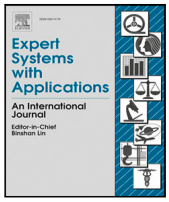
Expert Systems With Applications 235 (2024) 121082

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Personalized and adaptive neural networks for pain detection from multi-modal physiological features

Mingzhe Jiang [a,b,h], Riitta Rosio [c], Sanna Salanterä [c], Amir M. Rahmani [d,e], Pasi Liljeberg [f], Daniel S. da Silva [g], Victor Hugo C. de Albuquerque [g], Wanqing Wu [a,b,*]

[a] School of Biomedical Engineering, Sun Yat-sen University, Guangzhou, China
[b] Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[c] Department of Nursing Science, University of Turku, Turku, Finland
[d] School of Nursing, University of California, Irvine, USA
[e] Department of Computer Science, University of California, Irvine, USA
[f] Department of Computing, University of Turku, Turku, Finland
[g] Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza-CE, Brazil
[h] Guangzhou Institute of Technology, Xidian University, Guangzhou, China

ARTICLE INFO

Keywords:
Pain assessment
Multi-modal fusion
Pain sensitivity
Attention mechanism
Neural networks

ABSTRACT

Pain assessment is essential for pain diagnosis and treatment. Automating the assessment process from pain behaviors could be an alternative to self-report; however, inter-subject and time-dynamic differences in pain behaviors hinder pain recognition as generic patterns. To address this problem, we proposed a neural network method integrating pain sensitivity in personalized feature fusion and dynamic feature attention leveraging the Squeeze-and-Excitation block. Ablation results from our physiological pain data show that dynamic attention effectively improved prediction recall through soft physiological feature selection, and fusing pain sensitivity improved precision, yielding better F1-score together. By testing our trained models with external BioVid Heat Pain data, we observed better adaptivity to a different pain protocol with higher accuracy in time-continuous pain detection than simple neural networks. At last, we found our method outperformed SOTA works using the same public database in pain intensity classification and regression, reaching 84.58% accuracy in high pain detection with model pretraining.

1. Introduction

Pain is a sensory and affective experience with inter-subject differences, which makes self-reports from patients the first choice in pain assessment whenever it is available (Raja et al., 2020). Pain assessment usually represents evaluating the intensity, location, and duration of pain for acute pain caused by trauma, surgery, or acute medical disease. Among the evaluation dimensions, pain intensity is important in evaluating the effects of pain treatment, where multiple scales or tools were developed and tested for their effectiveness and consistency (Breivik et al., 2008). The most well-known scales are the visual analog scale (VAS) in 100 mm and the numerical rating scale (NRS) from 0 to 10 representing 'no pain' to 'worst pain imaginable'.

Self-report is not always available from a patient during sedation or from a special population, where objective tools such as pain behavior observation have to be used (Herr et al., 2019). The clinical experience with objective pain assessment further inspired automated assessment of pain, aiming at quantifying pain behaviors automatically and translating them to pain intensity. Most automated pain assessment studies formulate pain recognition as a pattern recognition problem and work on interpreting pain intensity from facial expressions (Lucey et al., 2012), head poses (Werner et al., 2017), body protective behaviors (Aung et al., 2016), physiological signals (Ledowski, 2019), and their multi-modal fusion (Jiang et al., 2019; Werner, Al-Hamadi, et al., 2019; Werner et al., 2014). Werner et al. have presented a (Werner, Lopez-Martinez, et al., 2019) exhaustive survey on the rationals, modalities, and techniques used in common for pain recognition, which are not repeated here. Physiological signals such as galvanic skin response (GSR), electrocardiography (ECG), and photoplethysmography (PPG)

* Correspondence to: No. 66, Gongchang Road, Guangming District, Shenzhen, Guangdong 518107, PR China.
E-mail addresses: mzjiang@xidian.edu.cn (M. Jiang), riitta.rosio@utu.fi (R. Rosio), sansala@utu.fi (S. Salanterä), a.rahmani@uci.edu (A.M. Rahmani), pasi.liljeberg@utu.fi (P. Liljeberg), danielssilva@alu.ufc.br (D.S. da Silva), victor.albuquerque@ieee.org (V.H.C. de Albuquerque), wuwanqing@mail.sysu.edu.cn (W. Wu).

https://doi.org/10.1016/j.eswa.2023.121082
Received 21 November 2022; Received in revised form 30 June 2023; Accepted 28 July 2023
Available online 5 August 2023

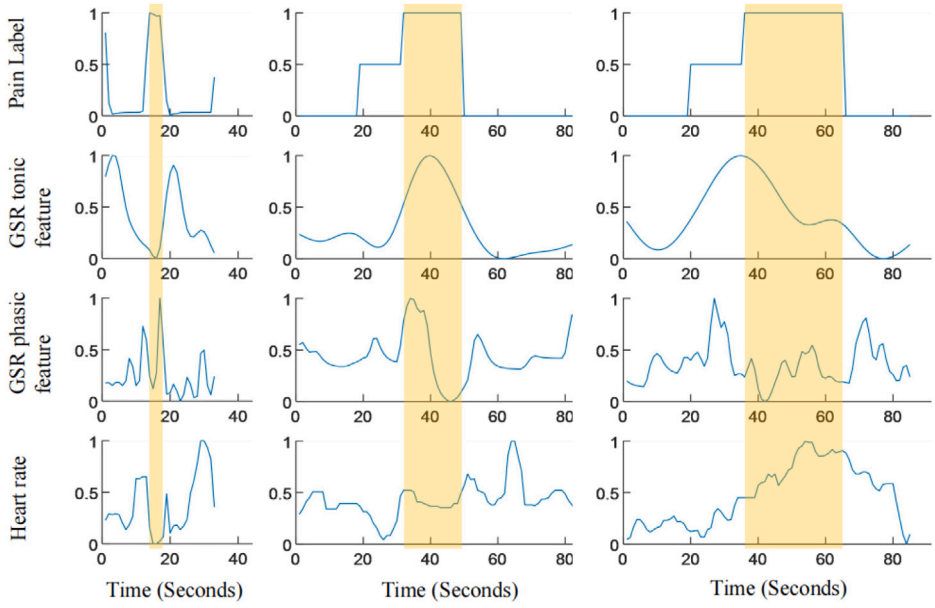


Fig. 1. Response delay and behavior differences with GSR components and heart rate in pain events of three duration lengths. Examples are from the BioVid heat pain database (Walter et al., 2013) and SpaExp dataset (Jiang et al., 2019).

reflecting sympathetic and parasympathetic activities induced by acute pain outperformed the pain behavior modality represented by facial expressions in their synchronized analysis (Kächele et al., 2017; Werner, Al-Hamadi, et al., 2019). This work focused on GSR and ECG signals due to their unobtrusive measurement and the need for improvements in recognition accuracy and resolution in comparison to the brain signals (Vijayakumar, Case, Shirinpour, & He, 2017).

The GSR signal, which was also mentioned as skin conductance or electrodermal activity was found well responsive to both short and long pain; while the ECG signal responds more to long-lasting pain. Although end-to-end solutions with neural networks are trendy with physiological signals with deep feature mining, the power of hand-crafted features still exists with the advantage of explainability (Wang et al., 2020; Zhang et al., 2021). Methods developed with the two physiological signals would also be easily expanded to scenarios with streaming signals from a bedside patient monitor inward or free-to-move scenarios with a wristband collecting GSR and PPG.

The complexity of pain and physiological response pattern to pain is one challenge in tracking pain continuously in time. For example, the response delay and recovery to pain stimulations with physiological features and pain duration could regulate physiological response patterns to pain, as shown in Fig. 1, which makes conventional feature selection from the whole dataset less adaptive to variations in patterns. Therefore, one motivation for this work was to employ a soft selection strategy for adaptive recognition of pain with unknown durations or physiological response behavior differences.

Another bottleneck to pain recognition was assuming a consistent pain response pattern across subjects, while individual differences exist in pain experience and behaviors. Customizing pain stimulation intensity according to personal pain threshold and pain tolerance is usually used to standardize subjective pain experience. Still, there are few clues on how to normalize pain behaviors. Lopez-Martinez, Rudovic, and Picard (2017b) observed a significant difference in age between feature clusters, which was then validated with the UNBC-McMaster shoulder pain expression archive in Liu et al. (2017) where age contributed more to pain recognition performance than gender and complexion when fusing with facial expression features. Inspired by these works, we propose fusing personal pain sensitivity for personalized pain detection. Pain sensitivity is a concretized measure of individual differences in pain perception and is possibly predictable (Jiang et al., 2022; Ruscheweyh, Marziniak, Stumpenhorst, Reinholz, & Knecht, 2009; Spisak, Kincses, Schlitt, Zunhammer, Schmidt-Wilcke, Kincses, & Bingel, 2020).

To sum up, we aim to optimize continuous pain detection from GSR and ECG signals by introducing a soft feature selection strategy and pain sensitivity representations for personalized recognition. Our main contributions are:

- We propose neural networks integrated with dynamic feature attention as soft feature selection in time-continuous prediction to deal with pain response pattern variations due to complex pain cases.
- We propose neural networks fusing with personal pain sensitivity representations for personalized recognition, and discussed whether pain sensitivity estimates were interchangeable with pain sensitivity measures for easy use.
- We proved the effectiveness of the proposed methods with our own and external pain datasets to the state-of-the-art and identified boundaries to use.

The rest of this paper is organized as follows: Section 2 introduces related work on personalized pain recognition and the attention mechanism background; Section 3 illustrates our proposed pain detection framework in detail; Section 4 describes the experiments designed to validate our proposed methods and their implementation details; Section 5 presents results and performance comparisons; Section 6 discusses on results and their implications, and Section 7 concludes the work.

2. Related work

2.1. Personalized pain recognition

Individual differences in pain have contained sustained attention, which was observed in pain sensitivity, susceptibility to developing painful disorders, and response to analgesic manipulations (Mogil, 2021). Pain researchers worked extensively on the explanations for such variability in genetics, genders, age, ethnics, personality traits, psychological traits, and possibly environmental factors. Among those, gender, age, and ethnicity were used as personal features (Liu et al., 2017), fusing frame-level predictions from facial expression images to get sequence-level self-reported pain intensity.

Previous efforts on personalized pain recognition were mostly motivated by grouping subjects with similar pain behaviors. For example, Kächele et al. (2017) applied personalized training, where the training set for a test subject was made from subjects with similar feature distribution. From a technical point of view, multitask learning is frequently used for implementing personalization principles. Lopez-Martinez et al. applied feature clustering to group the subjects and defined each cluster as one task in their work (Lopez-Martinez et al., 2017b). In their other work, each subject was defined as one task with a person-specific hidden layer trained before the output layer of each task (Lopez-Martinez & Picard, 2018b). With a similar principle, Rudovic et al. (2021) used federated learning to fine-tune local models for specific users. A different personalization motivation was proposed in Lopez-Martinez, Rudovic, and Picard (2017a) from pain facial expressions images, which leveraged the objective coding of pain based on facial action units. Additionally, Casti et al. (2020) used personalized feature selection in the proposed pain monitoring platform.

