Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 270 (2025) 4905–4916

Procedia Computer Science

www.elsevier.com/locate/procedia

29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025)

# MOVEXor: Motion-Video Attention Explainer for Low Back Pain Classification

Zebang Liu[a,*], Yulia Hicks[a], Liba Sheeran[b]

[a]*School of Engineering, Cardiff University, Cadiff CF24 3AA, United Kingdom*
[b]*School of Health Sciences, University of Southampton, SO17 1BJ, United Kingdom*

**Abstract**

Accurate classification of Movement Impairment (MI) and Motor Control Impairment (MCI) in non-specific low back pain (NSLBP) is essential for targeted rehabilitation but remains challenging due to subjective assessments and subtle movement differences. We present MOVEXor, a lightweight and explainable multi-modal framework that integrates spinal curvature images and motion-derived features through a modality-aware attention gating mechanism. MOVEXor achieves high classification performance (up to 97.5% accuracy) while offering transparent decision-making via Grad-CAM and Integrated Gradients (IG). Our analysis shows that the model focuses on physiologically meaningful movement phases, particularly minimal flexion angle, and relies heavily on motion stability for classification. The fused attention-based design outperforms static fusion methods, especially when handling noisy inputs. With minimal hardware requirements and real-time explainability, MOVEXor holds strong potential as a clinical decision-support tool for both in-clinic and remote settings, enabling objective, interpretable, and personalised rehabilitation exercise of LBP subgroups.



*Keywords:* Non-specific Low Back Pain (NSLBP); LBP Subgroups Classification; Explainable Artificial Intelligence (XAI); Multi-Modal Fusion; Motion Features Analysis.

## 1. Introduction

Low back pain (LBP) is the leading cause of disability worldwide, affecting up to 80% of individuals at some point in their lives and posing a major socioeconomic and healthcare burden globally [1]. Nearly 90% of LBP cases are categorised as non-specific low back pain (NSLBP), in which no clear anatomical or pathological cause can be identified [1]. Of all chronic pain conditions, it affects 37% of males and 44% of females in the UK. Within the National Health Service (NHS), more than £1000 million per year was spent in 1998 [5].

* Corresponding author. Tel.:+44(0)7731696509.
*E-mail address:* Liuz91@cardiff.ac.uk


10.1016/j.procs.2025.09.617

According to NICE guidelines [5], treatment for NSLBP emphasises exercise programmes tailored to patients' specific needs and abilities. To facilitate individualised care, a widely recognised Multidimensional Classification System (MDCS) categorises NSLBP into functional subgroups, notably including Movement Impairment (MI) and Motor Control Impairment (MCI). MI is typically characterised by reduced range of motion due to pain and avoidance behaviour, while MCI is characterised by unrestricted yet painful movement and pain-provoking behaviour [2].

However, the complexity and heterogeneity of NSLBP, coupled with subtle differences between MI and MCI, and the subjectivity of clinical assessment, make accurate classification challenging.

To improve objectivity and accuracy in LBP classification, recent research has explored the use of machine learning (ML) and computer vision techniques to automate assessment based on motion capture, video recordings, and patient-reported outcomes. Some research have employed kinematic features such as bending angles, angular velocity, and acceleration to classify LBP subtypes using feedforward neural networks [3]. Others have leveraged convolutional neural networks (CNNs) to analyse video frames of patients performing standard movement tasks [4]. While promising, these approaches often treat modalities independently or naively combine them, which may lead to suboptimal performance and poor model transparency.

In this paper, to address the clinical challenge of distinguishing Movement Impairment (MI) from Motor Control Impairment (MCI) in NSLBP, we propose MOVEXor, a novel and lightweight Multi-modal Attention Gating (MAG) network. Our goal is to develop a model that not only improves classification accuracy but also maintains a lightweight architecture suitable for clinical deployment, while providing clinically meaningful explanations through explainability-enhancing mechanisms.

Accurate classification between MI and MCI is crucial for guiding personalised rehabilitation strategies. However, existing approaches either rely on handcrafted features such as bending angles and patient-reported outcomes (PROMs) [7,8], or deep learning methods applied to video data [3,4], without effectively integrating multimodal cues. More critically, many models lack explainability, a key requirement in clinical contexts.

We present a lightweight and accurate framework, MOXVEor, a novel multi-modal attention-gating network for classifying NSLBP subgroups. It distinguishes MI and MCI with up to 97.5% accuracy, while also delivering strong clinical explainability. Through integrated multiple explainability analysis approaches, our model offers clear and actionable explanations aligned with expert reasoning.

## 2. Related Work

Early computational methods focused on kinematic parameters such as flexion angles and posture transitions extracted via motion capture. Sheeran et al. [6] utilised 3D repositioning posture data with a Dempster–Shafer classifier to separate NSLBP subgroups. Others explored wearable sensors: Laird et al. [8] showed that lumbar-pelvic kinematics exhibit distinguishable subgroups in both healthy and LBP populations, while Bacon et al. [18] developed an LBP classifier using inertial measurement units (IMUs).

