Multimodal automatic acute pain recognition using facial expressions and physiological signals

Jaleh Farmani[1], Alessandro Giuseppi[1], Ghazal Bargshady[2], and Raul Fernandez Rojas[2]*

[1] Dept. of Computer, Control and Management Engineering, University of Rome "La Sapienza", Rome, Italy
[2] Human-Centred Technology Research Centre, Faculty of Science and Technology, University of Canberra, Canberra, ACT, Australia

Abstract. Accurate and objective pain assessment is crucial for effective pain management. This paper proposes a novel multimodal deep learning framework for automatic pain detection using a hybrid architecture with feature-level fusion. The framework leverages multimodal data including facial expressions and physiological signals (EDA and ECG) from the BioVid Heat Pain database (Part A). The novel hybrid architecture consists of two streams as stream 1 employs an attention-based CNN-LSTM to extract features from facial expressions videos, capture temporal dependencies, and focus on relevant aspects of the video data, and stream 2 with an LSTM to capture temporal patterns in the physiological signals. The performance of the proposed model was examined in both unimodal and multimodal settings. In a binary classification task distinguishing No Pain from Severe Pain, electrodermal activity (EDA) outperformed all other single data sources, achieving high average accuracy (83.05% for 67 subjects and 82.69% for 87 subjects) and F1-scores (81.66 and 80.18, respectively) using k-fold cross-validation. Additionally, the multimodal setting (Video + EDA) achieved higher accuracy (84.15% for 67 subjects and 83.35% for 87 subjects) and F1-scores (82.86 and 82.36, respectively).

Keywords: Hybrid deep learning · Pain recognition · Multimodal Analysis · Facial expression · Physiological signals.

1 Introduction

Pain, a common experience yet a complex phenomenon, remains an active area of scientific investigation. The International Association for the Study of Pain (IASP) offers a foundational definition: "*an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage*" [13]. Pain can be broadly categorized into two main types based on its duration and characteristics according to [21]: acute and chronic pain. Acute pain, a sudden warning triggered by injury, typically heals within

* Corresponding author

weeks. Chronic pain, on the other hand, lasts more than twelve weeks, can arise from various conditions and can significantly impact an individual's quality of life.

Traditional pain assessment and detection methods often rely on self-reported measures, such as the numerical rating scale (NRS) or the visual analogue scale (VAS) [30], which on one hand can be prone to language barriers and limitations in communication, particularly for infants, patients under sedation, anesthesia, recovering from a stroke and those with cognitive impairments or advanced dementia. For these patients, caregivers rely on tools like the Behavioral Pain Scale (BPS) [16]. However, continuous monitoring with these tools can strain medical staff. Automatic pain assessment by providing more accurate, regular, and real-time options facilitate clinicians and patients' caregivers. It also can bring economical benefits for saving time of nurses to focus on more critical tasks. In contrast to traditional assessment tools, automatic systems may be more objective than a human observer and more sensitive to slight changes in pain levels than common manual annotation by humans [24].

On top of these, to have an accurate, automatic, and continuous pain detection system, utilising multiple sensing modalities offers significant advantages compared to relying on a single sensor. First of all, using multimodal approaches have the potential to achieve more accurate pain assessment [5, 8, 27]. Unimodal approaches moreover can be prone to several limitations, such as sensor failure, data quality issues or limited sensitivity [5].

This research presents the following contributions: 1) a novel hybrid framework combining CNN, LSTM, and attention blocks for robust pain recognition; 2) a multimodal approach with two streams—an attention-based CNN-LSTM for extracting spatio-temporal features from facial expression videos, and an LSTM for capturing temporal patterns in physiological signals (EDA and ECG); and 3) improved performance over previous studies, without relying on manual feature extraction, with potential applications in clinical settings to recognise pain in real time.

2 Related Work

Several studies (e.g., Werner *et al.* [27]) have explored facial expressions as a valuable source of information for automatic pain recognition, highlighting their potential as a clinical tool. Focusing on the dynamic nature of pain, Werner *et al.* [23] proposed facial activity descriptors to capture the dynamics of facial expressions during pain. In a follow up study [26], they explored the role of individual variability and facial expressiveness. Their results demonstrated that facial responses were less frequent during low-intensity pain and that their measures effectively identified highly expressive individuals (easier to classify for pain recognition) and those who might be less expressive (potentially leading to misclassified pain levels). This research highlights the importance of considering individual differences in facial expressiveness for developing robust pain recognition systems. Zhi and Wan [31] proposed Sparse Long Short-Term Mem-

ory Networks (SLSTM). Their approach incorporates a sparse representation technique within the LSTM network structure, aiming to reduce computational complexity for sequence processing. Ayral *et al.* [2] also proposed a method for efficient training of 3D CNNs, addressing the high computational cost. Their approach utilises Softmax Temporal Pooling to identify and focus on the most informative video segments during training. Othman *et al.* [14] addressed limited model generalizability in pain recognition by proposing cross-database validation with facial expressions. Their method leveraged temporal information by merging features from three video frames into a single image for classification using a transferred model.

