search and Selection Strategy

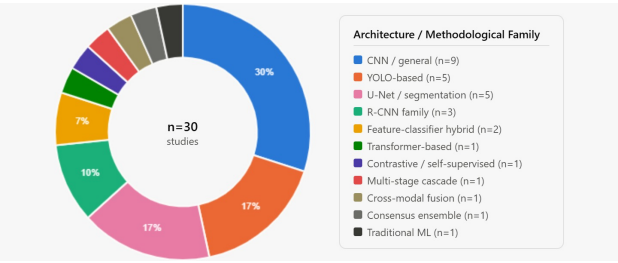
Parameter	Details
Databases	PubMed/MEDLINE, Scopus, Web of Science, IEEE Xplore
Search Date	Initial search: September 2025; updated through May 2026
Time Period	January 2018 – May 2026
Search Query	*(“lumbar spine” OR “intervertebral disc” OR “disc herniation” OR “lumbar spinal stenosis” OR “disc degeneration” OR “Modic changes” OR “foraminal stenosis”) AND (“magnetic resonance imaging” OR “MRI” OR “T2-weighted” OR “T1-weighted”) AND (“artificial intelligence” OR “deep learning” OR “convolutional neural network” OR “CNN” OR “transformer” OR “foundation models” OR “explainable AI” OR “transfer learning” OR “automated grading”)*
Inclusion Criteria	Studies were included if they: presented original peer-reviewed research published as journal articles or conference proceedings; proposed or evaluated AI-based methods for automated analysis of lumbar spine MRI; addressed one or more tasks, including detection, localization, segmentation, classification, ordinal severity grading, or quantitative/morphometric analysis; and reported quantitative results relevant to the evaluated task.
Exclusion Criteria	Duplicate publications; non-peer-reviewed studies; non-MRI studies; studies unrelated to lumbar disc abnormalities; non-English publications; and studies lacking quantitative performance evaluation.

2.2 Bibliometric Characteristics of Included Studies

To provide readers with a clearer overview of the field, the bibliometric characteristics of the 30 included studies were also summarized in terms of publication trends and the distribution of architectural/methodological families. **Figure 2a** shows an increase in publication output from 2021 onward, peaking in 2024 (n = 8), in line with the growing use of deep learning in medical imaging in recent years. **Figure 2b** summarizes the architectural distribution across the reviewed studies: conventional CNN-based/general architectures remained the most frequently used approach (30%), followed by YOLO-based detectors and U-Net/segmentation models (17% each), while advanced hybrid AI frameworks – feature–classifier hybrids, multi-stage cascades, cross-modal fusion, and consensus ensembles – accounted for a smaller but growing share (approximately 17%) of the reviewed literature, indicating that they are still less common than the more established approaches.



(a)



(b)

Figure 2: Bibliometric overview of the 30 included studies

3. DISC DISEASES AND MRI DATASETS

3.1 Anatomy and Pathophysiology of Lumbar Disc Diseases

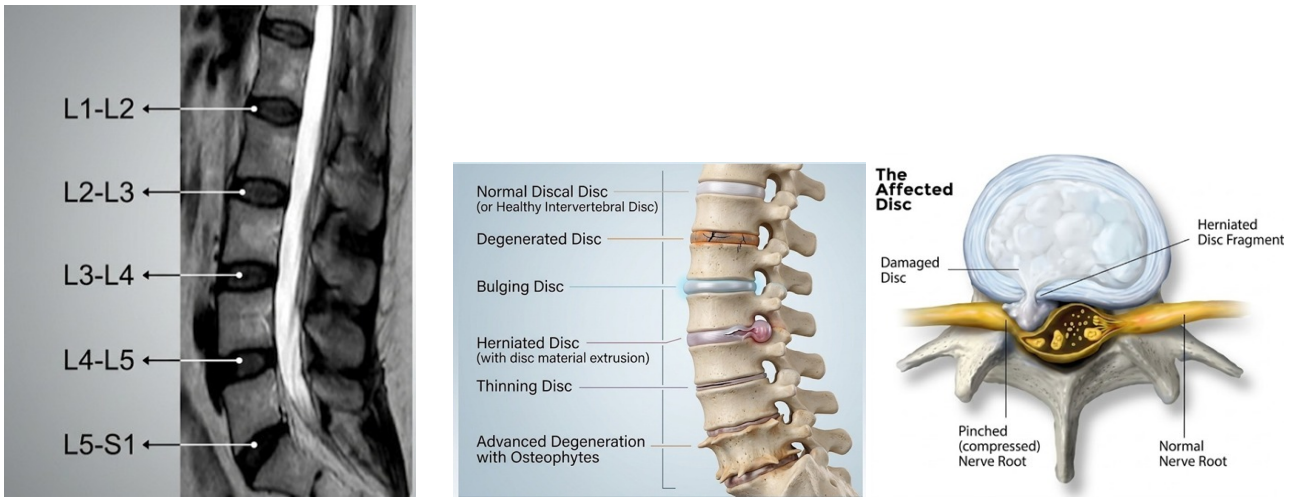
The lumbar spine is composed of five vertebral levels, designated L1–L2 through L5–S1, with an intervertebral disc positioned between each pair of adjacent vertebral bodies. These discs function as load-bearing structures and shock absorbers while permitting controlled movement and flexibility of the spinal column. Each intervertebral disc consists of a central gelatinous nucleus pulposus, a surrounding fibrocartilaginous annulus fibrosus, and superior and inferior cartilaginous endplates. The arrangement of the lumbar discs and their anatomical levels is illustrated by the sagittal T2-weighted magnetic resonance imaging shown in **Figure 3a** [14, 15]. Intervertebral disc degeneration is a progressive pathological process associated with aging, mechanical loading, genetic susceptibility, and biochemical changes within the disc. Degeneration commonly begins with the loss of water and proteoglycans in the nucleus pulposus. These changes reduce the disc's ability to distribute mechanical loads and may subsequently lead to annular fissures, loss of disc height, altered signal intensity on T2-weighted MRI, and progressive deformation of the disc space [16, 17]. The progression from a relatively healthy disc to advanced degeneration, including osteophyte formation, is represented in **Figure 3b**.

The principal degenerative conditions relevant to lumbar spine MRI analysis are described below.

- **Lumbar disc degeneration:** Lumbar disc degeneration is characterized by progressive dehydration of the nucleus pulposus, reduced disc height, and decreased signal intensity on T2-weighted MRI. In advanced stages, degeneration may be accompanied by annular fissures, endplate abnormalities, sclerosis, and osteophyte formation. These structural changes can alter spinal biomechanics and may contribute to axial low back pain or secondary neural compression [17, 18].
- **Disc bulging:** A disc bulge occurs when disc tissue extends diffusely beyond the normal margins of the vertebral endplates. According to the widely used nomenclature for lumbar disc pathology, a bulge generally involves more than 25% of the disc circumference and is usually associated with a relatively small extension beyond the apophyseal margins. Although a disc bulge may be asymptomatic, it can reduce the available space within the spinal canal or neural foramina and may contribute to nerve-root irritation when combined with other degenerative changes [3, 19].
- **Disc herniation:** Disc herniation refers to the focal displacement of disc material beyond the normal boundaries of the intervertebral disc space. It is commonly classified as protrusion, extrusion, or sequestration according to the relationship between the displaced material and the annulus fibrosus. A protrusion occurs when the base of the herniated material is wider than its outward extension and the outer annular fibers remain relatively contained. An extrusion occurs when the displaced material extends farther than its base or passes through a defect in the annulus.

Sequestration represents a more advanced form in which a fragment becomes detached from the parent disc and may migrate within the spinal canal. A schematic representation of disc herniation with nerve-root compression is shown in **Figure 3c** [3, 20, 21].

- **Lumbar spinal stenosis:** Lumbar spinal stenosis is characterized by a narrowing of the central spinal canal, lateral recesses, or neural foramina. The narrowing may result from a combination of disc degeneration, disc bulging or herniation, facet-joint hypertrophy, ligamentum flavum thickening, and osteophyte formation. Compression of the dural sac or exiting nerve roots may produce neurogenic claudication, radiating leg pain, sensory disturbance, weakness, or reduced mobility. Lumbar spinal stenosis is also an important cause of disability and a common indication for spinal surgery in older adults [22, 23].



(a) (b) (c)

Figure 3: Visual overview of lumbar spine anatomy and pathologie

3.2 Public Datasets for Lumbar Disc Disease

The successful application of machine learning and deep learning technologies depends on the availability of large and diverse datasets that have been appropriately annotated for training, validating, and assessing model performance [24]. The last few years have seen multiple lumbar spine MRI datasets and imaging systems developed for AI-based research on degenerative disorders of the lumbar spine. The different resources vary considerably in terms of imaging acquisition techniques, annotation methods, as well as types of clinical application, ranging from vast image collections to established benchmark datasets. The available publicly distributed datasets can be divided into three categories: 1) foundational imaging datasets and their subsequent derivations; 2) benchmark datasets; and 3) clinical datasets.

3.2.1 Foundational Lumbar Spine MRI Dataset and Derived Resources

The Lumbar Spine MRI Dataset introduced by Sudirman et al. is a widely used public resource for lumbar spine research. It contains 48,345 anonymized MRI slices from 515 patients with symptomatic low back pain. The images include sagittal and axial views of the lumbar spine and are stored in DICOM format. Most slices have a resolution of 320×320 pixels, although images from three studies have a resolution of 320×310 pixels. The dataset has supported the development of automated methods for intervertebral disc localization, lumbar spinal stenosis detection, disc segmentation, and degeneration assessment [25]. Several resources have been derived from this original collection. The Composite Lumbar Spine Mid-Sagittal Dataset contains 514 mid-sagittal MRI images with manual annotations and clinically relevant spinal measurements [26]. The annotations cover the vertebral bodies from L1 to L5 and the sacrum (S1), together with measurements such as vertebral height, lumbar lordotic angle, lumbosacral angle, spinal curvature, and vertebral dimensions. These data support the assessment of spinal alignment and deformity, as well as segmentation and quantitative morphometric analysis. The measurements may also be relevant to preoperative planning, although their clinical use requires appropriate validation. The Lumbar Spine MRI Annotation Dataset provides 515 T2-weighted mid-sagittal images selected from the original collection and resized to 384×384 pixels [27]. It includes manually produced segmentation annotations and measurements related to intervertebral disc height and degeneration. The resource was developed to support automated analysis of intervertebral disc height and Pfirrmann grading and can be used for supervised model development and performance evaluation. Together, the original dataset and its derived resources provide complementary data for lumbar

spine MRI research. The original collection offers a large number of axial and sagittal slices, whereas the derived datasets provide focused mid-sagittal images, anatomical annotations, quantitative measurements, and degeneration-related labels. These resources therefore support a range of tasks, including anatomical localization, segmentation, classification, severity grading, and quantitative analysis.

3.2.2 Large-Scale Public Benchmark Datasets

While foundational resources provide imaging data and manual annotations, large-scale benchmark datasets are important for reliable model evaluation and improved generalization across institutions. SPIDER [28] is a multi-institutional lumbar spine MRI benchmark resource that contains sagittal T1- and T2-weighted images from 218 patients collected from four hospitals. It includes expert-annotated labels for major anatomical structures, including vertebrae, intervertebral discs, and the spinal canal. The resource was developed for segmentation benchmarking and facilitates the development and comparison of automated lumbar spine analysis methods. The RSNA Lumbar Degenerative Imaging Spine Classification (LumbarDISC) resource is one of the largest publicly available lumbar spine MRI collections, including imaging studies from 2,697 patients collected from multiple institutions worldwide. It provides expert-annotated severity grades for spinal canal stenosis, right and left neural foraminal narrowing, and right and left subarticular stenosis, supporting automated classification and severity assessment of lumbar degenerative disease [29]. More recently, the LSpineSMRI resource (2025) expanded publicly available data by introducing more than 17,000 sagittal and axial lumbar spine MRI slices obtained from 640 patients. It includes structured lumbar spinal stenosis grading labels and facilitates the development of deep learning models for severity assessment and clinical decision support [30]. Collectively, these benchmark resources illustrate the evolution of lumbar spine MRI datasets from small datasets collected at individual centers to larger multi-institutional datasets with consistent annotations. They have played an important role in supporting the training, validation and evaluation of more advanced AI systems for lumbar spine analysis.