In recent years, researchers used machine learning to classify MI/MCI using features such as lumbar acceleration, bending angle, and PROMs. For example, Hartley et al. [3] employed feedforward neural networks on PROMs with some success. However, these models are often limited by unimodal inputs and lack explainability, making clinical trust difficult to establish.

Video-based deep learning has emerged as a promising approach to assess functional movement. Liu et al. [4] proposed SpineSighter, a CNN model classifying spinal function from video. Hartley et al. [3] demonstrated that combining video data with PROMs can significantly improve MI/MCI classification performance. Nonetheless, this research still treated modalities independently or via simple concatenation, without a learnable fusion mechanism or detailed explainability framework.

Multi-modal learning integrates heterogeneous information—such as vision, kinematics, and clinical scores—allowing models to leverage their complementary strengths. While prior works have used naive concatenation or early fusion, adaptive fusion mechanisms that dynamically weigh each modality based on its relevance to the task are still underexplored in LBP research. Attention-based fusion, in particular, has shown strong potential to dynamically weight informative modalities [19]. For medical explainability, Grad-CAM [9] and Integrated Gradients (IG) [10] are widely adopted for visualising CNN activations and feature contributions, respectively.

However, previous studies either used unimodal data or simple fusion techniques and often lacked interpretability,

and few studies fuse visual and feature-based interpretations into a unified visualisation space, whereas our work introduces a dynamically weighted multi-modal fusion with built-in explainability, which has not been explored in NSLBP classification before.

To this end, our study introduces MOVEXor, a lightweight, dynamically gated multimodal architecture that fuses CNN-based visual features (spinal curvature features) with motion features and PROMs. Unlike static methods, MOVEXor learns modality relevance on the fly, improving robustness to noisy or less informative inputs.

Compared with prior works, our framework emphasises not only classification accuracy but also explainability, clinical relevance, and efficiency, making it more suitable for deployment in real-world healthcare environments.

## 3. Dataset

This study used a previously described dataset of 83 patients with NSLBP (42 MI, 41 MCI, 47 females, mean age 44.7 years [SD=11.8, Range 22-76 years old]; height 170 cm [SD=9.9cm, Range 153cm-188cm]; mass 81.3kg [SD=16.7kg, Range 46kg-123kg]) [6], including motion data from a range of spinal functional assessment tasks that were classified according to consensus between two clinical experts [3]. These assessments included spine flexion, extension, side flexion, along with functional tasks like sit-to-stand, squat and both upright and slouched sitting postures, and were recorded using Vicon™ (Vicon, Oxford Metrics, UK), inertial measurement units (Xsens MVN, Xsens Technologies B.V., Netherlands), and videos (GoPro HERO, GoPro Inc., USA). Details of the specific exercise procedures, monitoring device parameters, participant demographics, and pain duration, as well as pretreatment steps, can be found in a previous publication [6].

The dataset also includes patient-reported outcomes (PROMs) that quantify LBP-related measures such as pain intensity, disability, and fear of movement, all of which have been fully described previously [3].

Because spinal flexion had the highest expert consensus (98%) for the classification of MI/MCI [3], this study focused on automatically classifying NSLBP patients by analysing their performance in spinal flexion. Two examples of MI and MCI patients are shown in Fig. 1.

(a) (b)

Fig. 1. Patients with NSLBP. (a) MCI and (b) MI.

## 4. Method

### *4.1. Overview*

Our framework classifies NSLBP patients into Movement Impairment (MI) and Motor Control Impairment (MCI) using multi-modal features derived from spinal curvature images, flexion movement, and patient-reported outcomes (PROMs). The overall approach consists of three main stages: (1) feature extraction from flexion videos and PROMs; (2) representation learning using a lightweight ResNet-18 for image encoding and feedforward layers for motion/PROM features; and (3) multi-modal attention gating, which adaptively fuses modalities based on their per-sample importance. To enhance transparency, we integrate Grad-CAM and Integrated Gradients (IG) for spatial and feature-level explanation. The following subsections detail each component.

### 4.2. Feature Extraction

We first performed feature extraction on the patient's spinal flexion video, including spinal curvature and motion feature. These features are used for MI/MCI classification, which have been shown to be effective for NSLBP-related classification [3, 4].

Individuals were side facing the camera while performing a forward flexion test, as shown in Fig. 2. We extracted these features from the video through human pose estimation (HPE) and represented them as mathematical features related to the spinal curvature angle ('$\theta$' in Fig. 2), which was calculated using the formula below:

$$\theta = \frac{\cos^{-1}(hn^2 + ah^2 - an^2)}{2 \cdot (hn \cdot ah)} \tag{1}$$

where $\cos^{-1}$ is Inverse function of cosine function, $hn$ is the line from hip to neck, $ah$ is the line from ankle to hip, $an$ is the line from hip to neck.