Patania *et al.* [15] proposed a novel approach using deep Graph Neural Networks (GNNs) and dense maps. Unlike methods that focus on individual features, their approach is to analyze the relationships between facial landmark points tracked on videos.

In addition to facial expressions, researchers are exploring physiological signals for pain assessment. Electrodermal Activity (EDA), a measure of skin conductance, shows promise due to its sensitivity to pain. Studies by Thiam *et al.* [19] and Lopez-Martinez *et al.* [12] found EDA outperformed other modalities like ECG and Heart Rate Variability (HRV) in pain classification and intensity prediction. Subramaniam and Dass [18] also achieved higher accuracy with EDA compared to ECG in their deep learning approach.

Numerous studies support the idea that combining multiple sensing technologies can improve automated pain assessment [5, 8, 27]. This is because pain is a multidimensional experience, and different modalities capture complementary information. Pioneering work by Werner *et al.* [25] demonstrated the effectiveness of combining video data with physiological signals (EDA, ECG, EMG) for pain detection. Subsequent research by Kächele *et al.* [9] and Lopez-Martinez *et al.* [11] further emphasized the benefits of multimodal fusion.

Recent advancements explore deep learning and address practical challenges. For instance, Kasaeyan Naeini *et al.* focused on energy-efficient pain recognition on wearable devices. Moreover, Gkikas *et al.* [8] introduced an efficient framework for pain assessment using video and heart rate data, prioritizing efficiency for wearable devices. These advancements showcase the continuous progress in developing robust and practical methods for automated pain recognition.

3 Methods

In this section, we will discuss about the employed pain database, the preprocessing steps applied to videos and signals and our hybrid deep learning model.

3.1 BioVid Heat Pain Database

All the experiments are conducted utilising the publicly accessible *BioVid Heat Pain Database* (Part A) [22]. It contains data collected from 90 volunteers across three age groups (18-35 years, 36-50 years and 51-65 years) . Pain was induced

using a thermode on the right arm at four different levels specific to each participant. Researchers recorded frontal head videos and physiological signals like electrodermal activity (EDA), electrocardiogram (ECG), and trapezius electromyography (EMG) during the experiment. Each pain level has 20 samples per subject, captured within a 5.5-second window with a 3-second delay after the pain stimulus began.

Despite its popularity, the BVDB has some limitations. There is inconsistency in the subject subsets researchers use in their experiments. The results obtained by the creators of the database in [26] show that some subjects do not react visibly to the stimulus and one of the main reason stated is that those subjects have (intentionally or unintentionally) reported their pain tolerance lower than what they could actually tolerate. Following these results, some studies exclude 20 participants who showed minimal reactions to pain [18]. While some other researchers have used the entire dataset. Furthermore, facial expressions in response to the highest pain level starts after 2 seconds [26], which leads to using different time window lengths when analyzing the data. For example Thiam *et al.* in [19] and [20] utilised a different segmentation by truncating the original time frame by 1 second, employing only 4.5 seconds of the samples with a shift from the elicitations' onset of 4 instead of 3 s. These challenges complicate the comparison of performances.

To mitigate these challenges and improve the generalisability of findings, this study conducts two experiments. The first experiment utilises the complete dataset with all 87 subjects and the full 5.5-second time windows. The second experiment uses a reduced dataset by removing 20 participants with minimal expressive responses and truncating the initial 2 seconds of each sample, resulting in a 3.5-second window.

3.2 Pre-processing

The pre-processing steps that have been used for the videos in our experiments are depicted in Fig.1. First of all, frames are extracted from each video sample. Then we detect and extract faces using RetinaFace [4] provided by Serengil and Ozpinar in [17]. After that images are resized to (227×227) to ensure consistency in input size for our model architecture. Finally, we normalize pixel intensities to be scaled between 0 and 1.

Moreover, for the experiments we used the preprocessed signals available in the dataset. The EDA signal has been filtered with a low-pass Butterworth filter, and ECG signal with a Butterworth band-pass filter (0.1-250 Hz). Furthermore, prior to the classification experiments, the sampling rate of the recorded physiological modalities was reduced to 25 Hz, to ensure that corresponding sample point and the frames across different modalities represent the same moment in time. After reducing the sampling rate, the data was pre-processed using min-max normalization.

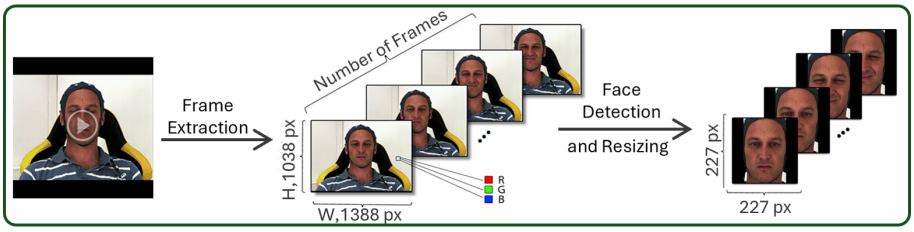

Fig. 1. Pre-processing steps applied to one video sample from the dataset.