3.2.3 Clinical and Morphometric Datasets

Clinical and morphometric databases also provide other types of quantitative data, such as measurements of spinal curvature, sagittal alignment, and biomechanical properties. These resources should be distinguished from image based datasets because they primarily describe clinical and anatomical characteristics rather than image content. The spinopelvic sagittal alignment dataset is one example. It contains data from 213 patients with lumbar degenerative disease and reports important spinopelvic parameters, including pelvic incidence, pelvic tilt, sacral slope, lumbar lordosis, and sagittal vertical axis. These measurements can support the assessment of spinal balance, prediction modelling, biomechanical analysis, and monitoring of disease progression [31]. **Table 2** summarizes the publicly available lumbar spine MRI datasets and related clinical resources discussed in this section.

Table 2: Summary of publicly available lumbar spine MRI datasets and related clinical resources for AI based analysis

Dataset	Year	Size	Modality	Tasks	Annotations	Main features	Reference
Lumbar Spine MRI Dataset by Sudirman et al.	2019	515	Lumbar spine MRI with sagittal and axial views	Segmentation, detection, and lumbar spinal stenosis detection	Ground truth segmentation masks and radiological annotations	Benchmark dataset for localization, segmentation, and degenerative disease detection	[25]
Composite Lumbar Spine Mid Sagittal Dataset	2021	514 images	Mid sagittal MRI	Morphometry and segmentation	Manual vertebral annotations and spinal measurements	Vertebral labels and spinal curvature measurements for alignment analysis	[26]
Lumbar Spine MRI Annotation Dataset for IVD Height and Pfirrmann Grade	2024	515	T2 weighted mid sagittal MRI	Segmentation, classification, and grading	Segmentation masks and computed measurements	Intervertebral disc height estimation and Pfirrmann degeneration grading	[27]
SPIDER Lumbar Spine Segmentation Benchmark Dataset	2024	218	T1 and T2 weighted sagittal MRI	Segmentation	Vertebral bodies, intervertebral discs, and spinal canal annotations	Multicenter benchmark for lumbar spine segmentation	[28]

Continued on next page

Table 2 continued from previous page

Dataset	Year	Size	Modality	Tasks	Annotations	Main features	Reference
RSNA Lumbar Degenerative Imaging Spine Classification Dataset	2026	2,697	MRI studies	Classification	Severity grades for spinal canal stenosis, neural foraminal narrowing, and subarticular stenosis	Large scale benchmark for degenerative disease classification and severity assessment	[29]
LSpineSMRI Dataset	2025	640	Sagittal and axial MRI	Severity assessment	Spinal stenosis grading labels	Dataset for automated lumbar spinal stenosis assessment	[30]
Spinopelvic Sagittal Alignment Dataset	2019	213	Clinical and radiographic measurements	Prediction and analysis	Spinopelvic parameters including PI, PT, SS, LL, and SVA	Quantitative dataset for spinal balance and sagittal alignment assessment	[31]

4. LITERATURE REVIEW

With recent advances in artificial intelligence (AI) and spinal imaging, many articles and reviews have been published on the use of machine learning and/or deep learning for analysis of lumbar spine MRIs. Although these reviews vary greatly in their objectives, methodologies, and clinical contexts, they demonstrate the continued development of AI-assisted lumbar spine MRI analysis from initial feasibility studies to applications with greater clinical relevance. Recent reviews increasingly focus on disease classification, image segmentation, severity grading, explainability, workflow integration, and the clinical implementation of machine learning and deep learning systems. The following discussion summarizes the representative systematic and narrative reviews published so far, presented in chronological sequence to illustrate the methodological evolution of AI-assisted lumbar spine MRI analysis. Early work focused on assessing the feasibility and diagnostic performance of AI-based approaches for lumbar spine MRI assessment. D'Antoni et al. [32] reviewed fifteen studies examining AI and computer-assisted diagnostic tools for chronic lower back pain. The review concluded that high classification performance was reported across imaging-related tasks, particularly for disc degeneration and disc herniation classification. However, the authors also highlighted several recurring limitations, including methodological diversity, limited external validation, and the absence of standardized reference criteria, which reduce the generalizability of the reported results. Later reviews shifted toward methodological developments and the growing use of deep learning techniques. Hussain et al. [13] presented a comprehensive review focusing on deep learning approaches for lumbar disc degenerative disease analysis. The review grouped published studies based on imaging methods, pre-processing procedures, databases, and deep learning architectures (i.e., convolutional neural networks, U-Net variants, and early transformer-based networks). Although the results were generally promising, several challenges remained, including small sample sizes, class imbalance, limited generalizability, poor model interpretability, and differences in evaluation protocols. Also in 2023, Compte et al. [12] assessed whether machine learning systems could achieve diagnostic performance similar to radiologists when evaluating disc degeneration, bulging, herniation, and Modic changes. Their systematic review and meta-analysis showed that AI models often performed similarly to radiologists on internal datasets. However, the authors also identified substantial bias, inconsistent labeling strategies, and limited external validation, showing a gap between promising experimental findings and routine clinical use. Other reviews focused on specific imaging tasks. Wang et al. [33] conducted a systematic review and meta-analysis of deep learning methods for lumbar intervertebral disc segmentation. The pooled results showed strong segmentation performance, with a Dice similarity coefficient of 0.900 and an IoU of 0.863. U-Net variants and attention-based segmentation models generally achieved better boundary delineation and greater robustness than earlier CNN-based methods, particularly in complex degenerative cases. However, considerable differences in pre-processing strategies, annotation quality, evaluation methods, and validation protocols limited direct comparisons across studies and made clinical integration challenging. Similarly, Wang et al. [34] evaluated AI-based methods for diagnosing lumbar spinal stenosis. Across 12 studies involving 15,044 patients, the pooled sensitivity, specificity, and AUC were 0.84, 0.87, and 0.92, respectively. Although these results suggest good diagnostic performance, results varied with dataset size, MRI acquisition protocols, class imbalance, and validation strategy. Models trained on larger multicenter datasets generally showed better generalizability and more consistent performance than those developed from small retrospective single-center cohorts. The authors also noted substantial methodological heterogeneity and highlighted the need for robust multicenter prospective external validation before routine clinical use. With the development of research, reviews focused more on specific clinical applications. Liawrungrueang et al. [35] reviewed the studies involving use of AI in the assessment of lumbar degenerative disc disease with diagnostic accuracies of 71.5% to 99%. In their review, they demonstrated the clinical potential of AI-based systems but also practical limitations such as retrospective study designs, heterogeneous annotation protocols, limited explainability, and limited integration into clinical workflows. More recently, Zhao et al. [36] published a narrative

review examining recent advances and challenges in AI-assisted MRI analysis for lumbar disc degeneration, including 17 studies published from 2014 to 2024, more recently. The review covered the transition from traditional machine learning methods to more recent architectures, such as CNNs, U-Nets, transformers, and hybrid models. Recent topics such as explainable AI, data augmentation, few-shot learning, workflow integration and regulatory requirements were also discussed. Unlike earlier reviews, which mainly focused on performance metrics, this work emphasized the broader issues affecting clinical implementation. Across these studies, the reviewed literature reflects the transition from conventional machine learning approaches to more advanced deep learning architectures, such as CNNs, U-Net variants, transformer-based models, and hybrid frameworks. Studies using attention-enhanced architectures, improved feature extraction pipelines, and data augmentation strategies generally achieved better segmentation and classification performance. Nevertheless, direct comparisons between studies remain difficult due to considerable differences in dataset composition, annotation quality, MRI acquisition protocols, evaluation metrics, and validation methodologies. Notably, many studies reporting high performance were based on relatively small retrospective single-center datasets with no strong external validation. This raises concerns about the generalizability of the model and its potential use in clinical practice. **Table 3** summarizes selected systematic and narrative reviews of AI-assisted lumbar spine MRI analysis. These reviews trace the field's progression from early feasibility studies to increasingly clinically oriented deep learning frameworks. Recent reviews have expanded their focus to disease grading, image segmentation, model explainability, workflow integration, and external validation, while continuing to identify challenges related to data heterogeneity, limited interpretability, and insufficient clinical evaluation. Overall, the evidence underscores the need for rigorous external validation, transparent and interpretable models, and methods that generalize across clinical settings to support the safe integration of automated analysis into lumbar spine MRI workflows.

Table 3: Overview of recent systematic and narrative reviews on AI based lumbar spine MRI analysis

Study and Year reference	Contribution	Strengths	Limitations
D'Antoni et al. 2022 [32]	Systematic review and meta analysis of machine learning and deep learning methods for evaluating lumbar spine MRI findings, including disc degeneration, herniation, bulge, and Modic changes.	Comprehensive evaluation of automated approaches, quantitative meta analysis, and comparison of multiple diagnostic tasks and algorithms.	High methodological heterogeneity, inconsistent reporting standards, limited external validation, and lack of standardized evaluation protocols.
Hussain et al. [13] 2023	Comprehensive review of deep learning methods for MRI based diagnosis of lumbar disc degenerative diseases, organized according to imaging modality, preprocessing, datasets, and network architecture, including two dimensional and three dimensional CNNs, U Nets, and transformers.	Strong technical taxonomy of deep learning pipelines and architectures. Disease specific MRI focus. Identifies class imbalance, limited datasets, domain shift, explainability, and inconsistent evaluation metrics as major challenges.	Primarily qualitative review without PRISMA based quantitative synthesis or meta analysis. Limited discussion of clinical deployment, external validation, and generalizability.
Compte et al. [12] 2023	Systematic review and meta analysis of 27 studies comparing machine learning based lumbar spine MRI assessment with radiologist performance for disc degeneration, bulge, herniation, and Modic changes.	Direct comparison of automated systems with radiologists. Includes pooled bivariate and multivariate analyses and demonstrates promising radiologist level performance on internal datasets.	High risk of bias in included studies, limited sample sizes, inconsistent ground truth labeling, scarce replication studies, and reduced performance under external validation.
Wang et al. [33] 2024	Systematic review and meta analysis of deep learning for lumbar intervertebral disc segmentation on MRI, based on 45 studies, with quantitative results synthesized from 16 studies.	Provides pooled quantitative benchmarks, including a Dice similarity coefficient of 0.900 and an intersection over union of 0.863. Examines architecture, dimensionality, data augmentation, and validation strategies.	High heterogeneity in preprocessing and evaluation protocols, limited external and multicenter validation, predominance of retrospective single center studies, and incomplete reporting of statistical performance measures.