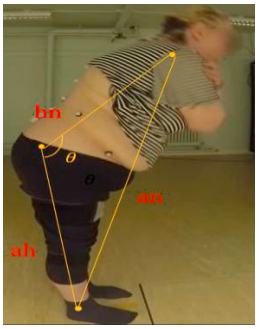



Fig. 2. Patients marked by Human Pose Estimation (HPE) model.

After obtaining the values of angle $\theta$ throughout the video, we calculated velocity and acceleration by taking the first and second derivatives, respectively. Then the motion features such as the mean, range, minimum, maximum, variance, standard deviation, stability time, depth variance, and repetition time variance of the angle θ, velocity and acceleration were calculated following the method described in detail in previous research [4].

Spinal curvature features are often used by physiotherapists when classifying NSLBP patients into different subgroups [2]. In this work, HPE and human instance segmentation (HIS) were used together to track and segment the human figure in the video. Once the patient reached a certain angle (from 100° to 180°, with an interval of 5°), a large number of masks that only reflect the back curvature were obtained by automatic cropping the human figure to represent the spinal curvature features. The bending masks of MI and MCI are shown in Fig. 3.

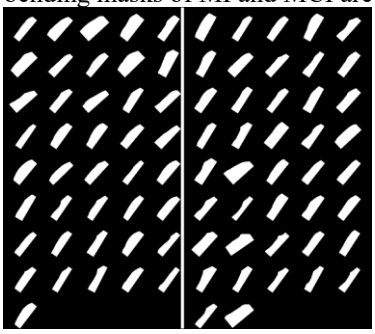


Fig. 3. The extracted masks of the back curve of all patients for the forward bending angle is 135°. The left group is MCI, another is MI.

### 4.3. MOVEor Design

The MOVEXor framework shown in Fig. 4 is proposed to classify NSLBP patients into MI and MCI subgroups and perform explainability analysis. MOVEXor integrates multimodal features, namely spinal curvature features,

motion features, and patient-reported outcome measures (PROMs). MOVEXor consists of three parallel data streams, namely the visual backbone for spinal curvature features, the motion feature encoder, and the PROM encoder. The visual backbone is used for spinal curvature features, while the motion and PROMs features are processed by their respective encoders and finally fused through a multi-modal attention gating (MAG) module.

For motion and PROMs features, each type of feature (e.g., angle) are passed through a BatchNorm layer, a linear layer, a LeakyReLU activation layer, and a dropout layer to normalize and project it into a shared embedding space. The purpose is to enhance feature stability, mitigate overfitting, and facilitate effective integration with visual features in later fusion stages.

The visual backbone is based on a ResNet-18 [11] pretrained on ImageNet to extract high-level back curvature features from each back mask to extract deep spatial features that capture curvature and postural pattern. The final classification layer is removed and global average pooling is applied to each mask. Each input mask is passed through the convolutional layers, and the resulting frame-level feature vectors are aggregated via a temporal average pooling operation. This enables the model to focus on back-related shape patterns, while keeping the visual representation compact and informative. ResNet-18 was selected due to its favorable balance between representational power and computational efficiency, which is particularly suitable for our limited dataset size. Deeper networks such as ResNet-50 did not show significant performance improvement in preliminary tests and risk overfitting in small-sample clinical data

The above three streams yield compact representations of the entire movement. To support flexible multi-modal integration, we introduce a multi-modal attention gating (MAG) mechanism that dynamically assigns weights to the spinal curvature and motion representations before concatenation and final prediction:

$$\alpha_{\text{image}}, \alpha_{\text{motion}}, \alpha_{\text{PROMs}} = softmax\left(W[f_{\text{image}} || f_{\text{motion}} || f_{\text{PROMs}}]\right) \quad (2)$$

where $f_{image}$, $f_{motion}$ and $f_{PROMs}$ are intermediate features from image and handcrafted streams.

The fused representation is then computed as:

$$f_{\text{fused}} = \alpha_{\text{image}} \cdot f_{\text{image}} + \alpha_{\text{motion}} \cdot f_{\text{motion}} + \alpha_{\text{PROMs}} \cdot f_{\text{PROMs}} \quad (3)$$

The MAG module allows the network to learn the relative importance of visual versus motion inputs on a per-patient basis, effectively focusing on the more informative modality. This dynamic weighting improves robustness to noise (e.g., inconsistent motion capture or video quality) and was found to boost classification accuracy (see Table 2, Table 3 and Fig. 5) compared to static fusion (e.g., simple concatenation of features).

This fusion strategy allowed MOVEXor to adaptively focus on the most informative modality per sample.