3.3 Proposed hybrid deep learning model

The block-diagram of the proposed multimodal system is depicted in Figure 2. As can be seen, the network is composed of two streams. Stream 1 (indicated as Visual Branch) consists of three sub-networks (Spatial, Temporal and Attention) and stream 2 (indicated as Physiological Branch) consists of a an LSTM network.To optimize each sub-network and ensure that their combination improves the performance of the system, we trained each sub-network individually. In the following sections, we will discuss about each sub-network and the implementation details used in our experiments.

Visual Branch:

Spatial sub-network: After pre-processing the frames, we utilise a CNN to perform the task of pain recognition, focusing on extracting spatial features from each frame within a video sequence. To initialize the parameters of our spatial sub-network, we utilise a pre-trained model from Albanie and Vedaldi [1], originally trained on the FER2013 database. All the layers of the model have been frozen except the last two fully connected layers which have been fine tuned. The model's architecture is shown in Table 1. After training this sub-network, we utilise it as a feature extractor. Specifically, we extract features from the FC-7 layer for subsequent stages of processing.

Temporal sub-network: After extracting the spatial features from frames, we group them into D segments of length T, where D will be the total number of videos such that $D = \frac{N}{T}$. We feed these sequences into a 2-layer LSTM architecture. LSTM is a variant of the recurrent neural network (RNN) architecture designed to capture long-term dependencies in sequential data. In our framework, the LSTM layers are responsible for capturing the temporal evolution of pain expression. Once the sequences have been processed by the LSTM layers, we obtain the hidden states of the second LSTM layer. These final hidden states encapsulate the learned spatio-temporal representations of the entire video and are subsequently passed to the attention mechanism block.

Table 1. Architecture and specifications of the CNN model.

Layer	Filter Size	Stride	Padding	Size	Activation
Input	-	-	-	$227 \times 227 \times 3$	-
Conv2d-1 (CONV2D)	11×11	4	-	$55 \times 55 \times 96$	ReLU
MaxPool2d-1 (MaxPooling)	3×3	2	-	$27 \times 27 \times 96$	-
Conv2d-2 (CONV2D)	5×5	1	2	$27 \times 27 \times 256$	ReLU
MaxPool2d-2 (MaxPooling)	3×3	2	-	$13 \times 13 \times 256$	-
Conv2d-3 (CONV2D)	3×3	1	1	$13 \times 13 \times 384$	ReLU
Conv2d-4 (CONV2D)	3×3	1	1	$13 \times 13 \times 384$	ReLU
Conv2d-5 (CONV2D)	3×3	1	1	$13 \times 13 \times 256$	ReLU
MaxPool2d-5 (MaxPooling)	3×3	2	-	$6 \times 6 \times 256$	-
FC-6 (Fully Connected)	-	-	-	4096	ReLU
FC-7 (Fully Connected)	-	-	-	4096	ReLU
FC-8 (Fully Connected)	-	-	-	2	Softmax

Attention sub-network: In this work, we used the attention mechanism proposed in [28], with different activation function. In neural networks, especially in the context of sequence modeling tasks, the attention mechanism allows the model to focus on different parts of the input sequence with varying degrees of attention. This is particularly useful when dealing with long sequences, where certain parts of the input may be more relevant than others.

In the first step, the Attention sub-network calculates scores (a_t) for each hidden state $H_t^{(2)}$ in the sequence. The scores are then passed through softmax function to compute attention weights. Then, we compute the context vector (attention vector). This is the output of the visual branch and is computed as a weighted sum of the LSTM output based on the attention weights:

$$V_k = \sum_{t=1}^{T} \alpha_t H_t^{(2)} \tag{1}$$

Where T is the number of frames in each video sample and h is the LSTM hidden size. The output of this layer (V_k) is passed to the fusion network to form our multimodal architecture.

Physiological Branch: The signals are first preprocessed and segmented into sequences, with each sequence containing consecutive data points from the signals P_k. These sequences are then fed into a separate LSTM network, which learns the temporal patterns and dependencies present in the physiological data. After processing the sequences through the LSTM layers, the final hidden state of the second LSTM layer is obtained which is denoted as $S_k \in \mathbb{R}^{T \times h}$, where T is the number of sample points in each sequence and h is the LSTM hidden size.