Continued on the next page

Table 3 continued from the previous page

Study and reference	Year	Contribution	Strengths	Limitations
Wang et al. [34]	2024	Systematic review and meta analysis of traditional machine learning and deep learning models for lumbar spinal stenosis diagnosis, based on 12 studies involving 15,044 patients.	Large pooled sample size, strong diagnostic performance with sensitivity of 0.84, specificity of 0.87, and area under the curve of 0.92. Includes methodological quality assessment using QUADAS 2.	Extremely high heterogeneity across studies, inconsistent reference standards, limited external validation, high risk of bias in several studies, and limited clinical applicability.
Liawrungrueang et al. [35]	2025	Systematic review of 20 studies applying automated MRI analysis to lumbar degenerative disc disease, including diagnostic and grading tasks.	Clinically oriented review connecting reported performance with potential clinical applications. Describes diagnostic and Pfirrmann grading approaches.	Predominantly retrospective studies, non standardized reference labels, limited interpretability, poor clinical workflow integration, and reliance on small heterogeneous datasets.
Zhao et al. [36]	2026	Narrative review of advances and challenges in automated MRI analysis for lumbar disc degeneration detection and classification, based on 17 studies published between 2014 and 2024.	Forward looking perspective that integrates technical progress, including CNNs, U Nets, and transformers, with clinical and ethical considerations. Discusses interpretability, data augmentation, and few shot learning.	Narrative rather than systematic review, selective coverage of 17 articles, no pooled results or quantitative synthesis, limited reproducibility, and possible publication bias.

5. MRI GRADING AND QUANTITATIVE ASSESSMENT

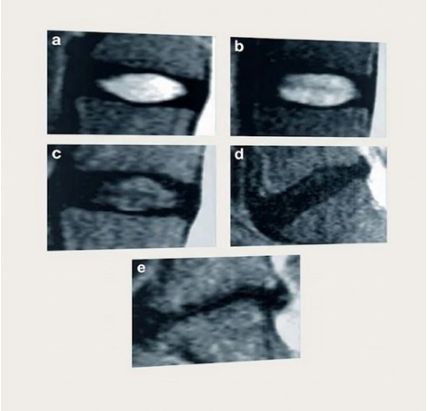
Standardized magnetic resonance imaging grading systems provide clinically interpretable reference labels for artificial intelligence frameworks. These labels are used to train and evaluate deep learning models for the classification, localization, segmentation, and severity grading of lumbar spine abnormalities. Because different grading systems emphasize different anatomical and pathological features, their definitions must be reported clearly when comparing AI studies.

5.1 Qualitative and Morphological Grading

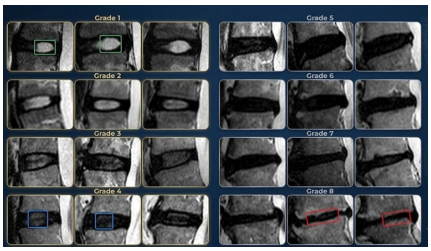
To reduce the subjectivity of manual radiological assessment, several ordinal and morphology-based grading systems have been developed for lumbar spine MRI.

- **Pfirrmann classification:** The Pfirrmann system is one of the most widely used methods for grading intervertebral disc degeneration on sagittal T2-weighted images. It uses a five-grade scale from Grade I to Grade V based on nucleus pulposus signal intensity, signal homogeneity, the distinction between the nucleus pulposus and annulus fibrosus, and disc height. Grade I generally represents a healthy disc with high and homogeneous T2-weighted signal intensity, whereas Grade V indicates advanced degeneration with marked signal loss and substantial disc-space narrowing. Representative examples of the standard Pfirrmann grades are shown in **Figure 4a** [18, 37]. Although the system is clinically useful, its interobserver reliability is not perfect, with reported interobserver and intraobserver kappa values of approximately 0.66 and 0.74, respectively [7].
- **Modified Pfirrmann systems:** Modified versions of the original Pfirrmann system have been proposed to provide more detailed discrimination between intermediate and advanced stages of disc degeneration. The eight-grade framework introduced by Griffith and colleagues incorporates additional distinctions in disc height, signal intensity, nucleus pulposus structure, and the integrity of the disc margins. These modifications can improve the description of progressive degeneration, particularly in older populations and in cohorts with a high prevalence of severe disease. Examples of modified Pfirrmann grading are presented in **Figure 4b** [38, 39].
- **Michigan State University classification:** The Michigan State University classification is used primarily to describe lumbar disc herniation on axial T2-weighted MRI. Herniation size is classified into Grades 1, 2, and 3 according to its extension relative to the inter-facet line. The system also describes the anatomical location of the herniation using three zones: Zone A for central herniation, Zone B for subarticular or paracentral herniation, and Zone C for foraminal or far-lateral herniation. The herniation-size grades are illustrated in **Figure 4c**, while the anatomical zones are shown in **Figure 4d** [40].

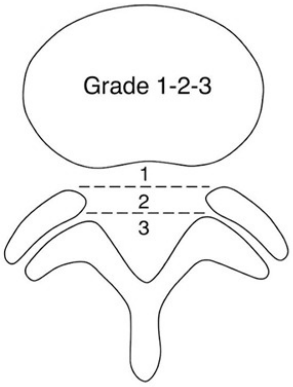
These grading systems provide clinically meaningful labels for AI-based lumbar spine MRI analysis, but they should not be treated as interchangeable outcomes. The Pfirrmann and modified Pfirrmann systems assess the severity of intervertebral disc degeneration, whereas the Michigan State University system characterizes the size and location of disc herniation. Consequently, an AI model trained to predict Pfirrmann grade addresses a different clinical task from a model trained to classify herniation size or anatomical zone. The use of standardized grading systems also supports more reliable comparison between AI studies. However, differences in reader expertise, MRI sequence, slice thickness, disease prevalence, labeling protocol, and the handling of borderline cases may affect the resulting ground-truth labels. In addition, ordinal grades contain an inherent degree of subjectivity, particularly when adjacent grades have similar imaging appearances. For this reason, AI studies should report the grading system used, the number and experience of annotators, the method used to resolve disagreements, and whether the final labels were assigned at the disc, vertebral level, patient, or examination level.



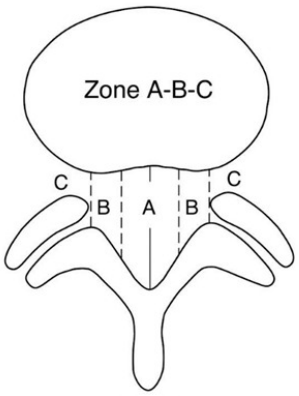
(a)



(b)



(c)



(d)

Figure 4: Established lumbar spine MRI grading frameworks

5.2 Quantitative MRI and Features Beyond Morphological Grading

Although conventional qualitative grading systems are clinically useful, they primarily describe macroscopic structural changes and may be insensitive to early biochemical alterations within the intervertebral disc. Changes such as reduced water and glycosaminoglycan content can occur before substantial disc-space narrowing or other advanced morphological abnormalities become visible on routine MRI. Qualitative grading may also provide limited information about associated imaging findings, including Modic endplate changes and high-intensity zones, which may be relevant to disc pathology and pain generation [41, 42]. Quantitative MRI techniques have therefore been investigated as complementary approaches for characterizing disc composition and degeneration. Methods such as $T_1\rho$ and T_2 mapping generate voxel-level relaxation maps from which quantitative imaging biomarkers can be derived. These techniques may provide indirect information about tissue hydration, proteoglycan content, and changes in the extracellular matrix that are not fully captured by conventional qualitative assessment [43, 44]. Quantitative measurements can consequently support the detection of subtle degenerative changes and improve the characterization of disease severity across different stages of disc degeneration. In addition to relaxation mapping, quantitative MRI analysis may incorporate disc height, volume, signal intensity, shape, endplate characteristics, and other morphometric features. These measurements can be extracted manually, semi-automatically,

or through deep learning-based segmentation and image-analysis pipelines. Their reliability depends on image quality, scanner characteristics, acquisition parameters, segmentation accuracy, and the consistency of the reference standard used for validation. The integration of conventional morphological grades with quantitative multi-parametric MRI features represents an important direction for computer-aided diagnosis in spinal radiology. A reproducible end-to-end pipeline could combine anatomical localization, disc segmentation, qualitative grading, quantitative feature extraction, and automated severity prediction. Such multimodal models may provide a more comprehensive representation of disc pathology than either visual grading or isolated quantitative biomarkers alone. However, their clinical usefulness requires standardized acquisition protocols, harmonized quantitative measurements, external validation across scanners and institutions, and prospective evaluation against clinically meaningful outcomes [8].

6. AI-BASED APPROACHES IN LUMBAR SPINE MRI

The use of artificial intelligence (AI) technologies for clinical auxiliary diagnosis has grown rapidly in the last decade. These methods have achieved good performance in medical image analysis, particularly when evaluating intervertebral disc abnormalities. AI-based approaches can improve diagnostic accuracy, extract clinically relevant features, and reduce errors related to observer variability [45]. Deep learning methods are commonly used in medical imaging and enable models to learn hierarchical features from raw data through multiple layers of nonlinear transformations. In medical image analysis, early layers typically learn basic image features including edges and textures, while deeper layers gradually learn higher-level representations, such as anatomical structures and pathological patterns [46].

6.1 AI Analytical Pipeline in Lumbar Spine MRI

Artificial intelligence (AI) has played an important role in spinal imaging by enhancing diagnostic precision and simplifying image analysis workflows across tasks such as image pre-processing, segmentation, and automated disease grading [47, 48]. As illustrated in the conceptual framework in **Figure 5**, a complete AI analysis workflow for lumbar spine MRI generally consists of several stages, from raw image acquisition to clinical decision support:

- **Low-Level pre-processing and High-Level Image Processing:** After MRI acquisition, several pre-processing steps are applied to reduce imaging artifacts and improve image quality. These steps include denoising, filtering, registration, bias-field correction, intensity standardization, and normalization [13]. The next step is to identify specific anatomical regions through region-of-interest (ROI) selection and patch extraction methods [49, 50]. Because many lumbar MRI datasets come from a single institution and may differ in scanner and vendor characteristics, data augmentation is often used during model training. Geometric and intensity transformations can increase data diversity and improve model robustness and generalizability [51, 52].
- **Core AI Tasks (Detection and Segmentation):** The next step focuses on the main AI tasks of detection and segmentation. Object detection algorithms predict bounding boxes to locate regions of interest in MRI scans [53]. These regions are then used as input to semantic segmentation networks to obtain objective and reproducible pixel-level measurements of disc height and spinal canal narrowing [54]. Recent architectures extend conventional U-Net models with attention mechanisms and Transformer-based components, which can improve boundary delineation in advanced degenerative disease, especially in cases such as Pfirrmann Grade V, where anatomical boundaries can be difficult to distinguish [55].
- **Classification, Explainability, and Clinical Decision Support:** The extracted morphological features and semantic representations are then provided to deep learning classifiers to identify the underlying pathology and assign lesion grades [56, 57]. Rather than relying only on end-to-end models, combining detection and segmentation with classification can provide more targeted feature representations and may improve model generalization [58, 59]. For clinical interpretation, post-hoc explainability methods such as Grad-CAM and saliency maps can highlight image regions that contribute most to the model's predictions. This pipeline can support several clinical applications, including automated report generation, personalized treatment planning, and long-term outcome prediction.

