British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 2, 2026 (pp. 68-86)
www.abjournals.org

FROM BLACK BOX TO CLINICAL TRUST: A CONCEPTUAL REVIEW OF EXPLAINABLE AND LIGHTWEIGHT DEEP LEARNING FOR SPINE DISEASE DETECTION AND SEGMENTATION

Eze Monday[1], Ebiesuwa Oluwaseun[2], Oyebola Akande[3], Okesola Kikelomo I.[4],

Ojo Abosede Ibironke[5], and Mgbeahuruike Emmanuel O.[6]

Emails:
[1]ezem@babcock.edu.ng; ; [2]ebiesuwao@babcock.edu.ng; [3]akandeo@babcock.edu.ng;
[4]okesolak@babcock.edu.ng; [5]abosedeojo@ogitech.edu.ng; [6]mgbeahuruikee@babcock.edu.ng

Cite this article:

Eze, M., Ebiesuwa, O., Oyebola, A., Okesola, K., Ojo, A., Mgbeahuruike, E. (2026), From Black Box to Clinical Trust: A Conceptual Review of Explainable and Lightweight Deep Learning for Spine Disease Detection and Segmentation. British Journal of Computer, Networking and Information Technology 9(2), 68-86. DOI: 10.52589/BJCNIT-MIZPEFHD

Manuscript History

Received: 2 Apr 2026

Accepted: 6 May 2026

Published: 25 Jun 2026

ABSTRACT: *Deep learning models have made significant contributions to the accurate detection of spinal anomalies. However, persistent challenges related to model transparency, explainability and computational demands continue to hinder the real-time deployment of AI-driven solutions in clinical settings. This review examines contributions from researchers on improving the explainability and transparency of deep learning models for spine image analysis, with the aim of identifying promising approaches and informing future research directions. Specifically, it explores the potential synergy between explainable AI techniques and lightweight models, with the expectation that such integration will yield models that are simultaneously accurate, interpretable, and clinically deployable. Concepts covered include deep learning architectures applied to spine imaging tasks such as classification and segmentation; lightweight model design strategies; and categories of explainability techniques including Grad-CAM, LIME, and attention mechanisms. Constraints to the full clinical adoption of AI solutions in spine imaging are also discussed. Key findings highlight several gaps in the research field. Dataset limitation issue, absence of standardised metrics for evaluating model interpretability, and challenges surrounding clinical acceptability. These gaps point to a research direction that calls for greater collaboration between healthcare professionals and AI researchers to develop spine imaging solutions that are both explainable and practically usable in clinical environments.*

KEYWORDS: Spine anomaly, lightweight, explainable- AI, deep learning, medical imaging.

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 2, 2026 (pp. 68-86)
www.abjournals.org

INTRODUCTION

Detection and segmentation are major procedures in spine image analysis to guide in anomaly detection, generating a treatment plan, and predicting patient outcomes. Spine issues range from different classes of spine diseases, such as scoliosis, spondylosis, and intervertebral disc degeneration, to fracture-related issues that are subtle and challenging to comprehend by the radiologists.

Precise vertebral segmentation is essential for tasks such as reliable Cobb angle calculation in scoliosis analysis, grading of disc degeneration, localisation of degenerative myelopathy and others. With the recent research trend in deep learning models, there has been impressive performance of detection algorithms in accurate identification of anomalies in the spine (Bin Ahmad et al., 2024), (Kong et al., 2022). As well as the implementation of different segmentation techniques (Z. Zhang et al., 2024) , (Hess et al., 2023), that precisely map out vertebrae, disc, and spinal canal structures. Using automated approach enhances diagnostic accuracy and promotes reproducibility to guide in clinical decision-making process, patient management, outcome prediction, disease monitoring and surgical planning (Li et al., 2021); consequently, streamlining clinical workflow (Maraş et al., 2022), (Lee et al., 2022). The integration of detection and segmentation models into clinical procedures creates a pathway for advancing precision medicine in spine healthcare.

Deep learning has emerged as a transformative tool in spinal imaging. Previous traditional approaches depend more on the expertise of the radiologist, making them subjective to observer variation, workload pressure, and the subtle nature of spinal anomalies (Yıldız Potter et al., 2024). However, segmentation-focused architectures have contributed greatly to the advancement in spinal imaging analysis. Some of the proposed architectures are Verdiff-Net (Z. Zhang et al., 2024), multi-tissue (Hess et al., 2023), and integration of lightweight models for real-time deployment (Liawrungrueang et al., 2023). Though these models are accurate, for example Transformer-based and CNN-based models, their interpretability and non-transparency of the reasoning process of these deep models remains a challenge (Haar et al., 2023) and makes it difficult to comprehend model's output (Sutradhar et al., 2025), (P. Chen et al., 2020). These limitations led to the restriction of radiologists and spine specialists in the adoption and acceptability of automated algorithms (Maraş et al., 2022); hence, the need for developing transparent models that balance performance with interpretability (Wang et al., 2024).

Grad-CAM and SHAP explainable techniques were explored in an attempt to enable model explainability through the mapping of specific features or regions in a medical image to model predictions. Despite this attempt, the model's underlying architecture remains inherently complex (Pereira et al., 2024), (Dindorf et al., 2023), establishing the need for lightweight models. The lack of resources in hospitals and mobile health systems is a major factor in the need for lightweight models in spine imaging. Many health care facilities do not have access to edge infrastructures and high computing resources required by most of the complex models (Basak et al., 2025).

Integrating lightweight models and explainable techniques is central to earning clinical trust in AI-assisted spine imaging. Early research approaches in this direction by (Liawrungrueang et al., 2023) and hybrid attention-based networks (H. Li et al., 2021) are attempts to bridge this gap, embedding attention maps or saliency features. Alongside this is the need for a standard

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

framework to evaluate explainability. Explainable AI (XAI) methods are transforming deep models from being a black box to a transparent tool which health professionals and regulatory bodies can monitor using known standards. Another area of concern is the vast volume of data involved in spine imaging procedures which can be cumbersome and time-consuming to process with traditional convolutional networks. Studies that employed YOLO-based frameworks (Liawrungrueang et al., 2023) and MobileNet-inspired designs highlighted the potential for rapid, low-latency inference, resulting in reduced processing time. Consequently, combining accurate lightweight models with explainability for real-time decision support in spine disease detection and segmentation will cover a broader gap in automated spine image analysis.

This review aims to provide a comprehensive overview of the state-of-the-art explainable and lightweight deep learning models for spine imaging. The reviewed studies demonstrate the effectiveness of deep learning techniques in tasks such as anomaly detection, segmentation of the different spine structures, grading of disc degeneration, and different categories of spine fractures. However, despite the high performance achieved by many of these models, a significant gap still persists between high-performing but closed models and the practical requirements of clinical adoption, where transparency and computational efficiency are essential. Therefore, this review examines how explainability techniques, including attention mechanisms, saliency maps, and prototype networks as well as lightweight models such as different variants of YOLO and MobileNet-based networks are being integrated to develop clinically reliable and computationally efficient systems for spine image analysis. Additionally, the review highlights existing challenges, which include performance trade-offs between accuracy and efficiency, limited dataset availability, and lack of standard explainability metrics for model evaluation. While identifying future research directions such as hybrid models that combine interpretability, efficiency, and scalability to support trustworthy AI adoption in spine imaging.