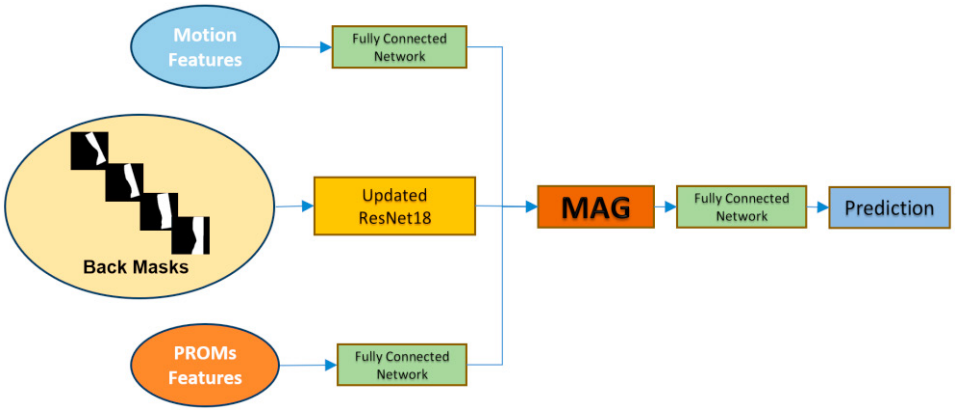



Fig. 4. MOVEXor framework

### 4.4. Evaluation

The proposed strategy is aimed at evaluation of the effectiveness of the MOVEXor in distinguishing MI/MCI subgroups in NSLBP patients using multi-modal features. The importance of combining spinal curvature features with other different types of features was also evaluated.

The model's generalisability was evaluated using 5-fold cross validation, which was a technique that maximises the use of model training data while ensuring model reliability and generalisation ability [12]. The data was divided into five equal parts, four of which were used for training and one for testing, and so on until each part was used as a test set. Finally, we took the average of the five tests. The averaged evaluation metrics included accuracy, sensitivity, specificity, and the F1 score. The detailed calculation methods of these indicators refer to previous study [4].

### *4.5. Explainability*

To improve transparency and trust, we adopted Grad-CAM [9] to identify spatially salient regions in video frames (including top-ranked informative frames) and IG [10] to quantify feature-wise importance within the input from feedforward stream (e.g., ranking top-k features).

### *4.6. Implementation details*

The proposed model was trained for classifying MI/MCI based on multi-modal features using a NVIDIA 4090 GPU. Binary cross-entropy loss was used as the loss function. Key training hyperparameters are shown in Table 1.

Table 1 – Hyperparameters

| | |
|---|---|
| **Optimizer** | **Adam** |
| **Learning Rate:** | **1e-4** |
| **Batch Size** | **16** |
| **Epoch** | **200** |
| **Number of Mask per patient** | **640** |

Key: Optimizer means the optimization algorithm; Learning Rate means the step size for updating model parameters. Batch Size: The number of samples processed before the model is updated. Epoch: One full pass through the entire training dataset. Number of Masks per Patient: The total number of masks selected for each patient during training and evaluation.

## 5. Results

### *5.1. Classification performance*

In this section, Table 2 shows the classification performance results obtained by the MOVEXor classification framework using different feature combinations.

Table 2 – Classification performance with multi-modal attention gating attention.

| Feature | Accuracy | Sensitivity (LF) | Specificity (HF) | F1 Score |
|---|---|---|---|---|
| 640 Images + Angle | **97.50%** | 97.50% | 98.00% | 0.9777 |
| 640 Images + Velocity | 94.12% | 100.0% | 90.00% | 0.9474 |
| 640 Images + Acceleration | 93.75% | 100.0% | 87.50% | 0.9412 |
| 640 Images + PROMs | 93.75% | 87.50% | 100.0% | 0.9333 |

The classification performance of the proposed MOVEXor model under various feature combinations is presented in Table 2. Overall, the combination of spinal curvature features (640 back mask images achieved the best results in the pre-experiment) with the bending angle achieved the best results, with an accuracy of 97.5%, sensitivity of 97.5%, specificity of 98.0%, and an F1 score of 0.978. This suggests that the bending angle is a highly discriminative feature for classifying movement impairment (MI) from motor control impairment (MCI).

In contrast, other speed-related features, such as combining "Images" with "Velocity" or "Acceleration", performed relatively poorly in accurately identifying MI/MCI, resulting in lower overall accuracy and F1 scores. This suggests that "speed" may have limited impact on classification, or even have a negative effect as the features become higher-order as shown in Table 2.

Table 3 – Classification Performance without multi-modal attention gating network.

| Feature | Accuracy | Sensitivity (LF) | Specificity (HF) | F1 Score |
|---|---|---|---|---|
| 640 Images + Angle | **86.77%** | 83.62% | 91.50% | 0.8717 |
| 640 Images + Velocity | **82.06%** | 76.20% | 89.50% | 0.8299 |
| 640 Images + Acceleration | **84.34%** | 83.59% | 85.33% | 0.8351 |
| 640 Images + PROMs | **81.91%** | 81.67% | 80.83% | 0.8180 |

To further assess the effect of the multi-modal attention gating (MAG) mechanism, we conducted an ablation study by disabling attention and fusing features via static concatenation. In this case, classification performance dropped across all settings. For example, the image + angle fusion without MAG achieved only 86.8% accuracy (F1 0.872), indicating that MAG plays a critical role in adaptively prioritising the most informative modality. The drop was more pronounced in configurations involving noisier features like velocity or acceleration, further reinforcing that static fusion is insufficient for NSLBP classification. The results of a direct comparison on "640 Images + Angle" between the two are shown in Table 3 and Fig. 5.

