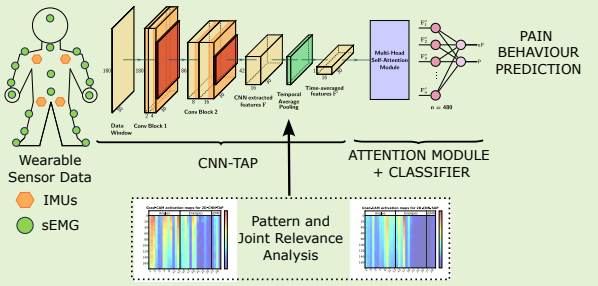
L-SFAN: Lightweight Spatially-focused Attention Network for Pain Behavior Detection

Jorge Ortigoso-Narro, Fernando Diaz-de-Maria *Member, IEEE*, Mohammad Mahdi Dehshibi *Senior Member, IEEE*, Ana Tajadura-Jiménez

***Abstract*— Chronic Low Back Pain (CLBP) afflicts millions globally, significantly impacting individuals' well-being and imposing economic burdens on healthcare systems. Detecting protective behavior is essential for effective chronic pain management, as it can help prevent pain aggravation and disability. To reduce this burden, we could leverage sensor information and AI techniques to facilitate at-home patient follow-ups. Precisely, by utilizing motion sensors and surface electromyography (sEMG) sensors, we can continuously monitor movement patterns and muscle activity. This paper introduces L-SFAN, a lightweight convolutional neural network (CNN) architecture that innovatively models both spatial and temporal dimensions of multivariate time series to detect protective behavior. L-SFAN uses 2D CNN to capture spatial patterns from the 2D matrix formed by the multivariate time series and uses self-attention to capture long-range temporal dependencies. Temporal average pooling is used to emphasize spatial patterns. On the EmoPain dataset, L-SFAN outperforms state-of-the-art methods while reducing the number of parameters up to 94%, making it lightweight and embedded systems friendly. The ablation study underscores the importance of jointly modeling spatial-temporal information. The competitive performance and efficiency of our proposed method demonstrate its practicality for accessible chronic pain monitoring.**



***Index Terms*— Motion sensors, sEMG sensors, Convolutional Neural Network, Global Average Pooling, Pain-related Behavior, Spatial Patterns, Self Attention.**

I. Introduction

CHRONIC low back pain (CLBP) is a debilitating condition affecting millions worldwide, with a significant impact on individuals' quality of life and substantial socio-economic costs [1], [2], [3]. Effective management of CLBP relies heavily on physical rehabilitation, typically involving self-management programs under the periodic supervision of medical professionals [4]. However, limited access to clinical facilities and the need for continuous monitoring present significant challenges in CLBP rehabilitation.

Emotional well-being plays a crucial role in the success of rehabilitation, especially for chronic pain patients, as emotional distress can intensify pain perception [5]. By moving rehabilitation to the comfort of home, patients benefit from a familiar environment that can reduce emotional stress and improve outcomes. CLBP patients frequently modify their movement patterns to avoid pain, leading to protective behaviors that can worsen disability and cause additional musculoskeletal problems [6]. Detecting these protective behaviors is challenging due to their subtle nature and individual variations. Current clinical assessment methods rely heavily on periodic in-person observations, which may miss important behavioral changes between visits. Therefore, developing automated methods for continuous monitoring of protective behaviors is essential for more effective intervention strategies [7]. This requires addressing several technical challenges, including processing multimodal sensor data, handling temporal variations in movement patterns, and achieving efficient computation for potential real-time applications.

To address this, the integration of motion sensors and surface electromyography (sEMG) sensors offers a promising solution for remote monitoring and intervention. Motion

This research has been funded through various sources. The Spanish Ministry of Science and Innovation (State Research Agency) provided partial funding through National Grant PID2020-118504GB-I00. Madrid Regional Government also provided partial funding through grants Y2020/NMT-6660 for the interdisciplinary project COMPANION-CM and REACT UE Grant IntCARE-CM for the Intelligent and Interactive Home Care System to mitigate the COVID-19 pandemic. The European Research Council (ERC) also provided funding through the European Union's Horizon 2020 research and innovation program with grant agreement No 101002711 for the BODYinTRANSIT project.

Jorge Ortigoso-Narro is affiliated with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid in Leganés, Spain (e-mail: jortigoso@tsc.uc3m.es).

Fernando Diaz-de-Maria is affiliated with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid in Leganés, Spain (e-mail: fdiaz@ing.uc3m.es).

Mohammad Mahdi Dehshibi is affiliated with the Department of Computer Science and Engineering, Universidad Carlos III de Madrid in Leganés, Spain, as well as the Unconventional Computing Laboratory at the University of the West of England in Bristol, U.K. (e-mail: mohammad.dehshibi@yahoo.com).

Ana Tajadura-Jiménez is affiliated with the Department of Computer Science and Engineering, Universidad Carlos III de Madrid in Leganés, Spain, as well as the UCL Interaction Centre, University College London in London, U.K. (e-mail: atajadur@inf.uc3m.es).

sensors capture detailed movement patterns, and sEMG sensors provide insights into muscle activity and performance. Previous research has demonstrated the effectiveness of these sensors in posture-related and joint position classification [8], [9] as well as for pain assessment [10]. These advances enable continuous, real-time assessment of physical function and muscular responses, facilitating personalized and adaptive therapeutic strategies that extend beyond conventional care environments [11].

Recent advancements in deep learning [12], [13] offer promising solutions for addressing these challenges. These contributions can potentially enable continuous, remote monitoring of patients and provide valuable insights to inform personalized rehabilitation strategies. However, existing models for protective behavior detection, such as Stacked-LSTMs [14], GRU-based recurrent neural networks [15], graph-based neural networks [16], and spatio-temporal attention networks [17], often struggle to effectively capture the complex spatio-temporal patterns associated with protective behavior. Moreover, many of these models are computationally expensive, limiting their applicability in resource-constrained settings [18].

To address these limitations, we propose L-SFAN, an end-to-end lightweight convolutional neural network (CNN) architecture designed to effectively model the spatial and temporal dimensions of multivariate time series data for detecting protective behavior. The lightweight design is crucial, as real-time processing of multimodal sensor data requires efficient computation. By using a 2D CNN to extract temporal and spatial (i.e., sensor placement) feature representations, we can identify which features are most relevant to our task. Additionally, incorporating self-attention mechanisms enables us to capture long-range dependencies in the data, which is essential for understanding complex protective behavior patterns.

Our main contributions and results can be summarized as follows:

1) L-SFAN achieves competitive performance with state-of-the-art methods while using significantly fewer parameters (8K compared to more than 100K in recent approaches). This demonstrates that protective behavior detection can be accomplished with more compact architectures without sacrificing performance.
2) Our systematic analysis highlights that spatial patterns—the relationships between different sensor outputs—are more informative than their temporal evolution for detecting protective behaviors. This finding provides new insights for the design of efficient architecture multimodal sensor data processing.