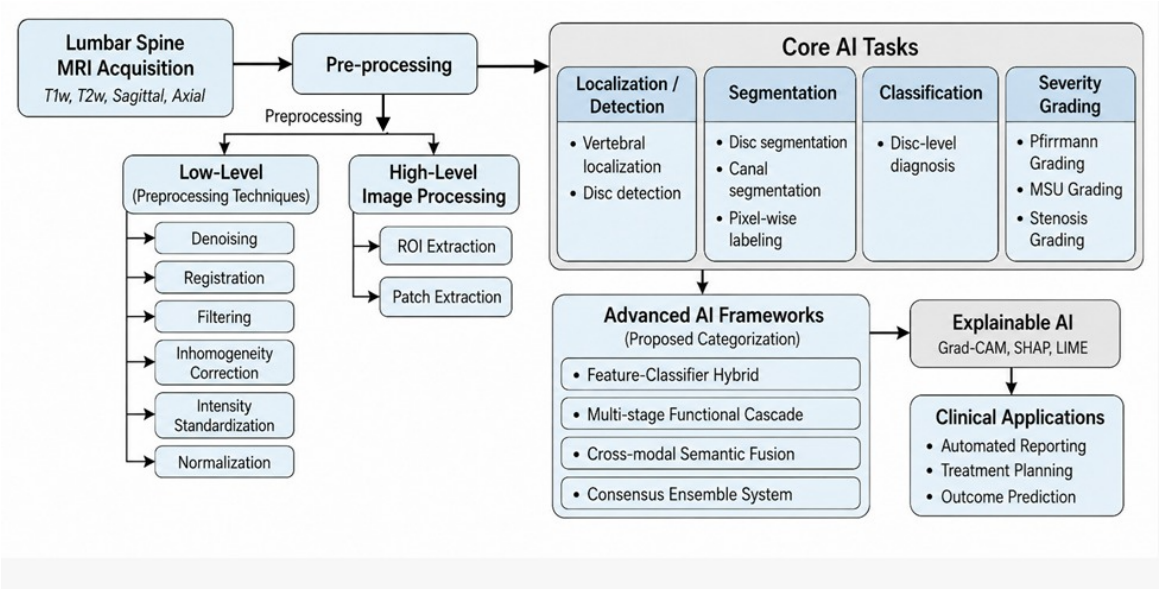


Figure 5: Conceptual framework illustrating the end-to-end AI analytical pipeline for lumbar spine MRI analysis

6.2 Methodological Challenges and Evidence Interpretation

A synthesis of the reviewed studies indicates that automated lumbar spine MRI methods often achieve strong results under the conditions evaluated. Several studies have reported accuracy, area under the curve, or Dice similarity values close to or above 0.90 [33, 34, 36]. However, these results should be interpreted within the context of the datasets, reference standards, and validation procedures used in each study. High performance on a carefully selected retrospective dataset does not necessarily demonstrate reliable performance in routine clinical practice. One important concern is the limited representativeness of many available datasets. A considerable number of studies use small retrospective cohorts collected at a single institution, often with limited variation in scanners, acquisition protocols, magnetic field strengths, and patient populations. Under these conditions, a model may learn characteristics associated with a particular institution or scanner rather than features that reflect the underlying pathology. Differences between 1.5 T and 3 T examinations, variations in image contrast, and scanner dependent noise can therefore affect performance when the method is applied to data from another clinical setting [13]. Data partitioning is another important source of potential bias. When images or slices from the same patient are distributed between the training and testing sets, the resulting evaluation may be overly optimistic because the two sets are not truly independent. Patient level separation is therefore preferable to slice level separation. Studies should also describe the method used to create the partitions, the number of patients and examinations in each subset, and whether images from the same examination or patient appear in more than one subset. The interpretation of reported performance measures is also affected by differences in study design. Accuracy, area under the curve, Dice similarity coefficient, and intersection over union measure different aspects of performance and should not be treated as interchangeable. Their values depend on class prevalence, the number of severity categories, slice selection, image preprocessing, segmentation criteria, and the quality of the reference annotations. Differences in radiologist experience and disagreement between readers may further influence the labels used for training and evaluation. For example, a high score obtained from a small and homogeneous dataset should not be compared directly with a lower score obtained from a heterogeneous multicenter cohort, a severely imbalanced dataset, or a task involving several ordinal severity grades. A larger numerical value does not necessarily indicate a better method. Reported metrics should instead be understood as estimates of performance under specific experimental conditions. Meaningful comparisons require similar datasets, reference standards, partitioning procedures, evaluation metrics, and clinical tasks.

6.3 Clinical Translation and Practical Considerations

The use of automated lumbar spine MRI analysis in clinical practice requires more than evidence of technical accuracy. The method must operate reliably on images acquired with different scanners and protocols, produce results that can be reviewed by radiologists, and fit within established reporting and image management procedures. Systems that require substantial computing resources or specialized hardware may be difficult to incorporate into routine radiology services, particularly where technical support and infrastructure are limited. Interpretability is also important for clinical acceptance. Radiologists and other clinical users need to understand which anatomical structures or image regions contributed to a classification or severity estimate. A result that is presented without supporting visual or quantitative information may be difficult to verify, especially in cases involving multiple degenerative findings. Clear visualization of the relevant

disc, vertebral level, herniation, stenosis, or endplate abnormality can assist clinical review and help identify incorrect predictions. Further evidence is needed before automated systems can be considered suitable for routine decision support. Prospective studies involving multiple institutions should assess performance in representative patient populations and should include different scanners, field strengths, acquisition protocols, and levels of disease severity. These studies should report confidence intervals, failure cases, subgroup performance, and agreement with qualified radiologists. Evaluation should also consider whether the system improves reporting consistency, reduces interpretation time, or changes clinical management in a beneficial manner. Regulatory, legal, and economic considerations must also be addressed. Clinical implementation may require approval by the relevant regulatory authority, clear definitions of responsibility when automated results are used, procedures for monitoring performance after deployment, and evidence that the benefits justify the cost of installation and maintenance. In addition, patient privacy and secure handling of medical images must be maintained throughout data collection, model development, evaluation, and clinical use [12, 35, 36]. At present, the available literature provides encouraging evidence of technical feasibility, but it remains insufficient to establish broad clinical effectiveness. Future work should therefore move beyond isolated performance scores and focus on independent validation, reproducibility, transparent reporting, clinically meaningful outcomes, and continuous evaluation after implementation. This approach would help determine whether automated lumbar spine MRI analysis can provide dependable support for radiological assessment without replacing professional clinical judgment.

6.4 Deep Learning Approaches

Deep learning (DL) has been widely adopted in lumbar spine MRI analysis because of its ability to learn imaging features directly from MRI scans without hand-crafting feature engineering. This has resulted in improved detection of clinically relevant patterns in tasks such as diagnosis, segmentation and disease grading [19].

6.4.1 Evolution of Early Frameworks to Targeted CNN Models (2018–2023)

Early studies showed that end-to-end pipelines could be used for lumbar spine MRI analysis by combining semantic segmentation networks with multi-task classifiers. Lu et al. [60], for example, used a U-Net backbone for localization followed by a ResNeXt-50 multi-task classifier. They also applied natural language processing (NLP) to extract reference labels from a large set of radiology reports, and the model achieved high diagnostic performance (AUC > 0.96). These results imply that the spatial limitations of early end-to-end multi-task models might be alleviated by increasing the amount of training data. In contrast, other early studies with smaller datasets showed limited generalizability. For example, Pan et al. [61] developed a multi-stage pipeline that combined Faster R-CNN with a ResNet classifier, while Niemeyer et al. [62] used ordinal regression for Pfirrmann grading. This approach is appropriate for the ordered structure of the Pfirrmann grading scale. Unlike conventional multi-class classification, ordinal regression takes the ordering of grades into account, with adjacent grades representing progressively greater levels of disc degeneration. Although both models achieved high internal accuracy and reduced interobserver variability, standard CNN architectures may perform less consistently across institutions and anatomical variations not represented in the training cohorts. From 2021 to 2023, studies focused more on CNN-based models for localized pathological detection, optimization, and real-time processing. During this period, one-stage object detectors were increasingly adopted, including YOLOv3 in the study by Tsai et al. [63] and YOLOv5 by Liawrungrueang et al. [64] for automated Pfirrmann grading. A major advantage of these one-stage architectures was their faster inference speed and higher screening efficiency, which makes them suitable for real-time clinical deployment. However, this computational efficiency is accompanied by certain limitations. Models using object detectors alongside lightweight classifiers, such as the optimized AlexNet proposed by Valarmathi and Nirmala Devi [21], as well as multi-stage frameworks developed by Sustersic et al. [65] and Shahzadi et al. [66], often achieved high performance on their evaluation datasets. Despite these encouraging results, their clinical applicability remains limited. Since these methods analyze localized 2D regions or focus on individual lumbar levels, they may overlook complex multi-level central or neural foraminal stenosis. Accordingly, their reported performance may not be maintained on external multicenter datasets.

6.4.2 Multi Task Systems, Advanced Architectures, and Multi View Fusion

The limitations of early single view and single task models encouraged the development of multi task learning, spatial attention, and multi view information fusion. A single sagittal or axial MRI slice may not fully represent the three dimensional anatomy and pathology of the lumbar spine. Recent studies have therefore combined information from different imaging planes to obtain a more complete representation of spinal structures and pathological findings. Qian et al. [67] used a three dimensional ResNet with spatial attention to jointly analyze sagittal and axial images, whereas Chen et al. [68] used Mask R CNN to develop a multi view fusion framework for disc and vertebral classification. The value of multi view analysis lies in the complementary information provided by different imaging planes. Sagittal images show disc position, vertebral alignment, disc height, and the overall distribution of degenerative changes. Axial images provide more detailed

information about the spinal canal, neural foramina, and the location of disc herniation. Combining these views may reduce false positive findings, particularly in cases involving asymmetric disc pathology. Multi stage systems that use explicit anatomical information have also shown promising results. The framework developed by Tumko et al. [4] and the joint localization and grading method proposed by Xu et al. [69] used anatomical localization to guide subsequent classification or grading. This arrangement can reduce the influence of irrelevant image regions, although an error during localization may affect all later stages. Recent work has combined multi view analysis with contrastive learning and methods that provide additional information about model predictions. Pan et al. [70] proposed a dual path contrastive learning framework with dynamic memory banks to learn detailed image representations from limited training data. Their results suggest that contrastive representation learning may improve classification compared with conventional convolutional networks and standard transformer models under the conditions examined. These findings should be interpreted in relation to the dataset size, reference labels, validation design, and whether independent external testing was performed. Related MRI research has also combined object detection with subsequent segmentation. Alabsi et al. combined YOLOv8 derived spatial prompts with the Segment Anything Model for automated brain tumor segmentation in MRI images [11]. Although this study examined brain MRI rather than lumbar spine disease, it provides a relevant methodological example of linking object detection with segmentation in an automated MRI workflow. Single stage detectors have also been developed for the rapid localization of lumbar abnormalities. Abed et al. [71] used YOLOv8m with polygon annotations for disc bulge localization and classification and reported a mean average precision of 93.7%. Polygon annotations may represent irregular disc boundaries more accurately than rectangular bounding boxes. However, the method was limited to binary classification and did not perform severity grading. Some recent frameworks have combined detection, classification, decision level fusion, and visual explanation to support clinical interpretation. Sayed et al. [72] proposed a three stage system that used YOLOv11 for object detection, attention enhanced convolutional networks for classification, decision level fusion, and Grad CAM visualization. Separating localization from classification may make it easier to review the anatomical basis of a prediction. However, the use of multiple models and processing stages may increase computational requirements, inference time, and the difficulty of integration into existing radiology workflows. Despite these developments, rigorous external validation remains uncommon. Many studies use retrospective datasets collected at a single institution, and the effects of differences in scanners, acquisition protocols, patient populations, and disease prevalence are not always examined. Consequently, high performance on an internal test set should not be interpreted as evidence of general clinical applicability. The studies summarized in **Table 4** were re-examined using the original publications to maintain consistency in the reported architectures, performance measures, and methodological contributions.