DEEP LEARNING IN SPINE DISEASE IMAGING: AN OVERVIEW

Clinical tasks in spine imaging

Deep learning techniques have been widely adopted in spine imaging for detection of abnormalities, lesions, and degenerative conditions. For example, (Maraş et al., 2022) demonstrated the use of CNN-based models for detecting osteoarthritic changes and loss of cervical lordosis in radiographs, while (H. Li et al., 2021) applied a hybrid attention CNN for accurate recognition of lumbar spine features. These automated detection approaches provide tools for the identification of pathology-specific patterns in Xray and MRI images, thereby reducing observer variation and enhancing efficiency in routine clinical workflows.

Segmentation tasks are another aspect of spinal analysis that deep models have contributed to. (Hess et al., 2023) proposed a multi-tissue segmentation approach for lumbar MRI, while (Li et al., 2021) introduced a multi-scale attention network for same task. Models for segmentation are vital for spine analysis tasks that have to do with disc grading, disease management, surgery planning, structural boundary identification, and tasks involving quantitative measurements.

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 2, 2026 (pp. 68-86)
www.abjournals.org

Major Deep Learning Architectures for Spine Analysis

Majority of the applications developed for spinal imaging are based on CNN architectures, and these remain the backbone for all other variants. From U-Net to RestNet and YOLO in both detection and segmentation tasks.

(Maraş et al., 2022) implemented an effective transfer learning with pre-trained CNN architectures (VGG-16) for the diagnosis of cervical anomalies from cervical radiographs. (H. Li et al., 2021) introduced a multi-scale attention network to enhance feature extraction in segmentation tasks from MRI spine images, adopting a U-Net-like architecture. Several YOLO variants were also employed for identification and grading of intervertebral disc degeneration. (Liawrungrueang et al., 2023) in their study, used the YOLOv5 architecture integrated with convolutional neural networks (CNNs) to automatically detect, classify, and grade lumbar intervertebral disc degeneration (IDD) based on the Pfirrmann grading system. Similarly, DenseNet and ResNet backbones have been incorporated for fracture prediction (Kong et al., 2022), reflecting the adaptability and robustness of CNN-based models to different spine analysis tasks.

In recent times, transformer-based and hybrid models have emerged as alternatives to CNNs, providing enhanced performance in capturing global contextual features. In the study of (Z. Zhang et al., 2024) Verdiff-Net, a framework comprising different modules for spine image analysis was introduced. It integrated conditional diffusion mechanisms with transformer blocks for accurate segmentation of multi-modal spinal images. In another study, (M. H. Guo et al., 2022) and (Zhang et al., 2024) used the multi-attention technique, built on a Transformer encoder model such as Swin-Transformer, to explore semantic segmentation of cancellous bone in the vertebral body. (Chen et al., 2024a) proposed a multi-scale hybrid attention CNN model called SymTC that combined a transformer-based module, incorporating a relative position embedding to enhance the capture of contextual dependencies within the image features. Thereby leveraging the strength of both architectures. Another hybrid approach introduced cascaded designs such as the YOLOv8 + Self-ONN model (Basak et al., 2025), illustrating the growing trend towards combining CNNs with other architectures to optimise both accuracy and efficiency.

Overall, CNNs remain dominant due to their efficiency and reliability; transformers offer powerful global reasoning but often at high computational cost, and resource demands, an important consideration as the field moves towards lightweight models suitable for clinical deployment.

Limitations of Conventional Models

Heavy computational demand is one of the biggest challenges with conventional deep learning models in spine imaging. It limits how easily they can be used in real-world clinical settings. Large CNNs such as ResNet or DenseNet have shown impressive results in detecting and classifying spinal conditions, but they require powerful hardware and long processing times that most hospitals, especially in resource-limited regions, do not have. (Kong et al., 2024) achieved strong results in fracture prediction with CNNs, but the models were far too computationally expensive for quick, bedside use. Similarly, (Zhang et al., 2024) developed Verdiff-Net, a cutting-edge segmentation framework that combines diffusion and transformer modules. While accurate, it comes with very high GPU requirements. A similar concern was reported by (Basak et al., 2025) with cascaded models for lumbar MRI, which improved

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

accuracy but at the cost of higher complexity. Establishing the need for future models to be lightweighted and still be able to deliver reliable performance without demanding advanced computing infrastructure.

Another limitation is the lack of transparency in how conventional models make their predictions, which reinforces the "black box" problem in clinical AI. Radiologists and surgeons often want to know why an algorithm flagged an abnormality or segmented a particular structure, but most current CNN and transformer-based models offer little explanation. In the work of (Maraş et al., 2022), CNNs were used to detect cervical osteoarthritic changes, but the outputs gave no insight into what features were driving the decision. (Li et al., 2021) added attention mechanisms to improve segmentation of lumbar MRI, and (Liu, et al., 2024) proposed a hybrid-attention-based CNN, yet in both cases the interpretability was still limited. Attention maps alone do not always provide explanations that clinicians can act on. This lack of explainability makes it difficult for doctors to trust or defend AI-generated results, highlighting the need for explainable AI that connects model reasoning with clinical judgement.

Several conventional deep models are not reproducible and less generalisable, which added to the low level of confidence in their decision process by carers. Authors in different studies, affirm that the performance of a model depends on the quality of the dataset used for training in the model (Lee et al., 2022). This claim was also established by (Liawrungrueang et al., 2023), in the study various YOLO-based models were proposed for detecting disc degeneration. The models achieved different levels of accuracy on different institution-based images. Consequently, it can be stated that such models are not reliable; if there is an occurrence of misclassification of a disease, such that it leads to irreparable damage or loss, who will take responsibility: the model developers, clinicians, or the health institution?

EXPLAINABLE DEEP LEARNING FOR SPINE IMAGING

Building trust in AI-assisted spinal imaging requires transparency, aside from accuracy. Hence, clinicians remain cautious about adopting AI solutions in practice due to its "black box" nature. Explainable deep learning aims to address this gap by providing interpretable outputs, for example, through the use of heatmaps, feature importance, or attention maps that highlight the anatomical regions influencing model decisions. When AI predictions are connected to clinically meaningful features, explainability will help radiologists to verify results and feel more confident to use these tools in practice.

Overview of Explainability Methods

Explainability in deep learning for spine imaging can be grouped into two broad categories: post-hoc methods and intrinsically interpretable approaches (van der Velden et al., 2022), (Patrício et al., 2024). Post-hoc methods are applied after training a model in order to visualize or approximate the reasoning behind the predictions. Techniques such as Class Activation Mapping (CAM) and Gradient-weighted Cam (Grad-CAM) have applied on Spine radiographs and MRI to highlight regions of interest, giving insight to clinicians on the structures influencing model's decision (Maraş et al., 2022). Other more advanced techniques such as are Local interpretable model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) provide feature-level interpretability and have been explored in broader medical

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

imaging tasks to quantify the contribution of input variables to model outputs (Sathyan et al., 2022). Respond-CAM is another extension of CAM. It has been reported to produce more table localization in MRI-based models, making it most relevant for segmentation of spinal lesions (Zhang et al., 2024). These post-hoc techniques are valuable because they can be applied to existing CNN and transformer architectures without redesigning the model, but their explanations are sometimes approximate and not clinically intuitive.

