Noninvasive and Objective Near Real-Time Detection of Pain Changes in Fluctuating Pain

Leonard Visser[1], Christian Büchel[1]

[1]Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Germany

Corresponding Author: Leonard Visser
Department of Systems Neuroscience
University Medical Center Hamburg-Eppendorf
20246 Hamburg, Germany
E-mail: visser.ninhamburg@gmail.com

Abstract

Chronic pain involves natural intensity fluctuations that patients cannot control, contributing to learned helplessness and functional impairment. Detecting spontaneous pain decreases could enable precisely timed interventions that enhance perceived control. However, existing research on objective pain assessment has focused primarily on estimating static intensity from short, predictable stimuli rather than detecting moment-to-moment changes in ongoing pain. This study investigated whether pain decreases can be detected objectively using easily obtainable physiological signals during fluctuating pain. We recorded multiple physiological signals (8-channel EEG, electrodermal activity, heart rate, pupil diameter, facial expressions) from 42 healthy participants (M_{age} = 26.2 years, SD = 5.1) during calibrated tonic heat pain stimulation on the left forearm with unpredictable intensity fluctuations (0–70 on visual analogue scale; twelve 3-minute trials). Temperature changes lasted 5–20 seconds. Using

minimal preprocessing suitable for real-time applications, we trained state-of-the-art deep learning models to classify pain decreases versus non-decreases from brief temporal windows on a held-out test set (9 participants). Combining electrodermal activity, heart rate, and pupil diameter yielded optimal classification performance using a transformer-based architecture (AUROC = 0.854, accuracy = 76.8%). Electrodermal activity emerged as the most informative single predictor. Continuous stream analysis demonstrated median detection latency of 5.75 seconds with 70.4% sensitivity, reducible to 4.25 seconds at the cost of increased false positives. Results indicate that electrodermal activity and heart rate enable straightforward practical deployment, while highly variable signals such as EEG and facial expressions require personalized fine-tuned models. These findings establish a basis for closed-loop interventions targeting spontaneous pain changes.

Summary

Deep learning models detect pain decreases using electrodermal activity, heart rate, and pupil diameter (AUROC=0.854, 5.75s latency), facilitating precisely timed interventions for fluctuating tonic pain.

Keywords

Tonic pain, Physiological signals, Electrodermal activity, Deep learning, Multivariate time series

Introduction

Expectations can effectively alleviate acute pain [11,52], though they show smaller effects in chronic pain patients [25]. Besides expectations, perceived control is another critical psychological factor in pain modulation [20]. However, chronic pain impairs patients' sense of control over their symptoms, contributing to functional disability and creating a sense of "learned helplessness" [39,45]. Restoring control over pain would offer therapeutic benefit [4]. However, chronic pain patients lack control over their pain by definition. As an alternative, the so-called "illusion of control" could be beneficial. The illusion of control refers to the tendency for individuals to perceive influence over outcomes even when the causal link between their actions and those outcomes is minimal or nonexistent [33].

This forms the basis of a recently proposed approach suggesting that patients may be able to regain at least some degree of control over their pain [10]. The core idea is that if one can detect a spontaneous reduction in their (naturally fluctuating) pain and then allow them to treat themselves while the pain diminishes, this could be perceived as a form of causal intervention. Given that the pain is already decreasing, a sham treatment (placebo intervention) would be sufficient. However, as a first step, this requires an instrument that can objectively and noninvasively track ongoing pain levels and in particular can reliably detect pain decreases.

Previous research on objective pain assessment has largely focused on brief, discrete, and static pain stimuli with clearly defined onsets and offsets, aiming to estimate absolute pain intensity from concurrently recorded physiological signals [17,18,21,36,38,49,50,53]. However, this approach does not adequately capture the moment-to-moment fluctuations characteristic of ongoing, tonic pain. This is particularly important because many chronic pain conditions exhibit natural fluctuations that lie outside the patient's control [16].

In this study, we investigated whether pain dynamics in healthy participants can be assessed objectively using various easily obtainable physiological measures. We focused on electrodermal activity (EDA), heart rate, pupil diameter, electroencephalogram (EEG), and facial expressions during controlled exposure to fluctuating, unpredictable tonic heat pain stimuli. Through systematic evaluation of different feature sets and deep learning model architectures, we then identified optimal approaches for detecting pain dynamics in particular

pain decreases. Finally, we validated the performance of our model through a continuous data stream analysis, where we applied a sliding window approach to approximate real-time application conditions.

Methods

Participants

Eligibility Criteria

Participants were eligible if they were healthy, right-handed, between 18 and 40 years of age, and had a body mass index (BMI) between 18 and 30 kg/m². All participants possessed full legal capacity to provide informed consent and demonstrated adequate German language proficiency for following complex instructions.

Exclusion criteria encompassed chronic pain conditions (Patient Health Questionnaire PHQ-15 [29], cutoff 10), elevated depressive symptoms (Beck Depression Inventory II BDI-II [22], cutoff 14), and other psychiatric, neurological, or cardiological conditions. Additional exclusions addressed factors that could interfere with physiological measurements: dermatological abnormalities on the left inner forearm (scarring, tattoos, skin conditions), distinctive hair characteristics (natural baldness, completely shaved head, dreadlocks, braided hairstyles), recent Botox facial treatments (within 6 months), facial tattoos or piercings, respiratory conditions (cold symptoms, pollen allergies), required glasses or hard contact lenses, pregnancy or breastfeeding, and regular medication use (except contraceptives).

Participant Characteristics

We enrolled 50 healthy right-handed participants with normal or corrected-to-normal vision (via soft contact lenses). Participants received monetary compensation for their time. Prior to participation, participants reported abstaining from pain medication and cannabis or similar substances for 24 hours, from alcohol for 12 hours, and from caffeine for 2 hours.

Participant Flow and Final Sample

Five participants were excluded from data analysis (due to inattentiveness, pain threshold issues, or BMI requirements). From the remaining 45 participants, 69 of a total of 540 trials were excluded for technical reasons (EDA sensor issues, thermode malfunctions, EEG record-

ing problems) or protocol violations as detailed in Figure 1. The final dataset thus comprised 471 valid trials from 42 participants (23 women; age range from 18 to 39 years; mean 26.2).

The Ethics Committee of the Medical Chamber Hamburg approved the study (procedure number 2023-101089-BO-ff), and all participants provided written informed consent. Data collection took place from June to December 2024 at the Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf.

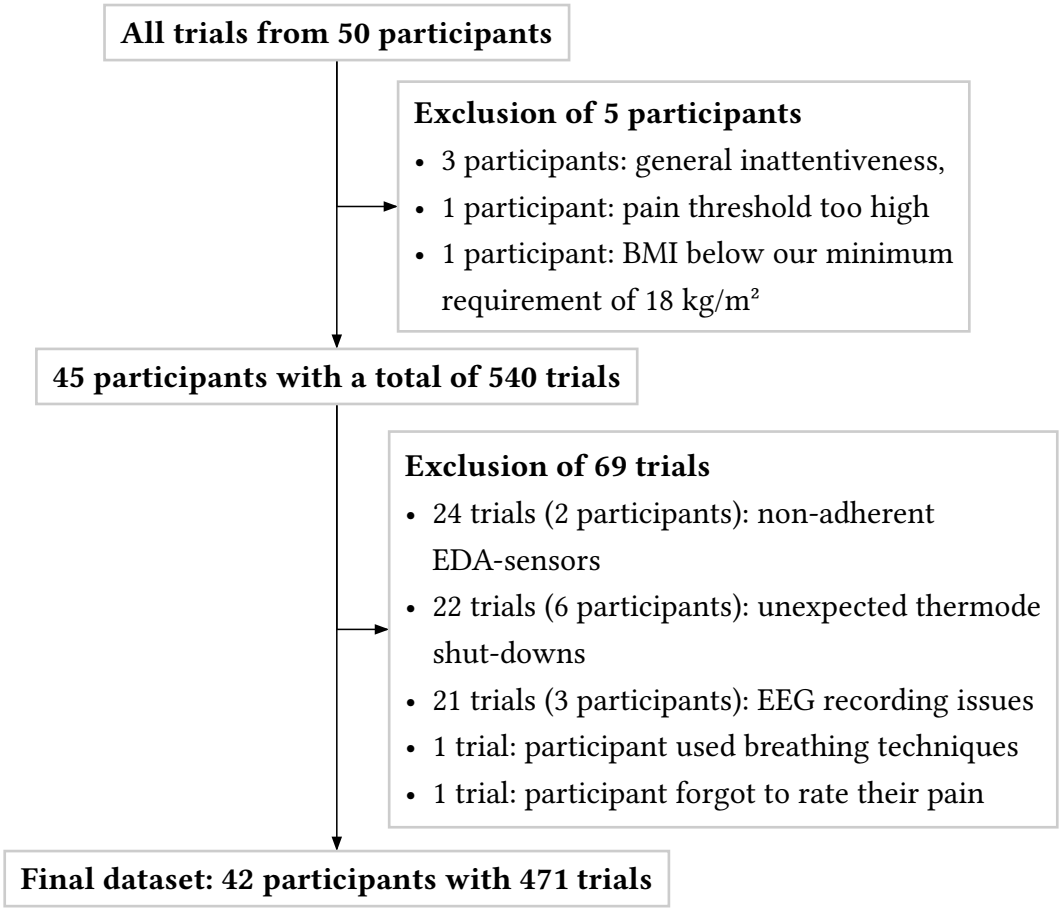

Figure 1: Flow diagram of participants and experimental trials. Exclusion criteria were applied in two stages: first at the participant level for eligibility requirements, then at the trial level for data quality and protocol adherence.