Table 4: Representative deep learning studies for lumbar spine MRI analysis

Reference	Year	Pathology	Task	Imaging modality	Key methodological contribution	Architecture	Performance	Limitations
[60]	2018	Spinal stenosis	Grading	Sagittal and axial T2	NLP based weak supervision with automated disc localization using U Net and spine curve fitting	U Net and multi input ResNeXt 50 with two dimensional axial and three dimensional sagittal branches	Class average accuracy of 80.4% for central stenosis and 78.1% for foraminal stenosis. Binary AUC of 0.983 for central stenosis and 0.961 for foraminal stenosis	Inter reader variability in NLP derived labels and reduced performance in severe scoliosis
[61]	2021	Disc abnormality including bulge and herniation	Localization and classification	Sagittal and axial T2	Four step automated pipeline integrating vertebral localization and geometric coordinate mapping	Faster R CNN and ResNet101 with cost sensitive learning	Localization accuracy of 100%. Three class accuracy of 92.7% at L1 L2, 84.4% at L2 L3, 92.1% at L3 L4, 90.4% at L4 L5, and 84.2% at L5 S1	Small dataset, image blurring, limited herniation samples at L2 L3, and morphological similarity between normal and bulged discs at L5 S1
[62]	2021	Disc degeneration based on Pfirrmann grade	Grading	Sagittal T2	Extended fractional Pfirrmann grading with evaluation of ordinal classification and regression strategies	CNN with ordinal classification and regression	Kappa of 0.92 and MAE of 0.082	Class imbalance in Grades I and V and no independent external validation

Continued on the next page

Table 4 continued from the previous page

Reference	Year	Pathology	Task	Imaging modality	Key methodological contribution	Architecture	Performance	Limitations
[63]	2021	Disc herniation	Detection	Sagittal T2	Targeted data augmentation for real time YOLOv3 detection	YOLOv3	Mean average precision of 92.4% and processing time below 3 seconds per image	Older YOLO version and restriction to sagittal view disc herniation detection
[65]	2022	Multiple types of disc herniation	Segmentation and classification	Sagittal and axial T2	U Net segmentation with CLAHE enhancement followed by DiscNet classification	U Net and DiscNet	Accuracy of 91% for sagittal images, 87% for axial images, and 80% for combined views	Small dataset, missing L1 to L3 levels, and no bootstrap validation
[66]	2023	Neural foraminal stenosis	Detection and grading	Axial T2	Multi region of interest strategy with data augmentation for customized CNN grading	Customized CNN	Accuracy of 97.71% with a single region of interest and 97.01% with multiple regions of interest	Single public dataset and need for additional data to improve generalization
[21]	2023	Disc herniation including prolapse and extrusion	Localization, classification, and severity assessment	Sagittal T2	EVGG16 based YOLOv2 localization with enhanced AlexNet for classification and severity assessment	YOLOv2 IVD with EVGG16 and enhanced AlexNet	Accuracy of 93.59% and precision of 97.62%	Limited to prolapse and extrusion and severity assessment based only on disc length
[64]	2023	Disc degeneration based on Pfirrmann Grades I to V	Disc detection and Pfirrmann grading	Sagittal T2	Real time YOLOv5 based detection and multi grade Pfirrmann classification	YOLOv5	Accuracy above 95% and mean average precision approximately 0.995	Class imbalance and failure to identify vertebral levels
[69]	2025	Disc herniation based on MSU classification	Localization and MSU grading	Axial T2	Simultaneous MSU localization and grading with Integrated Gradients based interpretability	Faster R CNN and ResNet50	Internal AC1 of 0.81 and external AC1 of 0.69	Reduced external performance because Zone C cases were underrepresented
[67]	2024	Disc herniation	Detection and segmentation	Sagittal and axial T2	Transformer based multimodal mutual attention for sagittal and axial feature fusion	U Net with three dimensional ResNet18 encoder and transformer	F1 score of 0.971 for sagittal images and 0.903 for axial images	Lower axial segmentation accuracy and loss of spatial information in U Net
[4]	2024	Spinal stenosis	Segmentation, detection, and grading	Sagittal and axial T2	Three stage framework with segmentation of 17 anatomical structures and mask guided RegNet classification	U Net, RegNetY800MF, and RegNetY32GF	AUC of 0.963 for central stenosis, 0.914 for subarticular stenosis, and 0.931 for foraminal stenosis	Labor intensive annotation and lack of consensus ground truth
[68]	2024	Intervertebral disc and vertebral diseases	Localization and classification	Sagittal and axial T2	Multi angle sagittal and axial view fusion for joint disc and vertebral classification	Mask R CNN and VGG16 for discs, with ResNet50 for vertebrae	Disc F1 score of 96.36% and vertebral F1 score of 85.27%	Small and imbalanced dataset and limited robustness because of missing axial slices

Continued on the next page

Table 4 continued from the previous page

Reference	Year	Pathology	Task	Imaging modality	Key methodological contribution	Architecture	Performance	Limitations
[71]	2025	Disc bulging	Localization and classification	Sagittal T2	Real time localization and classification using YOLOv8m with polygon annotations	YOLOv8m compared with YOLOv7	Mean average precision of 93.7%, precision of 90.5%, and F1 score of 88.6%	Binary classification only and no severity grading
[70]	2025	Disc herniation	Classification	Axial T2	Nearest neighbor dual path contrastive learning	Dual path contrastive learning with ResNet50	Accuracy of 98.34% and F1 score of 97.05%	High computational complexity and reliance on a single public dataset
[72]	2025	Disc herniation and PLID	Detection and classification	Sagittal and axial T2	YOLOv11 based region of interest detection with decision level fusion	YOLOv11, VGG19, and ResNet50	Accuracy of 98.09% and recall of 97.83%	Binary classification, no external validation, and limited dataset diversity

6.5 Presentative of Advanced AI Frameworks

The rapid growth of AI methods for lumbar spine MRI analysis has produced increasingly varied combinations of models, processing stages, and data sources. To support a consistent comparison, this review classifies the frameworks according to their main computational paradigm and information flow rather than relying solely on the names or structures of the neural networks involved. This distinction is important because a single pipeline may contain several deep learning components, making the boundary between an advanced deep learning system and a hybrid framework difficult to define. A pipeline is classified as an advanced deep learning system when its main components are neural networks, including YOLO-based detectors, convolutional neural networks (CNNs), Transformers, and U-Net variants. A framework is classified as hybrid when it combines deep learning with a different computational paradigm, such as conventional machine learning, language models, evolutionary or other optimization algorithms, or structured clinical knowledge. The presence of multiple neural networks alone is therefore not sufficient to define a framework as hybrid. Based on these criteria, the reviewed studies are organized into four practical categories:

1. **Deep learning feature–classifier hybrids**, which use a neural network to extract image features and a separate classifier to make the final prediction;
2. **Multi-stage functional cascades**, which divide the analysis into sequential stages, such as localization, detection, segmentation, classification, and severity grading;
3. **Cross-modal semantic fusion**, which combines imaging information with structured clinical data or other non-image sources; and
4. **Consensus ensemble systems**, which combine predictions from multiple models or processing streams to improve reliability and reduce dependence on a single model.

The categories describe the dominant design principle of each framework and are not intended to be mutually exclusive. A system that combines YOLO with another CNN, for example, remains within the deep learning category because both components use neural-network-based processing. By contrast, a system that combines a CNN with an SVM, an optimization algorithm, or structured clinical knowledge is classified as hybrid. This approach allows architectural complexity to be distinguished from genuine methodological heterogeneity and provides a clearer basis for comparing the reviewed studies.

6.5.1 Hybrid Deep Feature Extraction and Classical Classification

This group of methods combines pretrained convolutional neural networks (CNNs) for image feature extraction with conventional machine learning classifiers, including support vector machines (SVMs) and k-nearest neighbors (k-NN). The CNN generates a feature representation from each MRI image, which is subsequently supplied to the selected classifier. Ebrahim et al. [73] evaluated several pretrained CNN backbones followed by SVM classification. Sari and Aydin Atasoy [74] applied Grey Wolf Optimization (GWO) to select features before SVM classification. By retaining a smaller subset of the

CNN-generated features, GWO may reduce redundancy and noise. This reduction may have contributed to the slightly higher accuracy reported in comparison with the complete feature representation. However, the independent effect of GWO cannot be established without an ablation study using the same data partitions and evaluation protocol. Separating feature extraction from classification reduces the number of trainable parameters and prevents classification errors from being propagated through a frozen CNN backbone. This arrangement may help control overfitting when training data are limited. However, fixed feature representations may not capture the fine-grained patterns needed to distinguish clinically similar conditions, such as mild and moderate spinal stenosis. Performance may therefore decline on external datasets acquired at different institutions, particularly when scanner manufacturers, imaging protocols, or patient populations differ.

6.5.2 Multi-Stage Functional Cascades

Recent frameworks have increasingly used multi-stage decision systems that more closely resemble the clinical workflow followed by radiologists. These systems divide the diagnostic process into sequential modules, each of which performs a specific task prior to the ultimate diagnosis. Instead of looking at the entire MRI volume at once, the process follows a cascading structure. One example is the framework proposed by Zhang et al. [75], which consists of three sequential modules: object detection (Faster R-CNN) for region of interest (ROI) localization, semantic segmentation (U-Net) for identifying tissue boundaries, and downstream feature-based multi-layer perceptrons (MLP) for final ordinal grading. This decoupled structure may partly explain the framework's high accuracy. The downstream MLP classifier uses features from the segmented anatomical region rather than the entire MRI volume, which reduces the influence of irrelevant surrounding structures. This focused representation may help explain the high accuracy reported for diagnosis and grading. Methodologically, this functional cascading approach offers better transparency than conventional end-to-end models because clinicians can review the localization and segmentation results produced at each stage. However, these multi-stage systems also have an important limitation: errors can be carried from one stage to the next. For example, if the initial localization network fails to accurately identify a severely degenerated or anatomically displaced intervertebral disc, the resulting error may affect both the segmentation and the final classification stages. As a result, model performance may decline when the system is applied in routine clinical practice, where imaging findings are often more variable than those seen in curated research datasets.

6.5.3 Cross-Modal Semantic Fusion

To complement MRI image features with clinical information, the cross-modal approach combines different types of data. It brings together imaging features with clinical information to support more comprehensive analysis. For instance, Dong et al. [76] developed a method using a co-attention mechanism to integrate sagittal and axial MRI images, radiology reports, and structured spinal measurements, linking a CNN for image analysis with a BERT-based language model. The key aspect of this approach is the integration of imaging features with structured clinical data and radiology reports in a shared feature space. This allows clinical information to inform image feature learning and more closely reflects routine clinical decision-making. This integration may be useful when imaging findings are difficult to distinguish. When a disc shows features that fall between two adjacent severity grades, the co-attention mechanism combines visual information with clinical measurements and report-derived information. This can help the model use both imaging and clinical evidence when making the final prediction, potentially contributing to the consistent segmentation and multi-class classification results reported across the lumbar disorders evaluated. However, these systems are often limited by challenges in aligning different data modalities and by the limited availability of complete multimodal datasets. Inconsistencies between data sources, missing clinical information, and non-standardized report formats may reduce performance during external validation and limit generalizability across institutions.