Modality Fusion: To achieve robust pain classification, the network employs multimodal fusion in its final layer. This fusion combines the attention-based

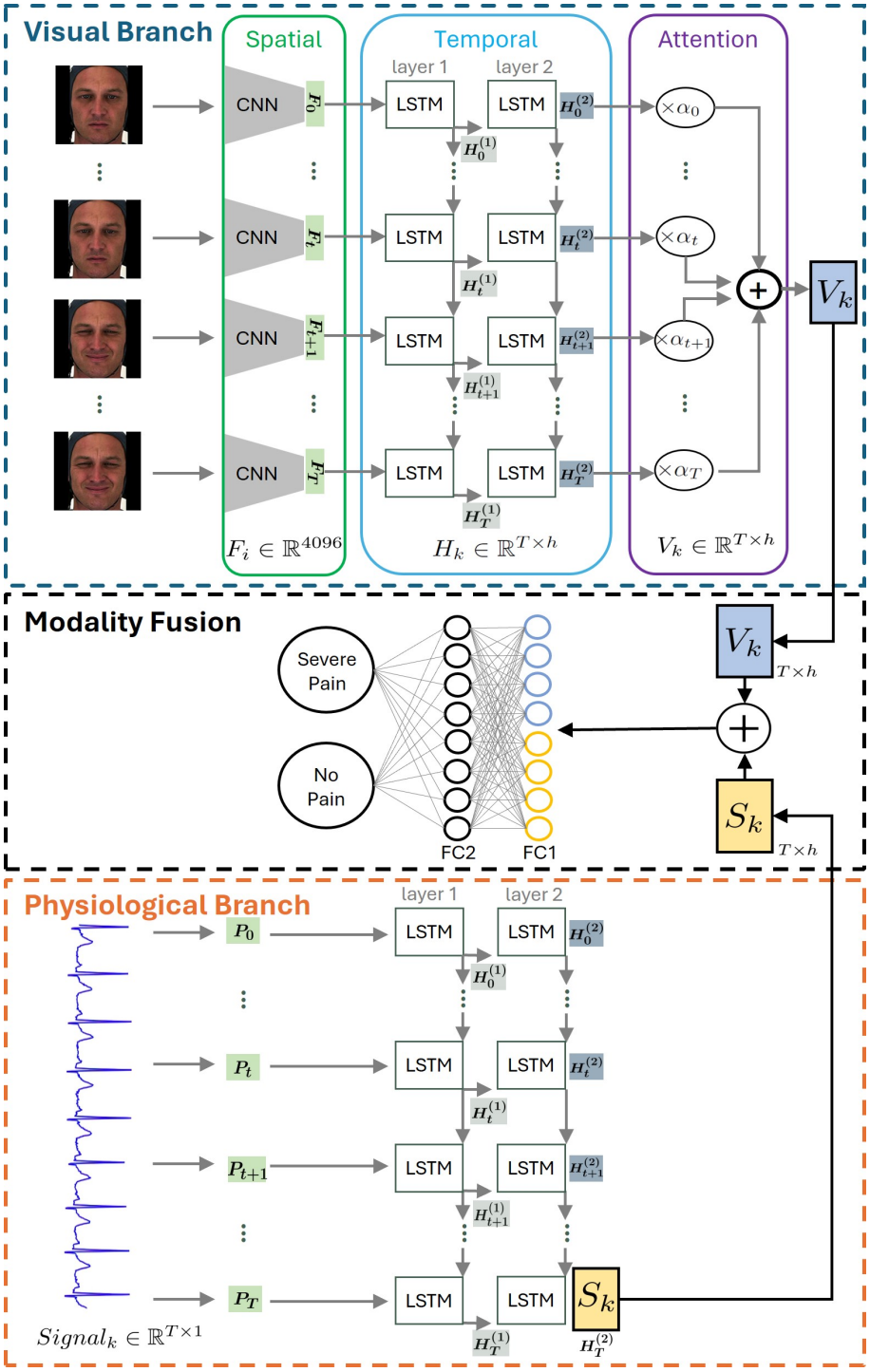

Fig. 2. Overview of the framework architecture

spatio-temporal features extracted from the video data stream (denoted as V_k) with the features extracted from the physiological signal stream (denoted as S_k). As discussed in Section 3.2, to ensure temporal alignment between video frames and physiological data, we downsampled the physiological signals to a sampling rate of 25 Hz. This sampling rate aligns with the frame rate of 25 frames per second (fps) of the video samples. By matching this rate, we guarantee that features extracted from each modality correspond to the same time window. The resulting fused feature vector is then fed into a fully-connected (FC) layer with a ReLU activation function. Finally, a second FC layer with a softmax activation function performs the final pain classification task.

3.4 Experimental Setup and Training Process

All experiments were conducted using an NVIDIA GeForce RTX 4080 GPU (Laptop variant) with 7424 CUDA Cores and a 13th Gen Intel(R) Core(TM) i7-13650HX CPU, supported by 16 GB of RAM. The experiments were implemented in Python 3.8.8 and PyTorch 2.1.2.

The model consists of a Visual branch, which utilizes a pre-trained CNN for feature extraction, followed by a two-layer LSTM and an attention block, and a Physiological branch that processes EDA and ECG signals using a separate two-layer LSTM. The total model size is approximately 50 million parameters. We employed a multi-step training approach, similar to the one adopted in [7], where each subnetwork (CNN, LSTM, attention) was trained individually, with the features extracted by each subnetwork passed sequentially to the next. This approach improves performance and scalability, allowing new modalities or network types to be integrated without the need to retrain the entire network, as required in end-to-end training. It also allows each subnetwork to be optimized using different hyperparameters according to the data type and model.