In summary, L-SFAN delivers not only strong performance and improved efficiency compared to existing methods but also emphasizes spatial patterns, which appear to be more relevant for detecting protective behaviors. Altogether, these properties enhance its potential for practical and accessible applications in chronic pain monitoring.

The rest of this paper is organized as follows: Section II reviews related work on protective behavior detection using the EmoPain dataset. Section III describes the proposed L-SFAN architecture in detail. Section IV presents the experimental setup, results, and analysis. Finally, Section V concludes the paper.

II. Related work

Initial research using the EmoPain dataset [19] pointed towards the utility of RNN architectures in detecting pain levels and recognizing protective behaviors, a premise further expanded by other researchers. Through a series of studies, Wang et al. transitioned from basic LSTM models [17] to more complex structures such as Dual-Stream LSTMs [14] and Stacked-LSTMs [20] for processing of body movement data along with data augmentation and segmentation window width approaches. This progression demonstrated the versatility of LSTMs and their ability to improve pain behavior detection by incorporating more complex and hierarchical temporal-focused structures. In [14], two sets of three LSTM layers were used to analyze the MoCap and sEMG data. Experiments demonstrated the potential for better architecture to accommodate different data types. Additionally, the impact of sliding window lengths on detection performance is discussed, with evidence that the choice should be based on knowledge of the dataset, as it is affected by the duration and complexity of the movement. In [20], recurrent networks with three LSTM units were explored. The experimental results revealed that the stacked-LSTM outperforms the dual-stream LSTM and CNN-based models.

Other RNN-based approaches aim to leverage different strategies to improve the performance of recurrent networks. Li et al. [21] introduced a novel anomaly detection-based system using an LSTM deep neural network with LSTM units to automatically assess chronic pain intensity and PBD from body movements within the sparse occurrences in the dataset. This method addresses the limitation of imbalanced expert-labeled data for pain estimation and proposes an anomaly detection-based approach to enhance the network's performance. The study also delved into joint training with exercise type labels, data balancing techniques, hierarchical classification for pain intensity estimation, and the cascaded estimation of protective behavior and pain intensity. Dehshibi et al. [15] discussed the challenges of analyzing biometric data, specifically in individuals with chronic pain, and proposed a method to classify pain levels and pain-related behavior. The proposed method involves a sparsely connected recurrent neural network ensemble with gated recurrent units and incorporates information-theoretic features to compensate for variations in the temporal dimension.

The inclusion of attention mechanisms was introduced in [17] where BodyAttentionNet (BANet) highlighted the importance of focusing on a subset of joint angles and using bodily attention mechanisms to capture the most informative temporal and body configurational cues characterizing specific movements and strategies. The proposed architecture demonstrates a substantial improvement in PBD performance and a significant reduction in the number of parameters compared to other LSTM-based architectures. The integration of Human Activity Recognition (HAR) with protective behavior detection

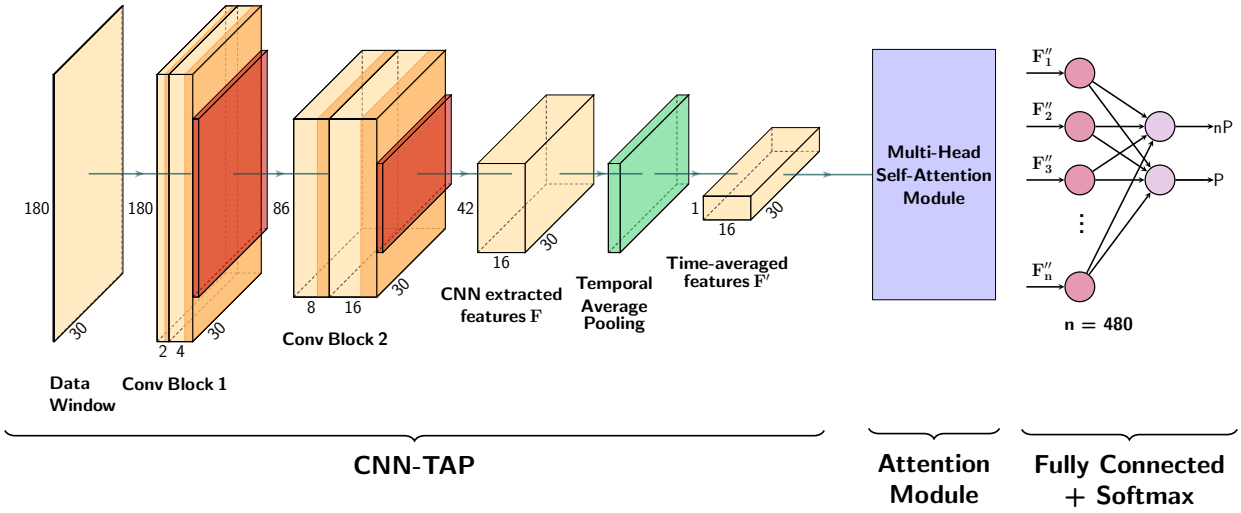


Fig. 1. The schematic of the proposed L-SFAN architecture for protective behavior detection. The 180×30 input matrix (13 **Joint Angles** + 13 **Joint Energies** from MoCAP IMUs + 4 sEMG outputs) is processed by a CNN-TAP backbone for feature extraction. A multi-head self-attention module further refines the features, which feed into a linear layer with softmax, providing probabilities for protective behavior (P) and its complement (nP).

for continuous data in chronic pain management was proposed in [16]. The authors suggested a hierarchical HAR-PBD architecture using a graph convolution network and an LSTM with a class-balanced focal categorical cross-entropy loss to alleviate class imbalances. This study highlights the potential applications of the proposed architecture in managing chronic pain. It also examined the limitations of current methods that depend on pre-segmented activities and offered deeper insights into the importance of incorporating information about spatial relationships.

Phan et al. [22] proposed a deep learning approach based on CNN and LSTM networks to extract multi-level context information from physiological signals to distinguish between pain and painlessness activities. A 1D CNN was used to extract spatial information where a pooling layer was responsible for dimension reduction. The extracted spatial information was then fed to a bidirectional LSTM network to process the temporal dimension of the data. After extracting the context vector, they used variant attention [23] after the spatio-temporal processing. Experimental results demonstrated that multi-level context information performs significantly better than uni-level context information for comprehensively analyzing spatio-temporal physiological signals.

The exploration into multi-level fusion approaches by Uddin and Canavan [22] and Phan et al. [24] introduced a nuanced understanding of pain behavior prediction. These authors presented an approach for pain estimation and PBD using a multimodal, multi-level fusion of movement data.

The inclusion of contextual information for personalized assessment was explored in [25], where the researchers leveraged information about the patient's current health condition to predict the pain level assessment. While this method explores the integration of a self-assessment framework, we find that its reliance on the patient's state beforehand makes it unsuitable for direct comparison with our work. Although prior knowledge of patient conditions could benefit the model, such information is not always available in real-world scenarios. Therefore, we excluded contextual information from our study to ensure a more realistic approach.