Although pain sensitivity has been long used as an interindividual pain difference indicator, it has not been considered to optimize pain recognition. Pain sensitivity is usually quantified through the quantitative sensory test (QST), where the lowest stimulus intensity that is perceived as pain is a person's pain threshold. However, such a test requires specific equipment until more and more clues indicate the predictability of pain sensitivity.

2.2. Personal pain sensitivity prediction

Strong evidence was reported recently from recent neuroimaging studies that personal pain sensitivity can be predicted by brain connectivity at the resting state (Spisak et al., 2020). Meanwhile, Hohenschurz-Schmidt et al. identified a three-way relationship between the autonomic nervous system represented by the low frequency of HRV, brain networks and difference in subjectively reported pain (Hohenschurz-Schmidt et al., 2020).

Before those brain studies, resting autonomic function represented by HRV frequency features was intermittently discussed regarding its possible correlation with personal pain sensitivity (Appelhans & Luecken, 2008; Tracy, Jarczok, et al., 2018; Tracy, Koenig, et al., 2018). Based on the above-mentioned prior knowledge, we expanded the representation of personal pain sensitivity and found representative resting HRV features for personal pain sensitivity prediction (Jiang et al., 2022). Predicting pain sensitivity and pain from different segments of physiological signals means no extra measurement will be required, which further motivated us to validate our assumption on the optimization of pain recognition via pain sensitivity representations.

2.3. From channel attention to dynamic feature attention

To adapt features extracted from fixed lengths of windows to various pain durations, our intuition was to apply a soft feature-selection strategy that can change dynamically, rather than finding the most representative features and discarding the rest. We assume the adaptivity to be enhanced from two aspects: (1) different window lengths for different features — our proposed feature extraction flow (Jiang, 2019) extracted physiological features from various lengths of sliding windows to ensure their interpretability, especially in terms of sympathetic and parasympathetic activities with HRV features; (2) the tonic and phasic components of GSR signals, which correspond to the slow and fast changes in sympathetic tone.

We found the squeeze-and-excitation attention module that was originally tested for object recognition in convolutional neural networks (CNN) (Hu, Shen, & Sun, 2018; Woo et al., 2018) might suit our expectations. It was used as convolutional blocks, separately calculating channel and spatial weights of CNN's middle output and then multiplying with it. An image data is a three-dimensional matrix ($H \times W \times C$), and its channel attention is obtained for dimension C. If considering a multi-feature times series as a three-dimensional matrix, its dimensions are 1 × series length × number of features, and the channel attention from an image is exactly the feature attention of a time series. In this way, we can dynamically derive feature attention along the time axis from a sliding window.

To summarize, efforts have been made in personal feature fusion and pain behavior clustering towards personalized pain recognition; but personal pain sensitivity, as a well-accepted indicator of inter-subject difference, has not been considered nor tested as a personal feature. This work addressed the gap by fusing measured or estimated pain sensitivity in neural networks as a personalized solution. In parallel, we converted channel attention for images into time-dynamic feature attention for time series in adapting neural networks to different pain events, which was rarely considered in fixed-length pain protocols.

3. The proposed pain detection framework

The proposed framework for time-continuous detection of pain from ECG and GSR signals is presented in Fig. 2. The processing flow includes feature extraction, personalized vector representation, dynamic feature attention calculation, deep feature extraction from neural networks and recognition.

In the framework, a dynamic feature attention module and personalization module were added to the base neural networks model to enhance recognition. The intuition of integrating dynamic feature attention to neural networks was adding soft feature selection dynamically to adapt to dominant feature change in pain response over time. The soft feature selection was implemented as feature weights calculated from features behaviors in the past T time period and their dot product with feature vector X_t before inputting to neural networks.

The personalization module extracts personal pain sensitivity information from ECG at rest state. Pain sensitivity information was treated as a personalized feature encoding inter-personal differences and was expected to boost recognition through personalized recognition.

In the methods validation part, we use the SpaExp database ((Fig. 3 shows one pain test example) to compare and find optimal options for base model architecture, personalized features and calculation. Then the public database BioVid was tested for vertical comparison to external studies using the same data.

3.1. Two pain datasets

We first developed and validated the proposed methods from our pain dataset, SpaExp, and then further tested them with a shared pain dataset, BioVid. Table 1 presents the basic information of two datasets regarding their shared features and differences. The pain was triggered experimentally by controllable intensity and duration in both datasets. ECG and GSR were continuously recorded throughout all pain and non-pain periods in both datasets, allowing methods to be tested on one another.

On the other hand, the two datasets are different in several aspects. First, SpaExp triggered tonic pain in general in terms of stimulation duration and allowed physiological recovery between two adjacent tests. In contrast, BioVid triggered phasic pain lasting for 4 s and repeated the test after 8–12 s. Second, the two datasets have a minor difference in pain definition, where BioVid defines four pain levels. As the goal of this work is pain detection, the four pain levels are all considered 'Pain'.

3.2. Processing and feature extraction of ECG and GSR

3.2.1. ECG processing

The processing of ECG signals includes denoising and R peak detection. We used a 10-point moving average filter to derive low-frequency motion artifacts and subtracted them from raw ECG signals. The locations of R peaks were then detected by peak detection using a sliding window based on an adaptive amplitude threshold and a fixed peak distance threshold.

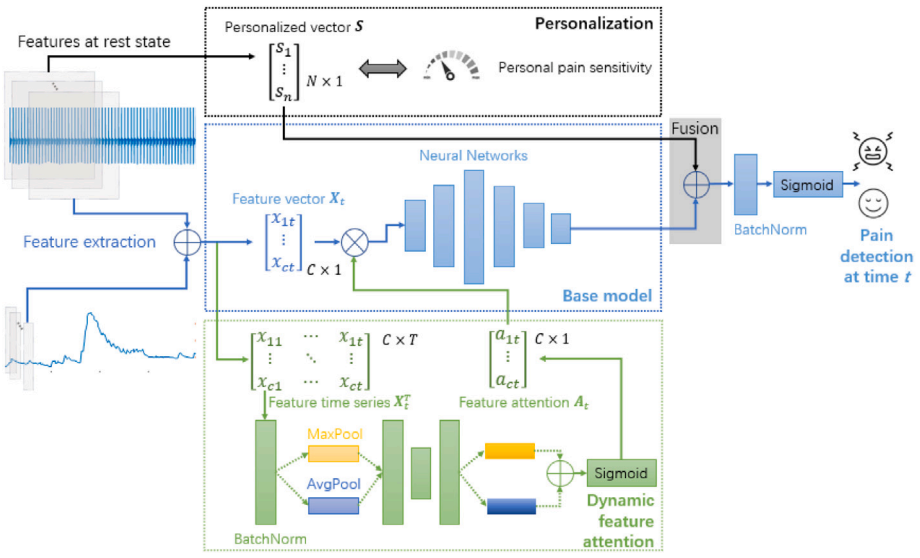


Fig. 2. The framework of the proposed pain detection method.

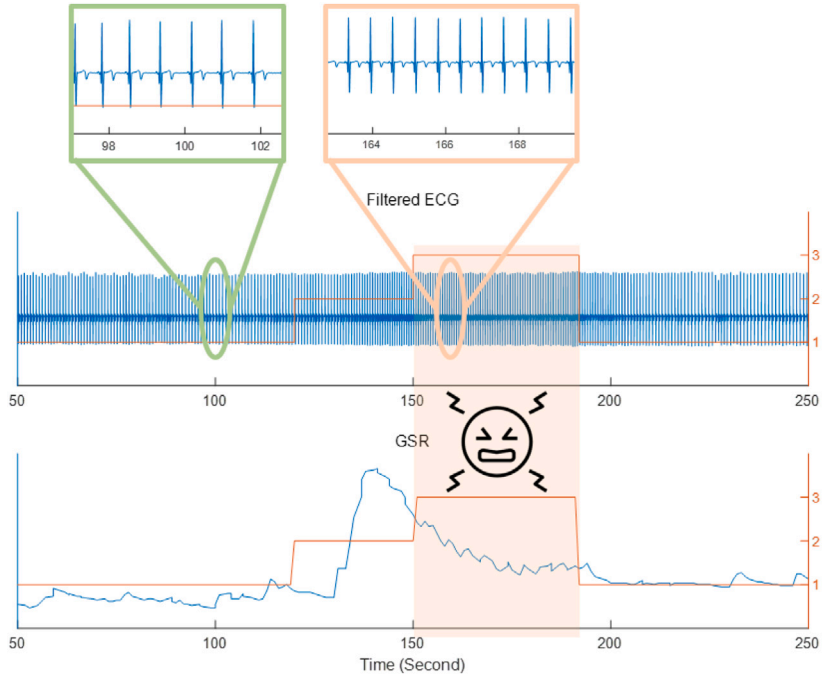


Fig. 3. One pain test from SpaExp. For the right Y axis, '1' denotes no stimulation, '2' denotes non-painful stimulation, and '3' denotes painful stimulation and indicates subjectively painful.

Table 1
Information of SpaExp and BioVid pain datasets.

	SpaExp	BioVid
Subject group	Healthy volunteers	Healthy volunteers
Pain stimulation	heat & electrical	heat
Stimulation duration	107.2 s (25.0) & 54.9 s (32.1)	4 s
Number of subjects included	30	87
Number of tests per subject	2 & 2	80 (20 × 4 levels)
Sampling rate of one-lead ECG	250 Hz	512 Hz
Sampling rate of GSR	1 Hz	512 Hz
Pain definition	Gradual-increasing pain stimulation from pain threshold to pain tolerance	Pain stimulation at pain threshold, pain tolerance and two levels between them

3.2.2. GSR processing

The processing of GSR is mainly its decomposition. We used cvxEDA (Greco et al., 2016) to decompose GSR signals into its tonic component – skin conductance level (SCL) and its phasic component – skin conductance responses (SCRs). The amplitude of GSR signals was z-score normalized before decomposition. Detailed information is from our

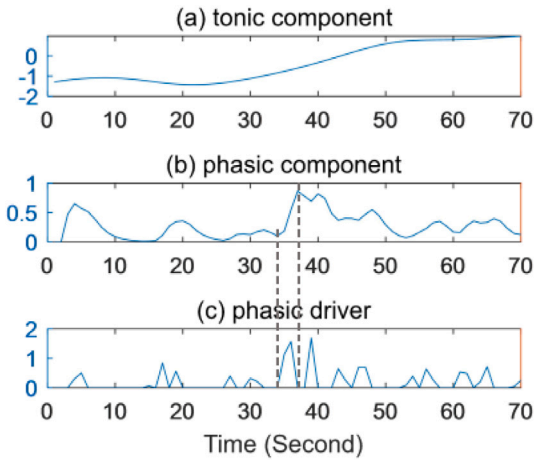


Fig. 4. Key components of GSR after decomposition. An original GSR signal = tonic component + phasic component + an additive Gaussian noise term.