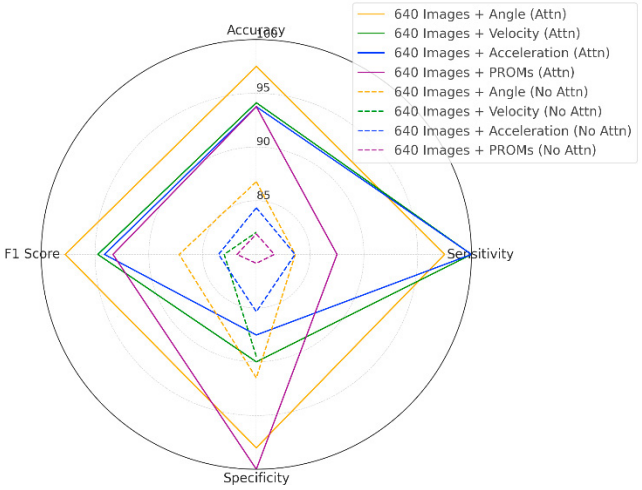



Fig. 5. Radar chart comparing classification performance across feature combinations, showing consistent improvements with MAG, especially for image + angle fusion.

These findings demonstrate two key insights: (1) among motion-derived metrics, bending angle consistently contributes the most reliable information for MI/MCI classification; and (2) the multi-modal attention gating (MAG) mechanism substantially improves model performance by dynamically weighting features on a per-case basis.

### *5.2. Explainability Results*

A major strength of MOVEXor lies in its explainable design, combining Grad-CAM and IG. These provide comprehensive, multi-level insight into the model's decision process.

(1) Local Explainability

(a) (b)

Fig. 6. (a) max Grad-CAM, (b) min Grad-CAM

As shown in Fig. 6, the Grad-CAM heatmap consistently highlights the thoraco-lumbar spine region during key motion phases. Attention usually peaks at the frames at the extremes of flexion (Fig. 6-a), and is lowest at the bottom of stance (Fig. 6-b).

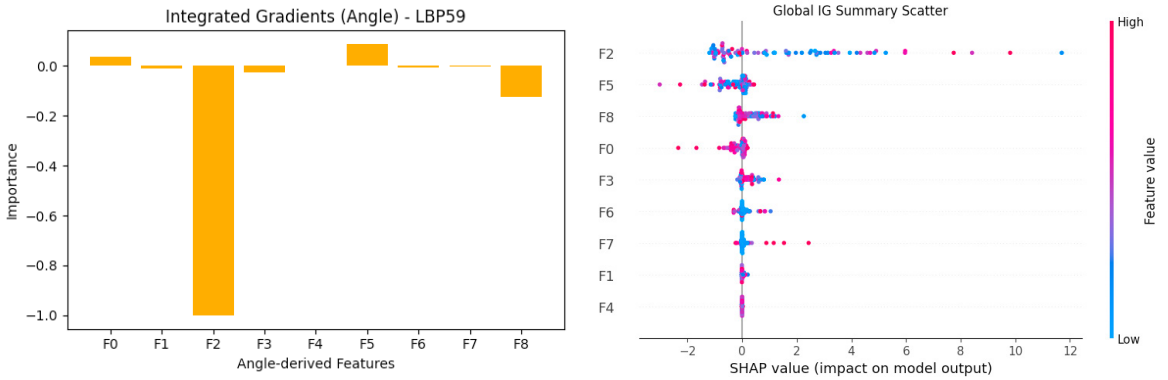
IG further complements local explainability on one LBP individual by ranking the motion features for each case. This confirms that MOVEXor is still heavily influenced by a single dominant feature even when all features are in effect. As shown in Fig. 7-a, in this case, F2 (Minimal full spine flexion angle) ends to decrease the probability of predicting the positive class, whereas F5 (Repetition Time) is the most influential feature that enhances classification accuracy.

(2) Global Explainability

We conducted a global explainability analysis on our dataset using Integrated Gradients (IG) to identify which motion features contributed most to classification decisions. As shown in Fig. 7-b, F2 (Minimal flexion angle) exhibited the strongest and most consistent positive attribution, indicating that patients with more limited flexion were more likely to be classified as **MI**. Red points on the right side of F2 reflect high feature values driving MI predictions, consistent with clinical reasoning.