Current recurrent architectures, while effective for temporal modeling, face computational challenges due to their architectural complexity. These models require a large number of parameters (see Table IV) and make design choices that impact computational efficiency. For example, Dual-Stream LSTMs process motion and EMG data separately, while BodyAttentionNet adds complexity through its attention mechanisms. These architectural limitations highlight the need for more efficient approaches to protective behavior detection.

III. Proposed Architecture

We use a sliding window technique on continuous recordings to extract 3-second sEMG and MoCAP data intervals. This duration effectively captures enough movement data to identify protective behaviors, balancing temporal sensitivity with model stability, as experimentally validated in [20]. These intervals, captured as 180×30 matrices, reflect a detailed compilation of 60 frames per second over 3 seconds, including sEMG readings, **Joint Angles**, and **Joint Energies** derived from anatomical positions of IMU sensors [19]. We also strategically omitted activity type considerations to ensure robust generalization across varied trials, focusing on detecting protective behaviors within each data window. Section IV further discusses the data preprocessing and representation.

The proposed architecture, named Lightweight Spatially-Focused Attention Network (L-SFAN), comprises two main parts: a 2D CNN feature extractor backbone and an attention mechanism to further refine the features for protective behavior detection. Fig. 1 depicts the proposed architecture, including details regarding the number of layers and dimensions of the activation maps.

A. L-SFAN – Feature Extraction Module

In this study, we introduce L-SFAN, an end-to-end lightweight 2D-CNN architecture specifically designed to address the unique challenges of identifying protective behavior in chronic pain using multivariate time series data. While 2D CNNs are commonly used in computer vision tasks [26], [27], [28], [29], [30], their ability to capture intricate patterns within two-dimensional data extends to other domains. Mainly, they are well-suited for modeling the complex relationships between multiple sensor modalities by inherently capturing spatial and temporal dependencies [31]. L-SFAN enables the model to learn discriminative features from sEMG and bodily-arranged motion sensor modalities for accurate protective behavior detection, which is crucial for understanding and managing chronic pain conditions.

We have configured our sensory data to form a 2D matrix of size 180×30. This specific data configuration incorporates temporal dynamics (i.e., 3 seconds of data at 60 Hz) and spatial information, including IMU and sEMG sensor locations and arrangements (i.e., Joint Angles and Joint Energies). For clarity, let us denote our time window of data as $\mathbf{X} = [s_1^{(k)}, s_2^{(k)}, \cdots, s_L^{(k)}]$, where $s_t^{(k)}$ represents a k-dimensional feature vectors at time step t over a time span of $0 < t < L$ with a window length of L.

Our sensory data naturally organizes as a 2D tensor (sensors vs. time), making it suitable for processing with 2D CNNs. Building on this data representation, we propose the Lightweight Spatially-Focused Attention Network (L-SFAN). The architecture's core component is its lightweight feature extractor backbone. This backbone consists of two basic convolutional blocks that learn hierarchical feature representations through convolutional and pooling layers. Each convolutional block comprises a sequence of two convolutional layers, each followed by a batch normalization layer and the corresponding activation function, and concludes with a max-pooling layer (see Fig. 2).

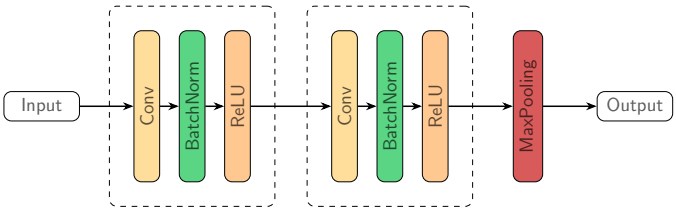


Fig. 2. The convolutional block used in the feature extractor module.

Convolutional layers employ a set of C filters (i.e., kernels) to extract distinctive local features or patterns from the input data and produce corresponding feature maps. Mathematically, the operation of a convolutional layer is expressed in Eq. 1.

$$Y_{i,j}^{(c)} = f\left(\sum_{m,n} X_{(i+m,j+n)} \cdot W_{m,n}^{(c)} + b^{(c)}\right), \qquad (1)$$

where $Y_{i,j}^{(c)}$ represents the feature map at spatial position (i, j) resulting from applying the c^{th} kernel, $X_{(i+m,j+n)}$ denotes the input data values within the kernel region, $W^{(c)}$ represents the weights associated with the c^{th} kernel, $b^{(c)}$ is the bias term for the c^{th} kernel, and f denotes a non-linear activation function. The convolutional operation enables the learning of local spatiotemporal patterns through learnable filters $W^{(c)}$, where each filter captures specific aspects of the sensor relationships and their temporal evolution. This localized processing is particularly effective for detecting characteristic movement patterns associated with protective behaviors.

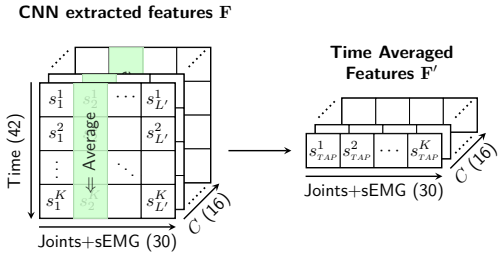


Fig. 3. Illustration of the global average pooling operation over the CNN extracted feature block.

The activation function used in this work is the Rectified Linear Unit (ReLU) [32] defined as $\mathrm{ReLU}(x) = \max(0, x)$)). Following this, a pooling layer is incorporated to reduce the feature map dimensionality, increase the receptive field, and enhance robustness against minor input variations.

We use a kernel size of $(3, 3)$ with zero-padding to maintain the spatial dimension of the feature maps. The zero-padding ensures the preservation of all 30 dimensions along the sensor (spatial) axis, which is crucial for retaining full resolution in the sEMG and joint data. To solely reduce the temporal dimension, we employ a max-pooling layer with a kernel size of $(2, 1)$ and stride of $(2, 1)$. Consequently, the feature extractor backbone transforms the input matrix $\mathbf{X} \in \mathbb{R}^{180\times30}$ into a feature tensor $\mathbf{F} \in \mathbb{R}^{16\times42\times30}$, where 16 represents the number of feature maps, 42 the temporal resolution, and 30 the spatial resolution.

In this work, we aim to capitalize on the potential benefits of preserving spatial resolution while analyzing temporal sensorial data. To achieve this, we propose using **Temporal Average Pooling (TAP)**. The pooling strategy, illustrated as CNN-TAP in Fig. 1, involves averaging the feature tensor $\mathbf{F}$ across the temporal dimension. This results in $\mathbf{F}' \in \mathbb{R}^{16\times30}$, which retains the spatial resolution, i.e., the sensor dimension. This preservation of spatial information is crucial for identifying which sensor inputs contribute most significantly to PBD.