Table 2
List of features, extracted from sliding windows with 1-second stride.

	Feature name	Window length	Scalar λ
GSR, phasic driver			
1	rise_time_avg	15 s	–
2	phasic_driver_num_pks	15 s	–
3	phasic_pks_amp_max	15 s	–
GSR, tonic component			
–	tonic_value	– only in SpaExp	–
4	tonic_max	15 s	–
5	tonic_avg	15 s	–
6	tonic_auc	15 s	–
7	tonic_std	15 s	–
GSR, phasic component			
8	phasic_auc	15 s	–
9	phasic_std	15 s	–
10	phasic_avg	15 s	–
11	phasic_max	15 s	–
ECG			
12	hr_mean (bpm)	5 s	$\frac{75}{\text{subject mean hr}}$
13	hr_median (bpm)	5 s	–
14	SDNN (ms)	30 s	$\frac{800}{\text{subject mean nn}}$
15	RMSSD (ms)	10 s	$\frac{800}{\text{subject mean nn}}$
–	LF (ms^2) 0.04–0.15 Hz	250 s	$(\frac{800}{\text{subject mean nn}})^2$

previous work (Syrjälä et al., 2019). Fig. 4 shows three key components for further feature extraction. The start and end of one pulse in the phasic driver correspond to a rise in the phasic component. Therefore, rise time, the number of SCRs and amplitudes were derived from the phasic driver rather than the phasic component.

3.2.3. Feature extraction

Features of ECG and GSR signals were extracted from sliding time windows in different lengths and identical stride lengths of 1 s, ensuring all features have the same time resolution of 1 Hz. Table 2 makes a full list of features extracted in this work. We discussed the choices of window lengths in Jiang (2019), where the main principles are as short and interpretable as possible, especially for the ultra-short-term heart rate variability. In SpaExp, GSR signals and their tonic component are at 1 Hz, which is the same as the feature resolution, and therefore was considered one additional feature.

3.2.4. Feature normalization

The amplitude range of GSR was normalized in its processing step. A structural correlation exists between heart rate and heart rate variability for ECG features. Hallstrom et al. (2004) introduced HRV re-scaling by mean HR to adjust this effect, enabling normalized HRV to reveal power more than reflecting the difference in HR. Therefore, we normalized HR and HRV to a personal average heart rate of 75 beats per minute (800 ms R-R interval). The mean HR (hr_mean), the standard deviation of normal R-R intervals (SDNN), root-mean-square of the successive differences (RMSSD), low-frequency power of R-R intervals (LF), and high-frequency power of R-R intervals (HF) were re-scaled. Corresponding scalars were listed in Table 2.

Table 3
NN base model for SpaExp database. B is the batch size, C is the feature dimension.

Layer	Input shape	Output shape	Activation function
Input	$[B, C]$	–	–
Batch norm	$[B, C]$	$[B, C]$	–
Hidden 1	$[B, C]$	$[B, 64]$	
Hidden 2	$[B, 64]$	$[B, 128]$	
Hidden 3	$[B, 128]$	$[B, 64]$	Sigmoid
Hidden 4	$[B, 64]$	$[B, 16]$	
Output	$[B, 16]$	$[B, 1]$	

3.3. Base model

Two neural network architectures were tested for their optimization possibilities and recognition performance, artificial neural networks as NN base and recurrent neural networks as RNN base. Another motivation for the comparison was to examine whether the attention module can provide information other than feature time-series dependencies learned by RNN models.

The NN base model for the SpaExp database was described in Table 3, where C was the feature dimension and it was 16 for SpaExp input data X_t. Sigmoid was the activation function of all hidden layers and the output layer with output ranging between 0 and 1.

RNN base models for the SpaExp database were with two unidirectional hidden layers and 128 dimensions in each layer with a 0.5 dropout rate. The output of the second hidden layer was dimension reduced to 1 by a fully-connected layer activated by ReLU. The output layer was activated by Sigmoid function for binary classification. The input feature series was X_t^T, and the prediction made at time t was considered as a many-to-one RNNs prediction. We tested three RNN unit types (Vanilla RNN, LSTM, and GRU) and chose one for further tests with dynamic feature attention module and personalization module.

3.4. Dynamic feature attention module

We propose adding dynamic feature attention module, which forms an automated and dynamic way of feature selection in order to enhance the sensitivity in continuous detection. We applied channel attention proposed in object detection (Hu et al., 2018; Woo et al., 2018). The concept of channel attention for computer vision was converted to feature attention in this work, where the 2D spatial information in figures corresponds to the 1D time series here and the RGB-channel vector in figures corresponds to the feature vector here.

For a feature time series $X_t^T \in \mathbb{R}^{C\times T}$, its feature attention $A_t \in \mathbb{R}^{C\times 1}$ was calculated via M_f, one or compound pooling along the time dimension. We tested two attentions in this work, and they were:

$$M_f^{avg}(X) = \sigma(MLP(AvgPool(X))) \quad (1)$$

$$M_f^{avgmax}(X) = \sigma(MLP(AvgPool(X)) + MLP(MaxPool(X))) \quad (2)$$

The MLP here was a shared network when multiple poolings were applied. It was composed of a hidden layer ($C \times 1$) followed by the ReLU activation function and a linear layer ($1 \times C$). The outputs of the MLP were added, if there were multiple, before being fed into a sigmoid function (σ). The parameter T was primarily set as 30-second and 6-second for SpaExp and BioVid, respectively to be comparable to pain-lasting time. We present the impact of T on NN optimization in Results section.

3.5. Personalization module

As illustrated in Fig. 2, the idea of personalization was in fusion with the representation of personal pain sensitivity. Personal pain sensitivity is measurable and can be predictable from physiological signal features in the resting state. Therefore, three forms of pain sensitivity representation were tested for pain recognition optimization. Further, as estimated pain sensitivity is more accessible from ECG recording, performance comparisons were made to check whether it could replace the measured one in the optimization.

The three pain sensitivity representations tested in this work were named as SensMeasure, SensPredict and SensHRV. SensMeasure was a pain sensitivity score composed of measures from quantitative tests. SensPredict and SensHRV were pain sensitivity estimations from resting heart rate variability analysis results in our previous work (Jiang et al., 2022).

SensMeasure: For the SpaExp database, pain threshold level, pain tolerance level, and pain intensity reported at pain tolerance were recorded through pain tests. Pain threshold or tolerance is the stimulus intensity starting to be perceived as painful or intolerably painful. The reported pain intensities were recorded on a 100 mm visual analog scale. We combined threshold level, tolerance level, and tolerance intensity from heat and electrical tests by adding their values after z-score standardization across subjects. For the BioVid database, pain threshold level and tolerance level of heat stimulation were accessible to construct a composite score.

SensPredict: SensPredict was only for the SpaExp database, which was a pain sensitivity score predicted from the regression of resting SDNN, RMSSD and LF (Jiang et al., 2022). Based on our previous results, the SensPredict score was linearly correlated with the SensMeasure score.

SensHRV: SensHRV was composed of SDNN, RMSSD, and LF values at resting state for predicting pain sensitivity score, assuming neural networks could decode pain sensitivity instead of regression.

Fig. 5 presents the full architecture of neural networks with an NN base. The hiddenlayer before the output on acts as a fusion layer in models with a personalized vector. The output layer contains 1 neuron activated by Sigmoid in models for binary classification and regression, which was changed into 5 neurons activated by Softmax in models for 5-class pain level classification.

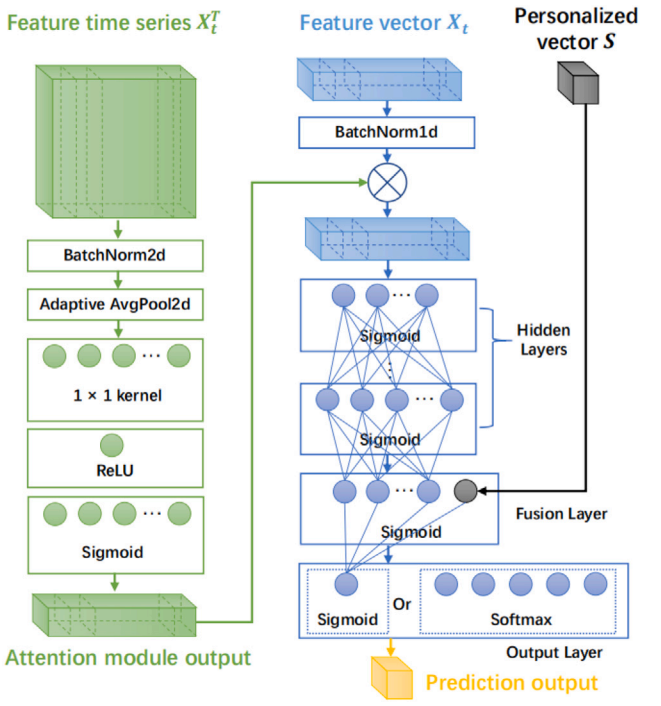


Fig. 5. The architecture of the full neural networks with NN base. 1 × 1 Sigmoid activated neuron for binary classification and regression output, 5 × 1 activated neuron for 5 pain level classification output.

4. Experiment design and implementation

4.1. Experiment design

Experiments were designed for two aims: to validate the proposed methods and to find the best-combined setting. Fig. 6 presents the roadmap of the overall experiment design. The input to models were X_t or X_t^T, composed of features from 1 to 15 in Table 2 plus tonic_value, as the dynamics are weak with features from long time windows (SDNN: 30 s).

The internal model comparison was conducted within the SpaExp dataset, where base models, attention types, and pain sensitivity representations were first compared independently and then jointly. In the sensitivity representations part, we also compared the performance of each representation with a random vector in the same dimension to check whether the representation brought information more than random noise to the model. Random noise was generated as random numbers between 0 and 100 with dimensions the same as the corresponding pain sensitivity representation, i.e., Random1: number of subjects × 1 for SensMeasure and SensPredict; Random3: number of subjects × 3 for SensHRV.