Intrinsic interpretability, on the other hand, refers to models designed with interpretability built into their architecture. Attention mechanisms are one of the most widely used intrinsic methods. (Li et al., 2021) used a multi-scale attention network for lumbar MRI segmentation; here, attention maps inherently highlight the most relevant anatomical regions during training. Similarly, (Liu, et al., 2024) incorporated hybrid attention modules into CNNs for vertebra segmentation, improving both performance and interpretability by explicitly modelling important spatial and channel-wise features. Beyond attention, prototype-based networks and keypoint-driven detection systems are also emerging as ways to make decisions more interpretable. These approaches allow models to explain in terms of "prototypes" or anatomical landmarks, which can align more closely with how clinicians reason about images.

Applications in Spine Disease

Deep learning explainability has been widely applied in detecting spinal lesions to support radiologists in the identification of abnormal regions. (Kim et al., 2022) developed a CNN-based framework for detecting lumbar spinal stenosis (LSS) in MRI; in the study, Grad-CAM visualisations helped localise suspicious areas, thereby improving the interpretability of the model's prediction. Using techniques like Grad-CAM, the researchers were able to visualise the specific regions and features within the radiographs that the AI model considered important for diagnosis, such as reduced disc height, hypertrophied facet joints, and short pedicles.

By providing clear visual explanations of the AI's decision-making process, the study enhanced transparency and trust in AI systems used in medical settings, which will help clinicians understand which radiographic features influence the AI's diagnosis, making it easier to validate and integrate these tools into clinical practice.

In the work of another author (Yaseen et al., 2024), a two-stage DL pipeline incorporates an object detection model to identify fracture regions and a custom attention-driving network to classify specific cervical. Respond–CAM, an explainable AI technique, generates visual explanations that show which parts of the spine the model focuses on when detecting fractures or identifying each vertebra. The activation maps are produced to interpret and visualize the trained model's focus areas during classification, not to influence the prediction process itself. By displaying these highlighted regions as heatmaps, doctors can clearly see how the system makes its decisions, which helps build trust in its results. (Chłąd & Ogiela, 2023) also emphasises how vision transformers can help make AI decisions more understandable by using attention heatmaps. These heatmaps act like visual guides, showing which parts of the CT scans the model is focusing on when making its diagnosis. For example, they highlight the bone areas that may be damaged, allowing doctors to see exactly what the AI is examining. These studies show how post-hoc and attention methods can make detection models more clinically transparent.

Another application area of explainable models/techniques in spine analysis is in studying degenerative conditions such as disc degeneration and scoliosis. (Addanki & Bala, 2025)

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

developed a 3D deep learning system that analyses spinal MRI to help detect early signs of spine degeneration. They also used Grad-CAM and LIME to show which spinal regions influenced the model's decisions, making the results clearer and more trustworthy for clinicians. In another study, a deep-learning system was developed that uses plain lumbar spine radiographs to detect lumbar spinal canal stenosis, a degenerative narrowing of the spinal canal that can compress nerve roots and lead to pain or mobility loss. The author incorporated Grad-CAM heatmaps to visualise the image regions (such as intervertebral joints and posterior discs) that the model used for its decision, thereby improving understanding of how the model arrives at its diagnosis (Suzuki et al., 2024).

A study by (Rhee et al., 2025) also focused on using advanced AI to accurately identify cervical canal narrowing on MRI scans. Employ Grad-CAM to visualise where the model was focusing. While (Guo et al., 2025) in their study addressed the challenge of automatically detecting and grading lumbar disc herniation (LDH) in MRI images and enhancing uniformity in diagnostic outcomes. To enhance the accuracy and efficiency of diagnosis, the researchers developed an improved deep learning model called GE-YOLOv8, which integrates advanced modules such as a gradient search (GS) module and efficient channel attention (ECA). The modules enhanced future learning, which enabled the model's ability to extract relevant features for intervertebral discs through gradient splitting. The approach reduced model computational complexity while the detection performance was improved. Though explainable techniques were not used explicitly in the study, the model design involved the integration of gradient search that could help in the detection of subtle anomalies. While ECA helped the model to focus on small region.

Vertebrae segmentation is one core aspect of spine image analysis that has also benefitted from intrinsic interpretability approaches. In segmentation, interpretability is embedded directly into the architecture. Although the primary goal of the study by (Huang et al., 2022) is accurate segmentation of vertebrae and intervertebral discs in spine MRIs, interpretability was also incorporated through a cross tri-attention mechanism. The attention module emphasises important vertebral and disc regions by highlighting the most relevant anatomical regions and features during segmentation, allowing clinicians to understand where the model is focusing and why certain labelling decisions are made. By making the feature-fusion process more transparent, the method supports trust and clinical usability alongside strong segmentation performance.

(Li et al., 2021) introduced a multi-scale attention network for lumbar spine MRI segmentation; in the study, explainability is introduced through attention maps that highlighted most relevant vertebral and disc boundaries, making the model's decision process to be transparent and interpretable. (Liu, et al., 2024) extend the work of the previous author by presenting a hybrid attention CNN that incorporated transformer-style modules to capture global dependencies while also producing interpretable attention outputs. The models in the references demonstrate how intrinsic approaches can provide accurate and clinically meaningful segmentation while reducing reliance on post-hoc explanation.

Strengths and Limitations

The adoption of deep learning (DL) in clinical settings has been hindered not by lack of accuracy alone, but significantly by the "black box" nature of many models, making it difficult for clinicians to understand how decisions are made (Markus et al., 2021).

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

When DL models provide explanations by integrating feature-attribution maps or saliency heatmaps, these will allow clinicians to understand how a model is making its decision. For example, attribution-technique frameworks applied to medical imaging highlight that when models point to clinically meaningful regions, clinicians view their outputs as more credible (Brima & Atemkeng, 2024).

Additionally, inclusion of explainability will aid in the comparison of system's results with clinician's reasoning and decision pattern. For example, if the model indicates a section of the spine image as an anomaly, and the same location has been identified by a clinician, it shows the model is "thinking" correctly. This likely behaviour will build medical professional's trust in model's results (Hulsen, 2023), when explanation differs, users can review and evaluate the results while staying involved in the decision-making process. This step will also help in detecting errors and oversight (Markus et al., 2021). Thereby enhancing patient safety and uphold ethical deployment, making explainable AI an enabler of AI adoption in clinical workflows while ensuring responsible deployment. Model decision explainability is a vital concept in assisting the correct usability of AI tools.

Though these explainable systems are helpful in healthcare, they also have their own set of problems. Visual explanation methods such as Grad-CAM and saliency maps are often used to understand how deep learning models make decisions lacks clear reliability. Few alterations in model specification, input data, or training examples can lead to significant changes in the heatmaps that will be generated, even when the model's predictions are just as accurate (Watson et al., 2022). Invariably, this lack of stability makes it hard to trust that the visual explanations truly reflect the model's reasoning, reducing their usefulness in clinical decision-making.

