m-PainAttnNet: Multimodal Transformer Encoder with Multiscale Deep Learning for Pain Classification using Physiological Signals

Chintalapudi Abhiroop
AI20BTECH11005

Nelakuditi Rahul Naga
AI20BTECH11029

Abstract

One of the important applications of AI in Health sector is 'The Classification of Pain Intensity'. Normally for a person it is hard to express how much pain he/she is experiencing, as it is quite difficult to differentiate between different levels of pain. That is especially so for those who are in perpetual pain. But by using Physiological signals namely EDA (ElectroDermal Activity), EMG (ElectroMyographic signals) and ECG (ElectroCardiographic signals); and training an AI model to recognise the intensity of pain from those signals, we can circumvent the above issue. There are a lot of AI models for Pain Intensity Classification. One of such models which gives reasonably good results is PainAttnNet model proposed by Zhenyuan Lu et al. [1]. PainAttnNet is an uni-modal model i.e; it provides classifications using only EDA signals. But using multiple signals can possibly yield better results compared to using only a single signal. In this paper we modified the PainAttnNet model architecture to create Multimodal-PainAttnNet (m-PainAttnNet) model. We trained and evaluated our model using three types of signals (EDA, ECG, EMG) obtained from the BioVid Heat Pain Database proposed by Walter et al. [2]. We also performed ablation study on our model extensively. The code for the implementation of m-PainAttnNet model can be accessed here: `https://github.com/BMI-Lab-IITH/project-final-Rahul27n`.

1. Introduction

It is estimated that approximately 37% of Indian middle-aged and older populations are often troubled with pain. Pain prevalence increases with age and is more common among older adults aged 75+ years [3].

Pain is a distressing emotion caused by intense or damaging stimuli. The International Association for the study of Pain defines pain as an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage.

Pain is always a personal experience that is influenced to varying degrees by biological, psychological, and social factors. Although pain usually serves an adaptive role, it may have adverse effects on function and social and psychological well-being. Verbal description is only one of several behaviors to express pain; inability to communicate does not negate the possibility that a human or a nonhuman animal experiences pain.

Pain can be of varying degrees. It might just be a dull ache, or a sharp spike. On the other hand, it might also be a gut wrenching, heart stabbing pain. Pain may not be continuous. It can occur only in particular times, or in a particular body part. Pain can be multidimensional. It can have different characteristics and affect different parts of a person's life. As such, multidimensional pain scales are quite useful and effective when used to assess complex or chronic pain.

There are many different ways to measure pain. One of the most simple and common ways is to use self-reporting scales. This includes using Numerical Rating Scales (NRS), Verbal Analog Scale (VAS), Categorical Scale and last but not least, Multidimensional tools which usually include questionnaires such as Brief pain inventory (BPI), McGill pain questionnaire (MPQ). However all these scales are subjective, and are not very reliable. In case of people with communication disorders (such as auditory disorder or aphasia), the above methods do not work at all.

We can circumvent this problem by training artificial intelligence based models to classify pain intensity level using physiological signals such as Electrodermal activity (EDA), Electrocardiography (ECG), Electromyography (EMG). EDA, also known as galvanic skin response (GSR), measures variations in skin conductance level (SCL), which is associated with sweat gland activity. Clinically, skin conductance is used as an indirect measure of pain. EDA consists of tonic (SCL) and phasic (skin conductance response, SCR) components, produced by sympathetic