In Fig. 7-c, the average IG attribution across all samples again highlights F2 as the most dominant feature, with F8 (Motion Stability) also showing moderate contributions. These results suggest that while range of motion is the primary factor, temporal and control-related features also influence decisions—especially in identifying MCI patterns. Overall, MOVEXor's behaviour aligns well with clinical expectations in NSLBP subgroup classification.

As shown in Fig. 8 and Fig. 9, global Grad-CAM analysis shows that Max CAM activations frequently occurred at bending angles between 100° –145° and during mid-to-late frames of the flexion task (frame indices 200–800). In contrast, Min CAM activations were often found at full flexion angles (170° –185°) and later frames. These results indicate that the model consistently focuses on key movement phases that differentiate MI and MCI, particularly limited or excessive bending patterns.



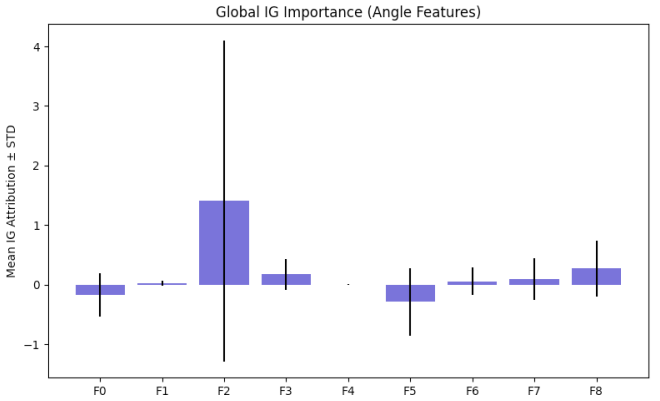



Fig. 7. Integrated Gradients (IG) analysis. (a) Integrated Gradients analysis, (b)Global IG Summary Scatter, (c)Global IG Importance bar chart. Key: F0 – Variance, F1 – Standard Deviation (SD), F2 – Minimal (full spine flexion angle), F3 – Range, F4 – Maximal (spine standing angle), F5 – Repetition Time Mean (RTM), F6 – Repetition Time Variance (RTV), F7 – Depth Variance (DV), F8 – Motion Stability (MS)

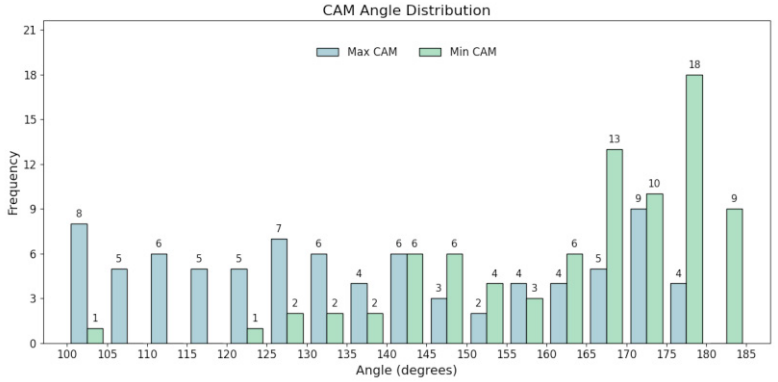



Fig. 8. Statistical analysis for max/min CAM angle

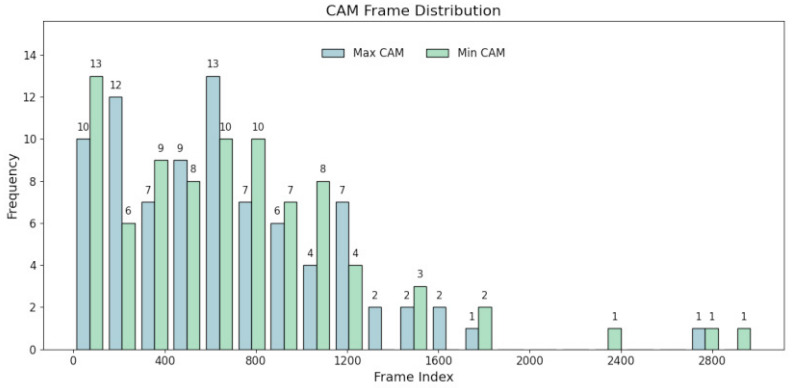



Fig. 9. Statistical analysis for max/min CAM frame

## 6. Discussion

### *6.1. Classification performance*

The results indicate that MOVEXor achieved high accuracy in classifying MI and MCI subgroups of NSLBP patients, with especially strong performance when combining image features with bending angle. This reinforces the clinical understanding that range of motion is a fundamental differentiator between these subtypes [2].

Importantly, the ablation results show that removing the attention mechanism (MAG) significantly impairs performance. This highlights that adaptive fusion is essential for handling patient variability. For example, some MI patients may exhibit clear visual stiffness, while others present more ambiguous patterns. A fixed fusion strategy treats all inputs equally, potentially diluting salient signals. In contrast, our MAG allows the model to dynamically prioritise modalities per individual—enhancing both robustness and explainability.