Reproducibility and reliability are other important concerns. Visual explanations can change depending on the dataset, imaging type, or even when the same model architecture is retrained (Watanabe et al., 2022). When making important medical decisions, it is hard to trust explanations that are not consistent. For example, a model might highlight a lesion in one analysis but focus on unrelated areas in another, not because of real clinical differences, but due to random variations in the model. This inconsistency makes it harder for clinicians to trust AI's diagnosis.

In critical situations, the reliability of explainable techniques gives rise to doubt about its decision dependability. Sometimes, visual explanations can mislead instead of aiding transparency (Ghassemi et al., 2021). Visual explanation techniques might emphasize areas that relate to the outcome, rather than the actual pathological basis of the prediction (Markus et al., 2021). Clinicians may interpret these outputs as indicators of sound reasoning, even when the model is utilizing unrelated associations (Wysocki et al., 2023). Lack of standard measures to evaluate accuracy, stability, and usefulness of explainable techniques and their outcomes is a major concern (Banerjee et al., 2022). Except sloid validation frameworks are established, explainability in deep learning will remain as an imperfect but valuable instrument for the creation of trustworthy and interpretable AI solution.

Lightweight Deep Learning for Spine Imaging

There has been considerably huge success in the application of deep learning in detecting and segmenting spine diseases. However, their full deployment in clinical settings is still limited due to the high computation requirement which requires significant memory and processing resources. This level of resources cannot be provided in every hospital. Lightweight deep

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

learning models could be a way out. With light weight approaches, there will be reduction in the number of parameters needed by AI models, which will lead to reduction in model's size, processing time, and the energy needed, while the model maintains its accuracy and reliability.

Strategies for Lightweight Model Design

Research focus on developing lightweight models is in three dimensions: developing strategies to compress models, developing efficient neural architectures, and methods for optimising hardware usage. The essence of developing a lightweight model is to enable deployment in a clinical environment while ensuring the model's accuracy.

To deploy models in mobile and edge devices, it is important that the size of the model and its complexity is reduced while its performance remains stable (Dantas et al., 2024), (Lyu et al., 2023). Model compression techniques are vital to achieve the nature of the expected model. Coupled with the trend in advancement in hardware architectures in the likes of CPU, GPU, and FPGA; which provides the ability to process large volumes of image data and makes medical image analysis more efficient and accessible (Alcaín et al., 2021), (Liu et al., 2025). Additionally, there was improvement in the efficiency at both algorithmic and systemic levels for the development of resource-efficient algorithmics for large foundation models (Xu et al., 2025).

Model compression approaches implemented to achieve lightweight models are: pruning, quantization, and knowledge distillation (Li et al., 2023). Pruning is the elimination of redundant weights to achieve a simplified network, thereby, reducing the number of parameters in a 3D CNN, while the Dice similarity in spine segmentation is kept at over 95% (Saeed et al., 2025). Quantization on the other hand converts high-precision computations into lower-precision formats. For example, a 32-bit operation to 8-bit operations. Eventually the memory usage is reduced while inference speed is improved with minimal or no loss in accuracy of the model. In the knowledge distillation approach, the learned representations of a large model "the teacher" transferred to a smaller model "the student", such that the lightweight model will be able to achieve performance close to the teacher's model, while less computational power will be required.

Another approach that can be implemented to achieve a lightweight model is to address it from the point of network design. That is designing the network to be a lightweight model. Some designs can follow a depth-wise separability of grouped convolutions to enhance computational efficiency. Examples of such design are used for MobileNet, ShuffleNet, and SqueezeNet. A MobileNet-based classifier was used for detecting spinal fractures, EfficientNet-Lite and TinyViT were made flexible for efficient deployment in clinical settings. Yolov5-Nano and YOLOv8-N in object detection maintained high accuracy in the identification of spine lesions and fractures in real-time inference. Lightweight U-Net adaptations, in the vein of Enet, Mobile-UNet, and UNet++, resulted in a competitive outcome for vertebrae and disc segmentation; model quality was not compromised, while computational cost was reduced.

The last approach in achieving lightweight models involves the model's optimisation for hardware. Inference pipelines were designed to align with the designs of the GPUs, TPUs, and embedded processors such as NVIDIA Jetson or Intel Movidius, giving researchers an opportunity to maximise computational efficiency and speed. This approach serves as the foundation and anchor for edge inference. In edge inference, there is direct data processing almost at the point of acquisition without an intermediary, thereby enabling patient's privacy

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

and reducing latency and bandwidth demand. Edge-optimised frameworks have been deployed to analyse CT and MRI scans with minimal delay and accurate outcomes within a clinically acceptable time frame.

Applications in Detection and Segmentation

MobileNet is a lightweight version of CNN architecture that has been deployed for fracture detection tasks on spine X-ray images, with 98% accuracy and low error rates (Goel & Singh, 2024). While 93% mean average precision was attained with adapted YOLO versions 5 and 8 for the detection of cervical spine fracture (Sutradhar et al., 2025). Similarly, (Yaseen et al., 2024) also implemented YOLOv8 large for the same task, achieving mAP50 of 93.5%. Ensemble approaches that integrated lightweight backbones (MobileViT and EfficientNet) with YOLO-based frameworks have achieved better performance on fracture detection accuracy with up to 93.4%. Most times, the ensemble approach outperforms human experts (Hsieh et al., 2024).

Segmentation is another task where lightweight models have demonstrated high performance in delineating spinal structures. Mobile Residual U-Net (MRU-net) combines MobileNetV2 with residual blocks for segmenting. The resulting model is suitable for integration into clinical workflow and for low-resource applications (Saeed et al., 2023). Additionally, improved U-Net variants incorporated attention mechanisms and multi-scale feature fusion that shows consistent gains over the traditional U-Net architecture, with a Dice coefficient between 0.85 and 0.95 across different image modalities, including ultrasound (Banerjee et al., 2022). The same thing applies to 3D U-Net and pruned UNet++ architectures with segmentation accuracy beyond 0.95 Dice similarity. Cutting inference time by 60% (Zhou et al., 2020).

Performance Trade-offs

From findings through the various studies in the literature, it has been established that lightweight segmentation architectures have high accuracy, though there are trade-offs in model capacity. Mobile Residual U-Net (MRU-Net), discussed by (Saeed et al., 2023), has a slight reduction in feature representation compared to heavier models. (Banerjee et al., 2022) incorporated an attention mechanism and multi-scale feature fusion, resulting in an enhanced U-Net which outperform the classical U-Net.

The substantial gains in speed and deplorability make lightweight design suitable for real-time clinical use and in low-resource environments.

Synergy: Explainable and Lightweight Models

Research focus in artificial intelligence models is now on integrating interpretability and efficiency. Existing lightweight models have been optimised to be deployable on edge devices with minimal computing resources. Lightweight models in the likes of MobileNet, EfficientNet, and SqueezeNet with fewer parameters are achieving comparable performance with larger models (Dantas et al., 2024). These models are enhanced with techniques such as attention mechanisms, explainability and interpretability techniques.

With attention modules, models are able to identify and focus on the most relevant regions in an input image, creating a channel to access the model's internal focus and decision process. According to (Guo et al., 2022), attention mechanisms can be classified into three categories:

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

spatial, channel, and temporal. This classification illustrates how attention enhances both performance and interpretability in visual recognition tasks. Integration of attention mechanism makes compact models become more explainable, such that humans can trace the influence of specific features on the model's outcome without need for extra post-hoc interpretation. This is a hybridisation approach that links both lightweight models' efficiency and built-in interpretability, aimed at delivering transparent and deployable AI.