With trained models, we first visualized attention outputs to validate the soft selection ability of the attention module. Next, to check the generalization ability of trained models, we tested models with the external dataset BioVid. Raw GSR and ECG signals in BioVid were processed the same as the steps described in Section 3. Models were trained and tested with the common 15 features in two datasets. Then we compared the performance of our methods to the state-of-the-art tested with the same BioVid data as a binary classification problem, a pain intensity classification problem, and a pain intensity regression problem separately. At last, we discussed how parameter T in the attention module impacts recognition.

4.2. Evaluation metrics

In the binary classification problem, 'Pain' was defined as the positive class and 'No pain' was defined as the negative class. The data in SpaExp was imbalanced, where the negative class is the majority class with a ratio of 9–10 to the minority class. F1 score was used as the measure of overall test accuracy, which is the harmonic mean combining precision and recall (sensitivity) as illustrated below:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall, Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

where TP, FP, FN, and TN are true positive, false positive, false negative, and true negative.

When comparing with performance in external studies, balanced accuracy was calculated in binary classification due to data imbalance.

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2} \tag{7}$$

Further, when comparing with external studies as a pain intensity regression problem, mean absolute error (MAE) and root-mean-square

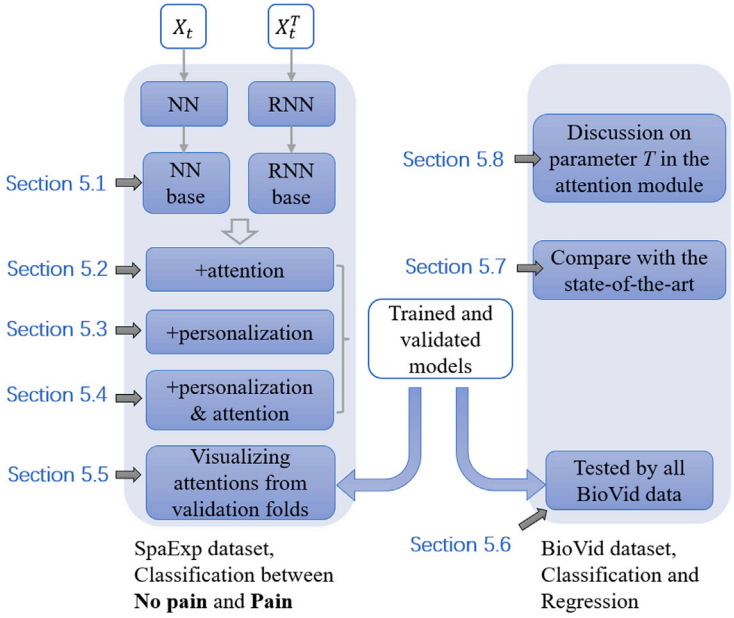


Fig. 6. Experiments roadmap.

error (RMSE). The lower the prediction errors, the better prediction performance. For n predictions $\hat{y}$ and their corresponding labels y:

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{9}$$

Additionally, R^2 and ICC(3,1) were used in pain recognition studies (Lopez-Martinez & Picard, 2018a) and (Lopez-Martinez et al., 2017b) as extra regression performance measures, and thus were included in this work as well for a comprehensive evaluation. R^2 score computes the coefficient of determination representing the proportion of variance of y has been explained. R^2 score 1 indicates perfect predictions; while 0 and negative scores indicate imperfect predictions.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{10}$$

The ICC is the intraclass correlation coefficient to assess the consistency of measures made by multiple raters in terms of intra-rater and inter-rater variability. ICC(3,1) is the two-way mixed single score, representing a number of fixed raters and a single measure of a single rater is considered. Ranging between 0 and 1, a higher value indicates a better inter-rater agreement.

4.3. Model training, validation, and testing

All models were trained, validated, and tested on a server with Intel Xeon CPU, 384 GB memory, and four NVIDIA Geforce RTX 2080Ti graphics cards. The models were implemented using the PyTorch 1.9.1 framework and Python 3.8.8.

4.3.1. Data partitioning for training, validation, and testing

All models were leave-subject-out cross-validated in this work. The SpaExp dataset was divided into a training set and a validation set in each fold. The size of each validation set was typically 1500, which was around 46000 for each training set.

The BioVid dataset was used for 4 purposes, and their sample sizes were:

- Testing data for binary classification of No pain and Pain: total size close to 132000 samples with an imbalanced class ratio of 3 to 4 in each subject;
- Data for single pain intensity detection as binary classification: typical sample sizes for training and testing were 3334 and 38;
- Data for 5 intensities classification and regression: typical sample sizes for training and testing were 8500 and 98;
- Data for 4 intensities regression: typical sample sizes for training and testing were 6800 and 79;

In the last three purposes, 20% samples of each training set were held out for model validation, where the best-validated model was kept for testing.

4.3.2. Training settings

In models built from SpaExp, network architectures described in Section 3 were used. Each model was trained using an Adam optimizer with default parameters and an initial learning rate of 0.01. In the total 100 epochs of training, the learning rate is multiplied by 0.1 after the first 50 epochs. For RNN models, the weights of the fully connected layer and MLP were initialized using kaiming initialization. The weights of RNN cells were initialized using an orthogonal initializer for a faster-converging speed (Yao et al., 2020). The dropout rate at the output of the second RNN layer was 0.5 to reduce overfitting. The class imbalance in the training dataset was adjusted to reach balance via weighted random sampling. Binary Cross Entropy was calculated from the batched output and labels in backpropagation, and the validation performance was checked every 10 training steps on whether the F1 score was improved. The batch size was equal to the size of each validation fold in within-dataset tests, which was 100 in cross-dataset tests. In within-dataset tests, results were reported as the average and standard deviation of the best validation performance in all folds.

In models built from BioVid, hidden layers with a 32-16-6 architecture were used instead considering its sample size. Adam optimizer was used. The learning rate was set as 0.01 for 50 epochs of training. Binary Cross Entropy loss was used for binary classification, Cross Entropy loss was used for multi-class classification, and Mean Squared Error loss was used for regression. The batch size equaled the size of each test fold. The validation performance was checked every 10 training steps on whether its accuracy or RMSE was improved. Best-validated models were next tested by each corresponding test fold, and results were reported as the average and standard deviation of test performance in all folds.

Table 4
SpaExp dataset, the performance of NN and RNN models, mean(std).

Model	Recall	Precision	F1 score
RUSBoost	0.827(0.228)	0.273(0.149)	0.394(0.180)
NN	0.762(0.185)	0.388(0.174)	0.495(0.170)
vanilla RNN	0.765(0.182)	0.455(0.177)	**0.547**(0.161)
GRU	0.781(0.168)	0.420(0.194)	0.517(0.162)
LSTM	0.753(0.206)	0.442(0.176)	0.532(0.159)

Table 5
SpaExp dataset, the performance of NN and RNN models with dynamic feature attention, mean(std)

Model	Recall	Precision	F1 score
NN base	0.762(0.185)	0.388(0.174)	0.495(0.170)
+ DynAvg	0.817⇑*(0.185)	0.388(0.186)	0.500⇑(0.182)
+ DynAvgMax	0.802⇑*(0.178)	0.387(0.174)	0.503⇑(0.171)
RNN base	0.765(0.182)	0.455(0.177)	0.547(0.161)
+ DynAvg	0.802⇑(0.189)	0.430(0.169)	0.538(0.161)
+ DynAvgMax	0.817⇑*(0.189)	0.422(0.185)	0.531(0.176)

⇑ denotes performance improvement to the base model.
⇑* denotes significant improvement (Wilcoxon signed-rank, $p < 0.05$).

Table 6
SpaExp dataset, the performance of NN and RNN model in fusion with pain sensitivity, mean(std).

Model	Recall	Precision	F1 score
NN base	0.762(0.185)	0.388(0.174)	0.495(0.170)
+ SensMeasure	0.677(0.219)	0.490⇑*(0.169)	0.545⇑*(0.164)
+ SensPredict	0.696(0.190)	0.509⇑*(0.171)	**0.562**⇑*(0.151)
+ SensHRV	0.681(0.209)	0.486⇑*(0.188)	0.546⇑*(0.174)
RNN base	0.760(0.206)	0.462(0.177)	0.549(0.165)
+ SensMeasure	0.792⇑(0.179)	0.458(0.174)	0.561⇑(0.171)
+ SensPredict	0.747(0.219)	0.502⇑*(0.197)	**0.562**⇑(0.163)

⇑ denotes performance improvement to the base model.
⇑* denotes significant improvement (Wilcoxon signed-rank, $p < 0.05$).

5. Results

5.1. Base models

Table 4 presents the performance of the NN and RNN models. In terms of the F1 score, RNN models better performed than the NN model, and they both outperformed the RUSBoost method used in our previous work (Jiang, 2019). RNNs show their advantage in detection precision (i.e., less wrongly detected pain). The vanilla RNN was with the highest F1 score and thus was used as the RNN base model in the following parts.

5.2. Dynamic feature attention

Dynamic feature attention was expected to improve pain detection sensitivity by increasing the impact of more sensitive features on pain state change. Table 5 shows the performance improved by the dynamic feature attention module. For NN models, both DynAvg (Eq. (1)) and DynAvgMax (Eq. (2)) brought significant improvement with recall and resulted in an improvement in the F1 score. For RNN models, the recall was improved in general, but precision was lowered so that the overall F1 score was not improved.

5.3. Fusion with personalized features

Table 6 presents the performance improvement by fusing each pain sensitivity representation. The F1 score of RNN base + SensHRV was inferior to its random control model and was therefore not presented in the Table. For NN models, all pain sensitivity representations improved the F1 score significantly by elevating the precision while sacrificing the recall. For RNN models, the F1 score was improved with SensMeasure and SensPredict but behaved differently on recall and precision. It is noticed that SensPredict improved NN base and RNN base models to the same F1 score (0.562).

5.4. Dynamic feature attention in combination with personalized features

Table 7 presents the performance of combined dynamic feature attention and personalized feature extended from Table 6. For NN models, all performance measures were generally improved by adding dynamic feature attention. For RNN models, the dynamic feature attention brought small oscillations in recall and precision in the opposite way, resulting in a slight drop in the overall F1 score. The best F1 score in Table 7 was 0.578(0.169) with NN+SensHRV+DynAvg, which was an 0.083 improvement to the NN base model.

5.5. Attention output visualization

To explain the behaviors of dynamic feature attention, we derived the output of the module from each validation fold and analyzed the weights.