Temporal average pooling aggregates information across the temporal dimension while preserving the spatial (sensor) dimension. This approach enables the final layers to focus on identifying relevant spatial patterns to make decisions (see Fig. 3).

B. L-SFAN – Self-Attention Module

The self-attention mechanism empowers deep learning models to effectively handle sequential data by selectively focusing on relevant input regions based on their similarity to other parts. Compared to RNNs, self-attention can capture long-range dependencies more effectively, avoids issues like

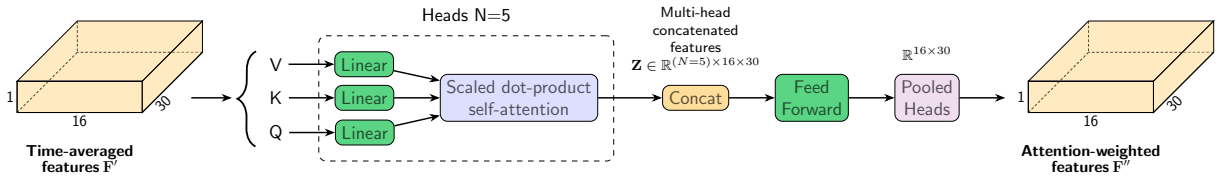


Fig. 4. Block diagram of the proposed multi-head attention module.

vanishing or exploding gradients, and enables parallel computation, making it highly advantageous for scalability and performance. Prior research has also demonstrated the benefits of incorporating attention mechanisms for protective behavior detection [17], [16]. In our case, we propose using self-attention to effectively capture the relevant interdependencies among the spatially-focused features extracted by our CNN-TAP backbone.

Formally, we calculate Queries (Q), Keys (K) and Values (V) with respect to the input sequence X using Eq. 2, where W_q, W_k and W_v represent learnable weight matrices.

$$\begin{aligned} Q &= W_q \cdot X \\ K &= W_k \cdot X \\ V &= W_v \cdot X \end{aligned} \tag{2}$$

Having calculated the query vector Q, we compute the affinity score. This score represents the similarity between the query and key, indicating the level of attention required for the corresponding value. The affinity score A for each input position is calculated using Eq. 3.

$$A = \frac{Q \cdot K^T}{\sqrt{d_k}}, \tag{3}$$

where $\sqrt{d_k}$ serves as a normalization factor that ensures the dot product result, $Q \cdot K^T$, is independent of the temporal dimension length ($d_k = 180$ in this study). Subsequently, a *softmax* function is applied to obtain normalized attention weights that sum to 1 for each query. Finally, the self-attention output is computed as a weighted sum of the value vectors, as described in Eq. 4.

$$Z = softmax\,(A) \cdot V \tag{4}$$

In the multi-head setting, this operation is repeated with different learned weight matrices ($W_q^{(i)}$, $W_k^{(i)}$, $W_v^{(i)}$) for each head $i = 1, \cdots, N$. The resulting multiple outputs are then concatenated and transformed linearly using a feedforward network. Unlike RNNs that process temporal information sequentially, self-attention computes direct relationships between all time points, enabling more efficient capture of long-range dependencies while maintaining computational parallelizability.

Given the pre-existing informative representation generated by the CNN-TAP feature extraction backbone, we opted not to use an additional embedding layer before the attention module. The number of attention heads ($N = 5$) was selected based on preliminary experiments and prior studies [33], [34] that demonstrate diminishing returns with additional heads in similar multivariate time series tasks. This choice balances computational efficiency with representative capacity, as each head can specialize in different aspects of the input patterns while maintaining a manageable parameter count. These heads feed a feedforward layer of dimension 30 with a dropout probability of 0.2. This process yields an enhanced feature tensor, $\mathbf{F}'' \in \mathbb{R}^{16\times 30}$, that leverages the self-similarities within our prior representation (see Fig. 4). Finally, we flattened the improved feature tensor $\mathbf{F}''$ to feed a linear layer producing an output dimension $\mathbf{h_{out}} = 2$ followed by a *softmax* function for binary classification for protective and non-protective behavior.

IV. Experiments and Results

A. EmoPain Dataset

The EmoPain dataset [19] contains sensor recordings from physical exercises performed by 25 healthy individuals and 22 CLBP patients. The mean age of the participants is 50.5 years, with 26 women and 21 men among the 47 subjects. The exercises were selected to mimic everyday movements such as "reaching forward", "bending down", "sit-to-stand", "stand-to-sit", and "one-leg-stand". These activities were executed at two difficulty levels, normal and difficult, with additional weights of 1 Kg and 2 Kg used for the "reaching forward" and "bending down" exercises to increase difficulty. For the "one-leg-stand", participants used both their preferred and non-preferred legs. Specific instructions were provided for "sit-to-stand" and "stand-to-sit" exercises but not for the others.

Data were recorded using 18 wearable inertial measurement units (IMUs) for motion capture and 4 sEMG sensors for muscle activation. The EmoPain dataset serves as the established benchmark in this domain, enabling direct comparison with state-of-the-art methods. Although it is a limited dataset, and additional data would be needed to improve generalization to a broader chronic low back pain (CLBP) population, we have carefully designed our approach to minimize potential overfitting (see Sections IV-C and IV-D).

B. Data preparation

Our data preprocessing strategy was meticulously designed to accommodate the pre-filtered nature of the EmoPain database, focusing on normalization and segmentation to maintain data integrity and fully leverage the end-to-end model's capabilities. Initially, Z-score normalization was applied to standardize the data across subjects, addressing variance discrepancies and ensuring uniformity in model input. Following normalization, we employed a sliding window approach, with a 75% overlap for segmenting the continuous MoCAP and