Another aspect of the synergy is the explainability approach. Explainability methods have increasingly been embedded within efficient or compressed models to balance transparency with computational practicality. This entails using post-hoc interpretability techniques to generate both local and global model explanations, such as Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive exPlanations (SHAP), and Gradient-weighted Class Activation Mapping) (Linardatos et al., 2021).

It is challenging to apply these techniques without sacrificing interpretability or accuracy. Findings from recent studies gave promising outcomes. In the work of (Ghose et al., 2025), a lightweight, attention-based stress-detection model for edge devices was developed. It integrated LIME to enable interpretability while maintaining computational efficiency.

Similarly, (Khan et al., 2025) proposed a modified MobileNetV2 architecture embedded with explainability modules, proving that compression and transparency can coexist in constrained environments. These results align with the wider trend towards the development of trustworthy AI systems, where every other feature, such as efficiency, interpretability, and fairness, is treated as a complimentary design objective (Teng et al., 2022), (Patnaik et al., 2025). Justification for the importance of designing explainable and lightweight models.

Several studies have integrated deep learning models and the visual explanation tools such as Grad-CAM, in clinical applications, resulting in efficient detection frameworks like YOLO. Also, with the emergence of transformers, CNN hybrids have enabled precise spine segmentation and classification using minimal parameters. This gave an improvement on both the computational efficiency and practical deployment feasibility in clinical settings.

Some of the studies that demonstrated how Grad-CAM, enhanced YOLO variants and lightweight transformers, are being applied in clinical spine image analysis are discussed in the following section.

(Liu, et al., 2024) introduced a hybrid transformer-convolutional neural network-based radiomics model for early detection of osteoporosis in CT scans. The model was employed for precise vertebral segmentation, demonstrating high correlation with manual segmentation with a Dice Similarity Coefficient of 0.968 for vertebral segmentation and 0.961 for trabecular compartment segmentation. The hybrid approach combines convolutional neural networks with transformer architecture to leverage the strength of both methods for efficient and interpretable segmentation. Transformer captures long-range dependencies and feature relationships, while CNN offers high accuracy.

In another study, (Chu et al., 2025) also implemented an interpretable method for diagnosis of lumbar disc herniation (LDH) using Vision Transformer (ViT) on CT images. ViT produces transparency through Grad-CAM visualisations. By including self-attention mechanisms, the model is able to capture spatial patterns while maintaining low parameters. Consequently,

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

making the model to be lightweight. Though there is still room for improvement in relation to the dataset and applicability in clinical settings.

The Spinal Context Transformer (SCT) is a lightweight transformer-based model introduced by (Windsor et al., 2022) used to analyse MRI scans. A lightweight 2-layered transformer encoder that combines visual embeddings from several vertebral and MR sequences with embeddings encoding the vertebral level and image modality enabled the model to utilise the full spinal column, enabling context-aware prediction. SCT supported explainability by integrating attention mechanisms and could adapt to varying sizes of input. It also used a 2D ResNet18 backbone for feature extraction, which highlighted relevant regions and contextual information that informed the model's decision. A lightweight and interpretable Transformer-based model for lumbar spine MRI segmentation that balances high accuracy with computational efficiency was proposed by (J. Chen et al., 2024b). The model employed effective data augmentation and robustness approaches to ensure reliable performance in resource-limited settings, supporting explainability and practical deployment in clinical environments.

YOLO is a deep learning model, though primarily developed for object detection, that has been combined with explainable AI techniques, such as Grad-CAM, to enhance interpretability while maintaining a lightweight and efficient design. (Yaseen et al., 2024) developed a two-stage deep learning approach for cervical spine fracture detection that incorporated Grad-CAM to enhance interpretability by visualising decision regions.

Recent studies on spine imaging analysis supported the integration of explainable and lightweight as hybrid models to enable highly accurate, automated, and transparent diagnosis while maintaining scalability and alignment with clinical standards (Mahasin et al., 2025). Implementing explainable architecture will ensure accountability and patient since safety the since clinical team will be able to reason with the decision process of the model and take necessary action when anomalies are observed. Furthermore, the inclusion of edge AI technology, which is vital for the deployment of these models, aids in reducing data privacy concerns through localised processing (Ehimah Obuse et al., 2024), (Mohanadas, 2025).

(Mastoi et al., 2025) stated how explainable federated learning enhanced the synergy by enabling collaborative training across multiple institutions without sharing raw data, thereby preserving data privacy and guarding against communication issues such as re-identification of data sources (Gaudio et al., 2023), (Mu et al., 2024) while providing diverse datasets that will improve model robustness. The decentralised approach of federated learning will allow lightweight models to be semantically enriched through ontologies and semantic aggregation, enhancing interpretability without necessarily increasing model complexity (Amato & Branco, 2025). Consequently, facilitating explainable AI-solutions that are both efficient and transparent, with dynamic adaptation to evolving health data characteristics (Muthalakshmi et al., 2024).

RESEARCH GAPS

Despite all the progress identified in this study and several others in literature, there are still numerous challenges that render reliability of AI-based solutions in clinical settings.

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

Data non-availability coupled with its quality is a major issue. The few existing data for spine diseases are small and imbalanced, with a lack of representation of different pathological conditions as well as shallow demographic patient data (Fraiwan et al., 2022). In addition to the data size is the scarcity of multi-centre and multimodal datasets that are a combination of MRI, CT, and clinical records (Küçükçiloğlu et al., 2024). Unavailability of such a combination limits the generalisability and richness of AI models (Lee et al., 2024). Models trained with unstandardised dataset are also opened to overfitting since the training data has a narrow distribution, eventually affecting the reliability of its diagnosis.

Lack of unified metrics for evaluating the explainability of AI models is another major issue affecting the research progress in spine image analysis (Ali et al., 2023). This makes it difficult to compare findings across studies. While in relation to model quality attributes accuracy, efficiency and, interpretability; researchers struggle with trade-offs. Hence, the need for a balance threshold.

Regulatory bodies are yet to deploy well -defined and robust framework to safeguard the use of AI models in clinical settings, thereby generating restrictions to full adoption (Goktas & Grzybowski, 2025). A major constraint faced is the pace at which this technology is evolving. The concerns of these regulatory bodies are about reliability, accountability, and disruption to standard clinical workflows (Hulsen, 2023), (Schmidt et al., 2024). To address this, researchers should focus on designing models that are interpretable, and that can involve a human-in-the-loop approach, in order for clinicians to contribute to the decision-making process.

FUTURE RESEARCH DIRECTIONS

Deductions from the literature established existing research gaps which can provide insight for future research directions. It is evident that there are several areas where further research can contribute significantly. The following directions are proposed: standardised benchmarks to evaluate explainable models for spine image analysis in order to enable fair model comparison; designing models to be both lightweight and interpretable; multimodal data sources should be considered; and improving predictive accuracy of the models.