Pain Stimulus

To evoke fluctuations in tonic pain, we used sinusoidal functions to create controlled temperature curves of 180 s length with unpredictable patterns. Each stimulus function consisted of 5 ascending and 5 descending segments with different amplitudes and periods (5 to 20 s). Individual segments were generated using half-cycle cosine waveforms following the equation

$$y(t) = A \times \cos(\pi f t) + y_0 - A,$$

where A is amplitude, f is frequency, and y_0 is the y-intercept. These segments were joined in an alternating up-and-down pattern (switching between positive and negative values) to form a smooth, continuous temperature curve of 145 s. To further prevent anticipation effects, three temperature plateaus were added: two plateaus of 15 s length into ascending segments, one plateau of 5 s length at the point of a local minimum for a total length of 180 s (see Suppl. Figure S1 for details). All pain stimuli included three uniform major pain decreases to facilitate our primary research focus on decreasing pain detection (see Figure 2 for an example).

Each participant received 12 unique 180 s stimuli (see Suppl. Figure S2 for a visualization of all stimulus curves), with pain intensities fluctuating between pain threshold and strong pain (70 out of 100 on a visual analogue scale (VAS)). While the stimulus patterns remained consistent across participants, we individually calibrated temperatures (see below for details) to each participant's pain threshold and VAS 70. The presentation order of the stimuli was randomized.

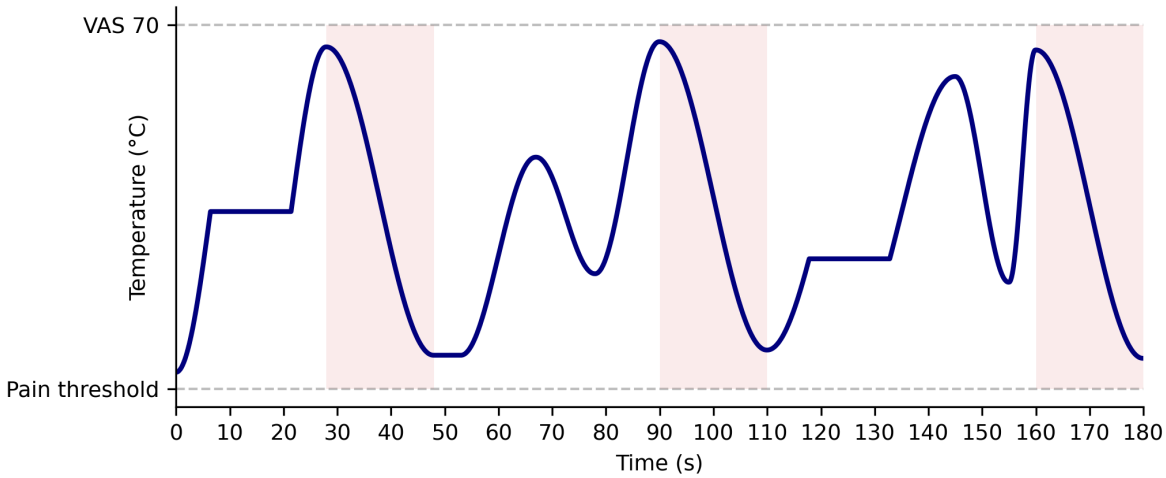

Figure 2: Example of a temperature curve for one trial. Each trial lasted 3 minutes with 12 distinct temperature curves throughout the experiment. The red patches highlight the uniform major decreasing intervals of which each stimulus had three. These intervals were used for the detection of pain decreases in the binary classification.

Procedure

We collected multiple psychophysiological measures along with online pain ratings while applying the heat stimuli. Psychophysiological measures included electrodermal activity (EDA), pupillometry, photoplethysmography for heart rate (HR), electroencephalogram (EEG), and facial expressions through video recordings. The experiment utilized a within-subject design and lasted approximately 2.5 hours per participant. All participants were tested individually by the same experimenter in a controlled laboratory environment with standardized illumination. Participants sat in a comfortable chair with a computer monitor at eye level, positioned approximately 65 cm away. The experimenter provided standardized instructions with limited interaction. The experiment consisted of five phases: (1) pre-experiment questionnaires, (2) experiment set-up, (3) pain calibration, (4) pain measurement, and (5) post-experiment questionnaires. These phases are outlined in the following sections.

Pre-Experiment Questionnaires

Following a brief introduction and informed consent procedure, participants completed psychological screening questionnaires: a general questionnaire assessing socioeconomic status, the BDI-II, and the PHQ-15. To enable comparison of mood states before and after the heat stimulation, participants also completed the Positive and Negative Affect Schedule (PANAS) [7].

Experiment Set-up

After completing the questionnaires, participants washed their hands without soap to optimize skin conductance electrode attachment. We then applied measurement equipment in the following sequence:

1. An 8-channel EEG system (Neuroelectrics Enobio 8) with a 500 Hz sampling rate was placed with electrodes at positions F3, F4, C3, CZ, C4, P3, P4, and OZ according to the international 10-20 system. We used Signal Gel electrode gel, with reference electrodes (DRL and CMS; Cardinal Health Kendall H124SG) placed on the previously cleaned left mastoid.
2. Electrodermal activity and photoplethysmography (PPG) for heart rate were recorded using a Shimmer3 GSR+ device with a 100 Hz sampling rate. EDA electrodes (3M Red Dot

2239) were attached to the upper and lower portions of the left palm below the little finger, and the PPG sensor was attached to the left earlobe.

3. Pupil diameter was measured using an eye tracker (Smart Eye AI-X, 60 Hz) mounted on the lower part of the monitor.
4. Facial expressions were recorded using a high-definition camera (Logitech C920s PRO, 1024 × 768 pixel resolution).

Signal quality was visually validated for all physiological measures, and eye tracker calibration was performed before proceeding to the pain calibration. Recording of all physiological measures commenced at this point and continued throughout the remainder of the experiment.

The heat pain stimulation was applied to six distinct skin patches on the participant's left forearm, organized in two parallel columns of three positions each according to a standardized template. To minimize habituation and sensitization effects, stimulation alternated between patches, maximizing the distance between consecutive stimulations. Each patch received two stimulations throughout the experiment. The stimulation order was counterbalanced across participants (see Suppl. Figure S3 for details on the counterbalancing scheme). All heat stimuli were applied using a Pathways thermode (Medoc Ltd., Israel) with an ATS thermode (30 × 30 mm^2).

Pain Calibration

We calibrated two individual pain temperatures: pain threshold and VAS 70. The heat stimuli lasted 9 s at the plateau with a 10°C/s rise rate from a 32°C baseline and randomized inter-trial intervals of 12–13 s. The protocol began with warm-up stimuli (39°C, 40°C) and pre-exposure temperatures (42°C, 43°C, 44.5°C) to assess initial pain sensation. The following pain threshold calibration employed recursive Bayesian updating with a Gaussian prior (μ = 43°C if pre-exposure pain reported or else μ = 44.5°C, σ = 1.75°C). Participants provided binary above/below pain threshold feedback for 6 trials, with progressively decreasing likelihood variance to reflect increasing estimation confidence. The threshold was set as the posterior distribution maximum. VAS 70 calibration used the same method with the prior centered at the pain threshold plus 2.5°C (σ = 0.65°C). To ensure participant safety, we set the maximum temperature to 48°C. We further enforced a minimum 1.5°C difference between threshold and VAS 70

to maintain a sufficient temperature range, adjusting the calibrated temperatures if necessary while preserving safety limits. The approximately 5-minute-long calibration yielded average temperatures of 44.5°C (*SD* = 1.5°C) for pain threshold and 46.8°C (*SD* = 0.8°C) for VAS 70 before any adjustments.

Pain Measurement

Participants received 12 unique heat pain stimuli, each lasting 3 minutes. The temperature fluctuated between the individually calibrated pain threshold and VAS 70 values. Throughout each trial, participants continuously rated their pain intensity by moving the computer mouse with their right hand on a digital, equiluminant VAS ranging from no pain ("Keine Schmerzen") to strong pain (70 VAS; "Sehr starke Schmerzen"). Participants were asked to fully concentrate on their pain experience and to avoid getting distracted. Starting from the baseline temperature of 32°C, we increased the temperature at a rate of 10°C/s until reaching each participant's individual pain threshold. After reaching the pain threshold, we introduced a pause of 1.5 s to display the interactive VAS rating scale, without recording ratings. Following this pause, we initiated the fluctuating tonic pain stimulus while simultaneously starting the recording of online pain ratings. Pain ratings were sampled at a frequency of 10 Hz. To increase motivational aspects, participants were rewarded with 0.50 € if the correlation between their rating and the applied temperature was at least 0.7 in a single trial. Between trials, we moved the thermode to the next skin patch. Throughout the pain measurement, all physiological signals were continuously monitored by the experimenter to ensure high data quality. The measurement lasted approximately 45 minutes.