For training the CNN, we used 30 epochs, a batch size of 64, and a learning rate of 1e-5, with the Adam optimizer and a MultiStepLR scheduler, with a milestone at 20 epochs and a decay rate of 0.1. The LSTM in the Visual branch was trained with two layers, 64 hidden units, sequence lengths of 138 for the full dataset and 88 for the subset, over 30 epochs, with a learning rate of 1e-4 and the Adam optimizer. The Physiological branch's LSTM was trained with two layers, 64 hidden units, and sequence lengths of 138 and 88, over 800 epochs, a batch size of 64, and a learning rate of 1e-3.

3.5 Evaluation

To evaluate the effectiveness of our approach, standard evaluation metrics were employed, including accuracy, F1-score, sensitivity, and specificity. Accuracy measures the overall correctness of the model's predictions, F1-score balances precision and recall, providing a harmonic mean useful for imbalanced datasets, sensitivity (also known as recall) measures the model's ability to correctly identify positive instances, and specificity evaluates the ability to correctly identify

negative instances. All experiments were conducted using 10-fold cross-validation to ensure robustness and generalisability.

4 Experimental Results

The pain detection experiments were conducted in binary classification which distinguishes between No Pain (BLN) and Severe Pain (PA4) utilising the video data, EDA and ECG of the *BioVid Heat Pain Database*(BVDB)(Part A) [22]. All the sub-networks and the multimodal framework are evaluated in 10-fold cross-validation. We have applied our proposed method on two different scenarios: 1) All 87 subjects were included, with a time window of 5.5 seconds, to allow for comparison with previous studies based on the full dataset. In total 480,240 frames and sample points and 3480 videos have been utilised. 2) Additionally, we report the performance of 67 subjects, with a time window of 3.5 seconds of the data. While there are fewer performance reports on this subset, we considered it important to include for comprehensive analysis. In this scenario, 235,840 frames and sample points and 2680 videos have been used.

The average performance of the proposed model in these two scenarios is presented in Table 2. Furthermore, Figure 3 depicts the average accuracy and the standard deviation achieved using different features from each modality and their combinations, evaluated through 10-fold cross-validation.

Table 2. The average performance of the proposed hybrid deep learning model on the **all available data (1)** and **truncated data (2)** of the BioVid database (Part A) in 10-fold cross-validation, classifying No Pain versus Severe Pain.

Type	**Modality**	**Accuracy (%)**		**F1-score (%)**		**Sensitivity (%)**		**Specificity (%)**	
		(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Vision-based	Frame level	51.44	62.24	49.74	60.73	40.23	51.71	63.92	71.71
	Video	**64.46**	**74.21**	**67.82**	**62.18**	**57.04**	**54.94**	**69.14**	**78.95**
Physiological-based	ECG	64.86	67.69	57.42	62.03	52.08	55.83	78.65	77.78
	EDA	**82.69**	**83.05**	**81.66**	**79.18**	**78.44**	**78.89**	**86.95**	**79.85**
Multimodal	EDA + Video	**83.35**	84.15	**82.36**	82.86	**80.15**	79.28	86.13	88.02
	ECG + Video	65.68	71.71	61.79	71.07	57.86	70.77	73.50	72.65
	EDA + ECG	81.08	**84.38**	79.13	**83.60**	73.16	**80.66**	**88.98**	87.97
	EDA + ECG + Video	81.43	84.04	79.62	82.97	76.68	78.28	84.18	**89.80**

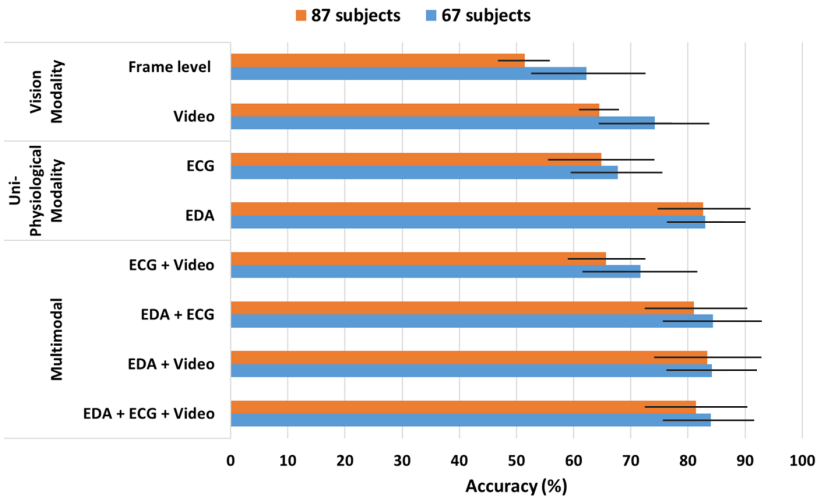

Fig. 3. Average accuracy obtained with different features from each modality and in combination evaluated in 10-fold cross-validation and classifying No Pain versus Severe Pain.