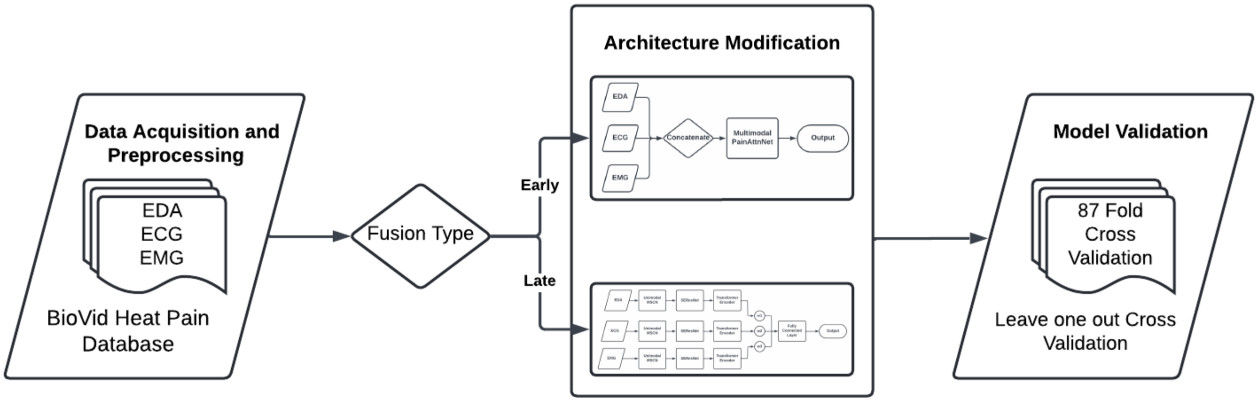

Figure 1. Overall Pipeline of m-PainAttnNet

neuronal activity. ECG records the heart's electrical activity to evaluate cardiac health and stress levels. EMG tracks muscle activity, detecting changes in muscle tension, while EEG monitors the electrical activity within the brain.

Most of the current state of the art models yield an accuracy of about 75-90% towards classification of pain intensity levels. Many of these models are restricted in their performance because they are usually not trained on multiple modalities. By using multiple signals to train the model, it is more probable to capture important (or) salient features in the data, thereby improving the model performance.

In this paper, we propose the Multimodal PainAttnNet model, abbrieviated as m-PainAttnNet. It is a mutltimodal extension of the PainAttnNet model developed by Zhenyuan Lu *et al.* [1]. m-PainAttnNet model can take multiple physiological signals such as EDA, ECG, EMG as inputs and provide pain intensity classification probabilities. m-PainAttnNet can perform two types of fusion - Early and Late. The basic building blocks of m-PainAttnNet remain same as that of PainAttnNet proposed in [1] - Multiscale Convolutional Neural Network (MSCN), Squeeze and Excitation Residual Network (SEResNet) and Transformer Encoder.

Our contributions include creating the Multimodal PainAttnNet class that performs multimodal fusion and modifying the architecture of PainAttnNet to support ablation study while using multiple physiological signals.

The overall pipeline of our m-PainAttnNet model is illustrated in figure 1. The model architecture for early and late fusion methods is explained in detail in Section 2.4.

2. Model Architecture

The model that we proposed in this paper is Multimodal PainAttnNet abbreviated as m-PainAttnNet. The model is a multimodal extension of PainAttnNet model which was proposed by Zhenyuan Lu *et al.* [1] [4]. We will discuss in brief the architecture of original PainAttnNet and then discuss the architecture of our proposed model.

2.1. Multiscale Convolutional Network (MSCN)

The idea of applying multiple time-scale convolutions to a signal and concatenating the outputs from various branches stemmed from this [5] paper. Multi-scale Convolutional Neural networks can be particularly useful for time-series mainly physiological signals (EDA, ECG, EMG etc.) as they are non-stationary. There are three branches that utilize short, medium and long windows to capture various intervals of the signals. The window size is essentially varied by varying the size of the kernel used for convolution with the signal. There are multiple max-pooling, batch norm, dropout and GeLU activation layers in each branch of the Multiscale Convolutional Network.

The outputs from each of the branch are fed into separate fully connected neural network layers that map each of the branch outputs to a fixed feature size. We chose it to be 75 (for each layer) in line with [1]. The outputs from the fully connected neural network layers are then concatenated and subsequently dropout layer is applied to the final concatenated output so as to reduce the model complexity and prevent overfitting.

For our MSCN model, we made sure that the model can support multiple modality signals. The fully connected neural network layers input size is appropriately decided based on the modality of the input signals. This is because the output size of the MSCN branches will change when the length (or) size of the input signal will change, which in turn changes with change in modality of the input. When the modality is one, the model boils down to the original MSCN model utilized in [1].

2.2. Squeeze and Excitation Residual Network (SEResNet)

The idea for SEResNet is inspired from similar model proposed in this [6] paper. SEResnet is used to extract the most salient information from the concatenated output of the MSCN branches. SEResNet operates in two stages - Squeezing and Excitation. Before performing these operations, the input from MSCN is transformed into a feature map using convolution (along with ReLU activation) operations.