5.5.1. Attention difference in heat and electrical tests

Our motivation for improving dynamic prediction was validated in the results. In the SpaExp dataset, each subject took two types of stimulation. The different experimental pain models lead to longer pain duration with heat tests than electrical ones (statistics in Table 1). Fig. 7 presents the average output of feature attention, and those were different between heat and electrical tests with a statistical difference. Behaviors of the dynamic feature attention were not identical in different models, considering NN + DynAvg in Fig. 7(a–c) and NN + DynAvgMax in Fig. 7(d). But both models paid more attention to fast changes (phasic component) of GSR in the electrical tests in Fig. 7(a, c, d), and NN + DynAvg paid more attention to slow changes (tonic component) in heat tests in Fig. 7(b).

5.5.2. The impact of personalized features

Results from both datasets show that the combination of dynamic feature attention and personalized SensHRV better performed than one with NN models. Figs. 8 gave examples showing changes in attention output after adding SensHRV. Comparing (c) and (d), shows that the use of SensHRV increased the difference among feature attention output weights. Comparing (a) and (b) shows that the use of SensHRV improved prediction precision. Meanwhile, we also observed the personalized features did not contribute equally in each subject.

As personal pain sensitivity mainly improved prediction precision, we checked the correlation between personal pain sensitivity measure (SensMeasure) and personal prediction precision in the SpaExp dataset (Fig. 9). It shows a positive linear correlation between personal SensMeasure and prediction precision in the base model, indicating subjects who are more sensitive to pain (smaller SensMeasure value), their 'No pain' state was more easily classified as 'Pain'. By fusing with personal pain sensitivity measures, such "wrong alarms" were corrected, especially with experimental tonic pain shown in Fig. 9(b).

5.6. Trained models tested with an external dataset

Cross-dataset tests were conducted with three binary classification models of pain detection, NN base, NN + DynAvg, and NN + SensMeasure + DynAvg. Models were trained and leave-subject-out validated with the SpaExp dataset and then tested with all subjects' data in the BioVid dataset. In the binary classification test, the four pain intensities were all labeled as pain.

We observed that pain was detected with a delay in external testing. We shifted the predictions to the left on the time axis and found a shift of −6 s reached the best F1 score and balanced accuracy. Test

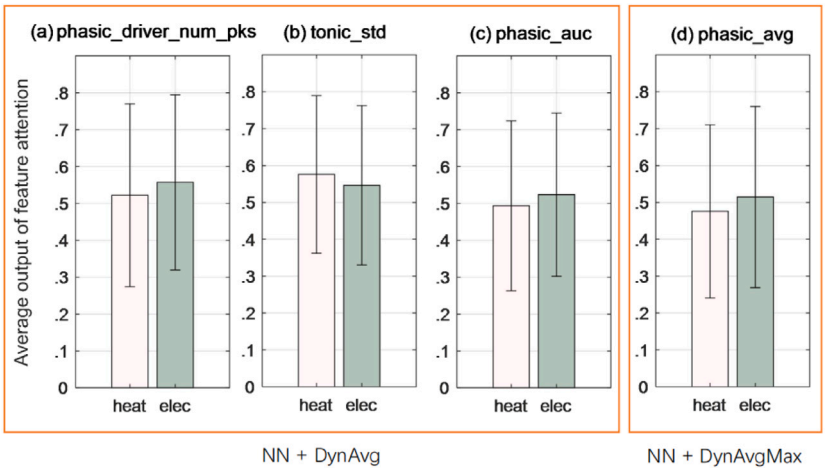


Fig. 7. SpaExp dataset, average attention with statistically significant difference (Wilcoxon signed-rank test, $p < 0.05$) between heat and electrical tests.

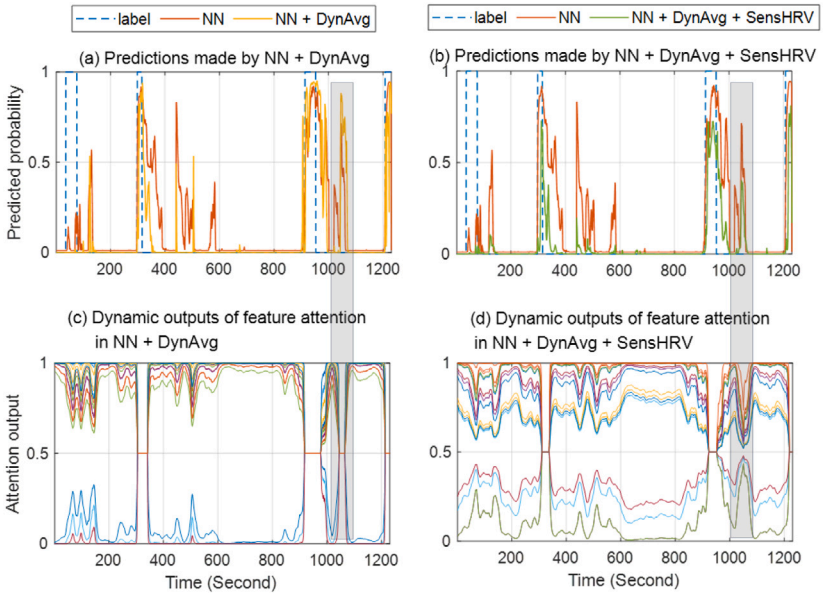


Fig. 8. SpaExp dataset, one subject prediction sensitivity/recall from 0.60 to 0.42, precision from 0.58 to 0.86.

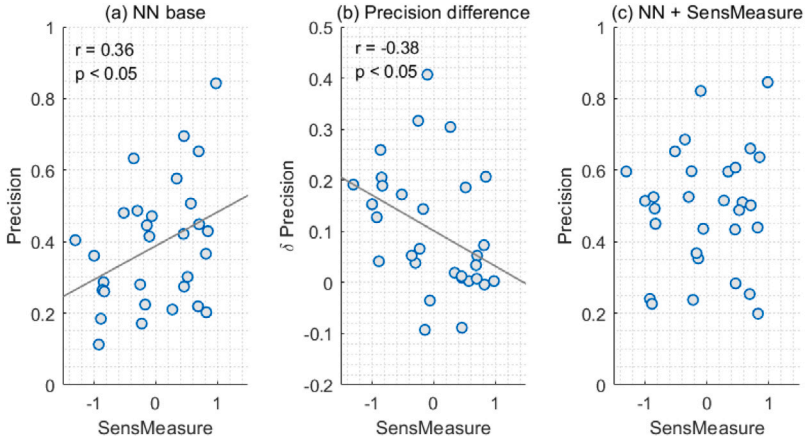


Fig. 9. SpaExp dataset, the distribution of personal prediction precision in 30 subjects, with Pearson's correlation between personal SensMeasure and personal prediction precision.

Table 7
SpaExp dataset, the performance of NN and RNN models with dynamic attention and personalized features, mean(std).

Model	Recall	Precision	F1 score	Balanced accuracy
NN+SensMeasure	0.677(0.219)	0.490(0.169)	0.545(0.164)	0.796(0.099)
NN+SensMeasure+DynAvg	0.688⇑(0.190)	0.516⇑(0.187)	0.565⇑(0.159)	0.806⇑(0.089)
NN+SensMeasure+DynAvgMax	0.706⇑(0.217)	0.521⇑(0.179)	0.563⇑(0.162)	0.805⇑(0.099)
NN+SensPredict	0.696(0.190)	0.509(0.171)	0.562(0.151)	0.803(0.088)
NN+SensPredict+DynAvg	0.696(0.179)	0.518⇑(0.197)	**0.569**⇑(0.170)	**0.807**⇑(0.094)
NN+SensPredict+DynAvgMax	0.682(0.187)	0.510⇑(0.170)	0.567⇑(0.165)	0.805⇑(0.092)
NN+SensHRV	0.681(0.209)	0.486(0.188)	0.546(0.174)	0.800(0.102)
NN+SensHRV+DynAvg	0.696⇑(0.201)	0.529⇑(0.192)	**0.578**⇑(0.169)	**0.812**⇑(0.103)
NN+SensHRV+DynAvgMax	0.700⇑(0.202)	0.518⇑(0.190)	0.568⇑(0.168)	0.810⇑(0.097)
RNN+SensMeasure	0.792(0.179)	0.458(0.174)	0.561(0.171)	0.843(0.097)
RNN+SensMeasure+DynAvg	0.786(0.196)	0.460(0.171)	0.559(0.166)	0.837(0.097)
RNN+SensMeasure+DynAvgMax	0.799(0.179)	0.452(0.176)	0.552(0.169)	0.843(0.096)
RNN+SensPredict	0.747(0.219)	0.502(0.197)	0.562(0.163)	0.827(0.108)
RNN+SensPredict+DynAvg	0.798(0.176)	0.460(0.186)	0.555(0.163)	0.844⇑(0.091)
RNN+SensPredict+DynAvgMax	0.771(0.195)	0.459(0.174)	0.555(0.167)	0.833⇑(0.098)

Table 8
Models built with SpaExp and tested with BioVid, predictions in the second level, without and with −6 s shift.

Tested model	F1 score	+shift	Balanced Acc	+shift
NN	0.084	0.259	0.455	0.561
+DynAtt	0.062	0.262	0.454	0.564
+SensMeasure+DynAtt	0.070	0.297	0.443	0.578

Table 9
The percentage (%) of successfully detected seconds after prediction shift.

Tested model	P1	P2	P3	P4	BLN
NN	11.9	17.9	23.9	44.6	87.3
+DynAtt	10.6	17.9	24.1	45.9	87.8
+SensMeasure+DynAtt	14.5	21.2	33.7	56.5	84.3

Table 10
The percentage (%) of successfully detected pain events after prediction shift.

Tested model	P1	P2	P3	P4
NN	17.5	26.4	39.2	59.2
+DynAtt	16.8	25.3	37.9	58.9
+SensMeasure+DynAtt	19.7	29.9	43.8	66.8

performances with three models without and with prediction shift were presented in Table 8. The models' generalization was proved as trained models could detect pain from unobserved features. The prediction delay could be explained by physiological response delay as presented in Fig. 1.

Moreover, the proposed method with dynamic attention and pain sensitivity fusion shows better generalization ability by achieving higher test performance. The solution with three HRV features as a pain sensitivity estimate was not transferable between the two datasets as explained in Jiang et al. (2022), and thus the test results were invalid as expected and were not presented.

We then looked into how well each pain intensity was detected in test results. Tables 9 and 10 present the accuracy of predicting each pain intensity at the second and event levels, respectively. They both show that higher intensity of pain was more likely detected by trained models.

5.7. Compare with the state-of-the-art

To date, the most widely tested pain dataset is the BioVid Heat Pain Database. The recognition problem was formulated as a binary classification of one intensity of pain against no pain baseline, multiple intensities classification, or pain intensity regression with four pain intensities or plus no pain. Pain data is imbalanced in nature. Most works balanced the five classes for training, validation, and testing. In this section, we first discussed how data selection impact classification with the same classifier or regressor.