Our initial experiment explored frame-level classification using CNNs. However, this approach (scenario 1) did not yield the expected results. The CNN models, including AlexNet, VGG16, and ResNet50, all failed to achieve accuracy exceeding chance level, regardless of whether trained from scratch or using pretrained models. This suggests the models may not have been able to capture the features of pain expression present in individual frames. In contrast, scenario 2 showed an approximately 11% improvement in performance simply by removing less expressive subjects and using half of the data. This could imply that the entire dataset may not be suitable for frame-based classification with simple CNNs. This result is also concluded in [3, 29]. However, by utilising the temporal information and the attention block, the model achieved a substantial improvement over frame-level classification (64.46% vs. 51.44% accuracy for all data, 74.21% vs. 62.24% accuracy for truncated data). This suggests that capturing temporal dynamics of facial expressions is crucial for pain detection. Among unimodal approaches, EDA achieved the highest accuracy (82.69% for all data, 83.05% for truncated data). This indicates that EDA provides valuable insights into pain state.

Combining modalities, particularly EDA and video, achieved the best overall performance (83.35% and 84.15% accuracy for all and truncated data, respectively). This combination not only achieved high accuracy but also balanced sensitivity and specificity. For example, EDA + Video in the all data scenario achieved a sensitivity of around 80% and a specificity of nearly 86%, demon-

strating a good balance between identifying those in pain and avoiding false positives.

4.1 Comparison of the performance with existing methods

In this section, we compare the results of our method with state-of-the-art studies using Part A of the BioVid dataset, evaluated with 67 and 87 subjects. Due to resource limitations, it was not possible to follow the leave-one-subject-out evaluation protocol used by most recent studies. Instead, we focused on comparisons with studies that used k-fold cross-validation, a well-established method that ensures statistical validity and generalizability. This allows for a fair assessment of our model's performance against both unimodal and multimodal approaches.

As shown in Table 3, our approach outperforms existing unimodal studies based on physiological signals and achieves the highest performance in the multimodal setting when combining features from EDA and video data (by accuracy 83.35%).

Table 3. Comparison of studies utilising BioVid in unimodal and multimodal settings and k-Fold Cross-Validation

Ref	**Modality**	**Validation Method**	Method	**67 subjects**	**87 subjects**
Werner *et al.* [25]	Video	10-FOLD CV	Facial landmarks, 3D distances, Rand. Forest	-	**76.60**
Othman *et al.* [14]	Video	5-FOLD CV	2D CNN (MobileNetV2)	-	67.90
Zhi *et al.* [31]	Video	LOGO (8-FOLD CV)	Sparse LSTM	-	61.70
Ayral *et al.* [2]	Video	8-FOLD CV	3D CNN (3D VGG)	-	68.12
Patania *et al.* [15]	Video	5-FOLD CV	Deep GNN	-	73.20
Our Approach	Frame-based	10-FOLD CV	2D CNN	62.24	51.44
Our Approach	Video	10-FOLD CV	Attention-based CNN LSTM	**74.21**	64.46
Werner *et al.* [25]'14	ECG	10-FOLD CV	Domain-specific features, Rand. Forest	-	64.00
Lopez *et al.* [11]'17	ECG	10-FOLD CV	Multi-task Neural Networks	-	62.50
Our Approach	ECG	10-FOLD CV	LSTM	**67.69**	**64.86**
Werner *et al.* [25]	EDA	10-FOLD CV	Domain-specific features, Rand. Forest	-	71.90
Lopez *et al.* [11]	EDA	10-FOLD CV	Multi-task Neural Networks	-	79.98
Subramaniam and Dass [18]	EDA	10-FOLD CV	CNN-LSTM	80.17	75.21
Our Approach	EDA	10-FOLD CV	LSTM	**83.05**	**82.69**
Werner *et al.* [25]	EDA, ECG, EMG, Video	10-FOLD CV	Facial landmarks, 3D distances, Rand. Forest	-	80.6
Werner *et al.* [25]	EDA, ECG, EMG	10-FOLD CV	Facial landmarks, 3D distances, Rand. Forest	-	75.6
Lopez *et al.* [11]	EDA, ECG	10-FOLD CV	Multi-task Neural Networks	-	82.75
Zhi *et al.* [32]	EDA, ECG, EMG, Video	5-FOLD CV	LSTM, 3D CNN	-	68.20
Naeini *et al.* [10]	EDA, ECG, EMG, Video	10-FOLD CV	2D CNN	-	74.00
Our Approach	EDA, Video	10-FOLD CV	Attention-based CNN LSTM	84.15	**83.35**
Our Approach	ECG, Video	10-FOLD CV	Attention-based CNN LSTM	71.71	65.68
Our Approach	EDA, ECG	10-FOLD CV	LSTM	**84.38**	81.08
Our Approach	EDA, ECG, Video	10-FOLD CV	Attention-based CNN LSTM	84.04	81.43

5 Discussion

Our experimental results demonstrate several key points. Firstly, using only physiological signals (EDA) achieved the highest performance among unimodal

approaches (Table 2, Figure 3). This suggests that physiological responses may provide a more objective and reliable measure of pain compared to visual cues, which can be subjective. Secondly, combining modalities led to improved performance compared to unimodal approaches. Specifically, the combination of EDA and video data yielded the best overall accuracy (83.35% for 87 subjects, 84.15% for 67 subjects). This indicates that the framework can effectively capture complementary information from different sources, leading to a more robust and comprehensive assessment of pain.