In the squeezing stage, this feature map is compressed using global average pooling. While in the excitation stage, the global averaged output is fed into various fully connected layers (along with ReLU and Sigmoid activations) to generate outputs that are used to scale the feature map obtained after convolution transformation of MSCN output.

The output from MSCN layer and the scaled feature maps obtained from the excitation layer are added together using a skip connection like in ResNet architecture. This added output is then ReLU activated to obtain the final output of SEResNet block.

2.3. Transformer Encoder

PainAttnNet utilizes the encoder part of the hugely popular transformer architecture that was proposed in this [7] landmark paper by Google.

Before the transformer encoder layer is employed, the output from SEResNet block is passed to Temporal Convolutional Network, the architecture of which is proposed in this [8] paper. It is generally used to process temporally varying data, which is the case with physiological signals like EDA, ECG and EMG. The main caveat in Temporal Convolutional Network is that the output at a particular time is only dependent on the past inputs, as the future inputs are masked. This is unlike normal convolution networks.

The output from the TCN layer is passed to the transformer encoder. Transformers employ Multi-Head Attention (MHA) to capture long distance dependencies in the input. Self Attention layers basically employ Query, Key and Value matrices (obtained from the same input) to obtain the attention scores. In multi-head attention, multiple attention heads are used which perform the attention operation (in a parallel fashion) on same input using various query, key and value matrices. The outputs from each of the attention heads are concatenated together and the concatenated output is linearly transformed to get the final output.

The transformer output is further processed using residual layers and fully connected layers (along with softmax layer at the end) to obtain the final pain intensity level classification probabilities.

2.4. Multimodal PainAttnNet

In our Multimodal PainAttnNet model, various types of signals are fused together using two different types of techniques inspired from this [9] paper. They are as follows:

1. **Early Fusion:** In this type of fusion, multiple signals are concatenated together before they are input to the PainAttnNet model. The concatenated output is then fed into the PainAttnNet model to get the classification results. The architecture of the model (particularly MSCN block) will adjust based upon the modality of the input. The early fusion model architecture is illustrated in figure 2.

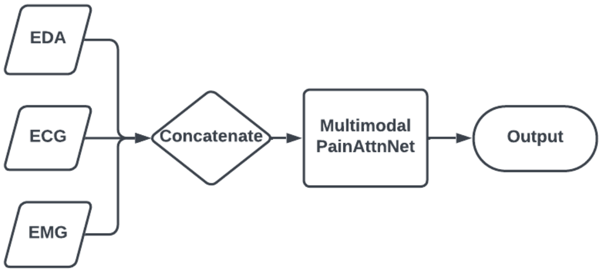

Figure 2. Early Fusion Architecture

2. **Late Fusion:** In this type of fusion, multiple unimodal PainAttnNet models are used to process multi-modal input. The number of models used is equal to the modality of the input, which must be greater than one. The outputs from the transformer encoder layer of each PainAttnNet model are linearly combined to get the combined transformer output. The combined transformer output is then flattened and passed to a fully connected layer to get the final classification probabilities. The late fusion model architecture is illustrated in figure 3.

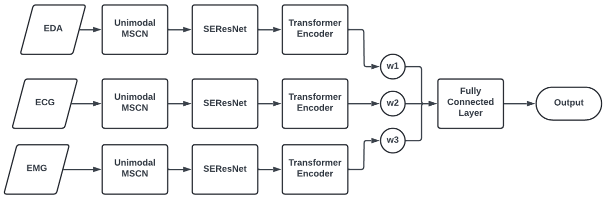

Figure 3. Late Fusion Architecture

3. Database

3.1. BioVid Heat Pain Dataset

For our experiments, we used the filtered Electrodermal Activity (EDA), Electrocardiogram (ECG) and Electromyogram (EMG) signals from Part A of BioVid Heat Pain Database (BioVid). It was created from scratch by Walter et al. [2]. One of earliest papers which used the BioVid database for pain recognition experiments is [10].