Post-Experiment Questionnaires

Following the pain measurement, participants completed the PANAS a second time and additional questionnaires: the Pain Catastrophizing Scale (PCS) [48] for pain catastrophizing, the Pain Vigilance and Awareness Questionnaire (PVAQ) [31] for pain attention and vigilance, the State-Trait Anxiety Inventory-Trait version (STAI-T-10) [46] for trait anxiety, and the Mindful Attention Awareness Scale (MAAS) [8] for dispositional mindfulness. Individual questionnaire results are provided in Suppl. Table S1 to Table S4. Participation in the exper-

iment was compensated with 37–43 €, based on the individual rating performance in the pain measurement task.

Data Acquisition

The experiment was programmed using the Python library Expyriment (version 0.11.0.dev1) [27], which managed stimulus presentation and online pain ratings. Synchronization and recording of all behavioral and psychophysiological data were coordinated by the iMotions software suite (version 10.0) [23] in conjunction with the Neuroelectrics Instrument Controller software (version 2.1.3.6) for EEG data. For further analyses and efficient data processing, we stored the CSV exports from iMotions in the SQL database system DuckDB (version 1.2.2) [44]. All original participant IDs were randomly reassigned to new, anonymized IDs.

Preprocessing

Our preprocessing pipeline was designed to support potential real-time application. Accordingly, we limited processing to causal transformations, i.e., methods that use only current and past data points. This constraint excluded more sophisticated preprocessing steps such as deconvolutions for EDA or eyeblink artifact removal via independent component analysis for EEG. For an additional exploratory data analysis, we utilized non-causal transformations to better preserve the temporal properties of the original signals.

In the final preprocessing step, we decimated all physiological signals except EEG data to a sampling rate of 10 Hz and merged them into a single feature matrix with equidistant time steps. EEG data were processed separately at a sampling rate of 250 Hz.

Electrodermal Activity

For the classification task, we used raw EDA data obtained directly from the Shimmer3 device. For the exploratory data analysis, we separated tonic (slow changes) and phasic (fast changes) components using a 5th-order Butterworth low-pass filter with a cutoff frequency of 0.05 Hz, implemented using Neurokit2 (version 0.2.10) [40].

Pupillometry

The pupillometric data underwent several preprocessing steps to improve signal quality. First, physiologically oriented thresholds were applied to remove pupil size measurements falling outside a 2–8 mm range. For the causal preprocessing, identified blink periods were filled

using forward propagation. For the exploratory data analysis, identified blink periods were extended by 120 ms on both sides to exclude measurement artifacts that frequently occur at the boundaries of blink events [28]. A median filter with a 1 s local window was then applied to remove partially detected blinks that remained in the signal. For the classification task, the median filter was implemented with forward padding, ensuring each output sample depended only on current and previous samples. For the exploratory data analysis, the filter was symmetric. Finally, the right and left pupil measurements were averaged into a single dilation measure.

Heart Rate

Stepwise heart rate data was acquired directly from the Shimmer3 device. To ensure physiological plausibility, we imposed an upper threshold of 120 beats per minute. For causal preprocessing, missing values were identified and filled using forward filling. For exploratory data analysis, we utilized linear interpolation to fill the gaps in the time series and applied a second-order Butterworth filter with a highcut of 0.8 Hz to mitigate abrupt transitions in the signal.

Facial Expressions

Over 30 different features were extracted using the Affectiva iMotions Facial Expression software [5], which has been demonstrated to reliably identify facial expressions with accuracy comparable to traditional electromyography (EMG) measurements [30]. Based on previous research on pain-relevant facial expressions [32,43], we focused our analysis on five key features: brow furrow, cheek raise, mouth open, nose wrinkle, and upper lip raise. These features were represented as probability values between 0 and 1, indicating the likelihood of the expression being present. The raw data were preprocessed causally using an exponential moving average smoothing with an alpha of 0.05; for non-causal preprocessing we used a second-order Butterworth filter with a highcut of 0.2 Hz.

Electroencephalography

We preprocessed EEG signals causally through a three-stage pipeline using SciPy (version 1.15.2) [51]. All filters were implemented as forward-only IIR filters. Data were first decimated from 500 Hz to 250 Hz utilizing a 6th-order Butterworth anti-aliasing filter. We then applied

a 4th-order Butterworth high-pass filter with a 0.5 Hz cutoff to remove slow drifts. Finally, second-order notch filters (quality factor of 30) at 50 and 100 Hz eliminated power line interference and its first harmonic. A detailed exploratory analysis of the EEG data was beyond the scope of the current study.

Exploratory Data Analysis

To examine relationships between physiological measures, we created grand-averaged, min-max scaled time series with a bin size of 0.1 s and 95% confidence intervals for each measure (pain ratings, tonic and phasic EDA, pupil diameter, heart rate, facial expressions) by averaging across all participants for each temperature curve. We calculated Pearson correlation coefficients between these grand-averaged time series to create two correlation matrices, one for the psychophysiological signals and one for the facial expressions. Additionally, we performed cross-correlation analyses to examine time-lagged relationships for the psychophysiological measures with the stimulation temperatures. To evaluate the reliability of these patterns on participant level, we calculated individual-level mean correlations across the 12 stimulus functions, then computed summary statistics (overall mean, standard deviation, minimum, and maximum) of these participant-level means across the entire sample. Correlations were averaged directly without Fisher z-transformation.

For all correlations, we excluded the first 20 s of each trial to ensure physiological equilibrium. This exclusion was necessary because the steep initial temperature ramp-up from baseline to pain threshold (>10°C increase in under 3 s) created artifacts unrelated to subsequent stimulation, particularly affecting low-pass filtered tonic components. The 20 s exclusion period was empirically determined from grand-averaged data, which showed stable responses after this timepoint, enabling reliable assessment of pain-related physiological dynamics.

Deep Learning Analysis

Sample Creation

We created 2826 samples of 7 s length for the binary classification task discriminating between pain decreases and non-decreases. For the non-decrease category, we utilized two types of

temperature patterns: increases and plateaus. Temperature increases consisted of the first 7 s of intervals with strictly increasing temperatures that contained no plateaus in them (see Figure 2). We excluded the first increasing temperature interval from each trial. Temperature plateaus were based on the 15 s periods of constant temperature stimulation, from which we extracted one 7 s sample starting at a 5 s offset. For pain decreases, we shifted the samples by 1 s from the start of a major decreasing interval to account for a latency in pain perception following the temperature reduction after the temperature peak. All samples were non-overlapping to ensure independence of observations.

Decreases followed a uniform temperature pattern due to the standardized curves of major decreases in the stimulus function. In contrast, pain increases exhibited more variability in amplitude and slope, and plateaus had different levels. Since more samples were available for non-decreases, we balanced the labels by down sampling the majority class to match the minority class (decreases). This ensured the models were trained on a balanced dataset, which is crucial for effective classification performance [9]. The final dataset comprised 2826 samples, with 1413 samples per class.

Data Splitting Strategy

To prevent participant-dependent data leakage, we implemented a group-based data splitting strategy. Participants were assigned exclusively to training, validation, or test sets in a 60:20:20 ratio. This resulted in 1615 samples (24 participants) for training, 626 samples (9 participants) for validation, and 585 samples (9 participants) for the test set. We standardized each feature independently via z-score normalization using parameters obtained from training samples only. For final model training, standardization parameters were computed from the combined training and validation sets.

Feature Sets

We systematically evaluated 11 distinct physiological feature sets for their predictive value in detecting pain changes. These sets consisted of (1) EDA, (2) HR, (3) pupil diameter, (4) EDA + HR, (5) EDA + pupil diameter, (6) EDA + HR + pupil diameter, (7) facial expressions, (8) facial expressions + EDA + HR + pupil diameter, (9) EEG, (10) EEG + EDA, and (11) all measures combined (EEG + facial expressions + EDA + HR + pupil diameter).

Model Architectures

For non-EEG signals (feature sets 1–8), we evaluated five distinct neural network architectures representing different modeling approaches: a multilayer perceptron (MLP) as a baseline reference [35], LightTS (MLP-based with different downsampling strategies) [56], TimesNet (CNN-based with period transformation) [54], PatchTST (transformer with patch-based tokenization) [42], and iTransformer (transformer with variate-based tokenization) [37].

For EEG data (feature sets 9–11), we employed a different set of architectures, since not all previously mentioned models scaled effectively to time series with higher temporal resolution. We evaluated an MLP, LightTS, iTransformer, and EEGNet—a convolutional neural network architecture specifically designed for EEG signal analysis [34].