Of importance is also real-world implementation and large-scale validation of the models. This will enable insight into the issues of usability, scalability, and potential to improve patient outcomes. Additionally, clinicians will be able to evaluate the model's confidence, promoting safer and more transparent decision-making. With these research directions, there will be a leading into a new generation of spine imaging models with accurate and efficient predictions that are, interpretable, privacy-preserving, and clinically trustworthy.

CONCLUSION

The transitional trend in AI "black box" models to transparent and trustworthy AI systems in spine image analysis signifies the importance of explainability in boosting clinician confidence and enhancing deployment in clinical workflows. This current direction requires coordinated multidisciplinary collaboration that aims at developing AI frameworks that are transparent, computationally efficient, and ethically robust.

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

Explainability is the foundation to building trust because it gives insight into the internal reasoning processes of AI models, thereby supporting clinicians in interpreting, verifying and validating automated decisions in spine diagnostics. The development of lightweight models contributes in parallel to scalability by delivering strong predictive performance with reduced computational burden while preserving interpretability, which is vital for adoption in clinical settings.

REFERENCES

Addanki, J., & Bala, K. (2025). A volumetric deep learning framework for spinal cord MRI classification with explainable AI integration. *2025 5th International Conference on Soft Computing for Security Applications (ICSCSA)*, 1903–1910. https://doi.org/10.1109/ICSCSA66339.2025.11171338

Alcaín, E., Fernández, P. R., Nieto, R., Montemayor, A. S., Vilas, J., Galiana-Bordera, A., Martinez-Girones, P. M., Prieto-de-la-Lastra, C., Rodriguez-Vila, B., Bonet, M., Rodriguez-Sanchez, C., Yahyaoui, I., Malpica, N., Borromeo, S., Machado, F., & Torrado-Carvajal, A. (2021). Hardware architectures for real-time medical imaging. *Electronics*, *10*(24), 3118. https://doi.org/10.3390/electronics10243118

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Amato, A., & Branco, D. (2025). SemFedXAI: A semantic framework for explainable federated learning in healthcare. *Information (Switzerland)*, *16*(6). https://doi.org/10.3390/info16060435

Banerjee, S., Lyu, J., Huang, Z., Leung, F. H. F., Lee, T., Yang, D., Su, S., Zheng, Y., & Ling, S. H. (2022). Ultrasound spine image segmentation using multi-scale feature fusion Skip-Inception U-Net (SIU-Net). *Biocybernetics and Biomedical Engineering*, *42*(1), 341–361. https://doi.org/10.1016/j.bbe.2022.02.011

Basak, P., Sarmun, R., Kabir, S., Al-Hashimi, I., Bhuiyan, E. H., Hasan, A., Khan, M. S., & Chowdhury, M. E. H. (2025). Machine-agnostic automated lumbar MRI segmentation using a cascaded model based on generative neurons. *Expert Systems with Applications*, *264*. https://doi.org/10.1016/j.eswa.2024.125862

Bin Ahmad, M. S. Z., Aziz, N. A. A., & Siong, L. H. (2024). Classification of spine abnormalities using deep learning. *International Exchange and Innovation Conference on Engineering and Sciences*, *10*, 998–1004. https://doi.org/10.5109/7323381

Brima, Y., & Atemkeng, M. (2024). Saliency-driven explainable deep learning in medical imaging: Bridging visual explainability and statistical quantitative analysis. *BioData Mining*, *17*(1). https://doi.org/10.1186/s13040-024-00370-4

Chen, J., Qian, L., Ma, L., Urakov, T., Gu, W., & Liang, L. (2024a). SymTC: A symbiotic Transformer-CNN net for instance segmentation of lumbar spine MRI. *Computers in Biology and Medicine*, *179*. https://doi.org/10.1016/j.compbiomed.2024.108795

Chen, P., Dong, W., Wang, J., Lu, X., Kaymak, U., & Huang, Z. (2020). Interpretable clinical prediction via attention-based neural network. *BMC Medical Informatics and Decision Making*, *20*(S3), 131. https://doi.org/10.1186/s12911-020-1110-7

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

Chłąd, P., & Ogiela, M. R. (2023). Deep learning and cloud-based computation for cervical spine fracture detection system. *Electronics (Switzerland)*, *12*(9). https://doi.org/10.3390/electronics12092056

Chu, Q., Wang, X., Lv, H., Zhou, Y., & Jiang, T. (2025). Vision transformer-based diagnosis of lumbar disc herniation with grad-CAM interpretability in CT imaging. *BMC Musculoskeletal Disorders*, *26*(1). https://doi.org/10.1186/s12891-025-08602-2

Dantas, P. V., Sabino da Silva, W., Cordeiro, L. C., & Carvalho, C. B. (2024). A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*, *54*(22), 11804–11844. https://doi.org/10.1007/s10489-024-05747-w

Dindorf, C., Ludwig, O., Simon, S., Becker, S., & Fröhlich, M. (2023). Machine learning and explainable artificial intelligence using counterfactual explanations for evaluating posture parameters. *Bioengineering*, *10*(5). https://doi.org/10.3390/bioengineering10050511

Ehimah Obuse, Noah Ayanbode, Emmanuel Cadet, Edima David Etim, & Iboro Akpan Essien. (2024). Edge AI solutions for real-time IoT device threat monitoring. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *10*(3), 996–1030. https://doi.org/10.32628/CSEIT25113576

Fraiwan, M., Audat, Z., Fraiwan, L., & Manasreh, T. (2022). Using deep transfer learning to detect scoliosis and spondylolisthesis from x-ray images. *PLOS ONE*, *17*(5), e0267851. https://doi.org/10.1371/journal.pone.0267851

Gaudio, A., Smailagic, A., Faloutsos, C., Mohan, S., Johnson, E., Liu, Y., Costa, P., & Campilho, A. (2023). DeepFixCX: Explainable privacy-preserving image compression for medical image analysis. *WIREs Data Mining and Knowledge Discovery*, *13*(4). https://doi.org/10.1002/widm.1495

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Ghose, D., Chatterjee, A., Balapuwaduge, I. A. M., Lin, Y., & Dash, S. P. (2025). Investigating lightweight and interpretable machine learning models for efficient and explainable stress detection. *Frontiers in Digital Health*, *7*. https://doi.org/10.3389/fdgth.2025.1523381

Goel, M. K., & Singh, G. (2024). Fracture detection using MobileNet model. *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, 1574–1579. https://doi.org/10.1109/ICUIS64676.2024.10866070

Goktas, P., & Grzybowski, A. (2025). Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy AI. *Journal of Clinical Medicine*, *14*(5), 1605. https://doi.org/10.3390/jcm14051605

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., Zhang, S. H., Martin, R. R., Cheng, M. M., & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. In *Computational Visual Media* (Vol. 8, Number 3, pp. 331–368). Tsinghua University. https://doi.org/10.1007/s41095-022-0271-y

Guo, Y., Huang, X., Chen, W., Nakamoto, I., Zhuang, W., Chen, H., Feng, J., & Wu, J. (2025). Deep learning-based automatic detection and grading of disk herniation in lumbar magnetic resonance images. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-10401-7

Haar, L. V., Elvira, T., & Ochoa, O. (2023). An analysis of explainability methods for convolutional neural networks. *Engineering Applications of Artificial Intelligence*, *117*, 105606. https://doi.org/10.1016/j.engappai.2022.105606