Table 11
Performance of binary classification and regression in six data selection cases, Lopez-Martinez and Picard (2018a) as performance reference.

Case	Binary classification, Logistic regression, accuracy (%)			
	BLN vs P1	BLN vs P2	BLN vs P3	BLN vs P4
Reference	54.57(10.90)	59.40(12.61)	66.00(14.83)	74.21(17.54)
A	50.92(2.00)	51.84(3.60)	51.96(3.32)	52.41(3.75)
B	49.99(0.06)	50.01(0.07)	50.10(0.40)	50.26(0.70)
C	54.14(7.57)	55.89(7.71)	59.43(10.21)	65.52(11.91)
D	66.60(13.13)	68.63(12.13)	70.36(13.44)	70.43(12.26)
E	64.86(12.31)	68.42(12.58)	73.99(13.62)	81.26(13.79)
F	68.62(12.35)	70.44(13.49)	74.18(14.28)	79.51(15.91)
	Regression [0, 4], SVR RBF			
	MAE	RMSE	R^2	ICC
Reference	1.11(0.14)	1.33(0.14)	0.11(0.19)	NA
C	1.14(0.14)	1.57(0.20)	−0.14(0.27)	0.23(0.19)
D	1.02(0.17)	1.45(0.20)	0.02(0.25)	0.33(0.20)
E	0.96(0.19)	1.18(0.20)	0.28(0.24)	0.49(0.24)
F	1.00(0.19)	1.22(0.21)	0.23(0.24)	0.41(0.27)

5.7.1. Data selection due to class imbalance

We present binary classification results with logistic regression and regression performance with SVM-RBF regressor from five cases defined below. Results are presented in Table 11.

- Case A: Imbalanced class, GSR signal, BLNs and Px were from the same test, feature at time t was extracted from t−15 - t second and was labeled based on stimulus at time t.
- Case B: Imbalanced class, GSR signal, BLNs and Px were from the same test, feature at time t was extracted from t−6 - t second and was labeled based on stimulus at time t.
- Case C: Balanced class, GSR signal, BLNs and Px were from the same test, BLNs: −6–0 s, Px: −2–4 s.
- Case D: Balanced class, GSR signal, BLNs and Px were from the same test, BLNs: −8–2 s, Px: 0–6 s.
- Case E: Balanced class, GSR signal, BLNs were from the first 20 no pain segments, BLNs: −8–2 s, Px: 0–6 s.
- Case F: Balanced class, GSR and ECG fusion (feature 1–3, 5–16 in Table 2), BLNs were from the first 20 valid no pain segments, BLNs: features at -2nd second, Px: features at 6th second.

Case A and B were to compare the impact of window length for feature extraction with the GSR signal. Case A represents data treatments applied in this work and the feature extraction window length was shortened from 15 to 6 s. Balanced accuracy was calculated for

Table 12
Performance comparison — binary classification.

Publication	Model and signals	BLN vs P4 Accuracy (%)
Kächele et al. (2015)	Random Forest (RF), Physiological+Video	83.1(NA)
Werner et al. (2017)	RF + facial descriptor, Video	72.4(NA)
Lopez-Martinez and Picard (2018b)	Person-specific multi-task NN, Physiological	82.75(1.86)
Wang et al. (2020)	RNN-ANN, Physiological	83.30(NA)
Ours	NN baseline, Physiological	80.00(15.18)
	+SensMeasure+DynAtt	82.94(13.89)
	+SensHRV+DynAtt	**83.79**(13.88)
Thiam et al. (2021)	Convolutional auto-encoder with a gated latent representation (DDCAE), Physiological	83.99(15.58)
	DDCAE + channel attention, Physiological	84.20(13.70)
Ours	NN+SensMeasure+DynAtt+pretraining	**84.58**(13.28)

Table 13
Performance comparison — five class classification.

Publication	Model and signals	Accuracy (%)
Thiam et al. (2021)	DDCAE + channel attention, Physiological	35.44(8.66)
Werner et al. (2017)	RF, Video	30.80(NA)
Kächele et al. (2016)	Find the most similar subjects and specialize classifier, Physiological+Video	best 40.48(NA)
Ours	NN baseline, Physiological	35.76(10.86)
	+SensMeasure+DynAtt	**38.29**(9.65)
	+SensHRV+DynAtt	**39.24**(8.65)

resample class cases. The longer window length performed slightly better and the differences in recognizing different pain intensities were subtle.

Case C and D were to compare how pain data definition around stimulus onset impact classification performance, where Case C labeled pain due to stimulus end and D due to start. Case D reached significantly better recognition considering physiological response latency to stimulus.

Case D and E were to compare how data selection or down-sampling with baseline data impacts classification performance, where Case D used BLN before each pain segment. In contrast, E used the same BLNs for the four intensities. Results show that the performance gap between recognition of a weaker and stronger pain is minor in case D.

Table 11 shows the data selection step could dramatically impact reported results. Data selected from Case E present better outcomes than other tested cases with the GSR signal in both classification and regression. Considering GSR and ECG fusion in this work, we deployed Case F for further external comparison. For Case F, each class had close to 20 samples per person, and thus nearly 8700 samples in total for 5 classes of 87 subjects.

We compared our method to state-of-the-art (SOTA) ones in three formulated problems: binary classification, 5-class classification, and regression. The feature time series length T was set to be 6-second. Results were reported as average and standard deviation performance of leave-subject-out cross-validation. In the training phase of each fold, 20% randomly selected training samples were used for model tuning, where the best-trained model of 100 epochs was tested by the left-out subject. The best model in the trained phase was chosen according to the highest overall accuracy for classification and the lowest RMSE for regression.

5.7.2. Performance comparison and discussion

Table 12 13 and 14 present the comparisons. We chose BLN vs P4 to compare binary classification performance as a higher pain intensity is more easily recognized and may be considered a performance ceiling in recognizing a single pain intensity. Our proposed methods are denoted as "NN+SensMeasure+DynAtt" and "NN+SensHRV+Dynatt". We present the performance of the basic NN model as the corresponding performance baseline.

Binary classification between high pain and no pain usually can achieve high accuracy, as shown in Table 12. Our baseline NN model performed inferior to most SOTA works while our method was comparable to models of hybrid modalities and hybrid neural network architecture. By comparing works using the random forest classifier (Kächele et al., 2015; Werner et al., 2017), it shows that the physiological signal modal contributed more than the facial expression video modal in recognizing short heat pain. In the aspect of models, deep learning solutions for physiological signals achieved better recognition than random forest in general. Lopez-Martinez and Picard (2018b) built multi-task neural networks and customized deep layers for each subject, which was cross-validated in ten folds and resulted in much smaller performance deviation. Both RNN and CNN architecture have been tested for deep feature extraction. Wang et al. (2020) fused selected 50 deep RNN features with 21 handcrafted features from GSR, ECG, and facial EMG signals, and had similar recognition accuracy to our proposed method. Thiam et al. (2021) proposed a CNN-based feature extraction and auto-encoder latent augmentation solution and achieved 84.2% accuracy, showing the effect of proper data augmentation. Therefore, we tried model pre-training with SpaExp database, froze the first two layers and tuned the rest parameters with BioVid database. Recognition accuracy was improved to 84.58% by pretraining[1]

In 5-class classification (Table 13), the NN baseline model performed similarly to the data augmentation solution with physiological signals, which was improved to an accuracy of 39.24%. Kächele et al. (2016) reported a slightly higher accuracy when finding subjects similar to the test one and specialized classifier in the training phase. The highest accuracy was met after traversing the number of similar subjects and recognition with the k nearest neighbor classifier, where the personalized training dataset constituted by 67 most similar subjects achieved the optimal results. By contrast, our proposed method reached similar personalized optimization from a different personalizing path.

Reported results of regression were either for pain intensities [1, 4] or plus the no pain baseline [0, 4] (Table 14). For pain intensity regression, the optimization was subtle with prediction errors close to the hybrid modality random forest solution (Kächele et al., 2015). The personalized solution integrating feature clustering and multi-task neural networks (Lopez-Martinez et al., 2017b) shows a relatively small prediction error as MAE. Comparatively, our proposed optimization was more effective when including no pain.

[1] Codes are available at https://github.com/jiangmz73/PainNet.

Table 14
Performance comparison — regression.

Publication	Model and signals	MAE	RMSE	R^2	ICC(3,1)
Regression [0, 4]					
Kächele et al. (2017)	Find similar subjects and specialize classifier, Physiological+Video	0.99(0.17)	1.16(0.18)	NA	NA
Lopez-Martinez and Picard (2018a)	LSTM-NN, Physiological	1.05(0.15)	1.29(0.16)	0.24(0.19)	NA
Thiam et al. (2021)	DDCAE, Physiological	0.97(0.19)	1.16(0.21)	NA	NA
Kächele et al. (2016)	Find similar subjects and specialize classifier, Physiological+Video	best 0.892	NA	NA	NA
Ours	NN baseline, Physiological	0.98(0.18)	1.17(0.19)	0.29(0.23)	0.44(0.24)
	+SensMeasure+DynAtt	**0.93**(0.18)	**1.12**(0.20)	**0.34**(0.23)	**0.50**(0.22)
	+SensHRV+DynAtt	**0.93**(0.19)	**1.12**(0.21)	**0.34**(0.23)	**0.51**(0.23)
Regression [1, 4]					
Kächele et al. (2015)	Random Forest, Physiological+Video	0.84(0.13)	0.98(0.14)	NA	NA
Lopez-Martinez et al. (2017b)	Personalized clusters multi-task NN, Physiological+Video	0.77(NA)	1.15(NA)	NA	0.31
Ours	NN baseline, Physiological	0.85(0.12)	1.00(0,13)	0.18(0.21)	0.31(0.22)
	+SensMeasure+DynAtt	**0.85**(0.13)	**1.00**(0.14)	**0.19**(0.22)	**0.33**(0.23)
	+SensHRV+DynAtt	**0.84**(0.13)	**1.00**(0.13)	**0.19**(0.21)	**0.34**(0.22)

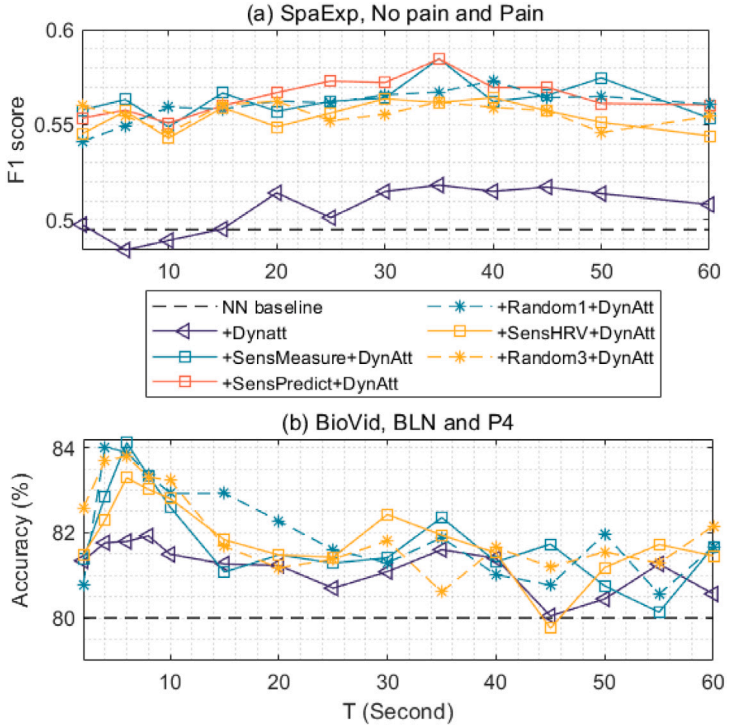


Fig. 10. Attention window length T and performance — binary classification.