6.5.4 Consensus Ensemble Systems

Unlike typical multi-stage systems, this approach is characterized by the use of multiple separately trained models that analyze the same task and combine their predictions to produce the final result. In practice, however, ensemble methods are often incorporated into broader multi-stage diagnostic systems. In a separate study, Sayed et al. [77] proposed an automated framework that uses YOLOv8 for localization together with an ensemble of CNN classifiers, including VGG, ResNet, and DenseNet, and several segmentation models, such as U-Net, LinkNet, and FPN. The ensemble can improve robustness by combining CNN backbones that capture complementary feature representations. This reduces the dependence of the final prediction on the performance of any single model. This may be useful when one framework handles several tasks simultaneously, as in the study by Sayed et al. [77], which combined detection, classification, and segmentation. Because different models are unlikely to make the same errors on the same cases, combining their outputs can improve the overall results across the three tasks. This reduces the chance that strong performance in one task comes at the cost of poorer performance in another. Although ensemble methods often achieve strong performance across different tasks,

their adoption in routine clinical practice is still limited because they require greater computational resources, longer inference times, and more complex deployment, which may limit their integration into routine hospital Picture Archiving and Communication System (PACS). Collectively, these approaches demonstrate how advanced AI frameworks for lumbar spine MRI analysis have progressed toward more integrated AI frameworks. Although each approach offers distinct advantages, none is currently able to achieve strong generalizability, computational efficiency, interpretability, and practical clinical implementation simultaneously. These remaining limitations point to several important directions for future research. **Table 5** summarizes representative advanced AI studies in lumbar spine MRI analysis based on the categorization framework proposed in this review.

Table 5: Representative advanced AI studies for lumbar spine MRI analysis based on the proposed categorization

Reference	Year	Primary AI Paradigm	Pathology	Task	Imaging Modality	Key Methodological Contribution	Architecture	Performance	Limitations
[73]	2024	Deep Learning Feature–Classifier Hybrid	Disc herniation	Classification	Axial T2	Deep feature concatenation from parallel frozen backbones (VGG19 + ResNet101) routed to classical classifiers	AlexNet, VGG16, VGG19, GoogleNet, InceptionV3, ResNet50, ResNet101, InceptionResNetV2, SqueezeNet; SVM, DNN, KNN, NB, Discriminant, DT	Acc.=98% (SVM on VGG19 + ResNet101 features)	Small dataset (164 images); single-center; no external validation
[74]	2024	Deep Learning Feature–Classifier Hybrid	Disc herniation	Binary classification	T1-weighted	Deep feature optimization using Grey Wolf Optimization (GWO) on ResNet50 descriptors	ResNet50, GWO, SVM, MLP	Acc.=99.42%; Precision=99.42%; Recall=99.42%; F1=99.42%	Binary classification only; single public dataset; no external validation
[75]	2024	Multi-Stage Functional Cascade	Disc herniation	Diagnosis + MSU + Pfirrmann + treatment advice	Sagittal T2 + Axial T2	Deconstructed functional pipeline for sequential object localization, boundary segmentation, and grading	Faster R-CNN, CNN, U-Net, MLP	Localization sensitivity=99.5%; Diagnosis Acc.=95.83%; MSU Acc.=95.0%; Pfirrmann Acc.=83.5%; Dice=95.0%; IoU=90.4%	Single-center study; potential systemic bias; no external validation; image-only input without clinical symptom integration
[76]	2024	Cross-Modal Semantic Fusion	Lumbar disorders (DDD, spinal stenosis, spondylolisthesis, etc.)	Segmentation + Multi-class classification	Mid-sagittal MRI	Early fusion of CNN-derived imaging features, quantitative spinal measurements, and clinical reports	CNN (feature extraction/segmentation) + BERT-based Spinal LLM	Dice$\geq$0.92; IoU$\geq$0.88; Acc./Pre./Rec./F1$\approx$0.90 across 61 disease combinations; externally validated on 514 cases	Potential performance variability across diverse populations; dependence on standardized pre-processing pipeline
[77]	2025	Consensus Ensemble System	PLID / Disc herniation	Detection + Classification + Segmentation	Sagittal T2	Ensemble-driven consensus framework embedded within a multi-stage functional cascade via parallel weighted CNN/segmentation models	YOLOv8; VGG16, VGG19, ResNet50, DenseNet121, InceptionV3, Xception; U-Net, LinkNet, FPN, PSPNet, DeepLabv3+	Det: mAP50=99.50%, 99.40%; Cls: Acc.=99.37%; Seg: Dice=95.74%, IoU=91.86%	Single-center dataset; no external validation; class imbalance (addressed via weighting); overfitting risk due to limited data; no cross-validation

Note. DDD: degenerative disc disease; DNN: deep neural network; DT: decision tree; FPN: feature pyramid network; GWO: Grey Wolf Optimization; IoU: intersection over union; KNN: k nearest neighbors; MLP: multilayer perceptron; MSU: Michigan State University classification; NB: naive Bayes; PLID: prolapsed lumbar intervertebral disc; PSPNet: pyramid scene parsing network; SVM: support vector machine; U Net: U Net convolutional network.

7. CHALLENGES AND FUTURE DIRECTIONS

Training data quality and heterogeneity are among the key factors affecting the performance of artificial intelligence (AI) models in lumbar spine MRI analysis. Limited dataset size, class imbalance, and variability across imaging protocols and MRI scanners can reduce model robustness and limit generalizability in clinical settings [78]. In addition, many AI systems trained on single-center datasets may not perform well on independent patient cohorts because of domain shift. This arises from differences in image characteristics across institutions, patient populations, MRI vendors, and acquisition protocols. To address the limited availability of training data and increase dataset diversity, researchers have explored conventional augmentation techniques together with generative approaches based on generative adversarial networks and diffusion models [79, 80]. These methods can introduce greater variability into the training data and may help reduce overfitting. However, their clinical value depends on whether the generated images preserve anatomically meaningful structures and reflect the pathological patterns observed in real patients. Synthetic images should therefore be assessed for anatomical fidelity, label consistency, image quality, and potential memorization before being incorporated into model development. In addition, it remains difficult to ensure that generated datasets represent the full range of disease severity, morphological variation, and imaging characteristics encountered in routine clinical practice. Although considerable progress has been made, a number of limitations continue to restrict the broader clinical use of deep learning techniques in lumbar spine MRI analysis. Current studies often use small and relatively homogeneous datasets, which limits their

use across different clinical settings. In addition, differences in evaluation methods and reporting standards make direct comparison between studies difficult. Another important challenge is the limited interpretability of many deep learning models, which may reduce clinician confidence and slow their adoption in clinical practice [81, 82]. Reducing the effects of domain shift requires moving from single-center internal validation to multicenter external validation studies. Testing deep learning models on independent prospective patient cohorts from different institutions is essential for demonstrating clinical generalizability and reliable performance across different clinical settings [68, 72, 75].

7.1 Explainable AI (XAI) and Trustworthy Systems

The adoption of deep learning models in routine spinal radiology is still limited by the well-known "black-box" problem. Although attention-guided CNNs and Transformer-based models often achieve high diagnostic performance, it can be difficult to understand how they arrive at a particular severity grade. In clinical applications such as lumbar spine MRI analysis, building clinician trust, demonstrating model reliability, and meeting regulatory requirements require greater use of Explainable AI (XAI) and more transparent AI systems.

7.1.1 Explainability Frameworks and Interpretability Mechanisms

To better understand how AI models reach their decisions, current lumbar spine imaging studies mainly use post-hoc explainability methods. These methods are generally divided into two groups: gradient-based visual explanation techniques and perturbation-based or game-theoretic methods:

- **Gradient-Based Saliency and Activation Mapping:** Frameworks such as Gradient-weighted Class Activation Mapping (Grad-CAM) proposed by Selvaraju et al. [83], Grad-CAM++ by Chattopadhay et al. [84], and Integrated Gradients by Sundararajan et al. [85] are commonly used to generate visual heatmaps directly on sagittal and axial MRI slices. For tasks such as automated Pfirrmann grading or disc herniation classification, these techniques estimate the gradients of the target score in relation to the final convolutional feature maps. These methods highlight the relevant anatomical regions such as the signal intensity loss within the nucleus pulposus or the focal boundaries of a herniated disc fragment compressing a nerve root that contributed most to the model's categorical inference.
- **Game-Theoretic and Perturbation Frameworks:** Shapley Additive exPlanations (SHAP) introduced by Lundberg and Lee [86] and Local Interpretable Model-agnostic Explanations (LIME) developed by Ribeiro et al. [87] offer a quantitative assessment of feature importance based on cooperative game theory and local perturbations. These frameworks selectively mask the chosen image patches, e.g., specific vertebral endplates or the region of the spinal canal, and the effect on diagnostic confidence is evaluated. This allows radiologists to assess whether the model is based on clinically relevant features e.g. osteophyte formation or ligamentum flavum hypertrophy , rather than image characteristics irrelevant for clinical diagnosis or scanner-related artifacts.

7.1.2 Bridging the Translational Gap: Clinical Trust and Radiologist Integration

The main goal of using XAI in lumbar spine analysis is to build clinical trust by addressing the key question of algorithmic reliability emphasized by Ribeiro et al. [87]. For radiologists in clinical practice, a high area under the receiver operating characteristic curve (AUC) or a superior Dice coefficient is insufficient without a verifiable diagnostic rationale. Visual heatmaps and regional feature attribution metrics serve as an objective "second-opinion" by linking model decisions to widely used radiological grading systems, such as the MSU and modified Pfirrmann classifications. By clearly highlighting the features underlying a specific classification, XAI methods can help reduce interobserver variability and support clinicians in identifying subtle or borderline degenerative changes that might otherwise be missed during busy clinical workflows. Furthermore, transparent model outputs can help clinicians detect potential algorithmic errors and the impact of domain shift, for example, when a model misclassifies a normal anatomical variant as severe stenosis. This allows clinicians to stay involved in clinical decisions and may contribute to safer patient care.

7.2 Future Research Directions

Future research on AI for lumbar spine MRI analysis may involve multimodal learning approaches that integrate MRI information with clinical, genetic and patient-specific information based on the cross-modal semantic fusion paradigm discussed above. Such approaches may enable a more comprehensive understanding of disease characteristics and support more individualized clinical decision-making. In parallel, self-supervised learning and foundation models are emerging as promising approaches for minimizing reliance on large annotated datasets, particularly for feature–classifier hybrid and multi-stage cascade architectures that currently depend on extensive manual annotation. Furthermore, wider adoption

of Explainable AI (XAI) techniques could enhance the clarity of AI-assisted decision-making across all four proposed framework categories and support the integration of these systems into real-world clinical settings.

8. CONCLUSION

Lumbar spine disorders, including intervertebral disc degeneration (IVDD), disc herniation (DH), and lumbar spinal stenosis (LSS), are common causes of low back pain and disability worldwide. Magnetic resonance imaging (MRI) is the main imaging method used to assess these conditions because it provides clear images of soft tissues. However, reviewing MRI scans can take considerable time, and differences in radiologists' interpretations may lead to variation in diagnosis. This review examines recent work on the use of artificial intelligence (AI) to analyze lumbar spine MRI scans. It covers lumbar anatomy, disease mechanisms, publicly available MRI datasets, commonly used grading systems, and the main steps involved in MRI analysis, including preprocessing, localization, detection, segmentation, classification, and severity assessment. Recent deep learning approaches are grouped into four categories: feature–classifier hybrids, multistage processing pipelines, cross-modal semantic fusion methods, and consensus-based ensemble systems. This organization provides a clearer basis for comparing the methods, strengths, and limitations reported across different studies. The studies reviewed indicate considerable progress in the automated analysis of lumbar spine MRI. High levels of diagnostic performance have been reported for localization, segmentation, classification, and severity grading. However, promising results in research settings do not necessarily translate into reliable clinical tools. The adoption of these methods remains constrained by differences among datasets, imbalanced classes, changes between imaging sites, inconsistent annotations, limited multicenter and external validation, and the difficulty of understanding how some deep learning models reach their conclusions. The lack of consistent evaluation and reporting standards also makes it difficult to compare findings across studies. Improving model accuracy alone will not be enough to ensure further progress. Greater attention is needed to the development of reliable and standardized systems that clinicians can understand and use in everyday practice. Future research should include prospective studies across multiple centers, consistent evaluation procedures, and clearer explanations of model predictions through explainable AI methods. Multimodal learning, self-supervised learning, and foundation models also deserve further investigation because of their potential to learn useful patterns from large and diverse medical imaging datasets. Addressing these issues could improve model performance across different clinical settings and support the safe use of AI in lumbar spine MRI assessment. Overall, this review brings together recent clinical and technical developments in AI-assisted lumbar spine MRI analysis. It also outlines the main methodological challenges and identifies research priorities for developing reliable systems that can move beyond experimental studies and closer to routine clinical practice.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally to the study conception and design. Material preparation, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Ethics declaration: not applicable. This study did not involve human participants or animals. Therefore, ethical approval and consent to participate are not applicable.