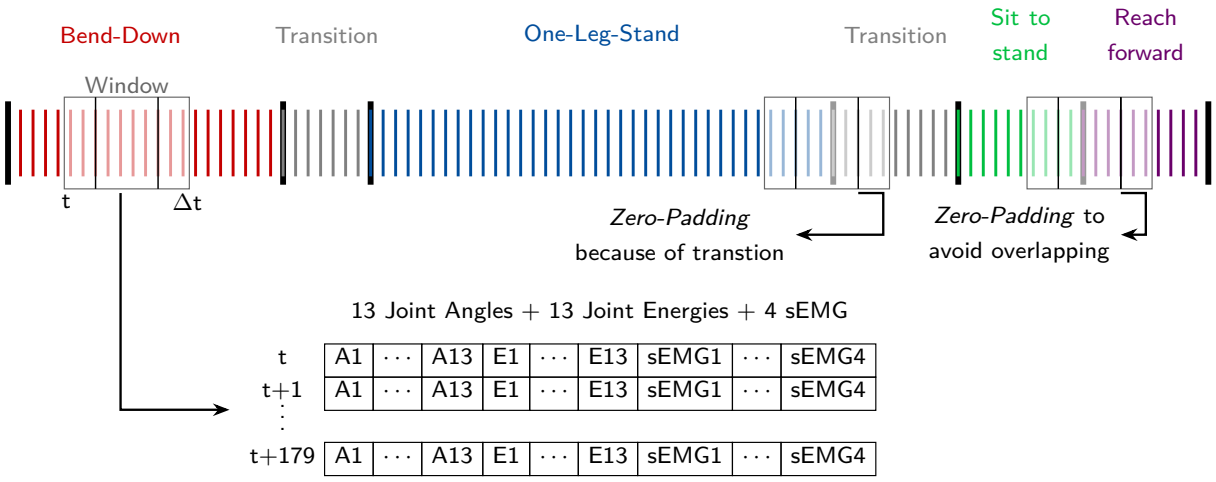


Fig. 5. The schematic of the sliding window segmentation technique for data preprocessing. Each window encapsulates a 3-second segment ($\Delta t = 3$ s) with a 75% overlap (hop size of 0.75 s). Within each window, 30 parameters are extracted per time step, encompassing 13 Joint Angles, 13 Joint Energies, and 4 sEMG values. Given the 60 Hz sampling rate, this translates to 180 distinct 30-dimensional vectors extracted per window. Zero-padding is applied if the 3-second window partially overlaps with a transition segment to ensure homogeneity within each window and avoid capturing transitions between activities. In this figure, we denote the transition activities as "Transition" segments.

sEMG data streams. This method segments the data into 3-second windows, equivalent to 180 timesteps at a 60 Hz sampling rate (see Fig. 5). Such segmentation facilitates a spatiotemporal representation of size 180×30, encompassing 13 Joint Angles, 13 Joint Energies, and 4 sEMG values at each timestep. To augment the dataset and strengthen the model's robustness against overfitting, we implemented random cropping with probabilities $p \in [0.05, 0.15]$ and jittering by adding Gaussian noise with standard deviations $\sigma \in [0.05, 0.15]$. Additionally, zero padding was used to ensure uniform window sizes across the dataset, particularly in scenarios where the sliding window could not capture 180 timesteps of data or when exercises overlapped. This comprehensive preprocessing approach, which follows that of Wang et al. [20], prepares the data and enhances the overall consistency of the dataset for the subsequent stages of model training and evaluation.

Following preprocessing and data augmentation, the initial dataset of 4,088 windows extracted from raw joint energies and angles (derived from IMU and sEMG recordings via a sliding window technique) was expanded to 28,616 instances (window-label pairs).

The EmoPain dataset poses several significant challenges, making it a valuable benchmark for human movement analysis and pain recognition. Its key characteristics are as follows: a small sample size, class imbalance, real-world multimodal nature (integrating sEMG and IMU data), and temporal dynamics (with continuous movement recordings and transitions between activities).

C. Evaluation Metrics

In line with established practices in analyzing the EmoPain dataset [21], [22], [24], we employ a multi-metric evaluation strategy to ensure a robust and comprehensive assessment of our model's performance, underlining its effectiveness in addressing the challenges posed by the EmoPain dataset. Specifically, we use the mean F1 score and the Matthews Correlation Coefficient (MCC) as primary metrics due to their effectiveness in handling class imbalances prevalent in this domain.

The mean F1 score, represented as Fm, harmonizes the balance between precision and recall, providing a holistic view of model accuracy across both positive and negative classes. It is computed using Eq. 5.

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP+FP}, \\ \text{Recall} &= \frac{TP}{TP+FN}, \\ \mathrm{Fm} &= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned} \tag{5}$$

The Matthews Correlation Coefficient (MCC), as defined in Eq. 6, offers a comprehensive measure that accounts for True and False Positives and Negatives (i.e., TP, FP, TN, and FN), making it particularly suitable for datasets with pronounced class imbalances.

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{6}$$

Both Fm and MCC require the selection of a specific threshold, which defines the operating point for these metrics. To mitigate potential bias from this selection and provide an evaluation independent of operational choices, we also include the Area Under the Precision-Recall Curve (AUC-PR) in our assessment framework. To mitigate overfitting, we carefully structured our experimental protocol using the "Leave-One-Subject-Out" (LOSO) cross-validation strategy [35]. This well-established method ensures that the model's performance is validated across unique test sets without overlap, thus minimizing any overfitting risk and providing a robust evaluation of generalization.

D. Implementation details

The experiments were conducted on a computer with an Intel Core i7 11700 CPU, 32GB of RAM, and a GeForce RTX3060 GPU. The PyTorch version 1.13.1 served as the deep learning framework. The proposed network was trained using the Adam optimizer [36] due to its efficient handling of sparse and noisy gradients with a fixed learning rate of $\eta = 0.001$ and a mini-batch size 40. Given the inherent class imbalance in protective behavior detection, we applied a weighted cross-entropy loss function, assigning each class a weight proportional to the inverse of its representation in the dataset. This approach mitigates the impact of imbalances by increasing the importance of underrepresented classes. The success of this strategy is evident in our balanced performance across evaluation metrics (see Section IV-E), particularly the Matthews Correlation Coefficient, which is robust to class imbalance. Table I summarizes the architecture details.

TABLE I
L-SFAN ARCHITECTURE DETAILS

Layer Type	Output Shape	Kernel Size	Stride	Activation
Input	180×30	-	-	-
Conv Block 1	$4 \times 180 \times 30$	3×3	1×1	ReLU
Max Pooling	$4 \times 86 \times 30$	2×1	2×1	-
Conv Block 2	$16 \times 86 \times 30$	3×3	1×1	ReLU
Max Pooling	$16 \times 42 \times 30$	2×1	2×1	-
TAP	16×30	-	-	-
Multi-Head Self-Attention	$5 \times 16 \times 30$	-	-	-
Feed-forward layer	16×30	-	-	-
Flatten	480	-	-	-
Linear	2	-	-	Softmax

E. Experimental results

In this section, we commence with an ablation study to evaluate the contributions of the key components of our proposed model. Firstly, we demonstrate the substantial performance improvement resulting from the use of temporal average pooling, highlighting the significance of spatial patterns in this context. Subsequently, we explore the contribution of the multi-head attention module. Finally, we thoroughly compare our proposed L-SFAN architecture and established state-of-the-art methods for protective behavior detection from multivariate time series data. This comparison serves to underscore the advantages of our approach and illustrate its potential impact.

1) Relevance of the Spatial and Temporal Patterns: The proposed CNN-based feature extractor produces a $16 \times 42 \times 30$ representation, where 16 denotes the number of feature maps, 42 indicates the temporal resolution, and 30 is the spatial resolution. This representation is then passed through a pooling layer, which performs aggregation to prepare the features for classification. The pooling operation can be performed over the temporal, spatial, or both dimensions, resulting in what we refer to as temporal average pooling (TAP), spatial average pooling (SAP), or spatio-temporal average pooling (STAP).

To determine the most effective approach for integrating information for protective behavior detection, we conducted an ablation study comparing the performance of the feature extractor across temporal, spatial, and spatio-temporal pooling methods. Table II summarizes the performance comparison of the 2D CNN architecture with different pooling strategies. The other aggregation methods included in this analysis are described next.