5.8. Parameter T in the attention module

Theoretically, the optimal length T would correspond to the time window length that best captures the change of physiological features. We checked the prediction performance of models when T was set differently from 2 to 60-second. Figs. 10 and 11 show that 4 and 6-second T attention learned physiological dynamics best with the BioVid dataset in binary classification, 5-class classification, and regression problems, which was 35-second with the SpaExp dataset binary classification.

The difference in optimal T reflects that the trainable attention module was able to learn the characteristics of dynamic change and for more accurate predictions. On the other hand, lengths incompatible with pain response patterns in data may bring adverse effects on optimization (e.g., $T < 15$ with SpaExp in Fig. 10(a) and $T > 30$ with BioVid regression in Fig. 11 (b–e)). Therefore, attention in selective length or multi-scale may help cope with more complex pain cases such as varied pain duration in data.

On top of the attention, fusion with a pain sensitivity vector or random vector before the output layer further improved predictions steadily across attention lengths and recognition problems. Personalized vectors represented by pain sensitivity measures and estimates contributed more than random vectors with long pain stimulation in SpaExp at 25, 30 and 35-second in Fig. 10. For pain intensity regression in BioVid at 4 and 6-second, pain sensitivity measures and estimates contributed more than random vectors; while the impact of personalized vectors was close to random vectors in classification. The improvement brought by fusion with a random vector may be explained by improved model generalization via injecting noise into the input and hidden layers which can be considered as data augmentation (Goodfellow, Bengio, & Courville, 2016). Additional improvement was made by SensMeasure, SensPredict, and SensHRV with long pain in SpaExp, which was SensMeasure with short pain binary classification and regression in BioVid.

To check the consistency in using SensMeasure and SensHRV, we checked Kendall's correlation between the performance curves and found significant trend correlations in F1 scores between SensMeasure+Dynatt and SensHRV+Dynatt (Kendall's $r = 0.47$, $p < 0.05$) and in precision (Kendall's $r = 0.37$, $p = 0.05$). Comparatively, correlations

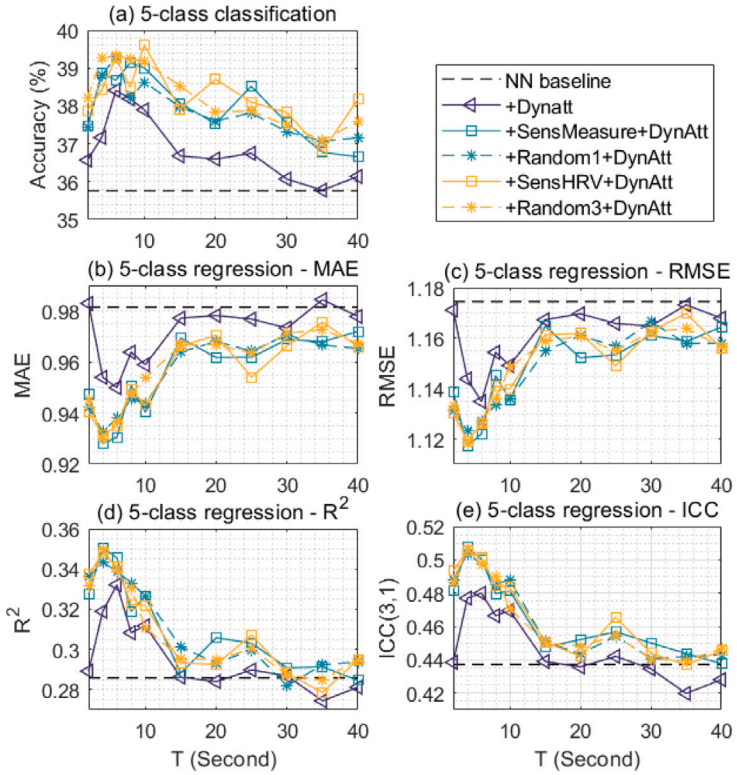


Fig. 11. Attention window length T and performance — BioVid, 5-class classification and regression.

were much weaker between sensitivity representation curves and random vector curves, showing consistency between fusing SensMeasure and SensHRV for long pain. While for short pain in BioVid, all accuracy curves with binary classification were correlated to Dynatt, showing the dominant impact of Dynatt rather than the late fusion.

The best-reached performance observed in this part was:

- F1-score: 0.585(0.169); SpaExp @35-second, SensMeasure+Dynatt
- Accuracy: 84.11(14.05)%; BioVid @6-second, BLN and P4, SensMeasure+Dynatt
- Accuracy: 39.61(8.69)%; BioVid @10-second, 5-class, SensHRV+Dynatt
- MAE: 0.93(0.18), RMSE: 1.12(0.20), R^2: 0.35(0.23), ICC(3,1): 0.51(0.23); BioVid @4-second, [0, 4] regression, SensMeasure+Dynatt

6. Discussion

We introduced and evaluated a new approach to automatic pain assessment. Two pain assessment datasets were involved in its development and evaluation, and results show the proposed approach outperforms state-of-the-art methods. The attention mechanism was leveraged as a soft feature selection for adaptive recognition of pain. In relatively complex conditions with variant pain durations, the feature attention dynamically along the time axis enhanced recognition sensitivity. Additionally, for simple neural networks and tonic pain, results show that fusing with pain sensitivity information before output improved prediction precision by rejecting false positives compared to models without personalization fusion.

The dynamic feature attention for multi-modal physiological time series was borrowed from image's channel attention due to its lightweight and easy-to-integrate into neural networks. We adopted the attention module to simple neural networks for time series attention in the context of pain detection. Adding the attention mechanism at the neural networks input end help with more sensitive detection and therefore is recommended in multi-modal fusion. The dynamics of attention were limited to a sliding time window with a fixed length, which could be expanded to multi-time scale attention for more complex pain cases in long-term monitoring and when fusing with more variant features.

It has been known that the way of cross-validation or data split for training and testing would impact the reported performance. The cross-dataset tests in this work also show the impact of data labeling and selection for models. Physiological response delay and recovery time to pain stimuli are comparable to very short pain stimulation, thus pain segment definition would impact recognition performance. Pain data is usually class imbalanced and is treated to be balanced especially in the training phase via resampling with minority class up-sampling, majority class down-sampling, or both. Our results show that choosing No Pain baselines differently in down-sampling also impacted the performance of the same recognition method, which may impact performance comparison across studies. In comparison, up-sampling or generation of minority class data could be less biased avoiding reporting overoptimistic performance. Synthesizing minority class data may leverage the classic SMOTE technique, or techniques with autoencoder architecture for data augmentation.

This work limits to experimental and nociceptive pain detection from ECG and GSR signals. The analysis concentrated on their dynamics in time, and therefore NN and RNN architectures were discussed in joint with attention in one dimension. In future work, the physiological signal modality can be ex- tended with EEG measurement (Chen et al., 2022) as a cross-reference where the explainability of pain processing can be enhanced. Also, with extended signal modality, the attention can be extended to more dimensions, for example, spatial and temporal (Jia et al., 2020) for interpretable analysis and recognition adapted to different pain cases.

7. Conclusion

Recognizing subjective pain from observable pain behaviors is challenging due to the complexity of pain and inexplicit patterns with

inter-individual differences. For continuous pain recognition from GSR and ECG recordings, we proposed solutions enabling neural networks to adapt to variations in pain duration and take in pain sensitivity information as a personalized factor. Within- and cross-datasets test results show that dynamic attention inspected from a proper window before the time of interest helped prompt pain pattern recognition performance with both long and short pain. Fusing pain sensitivity information further elevated pain detection performance more with long pain than short pain as a step towards personalized recognition. The proposed method was able to solve the challenge of pain complexity in part in pain pattern recognition, which is worth further validation with more abundant data.

CRediT authorship contribution statement

Mingzhe Jiang: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Visualization. **Riitta Rosio:** Conceptualization, Investigation, Data curation. **Sanna Salanterä:** Conceptualization, Investigation, Supervision. **Amir M. Rahmani:** Conceptualization, Investigation, Supervision. **Pasi Liljeberg:** Conceptualization, Investigation, Supervision. **Daniel S. da Silva:** Writing – review & editing. **Victor Hugo C. de Albuquerque:** Writing – review & editing. **Wanqing Wu:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgment

This work was supported by the Guangzhou Science and Technology Planning Project (202003000040).

References

Appelhans, B. M., & Luecken, L. J. (2008). Heart rate variability and pain : Associations of two interrelated homeostatic processes. *Biological Psychology*, *77*, 174–182. http://dx.doi.org/10.1016/j.biopsycho.2007.10.004.

Aung, M. S., Kaltwang, S., Romera-Paredes, B., Martinez, B., Singh, A., Cella, M., et al. (2016). The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal EmoPain dataset. *IEEE Transactions on Affective Computing*, *7*(4), 435–451. http://dx.doi.org/10.1109/TAFFC.2015.2462830.

Breivik, H., Borchgrevink, P.-C., Allen, S.-M., Rosseland, L.-A., Romundstad, L., Breivik Hals, E., et al. (2008). Assessment of pain. *British Journal of Anaesthesia*, *101*(1), 17–24. http://dx.doi.org/10.1093/bja/aen103.

Casti, P., Mencattini, A., Filippi, J., D'Orazio, M., Comes, M. C., Di Giuseppe, D., et al. (2020). A personalized assessment platform for non-invasive monitoring of pain. In *2020 IEEE international symposium on medical measurements and applications* (pp. 1–5). IEEE.