Interestingly, the inclusion of ECG data in some multimodal combinations (ECG + Video or ECG + EDA) resulted in decreased performance compared to using just one of these modalities. This result is also supported by other studies, such as [12]. Further investigation is needed to understand why ECG data might not be consistently beneficial for pain detection in this context. Potential reasons could be noise in the ECG signal or a lack of correlation between specific ECG features and pain state in this dataset.

There were also significant differences in performance between the two scenarios we evaluated (full data vs. a subset with higher quality data). This suggests that using cleaner, higher-quality data can lead to substantial improvements in accuracy, particularly when combining video with physiological modalities. This finding highlights the importance of data quality and careful data pre-processing in pain detection research.

In comparison to more modern networks like Transformers, our approach has certain advantages, including lower complexity, faster training, and the ability to perform well with less data. While Transformers are known for their strong feature extraction capabilities, our method achieves strong performance with fewer resources. In future work, we plan to compare the effectiveness of our method with transformer-based models to further assess the trade-offs and benefits in pain recognition tasks.

Looking forward, there are exciting opportunities to enhance this framework's performance. Integrating our data with more diverse datasets, such as the AI for Pain dataset [6], which includes PPG, respiration, and will soon incorporate fNIRS and EEG—key physiological markers of pain—could significantly improve accuracy and robustness. Additionally, exploring advanced architectures like Vision Transformers and employing GANs (Generative Adversarial Networks) for synthetic data generation to address data limitations could further strengthen the framework. This future work will also involve testing the model's generalization to other acute pain datasets, helping to ensure its robustness in broader pain recognition scenarios.

6 Conclusion

In conclusion, this paper presented a novel multimodal framework for automatic pain detection. Trained on the BioVid Heat Pain Database, our framework analyzes facial expressions, EDA, and ECG data to identify pain in patients. This

approach has the potential to alert medical staff promptly about a patient's condition based on either visual cues or physiological responses.

Our experiments demonstrated the effectiveness of the multimodal approach, achieving competitive accuracy (83.35% for video+EDA, 84.38% for EDA+ECG) compared to existing research. Notably, the framework exhibited strong generalizability to unseen data despite limitations in the dataset. Moreover, the system achieves high specificity (approximately 86% in EDA + video), indicating very few false alarms. This ensures that interventions are not mistakenly applied to non-painful conditions, optimizing resource allocation and minimizing unnecessary procedures. Furthermore, the model demonstrates a sensitivity of 80% in this setting, indicating a good ability to detect true pain cases. This balance between high specificity and acceptable sensitivity highlights the framework's potential for real-world applications.

References

1. Albanie, S., Vedaldi, A.: Learning grimaces by watching tv. arXiv preprint arXiv:1610.02255 (2016)
2. Ayral, T., Pedersoli, M., Bacon, S., Granger, E.: Temporal stochastic softmax for 3d cnns: An application in facial expression recognition. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3029–3038 (2021)
3. Benavent-Lledo, M., Mulero-Pérez, D., Ortiz-Perez, D., Rodriguez-Juan, J., Berenguer-Agullo, A., Psarrou, A., Garcia-Rodriguez, J.: A comprehensive study on pain assessment from multimodal sensor data. Sensors **23**(24), 9675 (2023)
4. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641 (2019)
5. Fernandez Rojas, R., Brown, N., Waddington, G., Goecke, R.: A systematic review of neurophysiological sensing for the assessment of acute pain. NPJ Digital Medicine **6**(1), 76 (2023)
6. Fernandez Rojas, R., Hirachan, N., Brown, N., Waddington, G., Murtagh, L., Seymour, B., Goecke, R.: Multimodal physiological sensing for the assessment of acute pain. Frontiers in Pain Research **4**, 1150264 (2023)
7. Giuseppi, A., Menegatti, D., Pietrabissa, A.: Identifying chaotic dynamics in noisy time series through multimodal deep neural networks. Machine Learning: Science and Technology **5**(3), 035059 (2024)
8. Gkikas, S., Tachos, N.S., Andreadis, S., Pezoulas, V.C., Zaridis, D., Gkois, G., Matonaki, A., Stavropoulos, T.G., Fotiadis, D.I.: Multimodal automatic assessment of acute pain through facial videos and heart rate signals utilizing transformer-based architectures. Frontiers in Pain Research **5**, 1372814 (2024)
9. Kächele, M., Thiam, P., Amirian, M., Schwenker, F., Palm, G.: Methods for person-centered continuous pain intensity assessment from bio-physiological channels. IEEE Journal of Selected Topics in Signal Processing **10**(5), 854–864 (2016)
10. Kasaeyan Naeini, E., Shahhosseini, S., Subramanian, A., Yin, T., Rahmani, A.M., Dutt, N.: An edge-assisted and smart system for real-time pain monitoring. In: 2019 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). pp. 47–52 (2019). https://doi.org/10.1109/CHASE48038.2019.00023