Hess, M., Allaire, B., Gao, K. T., Tibrewala, R., Inamdar, G., Bharadwaj, U., Chin, C., Pedoia, V., Bouxsein, M., Anderson, D., & Majumdar, S. (2023). Deep Learning for Multi-Tissue

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

segmentation and fully automatic personalized biomechanical models from BACPAC clinical Lumbar spine MRI. *Pain Medicine (United States)*, *24*, S139–S148. https://doi.org/10.1093/pm/pnac142

Hsieh, M.-H., Chang, C.-Y., & Hsu, S.-M. (2024). Accurate detection of fresh and old vertebral compression fractures on CT images using ensemble YOLOR. *Multimedia Tools and Applications*, *83*(41), 89375–89391. https://doi.org/10.1007/s11042-024-20355-z

Huang, M., Zhou, S., Chen, X., Lai, H., & Feng, Q. (2022). *Semi-supervised hybrid spine network for segmentation of spine MR Images*. http://arxiv.org/abs/2203.12151

Hulsen, T. (2023). Explainable Artificial Intelligence (XAI): Concepts and challenges in healthcare. In *AI (Switzerland)* (Vol. 4, Number 3, pp. 652–666). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/ai4030034

Khan, S., Siddiqui, F., & Ahad, M. A. (2025). Bridging efficiency and interpretability: Explainable AI for multi-classification of pulmonary diseases utilizing modified lightweight CNNs. *Image and Vision Computing*, *158*, 105553. https://doi.org/10.1016/j.imavis.2025.105553

Kim, T., Kim, Y. G., Park, S., Lee, J. K., Lee, C. H., Hyun, S. J., Kim, C. H., Kim, K. J., & Chung, C. K. (2022). Diagnostic triage in patients with central lumbar spinal stenosis using a deep learning system of radiographs. *Journal of Neurosurgery: Spine*, *37*(1), 104–111. https://doi.org/10.3171/2021.11.SPINE211136

Kong, S. H., Cho, W., Park, S. B., Choo, J., Kim, J. H., Kim, S. W., & Shin, C. S. (2024). A computed tomography–based fracture prediction model with images of vertebral bones and muscles by employing deep learning: Development and validation study. *Journal of Medical Internet Research*, *26*(1). https://doi.org/10.2196/48535

Kong, S. H., Lee, J. W., Bae, B. U., Sung, J. K., Jung, K. H., Kim, J. H., & Shin, C. S. (2022). Development of a spine X-Ray-based fracture prediction model using a deep learning algorithm. *Endocrinology and Metabolism*, *37*(4), 674–683. https://doi.org/10.3803/EnM.2022.1461

Küçükçiloğlu, Y., Şekeroğlu, B., Adalı, T., & Şentürk, N. (2024). Prediction of osteoporosis using MRI and CT scans with unimodal and multimodal deep-learning models. *Diagnostic and Interventional Radiology*, *30*(1), 9–20. https://doi.org/10.4274/dir.2023.232116

Lee, G. W., Shin, H., & Chang, M. C. (2022). Deep learning algorithm to evaluate cervical spondylotic myelopathy using lateral cervical spine radiograph. *BMC Neurology*, *22*(1). https://doi.org/10.1186/s12883-022-02670-w

Lee, S., Jung, J. Y., Mahatthanatrakul, A., & Kim, J. S. (2024). Artificial intelligence in spinal imaging and patient care: A review of recent advances. In *Neurospine* (Vol. 21, Number 2, pp. 474–486). Korean Spinal Neurosurgery Society. https://doi.org/10.14245/ns.2448388.194

Li, H., Luo, H., Huan, W., Shi, Z., Yan, C., Wang, L., Mu, Y., & Liu, Y. (2021). Automatic lumbar spinal MRI image segmentation with a multi-scale attention network. *Neural Computing and Applications*, *33*(18), 11589–11602. https://doi.org/10.1007/s00521-021-05856-4

Li, Z., Li, H., & Meng, L. (2023). Model compression for deep neural networks: A survey. *Computers*, *12*(3), 60. https://doi.org/10.3390/computers12030060

Liawrungrueang, W., Kim, P., Kotheeranurak, V., Jitpakdee, K., & Sarasombath, P. (2023). Automatic detection, classification, and grading of lumbar intervertebral disc degeneration using an artificial neural network model. *Diagnostics*, *13*(4). https://doi.org/10.3390/diagnostics13040663

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. In *Entropy* (Vol. 23, Number 1, pp. 1–45). MDPI AG. https://doi.org/10.3390/e23010018

Liu, J., Wang, H., Shan, X., Zhang, L., Cui, S., Shi, Z., Liu, Y., Zhang, Y., & Wang, L. (2024). Hybrid transformer convolutional neural network-based radiomics models for osteoporosis screening in routine CT. *BMC Medical Imaging*, *24*(1). https://doi.org/10.1186/s12880-024-01240-5

Liu, J., Zhou, Y., Cui, X., Jin, F., Suo, G., Xu, H., & Yang, J. (2024a). Multi-scale Hybrid attention convolutional neural network for automatic segmentation of lumbar vertebrae from MRI. *IEEE Access*, *12*, 77999–78013. https://doi.org/10.1109/ACCESS.2024.3407833

Liu, X., Dai, Z., Wang, Q., & Li, Z. (2025). Computing acceleration of medical image processing based on multi-accelerator heterogeneous systems. *ACM SIGAPP Applied Computing Review*, *25*(1), 16–24. https://doi.org/10.1145/3727257.3727259

Lyu, Z., Yu, T., Pan, F., Zhang, Y., Luo, J., Zhang, D., Chen, Y., Zhang, B., & Li, G. (2023). A survey of model compression strategies for object detection. *Multimedia Tools and Applications*, *83*(16), 48165–48236. https://doi.org/10.1007/s11042-023-17192-x

Mahasin, M. M., Naba, A., Widodo, C. S., & P. W., Y. Y. (2025). Explainable and lightweight machine learning model for cardiomegaly detection from chest XRay images. *2025 International Electronics Symposium (IES)*, 881–885. https://doi.org/10.1109/IES67184.2025.11161453

Maraş, Y., Tokdemir, G., Üreten, K., Atalar, E., Duran, S., & Maraş, H. (2022). Diagnosis of osteoarthritic changes, loss of cervical lordosis, and disc space narrowing on cervical radiographs with deep learning methods. *Joint Diseases and Related Surgery*, *33*(1), 93–101. https://doi.org/10.52312/jdrs.2022.445

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, *113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655

Mastoi, Q. U. A., Latif, S., Brohi, S., Ahmad, J., Alqhatani, A., Alshehri, M. S., Al Mazroa, A., & Ullah, R. (2025). Explainable AI in medical imaging: an interpretable and collaborative federated learning model for brain tumor classification. *Frontiers in Oncology*, *15*. https://doi.org/10.3389/fonc.2025.1535478

Mohanadas, S. (2025). Real-time diagnostics in critical care: AI for rapid decision-making and continuous monitoring. *International Journal of Computing and Engineering*, *7*(3), 1–22. https://doi.org/10.47941/ijce.2658