The data was acquired from 90 patients in ages 18-35, 36-50 and 51-65 with each group having 30 subjects, 15 males and females each. They were subjected to heat stimulus given through a thermode connected to right arm. In each trial, pain stimulation was administered to each participant for 25 minutes. Due to technical issues, 3 participants were excluded.

The authors of [2] calibrated each participant's pain threshold by progressively raising the temperature from the baseline room temperature. In each experiment, 5 temperatures – T_0, T_1, T_2, T_3 and T_4 – to induce five pain intensity levels from lowest to highest were determined. Each temperature stimulus was delivered 20 times for 4 seconds, with a interval of 8 to 12 seconds between each application.

EDA, ECG, and EMG signals were collected by the sensors at a sampling rate of 512Hz for a duration of 5.5 seconds. Therefore, the training sample of each signal type creates a input with dimensions given by $512 \times 5.5 \times 100 \times 87$.

3.2. Data Pre-processing

The signals data was provided in 100 csv files for each of the 87 participants. We have collated the data for each participant into a single numpy (npz) file. Thus the dataset consists of 87 npz files. Each numpy file consists of a dictionary that contains:

- Features - EDA, ECG and EMG signals
- Ordinally encoded labels (T_i is mapped to i) of size 100

Ordinal encoding has been used for the labels since the pain intensity level categorical labels are ordinal.

4. Results

4.1. Types of Tasks

We evaluated various combinations of the Multimodal PainAttnNet Model for different binary classification tasks. They are as follows:

1. **T_0 vs T_4:** This classification tries to differentiate between the two extremes of pain intensities. T_0 corresponds to the temperature at which the participant experiences no pain, while T_4 is the temperature at which the participant experiences intolerable pain.

2. **T_0 vs T_3:** This classification tries to differentiate between the first and fourth level of pain intensities. T_0 corresponds to the temperature at which the participant experiences no pain, while T_3 is the temperature at which the participant experiences very high amount of pain.

3. **T_0 vs T_2:** This classification tries to differentiate between the first and third level of pain intensities. T_0 corresponds to the temperature at which the participant experiences no pain, while T_2 is the temperature at which the participant experiences moderate amount of pain.

4. **T_0 vs T_1:** This classification tries to differentiate between the first two levels of pain intensities. T_0 corresponds to the temperature at which the participant experiences no pain, while T_1 is the temperature at which the participant just starts to experience pain.

4.2. Implementation Details

We have trained all the models for a duration of 100 epochs and evaluated them using leave one out cross validation. There are a total of 87 participants, so 87-fold cross validation was performed.

All the models, including Multimodal PainAttnNet were implemented using PyTorch. Other hyperparameters such as batch size, learning rate etc. were kept the same as in this [1] paper. Cross entropy loss (with Adam optimizer) was used to train all the models. In late fusion, the linear combination parameters (w_1, w_2, w_3) were initialized to all ones at the beginning of training.

We utilized a GPU-enabled machine to perform our training tasks, as they were computationally expensive. It was generously provided by MedInfo Lab at Indian Institute of Technology Hyderabad.

4.3. Evaluation Metrics

We are using the metrics Accuracy, Precision, Recall, Macro F_1 (MF_1) score and Cohen Kappa for evaluating the classification performance of our models. The metrics for early fusion and late fusion are shown in table 1 and table 2 respectively.

Early Fusion					
Tasks	Accuracy	Precision	Recall	MF_1	κ
T_0 vs T_4	97.931	97.935	97.931	97.931	95.862
T_0 vs T_3	97.126	97.126	97.126	97.126	94.253
T_0 vs T_2	96.810	96.816	96.810	96.810	93.620
T_0 vs T_1	97.155	97.162	97.155	97.155	94.310

Table 1. Performance of Multimodal PainAttnNet with Early Fusion

Late Fusion					
Tasks	Accuracy	Precision	Recall	MF_1	κ
T_0 vs T_4	99.023	99.024	99.023	99.023	98.046
T_0 vs T_3	97.988	98.005	97.988	97.988	95.977
T_0 vs T_2	96.753	96.763	96.753	96.753	93.506
T_0 vs T_1	96.954	96.954	96.954	96.954	93.908

Table 2. Performance of Multimodal PainAttnNet with Late Fusion

4.4. Ablation Study

We have also added ablation support to the multimodal PainAttnNet model. We have evaluated the performance of the following models as part of ablation study:

1. **MSCN:** The output from the multimodal MSCN block is flattened and then input into a fully connected layer to get the final classification results.