Our findings are consistent with, and extend upon, recent work [4], which demonstrated that motion-based features, particularly sagittal range of motion, were among the most informative for differentiating MI/MCI. Our MOVEXor, combines spinal curvature and motion features and achieves a higher performance (97.5% accuracy) with greater clinical feasibility.

Interestingly, adding velocity or acceleration features did not improve classification. This might be due to their greater variability and susceptibility to noise from pose estimation, especially in videos without marker-based tracking [13]. Moreover, high-order motion features may not consistently reflect subgroup-defining impairments. Velocity

fluctuations may occur in both MI (due to hesitancy) and MCI (due to poor control), making them less discriminative. This observation is aligned with earlier findings that qualitative movement quality—rather than just kinematic quantity—is critical in NSLBP classification [2,14].

### *6.2. Explainability*

A major strength of MOVEXor lies in the inclusion of the attention mechanism, which introduces flexibility in choosing feature importance. Additionally, combining Grad-CAM and provide comprehensive, multi-level insights into the model's decision-making process, opening up the black box and having great significance in actual clinical applications.

As shown in Fig. 6, Grad-CAM heatmaps consistently highlight the thoraco-lumbar region indicating that the model systematically attends to those anatomically relevant areas during the spinal flexion movement for classification. This is in line with previous research demonstrating distinct differences in lower thoracic and upper lumbar region across NSLBP subgroups [15]. The model's consistent focus on the thoraco-lumbar spine suggests that MOVEXor identifies motion-relevant regions in a physiologically meaningful way, aligning with prior findings that NSLBP patients exhibit altered movement coordination and segmental spinal control in the lumbar region [14,15].

In contrast, Fig. 6-b shows the same patient in a standing position, where Grad-CAM activations are markedly reduced. This reflects the model's ability to suppress attention when motion-related visual cues are minimal, aligning with clinical logic that neutral postures provide little discriminative information. The sharp contrast between activation levels in dynamic versus static phases reinforces that MOVEXor focuses on functional movement behaviour, rather than irrelevant static features, further enhancing its clinical explainability.

For global feature importance, as illustrated in Fig. 7, we analysed IG attributions across all patients. The minimal value exhibited the highest average IG value and largest variance, suggesting that it consistently played a significant role in the classification decision. Other angle-related features (e.g., motion stability, repetition time mean) also showed moderate contributions. This observation supports the view that the range of motion is the main discriminant factor, while supplementary features such as angular change or timing can also influence the decision process [2,3,14].

Further, we analysed the angle and frame index distributions where Grad-CAM activations reached their maximum and minimum, as shown in Fig. 8 and Fig. 9. Max-CAM activation occurred most frequently at maximum and mid-flexion angles (100° -145°) and mid-to-late frame indices (200-800), consistent with clinical times when the end range is reached or close to it. In contrast, Min-CAM activation clustered around full flexion angles (170° -185°) and late frames, suggesting that the model is finding less discriminative visual information when the patient is already fully upright. These distributions verify that the model's attention is not random but rather aligns with clinically meaningful stages of movement [16].

### *6.3. Clinical implications*

MOVEXor offers practical value as a decision-support tool in clinical practice, where MI/MCI classification is often subjective and its accuracy is dependent on physiotherapists' level of training and experience [20]. By providing objective, reproducible predictions from a simple side-view video, the system can assist both less experienced clinicians and telehealth scenarios [20].

In addition, the explainability outputs (IG table and Grad-CAM heatmap) also enhance patient and clinician trust—for example, showing the frame where motion stops early can help understanding patients' limitations and track improvement over time. Furthermore, modality-specific attributions can guide personalised exercise. Its lightweight architecture supports seamless integration into clinical workflows or even phone-based assessments.

### *6.4. Limitations*

This study has several limitations. First, the sample size (n = 83) limits its generalisability; future studies should include larger and more diverse cohorts (potentially several hundred) to robustly balidate the model's performance. Second, although the forward bending task was the most recognised task in classifying MI/MCI, the role of other functional tasks may have been overlooked.

### 6.5. Future work

Future research will extend MOVEXor in multiple directions. Our goal is to develop a multi-task model that supports classification across multiple NSLBP subgroups (e.g., flexion vs. extension patterns MCI). In addition, there is potential in using time-series-based modeling of input frames to enhance the impact of time and improve sensitivity to motion coordination. Beyond low back pain, there is also potential to explore MOVEXor in other musculoskeletal health conditions affecting the spine (e.g., scoliosis, spinal stenosis) and the hip (e.g., femoroacetabular impingement, hip osteoarthritis), broadening its clinical applicability. On the clinical side, MOVEXor could be integrated into deployable device and evaluated for its impact on treatment planning and patient outcomes.