Mu, J., Kadoch, M., Yuan, T., Lv, W., Liu, Q., & Li, B. (2024). Explainable federated medical image analysis through causal learning and blockchain. *IEEE Journal of Biomedical and Health Informatics*, *28*(6), 3206–3218. https://doi.org/10.1109/JBHI.2024.3375894

Muthalakshmi, M., Jeyapal, K., Vinoth, M., P S, D., Murugan, N. S., & Sheela, K. S. (2024). Federated learning for secure and privacy-preserving medical image analysis in decentralized healthcare systems. *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 1442–1447. https://doi.org/10.1109/ICESC60852.2024.10690003

Patnaik, N., Nayak, N., Agrawal, H. B., Khamaru, M. C., Bal, G., Panda, S. S., Raj, R., Meena, V., & Vadlamani, K. (2025). *Small vision-language models: A survey on compact architectures and techniques*. http://arxiv.org/abs/2503.10665

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

Patrício, C., Neves, J. C., & Teixeira, L. F. (2024). Explainable deep learning methods in medical image classification: A survey. *ACM Computing Surveys*, *56*(4), 1–41. https://doi.org/10.1145/3625287

Pereira, R. F. B., Helito, P. V. P., Leão, R. V., Bordalo Rodrigues, M., de Paula Correa, M. F., & Rodrigues, F. V. (2024). Accuracy of an artificial intelligence algorithm for detecting moderate-to-severe vertebral compression fractures on abdominal and thoracic computed tomography scans. *Radiologia Brasileira*, *57*. https://doi.org/10.1590/0100-3984.2023.0102

Rhee, W., Park, S. C., Kim, H., Chang, B.-S., & Chang, S. Y. (2025). Deep learning-based prediction of cervical canal stenosis from mid-sagittal T2-weighted MRI. *Skeletal Radiology*, *54*(10), 2067–2076. https://doi.org/10.1007/s00256-025-04917-2

Saeed, M. U., Bin, W., Sheng, J., & Saleem, S. (2025). 3D MFA: An automated 3D multi-feature attention based approach for spine segmentation using a multi-stage network pruning. *Computers in Biology and Medicine*, *185*, 109526. https://doi.org/10.1016/j.compbiomed.2024.109526

Saeed, M. U., Dikaios, N., Dastgir, A., Ali, G., Hamid, M., & Hajjej, F. (2023). An automated deep learning approach for spine segmentation and vertebrae recognition using computed tomography images. *Diagnostics*, *13*(16), 2658. https://doi.org/10.3390/diagnostics13162658

Sathyan, A., Weinberg, A. I., & Cohen, K. (2022). Interpretable AI for bio-medical applications. *Complex Engineering Systems*, *2*(4), 18. https://doi.org/10.20517/ces.2022.41

Schmidt, J., Schutte, N. M., Buttigieg, S., Novillo-Ortiz, D., Sutherland, E., Anderson, M., de Witte, B., Peolsson, M., Unim, B., Pavlova, M., Stern, A. D., Mossialos, E., & van Kessel, R. (2024). Mapping the regulatory landscape for artificial intelligence in health within the European Union. *Npj Digital Medicine*, *7*(1), 229. https://doi.org/10.1038/s41746-024-01221-6

Sutradhar, D., Fahad, N. M., Khan Raiaan, M. A., Jonkman, M., & Azam, S. (2025). Cervical spine fracture detection utilizing YOLOv8 and deep attention-based vertebrae classification ensuring XAI. *Biomedical Signal Processing and Control*, *101*. https://doi.org/10.1016/j.bspc.2024.107228

Suzuki, H., Kokabu, T., Yamada, K., Ishikawa, Y., Yabu, A., Yanagihashi, Y., Hyakumachi, T., Tachi, H., Shimizu, T., Endo, T., Ohnishi, T., Ukeba, D., Nagahama, K., Takahata, M., Sudo, H., & Iwasaki, N. (2024). Deep learning-based detection of lumbar spinal canal stenosis using convolutional neural networks. *The Spine Journal*, *24*(11), 2086–2101. https://doi.org/10.1016/j.spinee.2024.06.009

Teng, Q., Liu, Z., Song, Y., Han, K., & Lu, Y. (2022). A survey on the interpretability of deep learning in medical diagnosis. *Multimedia Systems*, *28*(6), 2335–2355. https://doi.org/10.1007/s00530-022-00960-4

van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, *79*, 102470. https://doi.org/10.1016/j.media.2022.102470

Wang, B., Wang, R., Chen, Z., Zhang, Q., Yuwen, W., & Liu, X. (2024). Improving vertebral diagnosis in computed tomography scans: A clinically oriented attention-driven asymmetric convolution network for segmentation. *Intelligent Medicine*, *4*(4), 239–248. https://doi.org/10.1016/j.imed.2024.02.002

Watanabe, A., Ketabi, S., Namdar, K., & Khalvati, F. (2022). Improving disease classification performance and explainability of deep learning models in radiology with heatmap generators. *Frontiers in Radiology*, *2*. https://doi.org/10.3389/fradi.2022.991683

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD

Watson, M., Awwad, B., Hasan, S., & Al Moubayed, N. (2022). *Agree to disagree: When deep learning models with identical architectures produce distinct explanations*. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 875-884).

Windsor, R., Jamaludin, A., Kadir, T., & Zisserman, A. (2022). *Context-Aware Transformers For Spinal Cancer Detection and Radiological Grading*. http://arxiv.org/abs/2206.13173

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, *316*. https://doi.org/10.1016/j.artint.2022.103839

Xu, M., Cai, D., Yin, W., Wang, S., Jin, X., & Liu, X. (2025). Resource-efficient algorithms and systems of foundation models: A survey. *ACM Computing Surveys*, *57*(5), 1–39. https://doi.org/10.1145/3706418

Yaseen, M., Ali, M., Ali, S., Hussain, A., Joo, M. Il, & Kim, H. C. (2024). Cervical spine fracture detection and classification using two-stage deep learning methodology. *IEEE Access*, *12*, 72131–72142. https://doi.org/10.1109/ACCESS.2024.3398061

Yıldız Potter, İ., Yeritsyan, D., Rodriguez, E. K., Wu, J. S., Nazarian, A., & Vaziri, A. (2024). Detection and localization of spine disorders from plain radiography. *Journal of Imaging Informatics in Medicine*, *37*(6), 2967–2982. https://doi.org/10.1007/s10278-024-01175-x

Zhang, Y., Shi, Z., Wang, H., Cui, S., Zhang, L., Liu, J., Shan, X., Liu, Y., & Fang, L. (2024). LumVertCancNet: A novel 3D lumbar vertebral body cancellous bone location and segmentation method based on hybrid swin-transformer. *Computers in Biology and Medicine*, *171*. https://doi.org/10.1016/j.compbiomed.2024.108237

Zhang, Z., Liu, T., Fan, G., Pu, Y., Li, B., Chen, X., Feng, Q., & Zhou, S. (2024). Verdiff-Net: A conditional diffusion framework for spinal medical image segmentation. *Bioengineering*, *11*(10). https://doi.org/10.3390/bioengineering11101031

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2020). UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, *39*(6), 1856–1867. https://doi.org/10.1109/TMI.2019.2959609

Article DOI: 10.52589/BJCNIT-MIZPEFHD
DOI URL: https://doi.org/10.52589/BJCNIT-MIZPEFHD