2. **MSCN + SEResNet (MSCN_SE):** The output from multimodal MSCN block is fed into the SEResNet block, which greatly reduces the size of MSCN output using the squeezing operation. The output from SEResNet block is then flattened and sent into a fully connected layer to get the final classification results.

3. **MSCN + Transformer (MSCN_TRAN):** The output from multimodal MSCN block is directly fed into the Transformer Encoder block, skipping the SEResNet block. The output from Transformer Encoder block is then flattened and sent into a fully connected layer to get the final classification results.

We chose not to train the MSCN + Transformer Encoder (MSCN_TRAN) model as the number of parameters of this model were significantly higher than the full PainAttnNet model itself, which might lead to overfitting. This was mainly because, SEResNet greatly reduces the feature dimension of MSCN output using the squeezing operation, but if we skip it, like in MSCN_TRAN, it leads to a large dimensional output being given by the transformer encoder. This leads to blowing up of parameters in the MSCN + Transformer Encoder model as we require a significantly larger fully connected layer to map the transformer encoder output to the classification probabilities.

The performance of MSCN and MSCN + SEResNet (MSCN_SE) models was then compared to the performance of early fusion (in **full** mode) m-PainAttnNet model. The results for T_0 vs T_4, T_0 vs T_3, T_0 vs T_2, T_0 vs T_1 are shown in figures 4, 5, 6 and 7 respectively.