We additionally implemented feature-level fusion ensemble models [3], combining separate PatchTST models for facial expressions and peripheral physiological signals (set 8), EEGNet for EEG with PatchTST for physiological signals (set 10), and EEGNet for EEG with PatchTST models for both facial expressions and physiological signals (set 11). Detailed architectural descriptions are provided in Suppl. Material C.

Model Training and Evaluation

The factorial design of 11 feature sets and their respective model architectures resulted in 53 distinct model-feature combinations. Following the standard train-validation-test split method, we conducted hyperparameter optimization for each combination using Bayesian-inspired search via the Optuna library [1] to maximize validation accuracy. After identifying the optimal configuration for each model architecture, we retrained the best-performing model for each feature set on the combined training and validation set before conducting a final evaluation on the previously unseen test set. By doing so, we maximized the training data available to each final model while preventing data leakage. This process yielded 11 final models, one optimal configuration for each feature set.

We evaluated overall test set performance using accuracy, area under the receiver operating characteristic curve (AUROC), precision, recall, F_1-score, and Matthews correlation coefficient (MCC), with pain decreases defined as positive cases. We generated receiver operating characteristic (ROC) curves to visualize discriminative performance across feature sets. Additionally,

we evaluated individual-level test accuracy for each participant across all feature sets to assess prediction consistency. Full details on hyperparameter search spaces, training procedures, and metric definitions are provided in Suppl. Material D.

Seed Stability Analysis

To assess the stability of results across different participant-based group splits, we performed a seed stability analysis. Using the best-performing model configuration, we evaluated performance across multiple random train-test splits (n = 200) produced by varying the random seed. For each split, we trained the model on the combined training and validation sets, then evaluated on the test set. We visualized the distribution of the test accuracies across all seeds using histograms for the feature sets EDA + HR and EDA + HR + pupil diameter.

Application to Continuous Data Streams

We applied the best-performing models to continuous data streams to detect pain decreases in fluctuating pain. Unlike the training set, where we used non-overlapping samples specifically selected from clearly defined temperature intervals, we created a new set of samples based on the full 180 s trials from participants in the test set. This approach better simulates real-world applications, where the model must continuously evaluate moment-to-moment pain dynamics. Using a sliding window approach with 7 s windows and a 250 ms step size between samples, we created 693 samples for each trial and participant in the test set. Samples were standardized with values obtained from the combined training and validation set. No additional training was performed. We converted the raw logits from the model's output into probability values using a sigmoid function and visualized them as confidence values for detecting pain decreases across time for each participant in a heatmap. We selected conservative classification thresholds to reduce false positives, as high precision would be crucial for ensuring that interventions are triggered primarily during genuine pain decreases. To determine the latency of the classifier, we calculated the median time and median absolute deviation (MAD) of the first detection of pain decreases in major temperature decreases. Detections occurring within 1 s after the start of a major temperature decrease were removed for physiological plausibility.

Results

Exploratory Data Analysis

Correlation analysis of grand-averaged data revealed systematic relationships between temperature curves, pain ratings, and physiological measures (see Figure 3). Temperature and pain ratings showed the strongest overall correlation (r = .89). Among physiological measures, pupil diameter demonstrated the highest correlation with temperature (r = .61), followed by heart rate (r = .42), tonic EDA (r = .39), and phasic EDA (r = .19). A visualization of the grand-averaged signals over all 12 temperature curves is provided in Suppl. Figure S4.

Regarding facial expressions, cheek raise showed a moderate correlation with temperature (r = .60), followed by mouth open (r = .39) and upper lip raise (r = .27). Nose wrinkle (r = .09) and brow furrow (r = .05) were only minimally correlated with temperature. All grand-averaged facial expression activations remained largely near baseline (normalized value of 0) throughout the full stimulation (for visualized grand-averages, see Suppl. Figure S5; for a correlation matrix, see Suppl. Figure S6).

Cross-correlation analysis revealed minimum lags behind temperature stimulation for pupil responses (M = 0.07 s, SD = 0.22 s), followed by tonic EDA (M = 0.95 s, SD = 0.62 s), pain rating (M = 1.01 s, SD = 0.34 s), heart rate (M = 1.76 s, SD = 1.36 s), and phasic EDA (M = 2.45 s, SD = 0.89 s). Additional lags for facial expressions can be found in Suppl. Table S5, and cross-correlation lags with pain ratings as reference are provided in Suppl. Table S6.

At the individual participant level, we observed consistent pain-temperature correlations (M = 0.77, SD = 0.05) and substantial variability in physiological response correlations. Pupil diameter showed the widest range of correlations with temperature, from r = -.18 to r = .61 (M = 0.26, SD = 0.20), closely followed by tonic EDA correlations ranging from r = -.09 to r = .67 per participant (M = 0.21, SD = 0.20). Heart rate ranged from r = -.15 to r = .53 (M = 0.12, SD = 0.16). Phasic EDA responses were minimal and consistent (M = 0.07, SD = 0.06). Suppl. Figure S7 provides a graphical representation of correlations for each participant. Facial expressions showed high variance across participants, with mean correlations ranging from r = -.37 to r

= .64, mostly distributed around $r = 0$ (see Suppl. Figure S8). Only two participants displayed moderate positive correlations (around $r = .40$) across all five analyzed facial expressions.

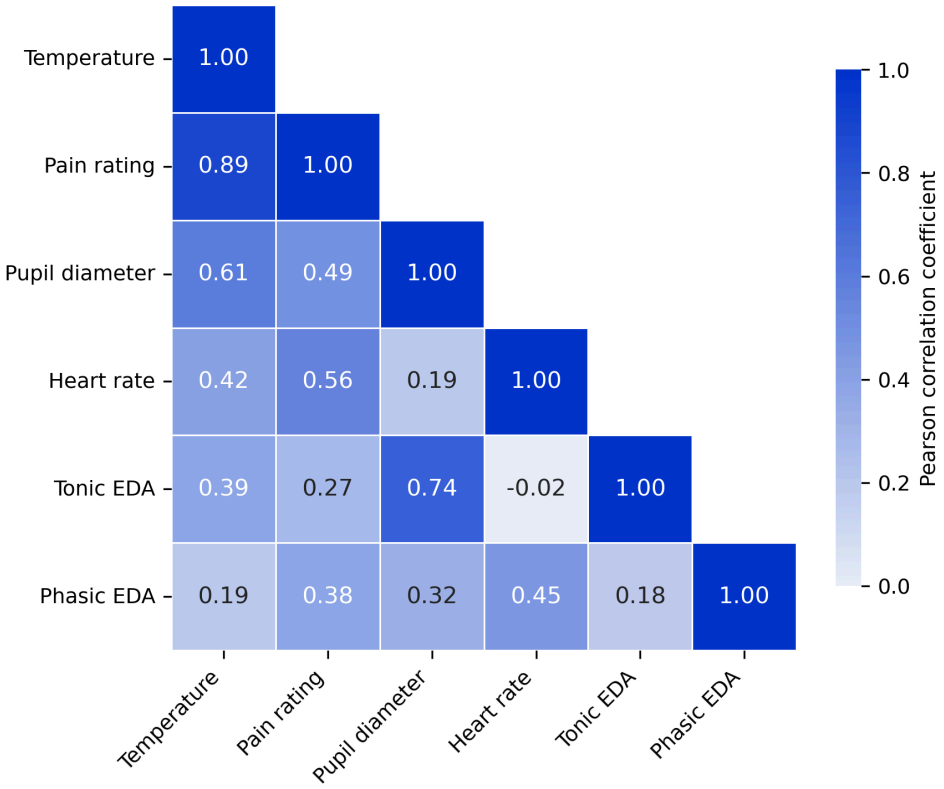

Figure 3: Correlation matrix of grand-averaged physiological measures, temperature curves, and pain ratings. The heatmap shows Pearson correlation coefficients between all measured variables, with darker blue indicating stronger positive correlations.

Deep Learning Analysis

Model Comparison Results

We evaluated 53 model-feature combinations, with performance metrics summarized in Table 1 and corresponding ROC curves shown in Figure 4. The EDA + HR + Pupil feature set achieved the highest AUROC of 0.854 (accuracy = 0.768) using the PatchTST architecture. The EDA + Pupil combination showed comparable performance with an AUROC of 0.851 (accuracy = 0.779). The third-best performance was achieved by the feature set containing EDA + HR + Pupil + facial expressions with an AUROC of 0.846 (accuracy = 0.747).

Individual physiological measures demonstrated varying discriminative capabilities. EDA emerged as the most informative single modality (AUROC = 0.805, accuracy = 0.762), followed by pupil diameter (AUROC = 0.738, accuracy = 0.667). Heart rate (AUROC = 0.557, accuracy

= 0.530), facial expressions (AUROC = 0.582, accuracy = 0.569), and EEG (AUROC = 0.555, accuracy = 0.535) demonstrated limited discriminative power at the group level.