11. Lopez-Martinez, D., Picard, R.: Multi-task neural networks for personalized pain recognition from physiological signals. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). pp. 181–184. IEEE (2017)
12. Lopez-Martinez, D., Picard, R.: Continuous pain intensity estimation from autonomic signals with recurrent neural networks. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 5624–5627. IEEE (2018)
13. Merskey, H.: Pain terms: a list with definitions and notes on usage. recommended by the iasp subcommittee on taxonomy. Pain **6**, 249–252 (1979)
14. Othman, E., Werner, P., Saxen, F., Al-Hamadi, A., Walter, S.: Cross-database evaluation of pain recognition from facial video. In: 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA). pp. 181–186 (2019). https://doi.org/10.1109/ISPA.2019.8868562
15. Patania, S., Boccignone, G., Buršić, S., D'Amelio, A., Lanzarotti, R.: Deep graph neural network for video-based facial pain expression assessment. In: 37th ACM/SIGAPP Symposium on Applied Computing. pp. 585–591 (2022)
16. Payen, J.F., Bru, O., Bosson, J.L., Lagrasta, A., Novel, E., Deschaux, I., Lavagne, P., Jacquot, C.: Assessing pain in critically ill sedated patients by using a behavioral pain scale. Critical care medicine **29**(12), 2258–2263 (2001)
17. Serengil, S.I., Ozpinar, A.: Hyperextended lightface: A facial attribute analysis framework. In: 2021 International Conference on Engineering and Emerging Technologies (ICEET). pp. 1–4. IEEE (2021). https://doi.org/10.1109/ICEET53442.2021.9659697, https://doi.org/10.1109/ICEET53442.2021.9659697
18. Subramaniam, S.D., Dass, B.: Automated nociceptive pain assessment using physiological signals and a hybrid deep learning network. IEEE Sensors Journal **21**(3), 3335–3343 (2020)
19. Thiam, P., Bellmann, P., Kestler, H.A., Schwenker, F.: Exploring deep physiological models for nociceptive pain recognition. Sensors **19**(20), 4503 (2019)
20. Thiam, P., Hihn, H., Braun, D.A., Kestler, H.A., Schwenker, F.: Multi-modal pain intensity assessment based on physiological signals: A deep learning perspective. Frontiers in Physiology **12**, 720464 (2021)
21. Turk, D.C., Melzack, R.: Handbook of pain assessment. Guilford Press (2011)
22. Walter, S., Gruss, S., Ehleiter, H., Tan, J., Traue, H.C., Werner, P., Al-Hamadi, A., Crawcour, S., Andrade, A.O., da Silva, G.M.: The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In: 2013 IEEE International Conference on Cybernetics (CYBCO). pp. 128–131. IEEE (2013)
23. Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S., Traue, H.C.: Automatic pain assessment with facial activity descriptors. IEEE Transactions on Affective Computing **8**(3), 286–299 (2016)
24. Werner, P., Al-Hamadi, A., Niese, R.: Comparative learning applied to intensity rating of facial expressions of pain. International Journal of Pattern Recognition and Artificial Intelligence **28**(05), 1451008 (2014)
25. Werner, P., Al-Hamadi, A., Niese, R., Walter, S., Gruss, S., Traue, H.C.: Automatic pain recognition from video and biomedical signals. In: 2014 22nd International Conference on Pattern Recognition. pp. 4582–4587. IEEE (2014)
26. Werner, P., Al-Hamadi, A., Walter, S.: Analysis of facial expressiveness during experimentally induced heat pain. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). pp. 176–180. IEEE (2017)

27. Werner, P., Lopez-Martinez, D., Walter, S., Al-Hamadi, A., Gruss, S., Picard, R.W.: Automatic recognition methods supporting pain assessment: A survey. IEEE Transactions on Affective Computing **13**(1), 530–552 (2019)
28. Xie, Y., Zhao, J., Qiang, B., Mi, L., Tang, C., Li, L.: Attention mechanism-based cnn-lstm model for wind turbine fault prediction using ssn ontology annotation. Wireless Communications and Mobile Computing **2021**, 1–12 (2021)
29. Yang, R., Tong, S., Bordallo, M., Boutellaa, E., Peng, J., Feng, X., Hadid, A.: On pain assessment from facial videos using spatio-temporal local descriptors. In: 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6. IEEE (2016)
30. Younger, J., McCue, R., Mackey, S.: Pain outcomes: a brief review of instruments and techniques. Current pain and headache reports **13**, 39–43 (2009)
31. Zhi, R., Wan, M.: Dynamic facial expression feature learning based on sparse rnn. In: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). pp. 1373–1377 (2019). https://doi.org/10.1109/ITAIC.2019.8785844
32. Zhi, R., Zhou, C., Yu, J., Li, T., Zamzmi, G.: Multimodal-based stream integrated neural networks for pain assessment. IEICE TRANSACTIONS on Information and Systems **104**(12), 2184–2194 (2021)