CONSENT FOR PUBLICATION

Consent to Publish declaration: not applicable.

DATA AVAILABILITY

No new primary clinical datasets were generated during this study. This work is based on a structured review and comparative analysis of previously published literature.

ACKNOWLEDGMENTS

The authors acknowledge the use of ChatGPT-5 for assistance in refining the manuscript. The authors reviewed and revised all assisted text and accept full responsibility for the accuracy, integrity, and originality of the final manuscript.

FUNDING

No funding.

DISCLOSURE STATEMENT

The authors declare no conflict of interest.

REFERENCES

[1] A. Raja'S, J. J. Corso, V. Chaudhary, and G. Dhillon, "Automatic diagnosis of lumbar disc herniation with shape and appearance features from mri," in *Medical Imaging 2010: Computer-Aided Diagnosis*, vol. 7624, pp. 364–372, SPIE, 2010.

[2] J. N. Katz and M. B. Harris, "Lumbar spinal stenosis," *New England Journal of Medicine*, vol. 358, no. 8, pp. 818–825, 2008.

[3] D. F. Fardon, A. L. Williams, E. J. Dohring, F. R. Murtagh, S. L. G. Rothman, and G. K. Sze, "Lumbar disc nomenclature: version 2.0: Recommendations of the combined task forces of the north american spine society, the american society of spine radiology and the american society of neuroradiology," *The spine journal*, vol. 14, no. 11, pp. 2525–2545, 2014.

[4] V. Tumko, J. Kim, N. Uspenskaia, S. Honig, F. Abel, D. R. Lebl, I. Hotalen, S. Kolisnyk, M. Kochnev, A. Rusakov, *et al.*, "A neural network model for detection and classification of lumbar spinal stenosis on mri," *European Spine Journal*, vol. 33, no. 3, pp. 941–948, 2024.

[5] R. U. Din, X. Cheng, and H. Yang, "Diagnostic role of magnetic resonance imaging in low back pain caused by vertebral endplate degeneration," *Journal of Magnetic Resonance Imaging*, vol. 55, no. 3, pp. 755–771, 2022.

[6] V. P. Grover, J. M. Tognarelli, M. M. Crossey, I. J. Cox, S. D. Taylor-Robinson, and M. J. McPhail, "Magnetic resonance imaging: principles and techniques: lessons for clinicians," *Journal of clinical and experimental hepatology*, vol. 5, no. 3, pp. 246–255, 2015.

[7] J. A. Carrino, J. D. Lurie, A. N. Tosteson, T. D. Tosteson, E. J. Carragee, J. Kaiser, M. R. Grove, E. Blood, L. H. Pearson, J. N. Weinstein, *et al.*, "Lumbar spine: reliability of mr imaging findings," *Radiology*, vol. 250, no. 1, pp. 161–170, 2009.

[8] J. Lin, H. Zhang, and H. Shang, "Convolutional neural network incorporating multiple attention mechanisms for mri classification of lumbar spinal stenosis," *Bioengineering*, vol. 11, no. 10, p. 1021, 2024.

[9] M. Khalifa and M. Albadawy, "Ai in diagnostic imaging: revolutionising accuracy and efficiency," *Computer Methods and programs in biomedicine update*, vol. 5, p. 100146, 2024.

[10] N. C. Lehnen, R. Haase, J. Faber, T. Rüber, H. Vatter, A. Radbruch, and F. C. Schmeel, "Detection of degenerative changes on mr images of the lumbar spine with a convolutional neural network: a feasibility study," *Diagnostics*, vol. 11, no. 5, p. 902, 2021.

[11] M. Alabsi, A. AQashlan, R. Mohamed, A. Taha, *et al.*, "Automated brain tumor segmentation via yolov8-derived spatial prompts for the segment anything model," *Computational Discovery and Intelligent Systems (CDIS)*, vol. 3, no. 2, pp. 150–163, 2026.

[12] R. Compte, I. Granville Smith, A. Isaac, N. Danckert, T. McSweeney, P. Liantis, and F. M. Williams, "Are current machine learning applications comparable to radiologist classification of degenerate and herniated discs and modic change? a systematic review and meta-analysis," *European Spine Journal*, vol. 32, no. 11, pp. 3764–3787, 2023.

[13] M. Hussain, D. Koundal, and J. Manhas, "Deep learning-based diagnosis of disc degenerative diseases using mri: a comprehensive review," *Computers and Electrical Engineering*, vol. 105, p. 108524, 2023.

[14] M. A. Adams and P. J. Roughley, "What is intervertebral disc degeneration, and what causes it?," *Spine*, vol. 31, no. 18, pp. 2151–2161, 2006.

[15] W. Mbarki, M. Bouchouicha, S. Frizzi, F. Tshibasu, L. B. Farhat, and M. Sayadi, "Lumbar spine discs classification based on deep convolutional neural networks using axial view mri," *Interdisciplinary Neurosurgery*, vol. 22, p. 100837, 2020.

[16] J. Dowdell, M. Erwin, T. Choma, A. Vaccaro, J. Iatridis, and S. K. Cho, "Intervertebral disk degeneration and repair," *Neurosurgery*, vol. 80, no. 3S, pp. S46–S54, 2017.

[17] S. Salamat, J. Hutchings, C. Kwong, J. Magnussen, and M. J. Hancock, "The relationship between quantitative measures of disc height and disc signal intensity with pfirrmann score of disc degeneration," *Springerplus*, vol. 5, no. 1, p. 829, 2016.

[18] C. W. Pfirrmann, A. Metzdorf, M. Zanetti, J. Hodler, and N. Boos, "Magnetic resonance classification of lumbar intervertebral disc degeneration," *spine*, vol. 26, no. 17, pp. 1873–1878, 2001.

[19] A. N. Laiwalla, A. Ratnaparkhi, D. Zarrin, K. Cook, I. Li, B. Wilson, T. Florence, B. Yoo, B. Salehi, B. Gaonkar, *et al.*, "Lumbar spinal canal segmentation in cases with lumbar stenosis using deep-u-net ensembles," *World Neurosurgery*, vol. 178, pp. e135–e140, 2023.

[20] E. Ebrahimzadeh, F. Fayaz, F. Ahmadi, and M. Nikravan, "A machine learning-based method in order to diagnose lumbar disc herniation disease by mr image processing," *MedLife Open Access*, vol. 1, no. 1, p. 1, 2018.

[21] G. Valarmathi and S. N. Devi, "Automatic localization and classification of intervertebral disc herniation using hybrid classifier," *Biomedical Signal Processing and Control*, vol. 86, p. 105291, 2023.

[22] J. N. Katz, Z. E. Zimmerman, H. Mass, and M. C. Makhni, "Diagnosis and management of lumbar spinal stenosis: a review," *Jama*, vol. 327, no. 17, pp. 1688–1699, 2022.

[23] D. S. Kreiner, W. O. Shaffer, J. L. Baisden, T. J. Gilbert, J. T. Summers, J. F. Toton, S. W. Hwang, R. C. Mendel, and C. A. Reitman, "An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spinal stenosis (update)," *The Spine Journal*, vol. 13, no. 7, pp. 734–743, 2013.

[24] F. Galbusera, G. Casaroli, and T. Bassani, "Artificial intelligence and machine learning in spine research," *JOR spine*, vol. 2, no. 1, p. e1044, 2019.

[25] S. Sudirman, A. Al Kafri, F. Natalia, H. Meidia, N. Afriliana, W. Al-Rashdan, M. Bashtawi, and M. Al-Jumaily, "Lumbar spine mri dataset," 2019.

[26] R. F. Masood, T. Hassan, M. Akram, I. Taj, M. Qureshi, and M. Khan, "Composite dataset of lumbar spine mid-sagittal images with annotations and clinically relevant spinal measurements," *Mendeley Data: London, UK*, 2021.

[27] F. Natalia, S. Sudirman, D. Ruslim, and A. Al-Kafri, "Lumbar spine mri annotation with intervertebral disc height and pfirrmann grade predictions," *PLoS One*, vol. 19, no. 5, p. e0302067, 2024.

[28] J. W. van der Graaf, M. L. van Hooff, C. F. Buckens, M. Rutten, J. L. van Susante, R. J. Kroeze, M. de Kleuver, B. van Ginneken, and N. Lessmann, "Lumbar spine segmentation in mr images: a dataset and a public benchmark," *Scientific Data*, vol. 11, no. 1, p. 264, 2024.

[29] T. J. Richards, A. E. Flanders, E. Colak, L. M. Prevedello, R. L. Ball, F. Kitamura, J. Mongan, M. Vazirabad, H.-M. Lin, A. Kendell, *et al.*, "The rsna lumbar degenerative imaging spine classification (lumbardisc) dataset," *Radiology: Artificial Intelligence*, vol. 8, no. 2, p. e250480, 2026.

[30] N. Delfan, M. Parvaresh, M. Shafieeha, and B. Moshiri, "Lspinesmri: A comprehensive dataset of non-contrast lumbar spine stenosis mri examinations," 2025.

[31] B. Zárate-Kalfópulos, F. Reyes-Tarrago, L. A. Navarro-Aceves, C. L. García-Ramos, A. A. Reyes-Sánchez, A. Alpízar-Aguirre, and L. M. Rosales-Olivarez, "Characteristics of spinopelvic sagittal alignment in lumbar degenerative disease," *World neurosurgery*, vol. 126, pp. e417–e421, 2019.

[32] F. D'Antoni, F. Russo, L. Ambrosio, L. Bacco, L. Vollero, G. Vadalà, M. Merone, R. Papalia, and V. Denaro, "Artificial intelligence and computer aided diagnosis in chronic low back pain: A systematic review," *International journal of environmental research and public health*, vol. 19, no. 10, p. 5971, 2022.

[33] A. Wang, C. Zou, S. Yuan, N. Fan, P. Du, T. Wang, and L. Zang, "Deep learning assisted segmentation of the lumbar intervertebral disc: a systematic review and meta-analysis," *Journal of Orthopaedic Surgery and Research*, vol. 19, no. 1, p. 496, 2024.

[34] T. Wang, R. Chen, N. Fan, L. Zang, S. Yuan, P. Du, Q. Wu, A. Wang, J. Li, X. Kong, *et al.*, "Machine learning and deep learning for diagnosis of lumbar spinal stenosis: systematic review and meta-analysis," *Journal of medical Internet research*, vol. 26, p. e54676, 2024.

[35] W. Liawrungrueang, J.-B. Park, W. Cholamjiak, P. Sarasombath, and K. D. Riew, "Artificial intelligence-assisted mri diagnosis in lumbar degenerative disc disease: a systematic review," *Global Spine Journal*, vol. 15, no. 2, pp. 1405–1418, 2025.

[36] P. Zhao and S. Zhu, "Advances and challenges in ai-assisted mri for lumbar disc degeneration detection and classification," *European Spine Journal*, vol. 35, no. 3, pp. 1291–1300, 2026.