The ensemble model combining all physiological measures without EEG achieved an AUROC of 0.846 (accuracy = 0.747), slightly lower than the optimal three-feature combination, while the full ensemble including EEG showed an AUROC of 0.832 (accuracy = 0.756). Across feature sets containing EDA, PatchTST consistently emerged as the best-performing architecture. For univariate EDA data, PatchTST achieved a validation accuracy 7.4 percentage points higher than the next-best architecture (see Suppl. Table S7 for validation set accuracies across all model-feature combinations). The iTransformer architecture proved most effective for EEG analysis. Hyperparameter optimization results can be found in Suppl. Table S9.

Table 1: Model comparison results across different feature sets

Feature set	Best model	Acc.	AUROC	Precision	Recall	F_1	MCC
EDA	PatchTST	0.762	0.805	0.806	0.700	0.750	0.530
HR	MLP	0.530	0.557	0.550	0.411	0.470	0.065
Pupil	iTransformer	0.667	0.738	0.675	0.663	0.669	0.333
EDA + HR	PatchTST	0.756	0.813	0.775	0.731	0.752	0.512
EDA + Pupil	PatchTST	0.779	0.851	0.778	0.791	0.785	0.559
EDA + HR + Pupil	PatchTST	0.768	0.854	0.762	0.788	0.775	0.535
Facial Expressions	PatchTST	0.569	0.582	0.593	0.481	0.532	0.143
Combined (w/o EEG)	PatchTST	0.747	0.846	0.726	0.805	0.764	0.496
EEG	iTransformer	0.535	0.555	0.538	0.593	0.564	0.069
EEG + EDA	Ensemble	0.732	0.793	0.733	0.741	0.737	0.463
Combined	Ensemble	0.756	0.832	0.781	0.721	0.750	0.513

Note: Metrics were computed at a classification threshold of 0.5 with a balanced test set. Abbreviations: Acc. = Accuracy; AUROC = Area Under ROC Curve; MCC = Matthews Correlation Coefficient.

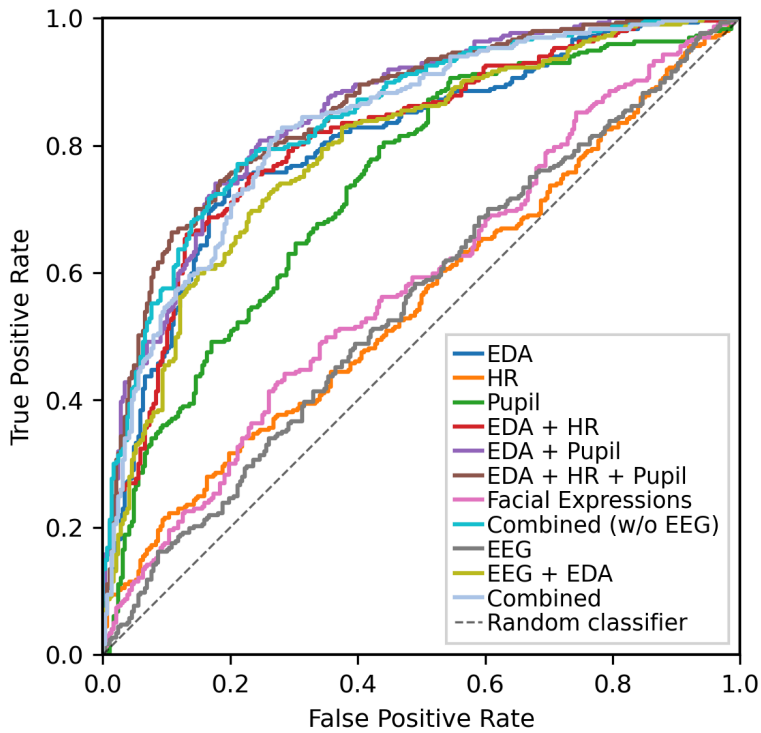

Figure 4: ROC curves comparing physiological feature sets. All classifiers performed above chance level (dashed diagonal line), with feature sets containing EDA data showing strong discriminative ability.

Seed Stability Results

Across 200 random participant-based train-test splits, the EDA + HR model achieved a mean test accuracy of 0.724 (*SD* = 0.030), while the EDA + HR + Pupil model achieved 0.744 (*SD* = 0.031) as shown in Figure 5. The test accuracies of our primary models (0.756 and 0.768, respectively) were one standard deviation above the mean across all random splits.

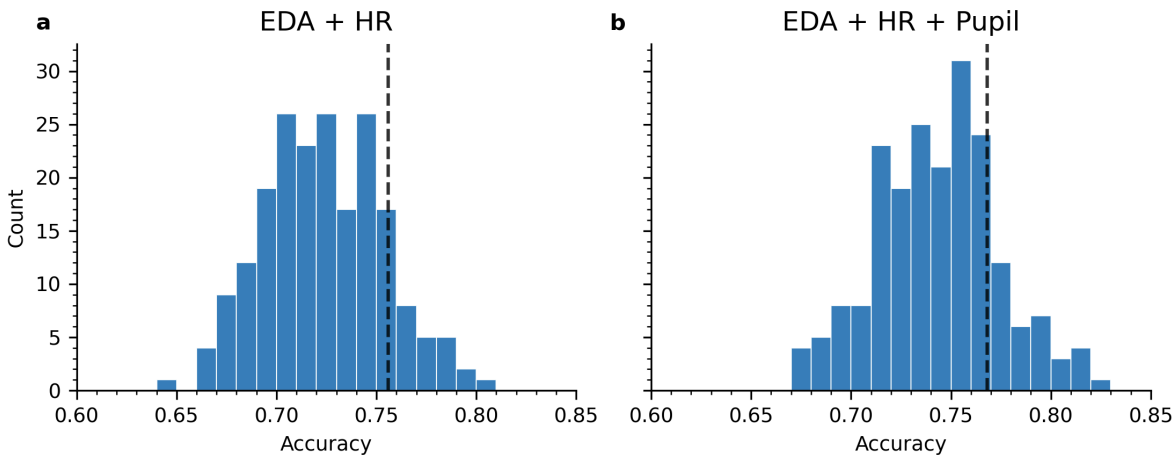

Figure 5: Seed stability analysis across 200 random train-test splits. Shown are the distribution of test accuracies for (a) EDA + HR and (b) EDA + HR + Pupil models. The dashed lines indicate primary model performance.

Interindividual Differences

Classification accuracy varied substantially across the nine test-set participants (see Figure 6). The EDA + Pupil feature set demonstrated the highest mean accuracy (M = 0.777), along with the widest interquartile range (IQR = 0.166) and the highest maximum value (0.922). Five participants achieved accuracies above 0.80 with this feature set, more than with any other feature combination.

Univariate EDA showed relatively consistent performance across participants (M = 0.760, IQR = 0.104), with seven of nine participants exceeding 0.70 accuracy. Pupillary responses yielded consistent but more moderate accuracy (M = 0.669, IQR = 0.127). Heart rate features produced the second-lowest mean accuracy (M = 0.534, IQR = 0.075) with pronounced variability: one participant achieved 0.778 accuracy while the remaining participants ranged between 0.493 and 0.618.

Facial expressions demonstrated limited discriminative capability (M = 0.570, IQR = 0.090) for most participants, with a maximum accuracy of 0.693. EEG also demonstrated limited discriminative capability (M = 0.529, IQR = 0.11), with only one participant exceeding 0.60 accuracy.

Notably, for EEG, HR, and facial expressions, classification accuracy sometimes fell below chance level. This was most prominent with EEG data, where one participant obtained an

accuracy of 0.409, the lowest accuracy overall. Combining EDA and pupil diameter with EEG and/or facial expressions stabilized performance across participants (Combined: $M = 0.748$, $IQR = 0.072$), though it did not improve maximum accuracy compared to simpler feature sets. No single feature set uniformly optimized classification across all participants, with individual-specific patterns evident throughout the classification. Note that individual participants in the test set had slightly different sample sizes (all being balanced), which are reported in Suppl. Table S10. Exact values of the boxplot diagram can be found in Suppl. Table S11.