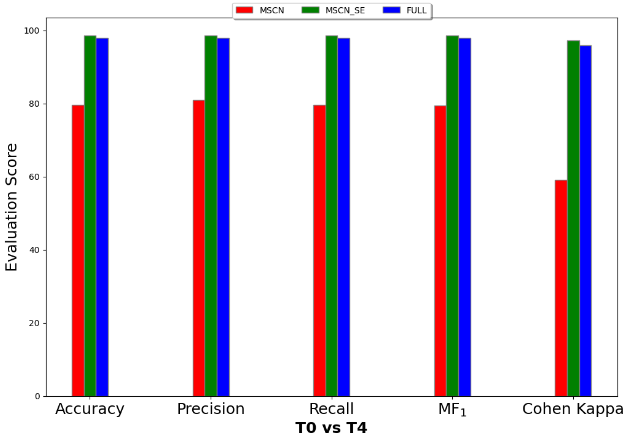

Figure 4. T_0 vs T_4 Ablation Study

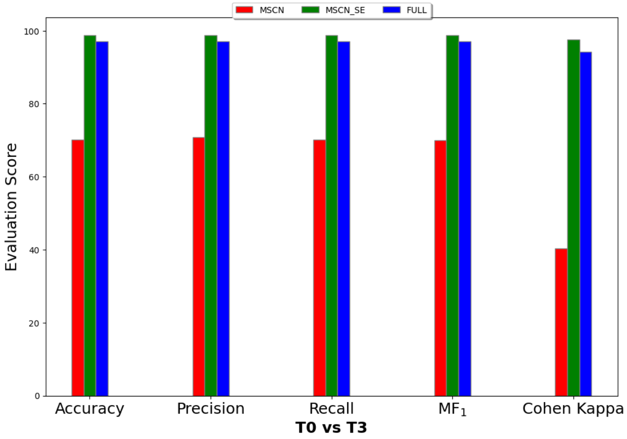

Figure 5. T_0 vs T_3 Ablation Study

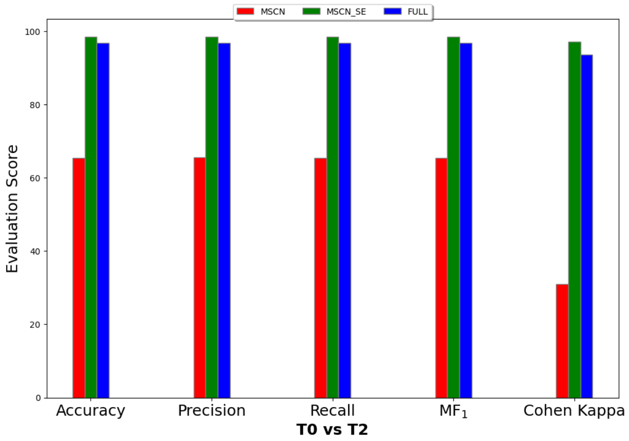

Figure 6. T_0 vs T_2 Ablation Study

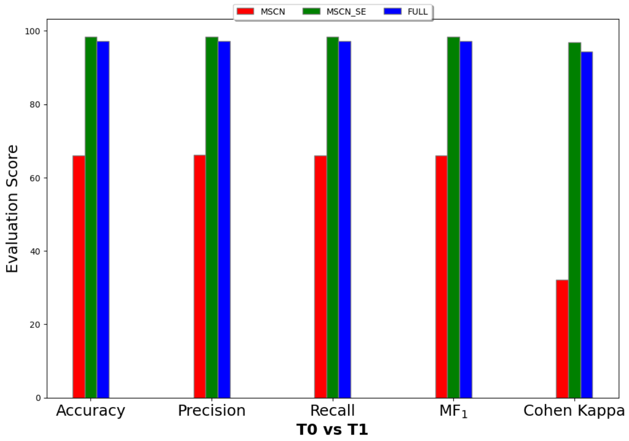

Figure 7. T_0 vs T_1 Ablation Study

5. Conclusions

Pain is a very common symptom of many diseases. It can be of many kinds. It is important to quantify the intensity of pain level that a person is experiencing in order to decide what further steps can be taken. Automated classification of pain intensity using Artificial Intelligence based models can be a plausible solution to this problem.

In this paper, we proposed Multimodal PainAttnNet (m-PainAttnNet) model that is a multimodal extension of the PainAttnNet model that was proposed in [1]. We utilized the BioVid Heat Pain Database for our studies. The model comprises of 3 basic architecture blocks - MSCN, SEResNet and Transformer Encoder. Multimodal PainAttnNet can take multiple types of signals such as EDA, ECG and EMG as inputs and classify between various pain intensity levels. m-PainAttnNet can perform two types of fusion - Early and Late. The performance metrics for both types of fusion were provided. We also performed ablation study on the model by comparing the performance of various blocks of the m-PainAttnNet model to the full model's performance. The results were reported using insightful bar plots.

The performance of our m-PainAttnNet model is exceeding the benchmarks reported by the authors of [1] on almost all the classification tasks. In future, we intend to test and improve the generalization capability of m-PainAttnNet on other pain classification datasets.

References

[1] Zhenyuan Lu, Burcu Ozek, and Sagar Kamarthi. Transformer encoder with multiscale deep learning for pain classification using physiological signals. *Frontiers in Physiology*, 14, December 2023. 1, 2, 3, 4, 6

[2] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Jun-Wen Tan, Harald C. Traue, Stephen Clive Crawcour, Philipp Werner, Ayoub Al-Hamadi, and Adriano de Oliveira Andrade. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. *2013 IEEE International Conference on Cybernetics (CYBCO)*, pages 128–131, 2013. 1, 4

[3] Amit Kumar Goyal and Sanjay K Mohanty. Association of pain and quality of life among middle-aged and older adults of india. *BMC Geriatrics*, 2022. 1

[4] PainAttnNet GitHub Code. 2

[5] Zhicheng Cui, Wenlin Chen, and Yixin Chen. Multi-scale convolutional neural networks for time series classification. *arXiv (Preprint)*, 2016. 2

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomezn, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems, volume 30*, 2017. 3

[8] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to ac tion segmentation. *Computer Vision–ECCV Workshops: Amsterdam, The Netherlands*, 2016. 3

[9] Patrick Thiam, Peter Bellmann Hans A. Kestler, and Friedhelm Schwenker. Exploring deep physiological models for nociceptive pain recognition. *Sensors 19, no. 20: 4503*, 2019. 3

[10] Philipp Werner, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald C. Traue. Automatic pain recognition from video and biomedical signals. *Proc. Int'l Conf. on Pattern Recognition (ICPR)*, 2014. 4