[37] X. Wei, L. Gengwu, C. Chao, L. Yifan, S. Shang, H. Ruixi, J. Yunhan, Z. Xiaodong, and L. Zhikun, "Correlations between the sagittal plane parameters of the spine and pelvis and lumbar disc degeneration," *Journal of orthopaedic surgery and research*, vol. 13, no. 1, p. 137, 2018.

[38] J. F. Griffith, Y.-X. J. Wang, G. E. Antonio, K. C. Choi, A. Yu, A. T. Ahuja, and P. C. Leung, "Modified pfirrmann grading system for lumbar intervertebral disc degeneration," *Spine*, vol. 32, no. 24, pp. E708–E712, 2007.

[39] Z. Soydan, E. Bayramoglu, D. U. Urut, A. C. Iplikcioglu, and C. Sen, "Tracing the disc: the novel qualitative morphometric mri based disc degeneration classification system," *JOR spine*, vol. 7, no. 1, p. e1321, 2024.

[40] L. W. Mysliwiec, J. Cholewicki, M. D. Winkelpleck, and G. P. Eis, "Msu classification for herniated lumbar discs on mri: toward developing objective criteria for surgical selection," *European spine journal*, vol. 19, no. 7, pp. 1087–1093, 2010.

[41] R. I. Riesenburger, M. G. Safain, R. Ogbuji, J. Hayes, and S. W. Hwang, "A novel classification system of lumbar disc degeneration," *Journal of Clinical Neuroscience*, vol. 22, no. 2, pp. 346–351, 2015.

[42] C. Aprill and N. Bogduk, "High-intensity zone: a diagnostic sign of painful lumbar disc on magnetic resonance imaging," *The British journal of radiology*, vol. 65, no. 773, pp. 361–369, 1992.

[43] M. A. Yoon, S.-J. Hong, C. H. Kang, K.-S. Ahn, and B. H. Kim, "T1rho and t2 mapping of lumbar intervertebral disc: correlation with degeneration and morphologic changes in different disc regions," *Magnetic resonance imaging*, vol. 34, no. 7, pp. 932–939, 2016.

[44] O. Togao, A. Hiwatashi, T. Wada, K. Yamashita, K. Kikuchi, C. Tokunaga, J. Keupp, M. Yoneyama, and H. Honda, "A qualitative and quantitative correlation study of lumbar intervertebral disc degeneration using glycosaminoglycan chemical exchange saturation transfer, pfirrmann grade, and t1-ρ," *American Journal of Neuroradiology*, vol. 39, no. 7, pp. 1369–1375, 2018.

[45] W. Yi, J. Zhao, W. Tang, H. Yin, L. Yu, Y. Wang, and W. Tian, "Deep learning-based high-accuracy detection for lumbar and cervical degenerative disease on t2-weighted mr images," *European Spine Journal*, vol. 32, no. 11, pp. 3807–3814, 2023.

[46] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[47] A. Piórkowski, R. Obuchowicz, A. Urbanik, and M. Strzelecki, "Advances in musculoskeletal imaging and their applications," *Journal of Clinical Medicine*, vol. 12, no. 20, 2023.

[48] R. Obuchowicz, M. Strzelecki, and A. Piórkowski, "Clinical applications of artificial intelligence in medical imaging and image processing—a review," *Cancers*, vol. 16, no. 10, 2024.

[49] S. Pang, C. Pang, L. Zhao, Y. Chen, Z. Su, Y. Zhou, M. Huang, W. Yang, H. Lu, and Q. Feng, "Spineparsenet: spine parsing for volumetric mr image by a two-stage segmentation framework with semantic image representation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 262–273, 2020.

[50] H.-D. Zheng, Y.-L. Sun, D.-W. Kong, M.-C. Yin, J. Chen, Y.-P. Lin, X.-F. Ma, H.-S. Wang, G.-J. Yuan, M. Yao, *et al.*, "Deep learning-based high-accuracy quantitation for lumbar intervertebral disc degeneration from mri," *Nature communications*, vol. 13, no. 1, p. 841, 2022.

[51] N. T. Singh, C. Kaur, A. Chaudhary, and S. Goyal, "Preprocessing of medical images using deep learning: A comprehensive review," in *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, pp. 521–527, IEEE, 2023.

[52] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, "A review of medical image data augmentation techniques for deep learning applications," *Journal of medical imaging and radiation oncology*, vol. 65, no. 5, pp. 545–563, 2021.

[53] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[54] S. Pang, Z. Su, S. Leung, I. B. Nachum, B. Chen, Q. Feng, and S. Li, "Direct automated quantitative measurement of spine by cascade amplifier regression network with manifold regularization," *Medical image analysis*, vol. 55, pp. 103–115, 2019.

[55] T. Martín-Noguerol, M. O. Miranda, T. J. Amrhein, F. Paulano-Godino, P. Xiberta, J. C. Vilanova, and A. Luna, "The role of artificial intelligence in the assessment of the spine and spinal cord," *European journal of radiology*, vol. 161, p. 110726, 2023.

[56] E. Tsalera, A. Papadakis, M. Samarakou, and I. Voyiatzis, "Feature extraction with handcrafted methods and convolutional neural networks for facial emotion recognition," *Applied Sciences*, vol. 12, no. 17, p. 8455, 2022.

[57] R. Ponnusamy, S. Sathyamoorthy, and K. Manikandan, "A review of image classification approaches and techniques," *International Journal of Recent Trends in Engineering & Research*, vol. 3, no. 3, pp. 1–5, 2017.

[58] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[59] B. Han and L. S. Davis, "Density-based multifeature background subtraction with support vector machine," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1017–1023, 2011.

[60] J.-T. Lu, S. Pedemonte, B. Bizzo, S. Doyle, K. P. Andriole, M. H. Michalski, R. G. Gonzalez, and S. R. Pomerantz, "Deep spine: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning," in *Machine learning for healthcare conference*, pp. 403–419, PMLR, 2018.

[61] Q. Pan, K. Zhang, L. He, Z. Dong, L. Zhang, X. Wu, Y. Wu, and Y. Gao, "Automatically diagnosing disk bulge and disk herniation with lumbar magnetic resonance images by using deep convolutional neural networks: method development study," *JMIR medical informatics*, vol. 9, no. 5, p. e14755, 2021.

[62] F. Niemeyer, F. Galbusera, Y. Tao, A. Kienle, M. Beer, and H.-J. Wilke, "A deep learning model for the accurate and reliable classification of disc degeneration based on mri data," *Investigative radiology*, vol. 56, no. 2, pp. 78–85, 2021.

[63] J.-Y. Tsai, I. Y.-J. Hung, Y. L. Guo, Y.-K. Jan, C.-Y. Lin, T. T.-F. Shih, B.-B. Chen, and C.-W. Lung, "Lumbar disc herniation automatic detection in magnetic resonance imaging based on deep learning," *Frontiers in Bioengineering and Biotechnology*, vol. 9, p. 708137, 2021.

[64] W. Liawrungrueang, P. Kim, V. Kotheeranurak, K. Jitpakdee, and P. Sarasombath, "Automatic detection, classification, and grading of lumbar intervertebral disc degeneration using an artificial neural network model," *Diagnostics*, vol. 13, no. 4, p. 663, 2023.

[65] T. Šušteršič, V. Ranković, V. Milovanović, V. Kovačević, L. Rasulić, and N. Filipović, "A deep learning model for automatic detection and classification of disc herniation in magnetic resonance images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6036–6046, 2022.

[66] T. Shahzadi, M. U. Ali, F. Majeed, M. U. Sana, R. M. Diaz, M. A. Samad, and I. Ashraf, "Nerve root compression analysis to find lumbar spine stenosis on mri using cnn," *Diagnostics*, vol. 13, no. 18, p. 2975, 2023.

[67] J. Qian, G. Su, X. Shu, K. Shen, B. Chen, and X. Wang, "Lumbar disc herniation diagnosis using deep learning on mri," *Journal of Radiation Research and Applied Sciences*, vol. 17, no. 3, p. 100988, 2024.

[68] K. Chen, L. Zheng, H. Zhao, and Z. Wang, "Deep learning-based intelligent diagnosis of lumbar diseases with multi-angle view of intervertebral disc," *Mathematics*, vol. 12, no. 13, p. 2062, 2024.

[69] Y. Xu, S. Zheng, Q. Tian, Z. Kou, W. Li, X. Xie, and X. Wu, "Deep learning model for grading and localization of lumbar disc herniation on magnetic resonance imaging," *Journal of Magnetic Resonance Imaging*, vol. 61, no. 1, pp. 364–375, 2025.

[70] D. Pan, Y.-X. Pan, H. Wang, Q.-J. Liu, and C.-M. Qiu, "Nearest-neighbor dual-path contrastive learning for lumbar disc herniation mri image classification," *IEEE Access*, vol. 13, pp. 92906–92920, 2025.

[71] M. Abed, Z. Al-Qaysi, and M. Suzani, "Improving lumbar disc bulging detection in mri spinal imaging: A deep learning approach," *Al-Salam Journal for Engineering and Technology*, vol. 4, no. 1, pp. 1–19, 2025.

[72] M. A. Sayed, A. Rahman, S. M. Nasif, S. Halder, A. Hossain, H. Ahmed, and M. A. Kadir, "Automated lumbar intervertebral disc identification and herniation detection in mr images using cascade cnn architecture," *Informatics in Medicine Unlocked*, vol. 55, p. 101648, 2025.

[73] M. Ebrahim, M. Alsmirat, and M. Al-Ayyoub, "Advanced disk herniation computer aided diagnosis system," *Scientific Reports*, vol. 14, no. 1, p. 8071, 2024.

[74] Y. Sarı and N. Aydın Atasoy, "Enhanced disc herniation classification using grey wolf optimization based on hybrid feature extraction and deep learning methods," *Tomography*, vol. 11, no. 1, p. 1, 2024.

[75] D. Zhang, J. Du, J. Shi, Y. Zhang, S. Jia, X. Liu, Y. Wu, Y. An, S. Zhu, D. Pan, *et al.*, "A fully automatic mri-guided decision support system for lumbar disc herniation using machine learning," *JOR spine*, vol. 7, no. 2, p. e1342, 2024.

[76] R. Dong, X. Cheng, M. Kang, and Y. Qu, "Classification of lumbar spine disorders using large language models and mri segmentation," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 343, 2024.

[77] M. A. Sayed, G. M. Rahman, M. S. Islam, M. A. Islam, J. Park, H. Ahmed, A. Hossain, and R. Shahrior, "Automatic detection, classification, and segmentation of sagittal mr images for diagnosing prolapsed lumbar intervertebral disc," *Scientific Reports*, vol. 15, no. 1, p. 593, 2025.

[78] F. Gao and M. Wu, "Deep learning-based denoised mri images for correlation analysis between lumbar facet joint and lumbar disc herniation in spine surgery," *Journal of healthcare engineering*, vol. 2021, no. 1, p. 9687591, 2021.

[79] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep learning approaches for data augmentation in medical imaging: a review," *Journal of imaging*, vol. 9, no. 4, p. 81, 2023.

[80] Y. Chen, X.-H. Yang, Z. Wei, A. A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, and Q. Guan, "Generative adversarial networks in medical image augmentation: a review," *Computers in Biology and Medicine*, vol. 144, p. 105382, 2022.

[81] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[82] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC medicine*, vol. 17, no. 1, p. 195, 2019.

[83] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[84] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847, IEEE, 2018.

[85] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*, pp. 3319–3328, PMLR, 2017.

[86] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[87] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, Association for Computing Machinery, 2016.