- **GAP** (Global Average Pooling): This method averages the feature map $\mathbf{F}$ across both the temporal and spatial dimensions, resulting in a single scalar vector $\mathbf{F}' \in \mathbb{R}^{16}$.
- **SAP** (Spatial Average Pooling): Here, $\mathbf{F}$ is averaged only along the spatial dimension, leading to a new feature map $\mathbf{F}' \in \mathbb{R}^{16\times 42}$ that retains the temporal resolution.
- **TAP** (Temporal Average Pooling): This method averages $\mathbf{F}$ across the temporal dimension, producing $\mathbf{F}' \in \mathbb{R}^{16\times 30}$ that preserves the spatial resolution.
- **STAP** (Spatio-Temporal Average Pooling): This approach combines the features obtained from both TAP and SAP. The resulting tensor is a concatenation of $\mathbf{F'}_{TAP} \in \mathbb{R}^{16\times 30}$ and $\mathbf{F'}_{SAP} \in \mathbb{R}^{16\times 42}$, resulting in $\mathbf{F}' \in \mathbb{R}^{16\times 30,16\times 42}$.

TABLE II
PERFORMANCE COMPARISON OF PROPOSED CNN WITH DIFFERENT AVERAGE POOLING STRATEGIES.

Feature Extractor	AUC	Fm	MCC
2D-CNN	0.676	0.673	0.257
2D-CNN-GAP	0.645	0.673	0.169
2D-CNN-SAP	0.699	0.688	0.299
2D-CNN-TAP	**0.819**	**0.752**	**0.481**
2D-CNN-STAP	0.786	0.725	0.414

As shown in Table II, emphasizing spatial information through temporal average pooling (2D-CNN-TAP) resulted in the highest performance, achieving an Area Under the Curve (AUC) of 0.819, F-measure of 0.752, and Matthews Correlation Coefficient (MCC) of 0.481. This observation aligns with the findings by Wang et al. [17], where incorporating a spatial attention module yielded superior results compared to a temporal attention module in their work.

Combining both spatial and temporal average pooling (2D-CNN-STAP) achieved the second-best performance. We hypothesize that this result primarily stems from including spatial information (2D-CNN-TAP) rather than their combined effect. This hypothesis is supported by the comparatively lower performance obtained when solely focusing on temporal patterns through spatial average pooling (2D-CNN-SAP).

Furthermore, the Grad-CAM visualization depicted in Fig. 6 illustrates the significance of the input data elements towards the model's output for a specific patient selected for illustration purposes. This visualization confirms the improved representation capabilities of 2D-CNN-TAP compared to 2D-CNN-SAP, as evidenced by the enhanced significance of the input data elements, which is quite compelling.

The L-SFAN architecture demonstrates competitive performance compared to state-of-the-art methods while maintaining a lightweight design. The ablation study highlights the importance of spatial patterns and the self-attention module in achieving optimal performance. However, the reliance on a single dataset and the potential limitations in generalizability should be considered when interpreting these results.

2) Multi-head Attention Module: To assess the contribution of the multi-head self-attention module, we incorporated it

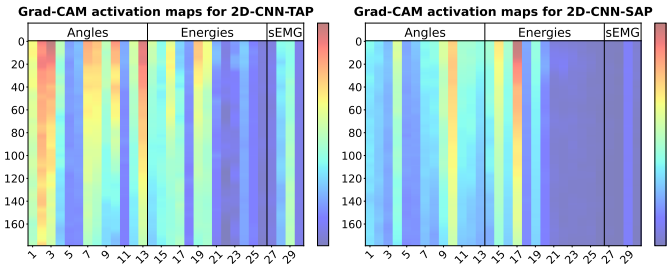


Fig. 6. The heatmap depicts the activation levels across the input data (180×30), with varying intensities indicating the significance of each input element towards the model's output. The graph on the left shows the activations for 2D-CNN-TAP, while the one on the right corresponds to 2D-CNN-SAP.

into the model using the best-performing feature extractor (i.e., 2D-CNN-TAP). The performance contribution of this subsystem is shown in Table III. As observed, the multi-head attention module led to a significant improvement in performance metrics.

TABLE III
PERFORMANCE COMPARISON OF 2D-CNN-TAP VS. L-SFAN, DEMONSTRATING THE IMPACT OF THE MULTI-HEAD ATTENTION MODULE.

Architecture	AUC	Fm	MCC	#Parameters
2D-CNN-TAP	0.819	0.752	0.481	2,582
L-SFAN	**0.849**	**0772**	**0.51**	8,282

In addition, Fig. 7 demonstrates the Grad-CAM [37] visualization for both architectures. In this case, we have used a time-aggregated version to focus solely on spatial patterns for the same patient. One can observe that the visualization is compelling, demonstrating how including the attention mechanism results in higher activation levels.

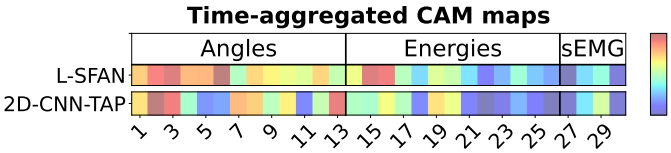


Fig. 7. Comparison of activation maps between L-SFAN (top) and 2D-CNN-TAP (bottom) architectures. The 2D maps were aggregated over time and normalized to highlight spatial patterns.

However, it is essential to acknowledge the trade-off involved. While the multi-head attention module demonstrably enhances performance, it comes at the cost of a nearly three-fold increase in trainable parameters, as shown in Table IV. This highlights the importance of considering both performance gains and computational complexity when selecting model components.

3) Comparison with the state-of-the-art: We conduct a comprehensive comparison of L-SFAN with state-of-the-art methods applied to the EmoPain dataset, encompassing various approaches such as LSTM-based models [38], [14], [20], graph-based neural networks [16], autoencoder-based methods [39], [40], [15], and attention mechanisms [17]. This rigorous evaluation aims to showcase L-SFAN's effectiveness and potential advantages in performance, efficiency, and interpretability. Table IV presents the comparative analysis results, highlighting key metrics and the number of parameters for each method.

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED MODEL WITH STATE-OF-THE-ART METHODS ON THE EMOPAIN DATASET.