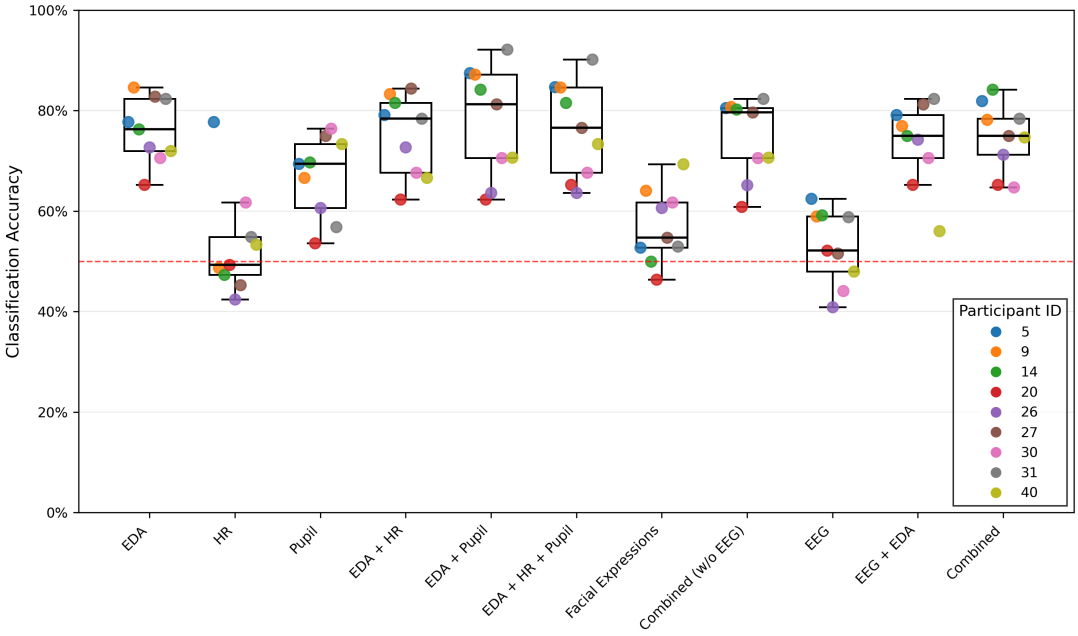

Figure 6: Classification accuracies across feature sets for individual test set participants. Box plots show the distribution of accuracies for n=9 participants, with individual participant data points overlaid. The red dashed line indicates chance level (50%).

Application to Continuous Data Streams

Application of the PatchTST model with EDA + HR + Pupil features to fluctuating temperature curves at a conservative classification threshold of 0.90 produced the prediction heatmap shown in Figure 7. The model achieved a sensitivity of 70.37% for major temperature decreases, detecting 209 out of 297. The median detection latency was 5.75 s (MAD = 1.50 s) within the 20-second windows of temperature decrease. Lowering the threshold to 0.5 decreased median latency to 4.25 s (MAD = 1.25 s), illustrating a sensitivity-specificity trade-off with improved detection speed at the cost of increased false positive classifications. The predicted pain decreases demonstrated strong temporal alignment with actual temperature decreases

across most participants, identifying temperature decreases that were both shorter in duration and smaller in magnitude than the temperature decreases presented during model training. Temperature plateau regions occasionally resulted in false positive classifications. Participant 40 exhibited almost no positive predictions at the conservative threshold level. Corresponding heatmaps for the other 10 feature sets are presented in Suppl. Figure S9 through Figure S18.

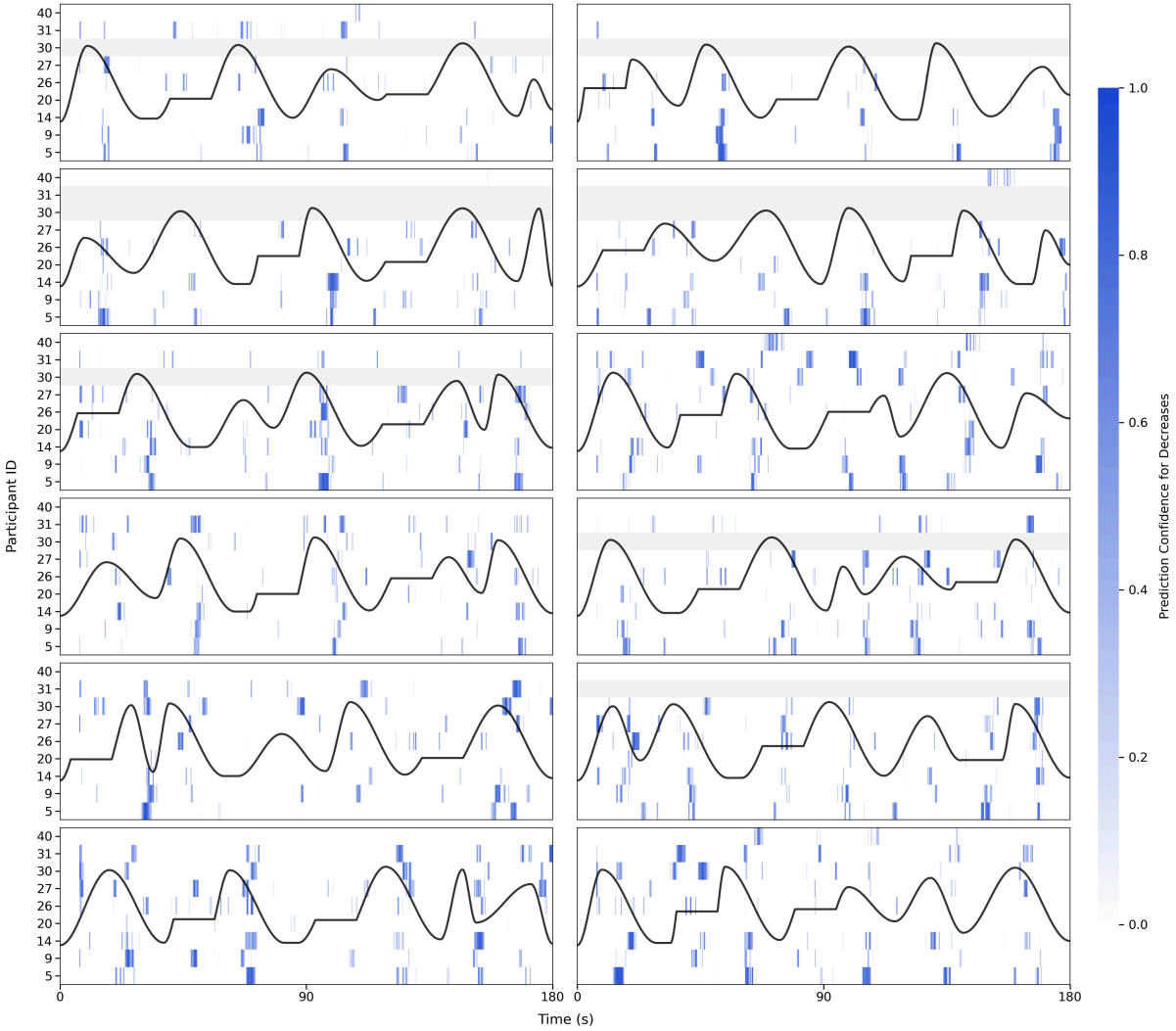

Figure 7: Continuous pain decrease prediction heatmap using EDA, HR, and pupil diameter. Applying the trained model on continuous temperature curves for 9 participants from the test set demonstrates the model's real-time pain detection capabilities. The model used EDA + HR + Pupil features with a classification threshold of 0.90. The black line shows the temperature curve in each trial, with each row representing one participant. Color intensity indicates prediction confidence for pain decrease, with darker colors showing higher confidence. Missing trials are displayed in light grey.

Discussion

This study investigated whether objective detection of pain decreases is feasible using readily obtainable physiological signals in healthy participants. Results demonstrate that pain decreases can be detected in real-time with substantial accuracy using a combination of electrodermal activity, pupillary responses, and heart rate, with electrodermal activity being the most informative predictor. Despite considerable interindividual variability indicating the need for personalized approaches, these findings provide proof of concept for noninvasive and objective pain monitoring systems.

Main Findings

Detection of Pain Dynamics

This study advances previous research by examining pain dynamics in tonic, fluctuating stimulation rather than static intensity estimation in phasic pain [17,18,21,38,49,50,53]. Our three minute heat pain stimulations—randomly fluctuating between pain threshold and higher intensities (VAS 70)—more closely approximate chronic pain conditions where individuals experience spontaneous pain fluctuations [16]. By targeting pain changes (first derivative of pain intensity) rather than absolute intensity levels, we may have addressed a more tractable classification problem than absolute intensity estimation. The classifier's ability to also detect pain decreases that were shorter and were based on smaller temperature differences than those used for training further demonstrates generalizability.

Multimodal Performance

The combination of electrodermal activity, pupillometry, and heart rate achieved the best classification performance in the model comparison. Electrodermal activity emerged as the most informative single predictor, consistent with its established sensitivity to nociceptive stimulation via direct sympathetic nervous system activation [17,47]. Pupillary responses demonstrated second-highest discriminative capability, reflecting sympathetic activation through dilator pupillae muscle contraction [13]. Heart rate variability captured sympathico-vagal interactions [41], with noxious stimulation triggering blood pressure increases and subsequent heart rate acceleration [13].

In the model comparison, the PatchTST architecture [42] strikingly outperformed other architectures by over 5 percentage points in accuracy for feature sets containing EDA data (see Suppl. Table S7). The patch-based tokenization of temporal segments aligned optimally with raw EDA data without requiring additional feature engineering. The use of only causal preprocessing (restricting each sample to past and present information) differs from prior work employing non-causal steps for pain intensity estimation [17,18,38,49]. Nevertheless, our minimally processed pipeline achieved strong performance suitable for real-time deployment, which we further validated through a continuous data stream analysis showing stable detection throughout extended periods of fluctuating pain.