Chen, D., Zhang, H., Kavitha, P. T., Loy, F. L., Ng, S. H., Wang, C., et al. (2022). Scalp EEG-based pain detection using convolutional neural network. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *30*, 274–285. http://dx.doi.org/10.1109/TNSRE.2022.3147673.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Regularization for deep learning. In *Deep learning* (pp. 216–261). MA, USA: MIT press Cambridge.

Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., & Citi, L. (2016). CvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, *63*(4), 797–804. http://dx.doi.org/10.1109/TBME.2015.2474131.

Hallstrom, A. P., Stein, P. K., Schneider, R., Hodges, M., Schmidt, G., & Ulm, K. (2004). Structural relationships between measures based on heart beat intervals: Potential for improved risk assessment. *IEEE Transactions on Biomedical Engineering*, *51*(8), 1414–1420. http://dx.doi.org/10.1109/TBME.2004.828049.

Herr, K., Coyne, P. J., Ely, E., Gélinas, C., & Manworren, R. C. (2019). Pain assessment in the patient unable to self-report: Clinical practice recommendations in support of the ASPMN 2019 position statement. *Pain Management Nursing*, *20*(5), 404–417. http://dx.doi.org/10.1016/j.pmn.2019.07.005.

Hohenschurz-Schmidt, D. J., Calcagnini, G., Dipasquale, O., Jackson, J. B., Medina, S., O'Daly, O., et al. (2020). Linking pain sensation to the autonomic nervous system: The role of the anterior cingulate and periaqueductal gray resting-state networks. *Frontiers in Neuroscience*, *14*(February), http://dx.doi.org/10.3389/fnins.2020.00147.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Jia, Z., Lin, Y., Cai, X., Chen, H., Gou, H., & Wang, J. (2020). SST-EmotionNet: Spatial-spectral-temporal based attention 3D Dense Network for EEG Emotion Recognition. In *MM 2020 - Proceedings of the 28th ACM international conference on multimedia, no. October* (pp. 2909–2917). http://dx.doi.org/10.1145/3394171.3413724.

Jiang, M. (2019). *Automatic pain assessment by learning from multiple biopotentials* (Ph.D. thesis), (248), University of Turku, URL: http://urn.fi/URN:ISBN:978-952-12-3889-5.

Jiang, M., Mieronkoski, R., Syrjälä, E., Anzanpour, A., Terävä, V., Rahmani, A. M., et al. (2019). Acute pain intensity monitoring with the classification of multiple physiological parameters. *Journal of Clinical Monitoring and Computing*, *33*(3), 493–507. http://dx.doi.org/10.1007/s10877-018-0174-8.

Jiang, M., Wu, W., Wang, Y., Rahmani, A. M., Salanerä, S., & Liljeberg, P. (2022). Personal pain sensitivity prediction from ultra-short-term resting heart rate variability. In *2022 44th annual international conference of the IEEE engineering in medicine & biology society* (pp. 1137–1140).

Kächele, M., Amirian, M., Thiam, P., Werner, P., Walter, S., Palm, G., et al. (2017). Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evolving Systems*, *8*(1), 71–83. http://dx.doi.org/10.1007/s12530-016-9158-4.

Kächele, M., Thiam, P., Amirian, M., Schwenker, F., & Palm, G. (2016). Methods for person-centered continuous pain intensity assessment from bio-physiological channels. *IEEE Journal of Selected Topics in Signal Processing*, *10*(5), 854–864. http://dx.doi.org/10.1109/JSTSP.2016.2535962.

Kächele, M., Thiam, P., Amirian, M., Werner, P., Walter, S., Schwenker, F., et al. (2015). Multimodal data fusion for person-independent, continuous estimation of pain intensity. In *Multiple classifier systems* (pp. 275–285).

Ledowski, T. (2019). Objective monitoring of nociception: a review of current commercial solutions. *British Journal of Anaesthesia*, *123*(2), e312–e321. http://dx.doi.org/10.1016/j.bja.2019.03.024.

Liu, D., Peng, F., Shea, A., & Picard, R. (2017). DeepFaceLIFT: Interpretable personalized models for automatic estimation of self-reported pain. *Journal of Machine Learning Research*, *66*, 1–16, arXiv:1708.04670.

Lopez-Martinez, D., & Picard, R. (2018a). Continuous pain intensity estimation from autonomic signals with recurrent neural networks. In *Annual international conference of the IEEE engineering in medicine and biology society* (pp. 5624–5627). http://dx.doi.org/10.1109/EMBC.2018.8513575.

Lopez-Martinez, D., & Picard, R. (2018b). Multi-task neural networks for personalized pain recognition from physiological signals. In *2017 7th international conference on affective computing and intelligent interaction workshops and Demos* (pp. 181–184). http://dx.doi.org/10.1109/ACIIW.2017.8272611, arXiv:1708.08755.

Lopez-Martinez, D., Rudovic, O., & Picard, R. (2017a). Personalized automatic estimation of self-reported pain intensity from facial expressions. In *IEEE computer society conference on computer vision and pattern recognition workshops, no. 2017-July* (pp. 2318–2327). http://dx.doi.org/10.1109/CVPRW.2017.286, arXiv:1706.07154.

Lopez-Martinez, D., Rudovic, O., & Picard, R. (2017b). Physiological and behavioral profiling for nociceptive pain estimation using personalized multitask learning. In *31st conference on neural information processing systems, no. i* (pp. 1–6). arXiv:1711.04036.

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., Chew, S., & Matthews, I. (2012). Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database. *Image and Vision Computing*, *30*(3), 197–205. http://dx.doi.org/10.1016/j.imavis.2011.12.003.

Mogil, J. S. (2021). Sources of individual differences in pain. *Annual Review of Neuroscience*, *44*, 1–25.

Raja, S. N., Carr, D. B., Cohen, M., Finnerup, N. B., Flor, H., Gibson, S., et al. (2020). The revised international association for the study of pain definition of pain: concepts, challenges, and compromises. *Pain*, *161*(9), 1976–1982. http://dx.doi.org/10.1097/j.pain.0000000000001939.

Rudovic, O., Tobis, N., Kaltwang, S., Schuller, B., Rueckert, D., Cohn, J. F., et al. (2021). *Personalized federated deep learning for pain estimation from face images*. Association for Computing Machinery, URL: http://arxiv.org/abs/2101.04800.

Ruscheweyh, R., Marziniak, M., Stumpenhorst, F., Reinholz, J., & Knecht, S. (2009). Pain sensitivity can be assessed by self-rating: Development and validation of the pain sensitivity questionnaire. *Pain*, *146*(1–2), 65–74. http://dx.doi.org/10.1016/j.pain.2009.06.020.

Spisak, T., Kincses, B., Schlitt, F., Zunhammer, M., Schmidt-Wilcke, T., Kincses, Z. T., et al. (2020). Pain-free resting-state functional brain connectivity predicts individual pain sensitivity. *Nature Communications*, *11*(1), http://dx.doi.org/10.1038/s41467-019-13785-z.

Syrjälä, E., Jiang, M., Pahikkala, T., Salanterä, S., & Liljeberg, P. (2019). Skin conductance response to gradual-increasing experimental pain. In *Proceedings of the annual international conference of the IEEE engineering in medicine and biology society* (pp. 3482–3485). http://dx.doi.org/10.1109/EMBC.2019.8857776.

Thiam, P., Hihn, H., Braun, D. A., Kestler, H. A., & Schwenker, F. (2021). Multi-modal pain intensity assessment based on physiological signals: A deep learning perspective. *Frontiers in Physiology*, *12*(September), http://dx.doi.org/10.3389/fphys.2021.720464.

Tracy, L. M., Jarczok, M. N., Ellis, R. J., Bach, C., Hillecke, T. K., Thayer, J. F., et al. (2018). Heart rate variability and sensitivity to experimentally induced pain: A replication. *Pain Practice*, *18*(5), 687–689. http://dx.doi.org/10.1111/papr.12652.

Tracy, L. M., Koenig, J., Georgiou-Karistianis, N., Gibson, S. J., & Giummarra, M. J. (2018). Heart rate variability is associated with thermal heat pain threshold in males, but not females. *International Journal of Psychophysiology*, *131*(January), 37–43. http://dx.doi.org/10.1016/j.ijpsycho.2018.02.017.

Vijayakumar, V., Case, M., Shirinpour, S., & He, B. (2017). Quantifying and characterizing tonic thermal pain across subjects from EEG data using random forest models. *IEEE Transactions on Biomedical Engineering*, *64*(12), 2988–2996. http://dx.doi.org/10.1109/TBME.2017.2756870.

Walter, S., Gruss, S., Ehleiter, H., Tan, J., Traue, H. C., Werner, P., et al. (2013). The BioVid heat pain database: Data for the advancement and systematic validation for an automated pain recognition system. In *IEEE international conference on cybernetics* (pp. 128–131).

Wang, R., Xu, K., Feng, H., & Chen, W. (2020). Hybrid RNN-ANN based deep physiological network for pain recognition. In *2020 42nd annual international conference of the IEEE engineering in medicine & biology society* (pp. 5584–5587). IEEE.

Werner, P., Al-Hamadi, A., Gruss, S., & Walter, S. (2019). Twofold-multimodal pain recognition with the X-ITE pain database. In *2019 8th international conference on affective computing and intelligent interaction workshops and demos* (pp. 290–296). IEEE, http://dx.doi.org/10.1109/ACIIW.2019.8925061.

Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S., & Traue, H. C. (2017). Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing*, *8*(3), 286–299. http://dx.doi.org/10.1109/TAFFC.2016.2537327.

Werner, P., Al-Hamadi, A., Niese, R., Walter, S., Gruss, S., & Traue, H. C. (2014). Automatic pain recognition from video and biomedical signals. *Proceedings - International Conference on Pattern Recognition*, 4582–4587. http://dx.doi.org/10.1109/ICPR.2014.784.

Werner, P., Lopez-Martinez, D., Walter, S., Al-Hamadi, A., Gruss, S., & Picard, R. W. (2019). Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, *13*(1), 530–552. http://dx.doi.org/10.1109/TAFFC.2019.2946774.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision* (pp. 3–19).

Yao, Q., Wang, R., Fan, X., Liu, J., & Li, Y. (2020). Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network. *Information Fusion*, *53*(June 2019), 174–182. http://dx.doi.org/10.1016/j.inffus.2019.06.024.

Zhang, X., Jiang, M., Wu, W., & de Albuquerque, V. H. C. (2021). Hybrid feature fusion for classification optimization of short ECG segment in IoT based intelligent healthcare system. *Neural Computing and Applications*, 1–15. http://dx.doi.org/10.1007/s00521-021-06693-1.