Methods	AUC	Fm	MCC	#Parameters
State-of-the-art methods				
Stacked-LSTM [14]	0.622	0.662	0.192	25,154
Dual-Stream LSTM [20]	0.618	0.671	0.238	16,258
BANet [17]	0.663	0.668	0.247	2,131
MiMT [38]	0.648	0.722	0.327	1,038
LSTM+GCN [16]	0.690	0.731	0.365	93,414
Sparse VAE [39]	0.641	0.658	0.312	64,572
Gaussian VAE [40]	0.622	0.631	0.185	111,826
GRU RNN [15]	**0.855**	0.765	0.411	140,352
Proposed methods				
2D-CNN-TAP	0.819	0.752	0.481	2,582
L-SFAN	0.849	**0.772**	**0.51**	8,282

Table IV shows that the proposed L-SFAN architecture achieves the highest Matthews correlation coefficient (MCC) of 0.510 and F1 score (Fm) of 0.772 among all the compared methods. While the proposed model in [15] achieves slightly higher AUC (0.855), it comes at the expense of a significantly higher number of parameters (140,352 compared to 8,282 for L-SFAN). This highlights the efficiency and effectiveness of the proposed architecture, particularly in the context of a limited dataset where overfitting can be a significant concern. We attribute the superior performance of L-SFAN to two key factors:

1) **Effective Spatial Pattern Extraction:** The emphasis on spatial patterns through temporal average pooling (TAP), along with the synergistic integration of CNNs and self-attention, enables the model to efficiently capture the spatial relationships crucial to our task. To further illustrate the efficacy of the proposed model in preserving crucial spatial information for protective behavior detection, we used Grad-CAM [41] to visualize the activation levels across the 13 joint angles, 13 joint energies, and 4 sEMG channels averaged over the temporal dimension (see Fig. 6). As previously discussed, this analysis highlighted how the data representation generated by L-SFAN enhances the significance of the input data elements for the task, making it more effective.
2) **Reduced Model Complexity:** The lightweight design of the L-SFAN architecture, with significantly fewer trainable parameters compared to most state-of-the-art methods (except BANet [17] and MiMT [38]), substantially mitigates the risk of overfitting, especially when dealing with datasets like EmoPain that have a limited number of samples (see Fig. 8).

The model's inference time and memory consumption were evaluated on the specified platform (see Section. IV-D) to highlight its lightweight design. For reliable measurements, results were averaged over 10,000 runs. Results, including mean inference time $\pm$ standard deviation and memory usage, are provided in Table V.

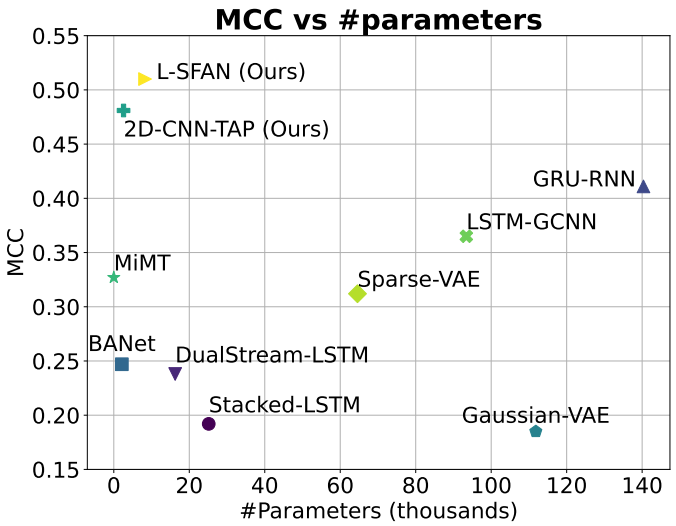


Fig. 8. Performance comparison of various state-of-the-art models in terms of MCC versus the number of trainable parameters.

TABLE V
COMPARISON OF THE COMPUTATIONAL RESOURCES USED DURING INFERENCE FOR L-SFAN AND 2D-CNN-TAP.

Architecture	Inference Time (ms)	Peak GPU Memory (MB)
L-SFAN	0.74 ± 0.03	16.28
2D-CNN-TAP	0.56 ± 0.02	16.23

V. Discussion and Conclusion

The L-SFAN architecture presents an efficient approach to protective behavior detection by combining 2D CNNs and self-attention mechanisms for multivariate time series analysis. Our comprehensive evaluation demonstrates that L-SFAN achieves competitive performance with an Area Under the Precision-Recall Curve of 84.9%, an F1-measure of 77.2%, and a Matthews Correlation Coefficient of 0.51 while requiring only 8,282 trainable parameters – a significant reduction compared to existing approaches that often exceed 100,000 parameters.

Our ablation studies revealed the significance of spatial pattern preservation through temporal average pooling in protective behavior detection. This analysis showed that spatial relationships between sensors provide more discriminative information than temporal evolution. We enhanced the architecture through self-attention, which further refined feature representations. While this addition moderately increased computational complexity, its benefits were clear. We validated these architectural choices through systematic experimentation and analysis.

The experimental evaluation followed rigorous protocols, employing Leave-One-Subject-Out cross-validation to ensure a robust assessment of generalization capabilities. The EmoPain dataset, while limited in size, provided a standardized benchmark that enabled direct comparison with state-of-the-art methods. Our data augmentation strategy and careful consideration of class imbalance through weighted loss functions helped address the dataset's inherent constraints.

The lightweight design and non-recursive nature of L-SFAN suggest potential advantages for resource-constrained environments. Our performance analysis showed an average inference time of 0.74 ± 0.03 ms and peak GPU memory usage of 16.28 MB, indicating promising efficiency characteristics. While these results suggest potential suitability for future development toward monitoring systems and wearable applications, significant research would be needed to validate such deployments.

This study focused on establishing the theoretical foundations and experimental validation of L-SFAN, demonstrating strong performance and computational efficiency compared to existing methods. While these results show promise for protective behavior detection, we acknowledge that bridging the gap between research findings and clinical applications requires substantial additional work. Future research could extend these findings by evaluating L-SFAN's performance on additional datasets and further investigating the relationship between spatial and temporal patterns. The demonstrated balance between performance and computational efficiency contributes meaningfully to the development of AI-based approaches in chronic pain research.

References

[1] R. J. Yong, P. M. Mullins, and N. Bhattacharyya, "Prevalence of chronic pain among adults in the United States," *PAIN*, vol. 163, no. 2, pp. e328–e332, 2022.

[2] R. Roy, S. Galán, E. Sánchez-Rodríguez, M. Racine, E. Solé, M. P. Jensen, and J. Miró, "Cross-National Trends of Chronic Back Pain in Adolescents: Results From the HBSC Study, 2001-2014," *The Journal of Pain*, vol. 23, no. 1, pp. 123–130, 2022.

[3] A. Fayaz, P. Croft, R. M. Langford, L. J. Donaldson, and G. T. Jones, "Prevalence of chronic pain in the UK: a systematic review and meta-analysis of population studies," *BMJ Open*, vol. 6, no. 6, p. e010364, 2016.

[4] A. Singh, N. Bianchi-Berthouze, and A. C. Williams, "Supporting everyday function in chronic pain using wearable technology," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI'17. Association for Computing Machinery, 2017, pp. 3903—-3915.

[5] C. K. B. Paul J. Watson and C. J. Main, "Evidence for the role of psychological factors in abnormal paraspinal activity in patients with chronic low back pain," *Journal of Musculoskeletal Pain*, vol. 5, no. 4, pp. 41–56, 1997.