Individual Variability

Despite these promising multimodal results, substantial interindividual differences across physiological measures represent a critical finding for clinical translation [13]. Variability patterns differed across modalities: EDA and pupillary responses demonstrated relative stability, whereas facial expressions, heart rate, and EEG exhibited high variability with some participants performing even below chance levels, warranting separate discussion.

Facial expression variability aligns with findings from a systematic review by Kunz et al. [32], which states that pain-related action units are combined differently across individuals rather than being uniformly expressed. Additionally, some participants' pain responses correlated positively with temperature changes, while others showed negative correlations, further complicating universal model development. Heart rate showed modest group performance with one notable exception achieving EDA-level accuracy, potentially influenced by breathing patterns under experimental pain [15].

EEG exhibited the greatest variability among physiological measures, with classification accuracies showing equivalent distances from chance performance for the highest and lowest performing participants. These findings align with previous research documenting the absence of consistent EEG-based pain biomarkers across individuals [58,59]. Although enhanced delta and gamma band power represent robust neurophysiological findings in pain perception [57,59], patterns most relevant for pain decreases remain unclear. The high interindividual variability ultimately contributed to the modest group-level classification performance for

EEG, as models trained on diverse response patterns do not inherently generalize well across individuals. Universal approaches may have fundamental limitations and personalized models seem to be more appropriate for pain detection via EEG signals, as shown by Aurucci et al. [2] for the real-time detection of neuropathic pain in patients.

Limitations & Clinical Translation

Generalizability

This study used healthy participants experiencing experimentally induced pain in controlled laboratory conditions, limiting generalizability to clinical pain populations. Chronic pain patients show different EEG signatures to painful stimulation compared with healthy participants [19], and clinical pain encompasses psychological, social, and pathophysiological factors beyond nociceptive processing. However, establishing proof of concept in healthy participants provides necessary groundwork for future clinical investigations where physiological heterogeneity may be even greater.

Temporal Constraints

This study was designed to examine whether a device can reliably detect decreases in pain. Once such a decrease is identified, the device would offer the participant a treatment option —an option that would be perceived as effective because the pain is already declining. This could re-establish a sense of control, even though the causal order is reversed (the decrease is detected first, and the treatment follows). Through the illusion of control, this sequence may nevertheless be experienced as a genuine instance of self-initiated pain relief [10].

For this approach, timing is crucial: patients must be offered a treatment option as soon as possible after a spontaneous pain decrease, allowing them to attribute the relief to their own actions rather than to uncontrollable fluctuations. This requires minimal detection delays to preserve temporal coupling between a potential treatment and pain relief. Given EDA's role as the primary predictor, physiological latency constraints imposed by sweat gland response times affect the detectability of skin conductance changes [6]. In our data, phasic EDA showed reasonable delays of approximately 2.5 seconds according to a cross-correlation lag analysis, and pain decreases were detected at a median of 5.75 seconds.

In an additional analysis not reported in the results section, we systematically evaluated shorter sample windows (4, 5, and 6 seconds) to potentially reduce detection latency. Results revealed a decrease in accuracy with shorter durations, without improvements in detection speed.

Classification Accuracy

Our models demonstrated reasonable classification accuracy with variable false positive rates across participants and measures. False positive detections occurred particularly during temperature plateaus. However, these apparent misclassifications as non-decreases often coincided with subjective pain decreases (visible in grand-averaged data; see Suppl. Figure S4), suggesting that models correctly detected habituation-related pain reductions rather than simply tracking the temperature-based ground truth labels.

Furthermore, these classification errors need not undermine therapeutic efficacy. False positive detections would result in participants administering a treatment during moments when pain does not decrease. If, however, other actual pain decreases are correctly detected, this could lead to a partial reinforcement schedule, which could produce even more robust and extinction-resistant learning than continuous reinforcement [14], potentially enhancing expectation effects [55]. This would also allow for less strict classification thresholds with the additional benefit of lower latencies by prioritizing sensitivity over specificity.

Although multimodal ensembles theoretically outperform single modalities through noise reduction, results revealed nuanced patterns. The addition of EDA and pupil diameter signals consistently improved performance, whereas facial expressions, heart rate, and EEG provided no benefits. Full ensemble performance was inferior to optimal subsets, suggesting trade-offs between information gain and noise accumulation and supporting personalized, fine-tuned approaches consistent with precision medicine principles [24].

Future Implementation

Results indicate that while the best-performing pain detection combines EDA, HR, and pupil signals—potentially enhanced by personalized, fine-tuned EEG and facial expression data—the most pragmatic approach uses only EDA and HR measurements. Despite high univariate pupillometry performance, real-world implementation encounters constraints from ambient

light variations and screen usage. Although EEG could potentially improve latency and accuracy, its application has two drawbacks, namely the need for fine-tuned models and the impractical equipment required for measurement. Facial expression analysis faces similar challenges with specialized cameras or electrodes. These practical constraints make EDA and heart rate the optimal candidates for accessible, continuous pain monitoring in everyday settings.

EDA and heart rate monitoring can be implemented using commercially available wearable devices [12]. For signal processing, the PatchTST architecture demonstrated superior performance for EDA-based detection, making it the optimal choice for real-time implementation. Heart rate thresholds (e.g., limiting classification to under 100 bpm) can restrict detection to resting states where physiological pain responses are most discriminable, while additional accelerometry data would enable context-aware detection by filtering motion-related artifacts. This setup could utilize a single device worn on one arm and enables straightforward practical deployment, making it suitable for everyday use. In the domain of pain versus no-pain detection, previous work has demonstrated comparable results for the real-time detection of experimental pain in healthy participants using EDA only [26], further supporting the feasibility of this pragmatic approach.

This study demonstrated that the objective, near real-time detection of pain decreases is achievable using easily obtainable physiological signals. Clinical validation remains essential for realizing the therapeutic potential of these findings in preventing pain chronification through precisely timed interventions.

Acknowledgements

The authors thank all participants for participating in the study. This work was supported by the European Research Council (ERC) under the grant ERC-AdG-883892-PainPersist. The authors have no conflicts of interest to declare. AI tools (Claude, ChatGPT) assisted with manuscript editing and code generation. The authors reviewed and take full responsibility for all content. All data will be made available in a forthcoming data publication. In the meantime,

data will be available upon reasonable request to the corresponding author. All code is publicly available at: https://github.com/visserle/pain-measurement.

References

[1] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. In:. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019. pp. 2623–2631.

[2] Aurucci GV, Preatoni G, Damiani A, Raspopovic S. Brain-Computer Interface to Deliver Individualized Multisensory Intervention for Neuropathic Pain. Neurotherapeutics 2023;20:1316–1329.

[3] Baltrušaitis T, Ahuja C, Morency L-P. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 2019;41:423–443.

[4] Bandura A. Human Agency in Social Cognitive Theory. American Psychologist 1989;44:1175–1184.

[5] Bishay M, Preston K, Strafuss M, Page G, Turcot J, Mavadati M. AFFDEX 2.0: A Real-Time Facial Expression Analysis Toolkit. In:. 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG).2023. pp. 1–8.

[6] Boucsein W. Electrodermal Activity, 2nd Ed. New York, NY, US: Springer Science + Business Media, 2012.

[7] Breyer B, Bluemke M. Deutsche Version der Positive and Negative Affect Schedule PANAS (GESIS Panel). Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften, 2016.

[8] Brown KW, Ryan RM. Mindful Attention Awareness Scale. 2011.

[9] Buda M, Maki A, Mazurowski MA. A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. Neural Networks 2018;106:249–259.

[10] Büchel C. The Role of Expectations, Control and Reward in the Development of Pain Persistence Based on a Unified Model. eLife 2023;12:e81795.

[11] Büchel C, Geuter S, Sprenger C, Eippert F. Placebo Analgesia: A Predictive Coding Perspective. Neuron 2014;81:1223–1239.

[12] Chen J, Abbod M, Shieh J-S. Pain and Stress Detection Using Wearable Sensors and Devices—A Review. Sensors 2021;21:1030.

[13] Cowen R, Stasiowska MK, Laycock H, Bantel C. Assessing Pain Objectively: The Use of Physiological Markers. Anaesthesia 2015;70:828–847.

[14] Ferster CB, Skinner BF. Schedules of Reinforcement. East Norwalk, CT, US: Appleton-Century-Crofts, 1957.

[15] Forte G, Troisi G, Pazzaglia M, Pascalis VD, Casagrande M. Heart Rate Variability and Pain: A Systematic Review. Brain Sciences 2022;12:153.

[16] Foss JM, Apkarian AV, Chialvo DR. Dynamics of Pain: Fractal Dimension of Temporal Variability of Spontaneous Pain Differentiates between Pain States. Journal of Neurophysiology 2006;95:730–736.

[17] Geuter S, Gamer M, Onat S, Büchel C. Parametric Trial-by-Trial Prediction of Pain by Easily Available Physiological Measures. Pain 2014;155:994–1001.