## 7. Conclusion

In this study, we proposed MOVEXor, a lightweight and explainable multi-modal framework for classifying NSLBP patients into movement impairment (MI) and motor control impairment (MCI). By integrating visual features from video with motion features through a modality-aware attention gating mechanism, MOVEXor achieved high classification accuracy while offering clinically meaningful explanations via Grad-CAM and Integrated Gradients (IG). Our analysis demonstrates that the model focuses on physiologically relevant phases and features, particularly range of motion, and adapted its modality reliance per patient. With minimal input requirements and strong explainability, MOVEXor holds promise as a practical, trustworthy tool for supporting clinical decision-making, personalised rehabilitation planning, and patient engagement in both in-clinic and remote care contexts.

## References

[1] Maher, C., Underwood, M., & Buchbinder, R. (2017). Non-specific low back pain. The Lancet, 389(10070), 736–747. https://doi.org/10.1016/S0140-6736(16)30970-9

[2] O'Sullivan P. Diagnosis and classification of chronic low back pain disorders: Maladaptive movement and motor control impairments as underlying mechanism. Manual Therapy. 2005;10(4):242–55. pmid:16154380

[3] Hartley, Thomas, et al. "BACK-to-MOVE: Machine learning and computer vision model automating clinical classification of non-specific low back pain for personalised management." Plos one 19.5 (2024): e0302899.

[4] Liu, Z., et al. (2024). SpineSighter: an AI-driven approach for automatic classification of spinal function from video. Procedia Computer Science, 246, 3977–3989. https://doi.org/10.1016/j.procs.2024.09.172

[5] NICE. National Institute of Health and Care Excellence, Low back pain and Sciatica over 16s: Assessment and Management (NICE guideline NG59)2016 (last updated 2020). https://www.nice.org.uk/guidance/ng59.

[6] Sheeran L, Robling M. Spinal function assessment and exercise performance framework for low back pain[C]//Orthopaedic Proceedings. Bone & Joint, 2019, 101(SUPP_9): 40-40.

[7] Sheeran, L., Sparkes, V., Whatling, G., Biggs, P., & Holt, C. (2019). Identifying non-specific low back pain clinical subgroups from posture tasks using a novel Dempster–Shafer classifier. Clinical Biomechanics, 70, 237–244.

[8] Laird, R. A., Keating, J. L., & Kent, P. (2018). Subgroups of lumbo-pelvic flexion kinematics in LBP. BMC Musculoskeletal Disorders, 19(1), 309.

[9] Selvaraju, R. R. et al. (2017). Grad-CAM: Visual Explanations via Gradient-based Localization. ICCV.

[10] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. International Conference on Machine Learning (ICML), pp. 3319–3328.

[11] He K., Zhang X., Ren S., Sun J. (2016) Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[12] Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. Journal of Machine Learning Research, 5(Sep), 1089-1105.

[13] Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience, 21(9), 1281–1289. https://doi.org/10.1038/s41593-018-0209-y

[14] Vibe Fersum K, O'Sullivan P, Skouen JS, Smith A, Kvåle A. Efficacy of classification-based cognitive functional therapy in patients with non-specific chronic low back pain: a randomized controlled trial. Eur J Pain. 2013 Jul;17(6):916-28. doi: 10.1002/j.1532-2149.2012.00252.x. Epub 2012 Dec 4. PMID: 23208945; PMCID: PMC3796866.

[15] Hemming, R., Sheeran, L., van Deursen, R. et al. Non-specific chronic low back pain: differences in spinal kinematics in subgroups during functional tasks. Eur Spine J 27, 163–170 (2018). https://doi.org/10.1007/s00586-017-5217-1

[16] Falla, D. L., Gizzi, L., Tschapek, M., Erlenwein, J., & Petzke, F. (2014). Reduced task-induced variations in the distribution of activity across back muscle regions in individuals with low back pain. Pain, 155(5), 944–953. https://doi.org/10.1016/j.pain.2014.01.027

[17] Luomajoki, H., Kool, J., de Bruin, E. D., & Airaksinen, O. (2008). Movement control tests of the low back: Evaluation of the difference between patients with low back pain and healthy controls. BMC Musculoskeletal Disorders, 9, 170. https://doi.org/10.1186/1471-2474-9-170ResearchGate+4PMC+4PubMed+4

[18] Z. Bacon, Y. Hicks, M. Al-Amri, and L. Sheeran, 'Automatic Low Back Pain Classification Using Inertial Measurement Units: A Preliminary Analysis', Procedia Computer Science, vol. 176, pp. 2822–2831, 2020, doi: 10.1016/j.procs.2020.09.272.

[19] Tadas Baltrušaitis, Chaitanya Ahuja and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 2018.

[20] K. V. Fersum, P. O'Sullivan, A. Kvåle, and J. Skouen, "Inter-examiner reliability of a classification system for patients with non-specific low back pain," Manual Therapy, vol. 14, no. 5, pp. 555-561, 2009, doi: 10.1016/j.math.2008.08.003.