[6] R. R. Austin, O. Ang, A. Haley, L. Hanson, D. Kennedy, H. Mendenhall, C. Schulz, D. Thorpe, and R. Evans, "Examining resilient pain behaviors for chronic low back pain: A scoping review," *Pain Management Nursing*, vol. 25, no. 4, pp. 417–424, 2024.

[7] F. Hens, M. M. Dehshibi, L. Bagheriye, M. Shahsavari, and A. Tajadura-Jiménez, "Stal: Spike threshold adaptive learning encoder for classification of pain-related biosignal data," 2024. [Online]. Available: https://arxiv.org/abs/2407.08362

[8] X. Xi, W. Jiang, X. Hua, H. Wang, C. Yang, Y.-B. Zhao, S. M. Miran, and Z. Luo, "Simultaneous and Continuous Estimation of Joint Angles Based on Surface Electromyography State-Space Model," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 8089–8099, 2021.

[9] B. Leelakittisin, M. Trakulruangroj, S. Sangnark, T. Wilaiprasitporn, and T. Sudhawiyangkul, "Enhanced Lightweight CNN Using Joint Classification With Averaging Probability for sEMG-Based Subject-Independent Hand Gesture Recognition," *IEEE Sensors Journal*, vol. 23, no. 17, pp. 20 348–20 356, 2023.

[10] S. D. Subramaniam and B. Dass, "Automated nociceptive pain assessment using physiological signals and a hybrid deep learning network," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3335–3343, 2021.

[11] J. W. Vlaeyen and S. J. Linton, "Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art," *Pain*, vol. 85, no. 3, pp. 317–332, 2000.

[12] M. Ashtari-Majlan, A. Seifi, and M. M. Dehshibi, "A Multi-Stream Convolutional Neural Network for Classification of Progressive MCI in Alzheimer's Disease Using Structural MRI Images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3918–3926, 2022.

[13] M. M. Dehshibi, B. Baiani, G. Pons, and D. Masip, "A Deep Multimodal Learning Approach to Perceive Basic Needs of Humans From Instagram Profile," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 944–956, 2023.

[14] C. Wang, T. A. Olugbade, A. Mathur, A. C. De C. Williams, N. D. Lane, and N. Bianchi-Berthouze, "Recurrent network based automatic detection of chronic pain protective behavior using mocap and semg data," in *Proceedings of the ACM International Symposium on Wearable Computers*. Association for Computing Machinery, 2019, pp. 225—-230.

[15] M. M. Dehshibi, T. Olugbade, F. Diaz-de Maria, N. Bianchi-Berthouze, and A. Tajadura-Jimenez, "Pain Level and Pain-Related Behaviour Classification Using GRU-Based Sparsely-Connected RNNs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 3, pp. 677–688, 2023.

[16] C. Wang, Y. Gao, A. Mathur, A. C. De C. Williams, N. D. Lane, and N. Bianchi-Berthouze, "Leveraging Activity Recognition to Enable Protective Behavior Detection in Continuous Data," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 2, 2021.

[17] C. Wang, M. Peng, T. A. Olugbade, N. D. Lane, A. C. D. C. Williams, and N. Bianchi-Berthouze, "Learning bodily and temporal attention in protective movement behavior detection," 2019.

[18] J. Castaño, S. Martínez-Fernández, X. Franch, and J. Bogner, "Exploring the Carbon Footprint of Hugging Face's ML Models: A Repository Mining Study," in *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2023, pp. 1–12.

[19] M. S. H. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A. C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. J. Watson, A. C. d. C. Williams, M. Pantic, and N. Bianchi-Berthouze, "The Automatic Detection of Chronic Pain-Related Expression: Requirements, Challenges and the Multimodal EmoPain Dataset," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 435–451, 2016.

[20] C. Wang, T. A. Olugbade, A. Mathur, A. C. D. C. Williams, N. D. Lane, and N. Bianchi-Berthouze, "Chronic pain protective behavior detection with deep learning," *ACM Transactions on Computing for Healthcare*, vol. 2, no. 3, pp. 1–24, 2021.

[21] Y. Li, S. Ghosh, and J. Joshi, "Plaan: Pain level assessment with anomaly-detection based network," *Journal on Multimodal User Interfaces*, vol. 15, no. 4, pp. 359–372, 2021.

[22] M. T. Uddin and S. Canavan, "Multimodal multilevel fusion for sequential protective behavior detection and pain estimation," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 844–848.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[24] K. N. Phan, N. K. Iyortsuun, S. Pant, H.-J. Yang, and S.-H. Kim, "Pain recognition with physiological signals using multi-level context information," *IEEE Access*, vol. 11, pp. 20 114–20 127, 2023.

[25] M. T. Uddin, G. Zamzmi, and S. Canavan, "Cooperative Learning for Personalized Context-Aware Pain Assessment From Wearable Data," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 11, pp. 5260–5271, 2023.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012, pp. 1–9.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 3431–3440.

[28] M. M. Dehshibi and D. Masip, "BEE-NET: A deep neural network to identify in-the-wild Bodily Expression of Emotions," pp. 1–10, 2024.

[29] M. M. Dehshibi, M. Ashtari-Majlan, G. Adhane, and D. Masip, "ADVISE: ADaptive feature relevance and VISual Explanations for convolutional neural networks," *The Visual Computer*, pp. 1–13, 2023.

[30] M. Ashtari-Majlan, M. M. Dehshibi, and D. Masip, "Deep Learning and Computer Vision for Glaucoma Detection: A Review," pp. 1–20, 2023.

[31] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162–169, 2017.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. PMLR, 2015, pp. 448–456.

[33] H. Wang and M. Tu, "Enhancing attention models via multi-head collaboration," in *2020 International Conference on Asian Language Processing (IALP)*. IEEE, 2020, pp. 19–23.

[34] J. Li, X. Wang, Z. Tu, and M. R. Lyu, "On the diversity of multi-head attention," *Neurocomputing*, vol. 454, pp. 14–24, 2021.

[35] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognition*, vol. 48, no. 9, pp. 2839–2846, 2015.

[36] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015, pp. 1–15.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[38] T. Olugbade, N. Gold, A. C. d. C. Williams, and N. Bianchi-Berthouze, "A Movement in Multiple Time Neural Network for Automatic Detection of Pain Behaviour," in *International Conference on Multimodal Interaction*. Association for Computing Machinery, 2021, pp. 442—-445.

[39] L. Antelmi, N. Ayache, P. Robert, and M. Lorenzi, "Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data," in *Proceedings of the 36th International Conference on Machine Learning*. Proceedings of Machine Learning Research, 2019, pp. 302–311.

[40] Guo, Yifan and Liao, Weixian and Wang, Qianlong and Yu, Lixing and Ji, Tianxi and Li, Pan, "Multidimensional Time Series Anomaly Detection: A GRU-based Gaussian Mixture Variational Autoencoder Approach," in *Proceedings of The 10th Asian Conference on Machine Learning*. Proceedings of Machine Learning Research, 2018, pp. 97–112.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 618–626.