[18] Gouverneur P, Badura A, Li F, Bieńkowska M, Luebke L, Adamczyk WM, Szikszay TM, Myśliwiec A, Luedtke K, Grzegorzek M, Piętka E. An Experimental and Clinical Physiological Signal Dataset for Automated Pain Recognition. Scientific Data 2024;11:1–13.

[19] Gozzi N, Preatoni G, Ciotti F, Hubli M, Schweinhardt P, Curt A, Raspopovic S. Unraveling the Physiological and Psychosocial Signatures of Pain by Machine Learning. Med 2024;5:1495–1509.e5.

[20] Habermann M, Strube A, Büchel C. How Control Modulates Pain. Trends in Cognitive Sciences 2025;29:60–72.

[21] Haque MA, B. Bautista R, Noroozi F, Kulkarni K, Laursen CB, Irani R, Bellantonio M, Escalera S, Anbarjafari G, Nasrollahi K, Andersen OK, Spaich EG, Moeslund TB. Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities: 13th IEEE International Conference on Automatic Face and Gesture Recog-

nition, FG 2018. Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018 2018:250–257.

[22] Hautzinger M, Keller F, Kühner C. Beck Depressions-Inventar (BDI-II). Harcourt Test Services, 2006.

[23] iMotions. iMotions (10). 2024.

[24] Jameson JL, Longo DL. Precision Medicine — Personalized, Problematic, and Promising. New England Journal of Medicine 2015;372:2229–2234.

[25] Kaptchuk TJ, Hemond CC, Miller FG. Placebos in Chronic Pain: Evidence, Theory, Ethics, and Use in Clinical Practice. BMJ 2020:m1668.

[26] Kong Y, Posada-Quintero HF, Chon KH. Real-Time High-Level Acute Pain Detection Using a Smartphone and a Wrist-Worn Electrodermal Activity Sensor. Sensors 2021;21:3956.

[27] Krause F, Lindemann O. Expyriment: A Python Library for Cognitive and Neuroscientific Experiments. Behavior Research Methods 2014;46:416–428.

[28] Kret ME, Sjak-Shie EE. Preprocessing Pupil Size Data: Guidelines and Code. Behavior Research Methods 2019;51:1336–1342.

[29] Kroenke K, Spitzer RL, Williams JBW. The PHQ-15: Validity of a New Measure for Evaluating the Severity of Somatic Symptoms. Psychosomatic Medicine 2002;64:258–266.

[30] Kulke L, Feyerabend D, Schacht A. A Comparison of the Affectiva iMotions Facial Expression Analysis Software With EMG for Identifying Facial Expressions of Emotion. 2020.

[31] Kunz M, Capito ES, Horn-Hofmann C, Baum C, Scheel J, Karmann AJ, Priebe JA, Lautenbacher S. Psychometric Properties of the German Version of the Pain Vigilance and Awareness Questionnaire (PVAQ) in Pain-Free Samples and Samples with Acute and Chronic Pain. International Journal of Behavioral Medicine 2017;24:260–271.

[32] Kunz M, Meixner D, Lautenbacher S. Facial Muscle Movements Encoding Pain—a Systematic Review. PAIN 2019;160:535.

[33] Langer EJ. The Illusion of Control. Journal of Personality and Social Psychology 1975;32:311–328.

[34] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces. Journal of Neural Engineering 2018;15:56013.

[35] LeCun Y, Bengio Y, Hinton G. Deep Learning. Nature 2015;521:436–444.

[36] Lin Y, Xiao Y, Wang L, Guo Y, Zhu W, Dalip B, Kamarthi S, Schreiber KL, Edwards RR, Urman RD. Experimental Exploration of Objective Human Pain Assessment Using Multimodal Sensing Signals. Frontiers in Neuroscience 2022;16.

[37] Liu Y, Hu T, Zhang H, Wu H, Wang S, Ma L, Long M. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. 2024.

[38] Lopez-Martinez D, Picard R. Continuous Pain Intensity Estimation from Autonomic Signals with Recurrent Neural Networks. In:. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Honolulu, HI: IEEE, 2018. pp. 5624–5627.

[39] Maier SF, Seligman MEP. Learned Helplessness at Fifty: Insights from Neuroscience. Psychological Review 2016;123:349–367.

[40] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SHA. NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing. Behavior Research Methods 2021;53:1689–1696.

[41] Malik M. Heart Rate Variability. Annals of Noninvasive Electrocardiology 1996;1:151–181.

[42] Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A Time Series Is Worth 64 Words: Long-term Forecasting with Transformers. 2023.

[43] Prkachin KM. The Consistency of Facial Expressions of Pain: A Comparison across Modalities. Pain 1992;51:297–306.

[44] Raasveldt M, Mühleisen H. DuckDB: An Embeddable Analytical Database. In:. Proceedings of the 2019 International Conference on Management of Data. SIGMOD '19. New York, NY, USA: Association for Computing Machinery, 2019. pp. 1981–1984.

[45] Samwel HJA, Evers AWM, Crul BJP, Kraaimaat FW. The Role of Helplessness, Fear of Pain, and Passive Pain-Coping in Chronic Pain Patients. The Clinical Journal of Pain 2006;22:245.

[46] Spielberger CD, Gorsuch RL. Manual for the State-Trait Anxiety Inventory (Form Y):" Self-Evaluation Questionnaire". (No Title) 1983.

[47] Storm H. Changes in Skin Conductance as a Tool to Monitor Nociceptive Stimulation and Pain. Current Opinion in Anesthesiology 2008;21:796.

[48] Sullivan MJL, Bishop SR, Pivik J. The Pain Catastrophizing Scale: Development and Validation. Psychological Assessment 1995;7:524–532.

[49] Thiam P, Hihn H, Braun DA, Kestler HA, Schwenker F. Multi-Modal Pain Intensity Assessment Based on Physiological Signals: A Deep Learning Perspective. Frontiers in Physiology 2021;12.

[50] Velana M, Gruss S, Layher G, Thiam P, Zhang Y, Schork D, Kessler V, Meudt S, Neumann H, Kim J, Schwenker F, André E, Traue HC, Walter S. The SenseEmotion Database: A Multimodal Database for the Development and Systematic Validation of an Automatic Pain- and Emotion-Recognition System. In: Schwenker F, Scherer S, editors. Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction. Cham: Springer International Publishing, 2017. pp. 127–139.

[51] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 2020;17:261–272.

[52] Wager TD, Atlas LY. The Neuroscience of Placebo Effects: Connecting Context, Learning and Health. Nature Reviews Neuroscience 2015;16:403–418.

[53] Walter S, Gruss S, Ehleiter H, Tan J, Traue HC, Werner P, Al-Hamadi A, Crawcour S, Andrade AO, Moreira da Silva G. The Biovid Heat Pain Database Data for the Advancement and Systematic Validation of an Automated Pain Recognition System. In:. 2013 IEEE International Conference on Cybernetics (CYBCO).2013. pp. 128–131.

[54] Wu H, Hu T, Liu Y, Zhou H, Wang J, Long M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. 2023.

[55] Yeung STA, Colagiuri B, Lovibond PF, Colloca L. Partial Reinforcement, Extinction, and Placebo Analgesia. Pain 2014;155:1110–1117.

[56] Zhang T, Zhang Y, Cao W, Bian J, Yi X, Zheng S, Li J. Less Is More: Fast Multivariate Time Series Forecasting with Light Sampling-oriented MLP Structures. 2022.

[57] Zhang ZG, Hu L, Hung YS, Mouraux A, Iannetti GD. Gamma-Band Oscillations in the Primary Somatosensory Cortex—A Direct and Obligatory Correlate of Subjective Pain Intensity. The Journal of Neuroscience 2012;32:7429–7438.

[58] Zhao R, Su Q, Song Y, Yang Q, Wang S, Zhang J, Qin W, Yu C, Liang M. Brain-Activation-Based Individual Identification Reveals Individually Unique Activation Patterns Elicited by Pain and Touch. NeuroImage 2022;260:119436.

[59] Zis P, Liampas A, Artemiadis A, Tsalamandris G, Neophytou P, Unwin Z, Kimiskidis VK, Hadjigeorgiou GM, Varrassi G, Zhao Y, Sarrigiannis PG. EEG Recordings as Biomarkers of Pain Perception: Where Do We Stand and Where to Go?. Pain and Therapy 2022;11:369–380.

Supplementary Materials

Contents

Supplementary Material A: Experimental Design and Procedures 35
Supplementary Material B: Participant Characteristics 37
Supplementary Material C: Deep Learning Model Architectures 43
Multilayer Perceptron 43
LightTS 43
TimesNet 43
PatchTST 44
iTransformer 44
EEGNet 45
Ensemble Models 45
Supplementary Material D: Hyperparameter Optimization, Model Training and Evaluation 46
Hyperparameter Optimization 46
Model Training 46
Model Evaluation 47
Supplementary Material E: Exploratory Data Analysis 48
Supplementary Material F: Model Comparison and Optimization Results 52
Supplementary Material G: Model Inference Visualizations 57
Supplementary References 62

Supplementary Material A: Experimental Design and Procedures

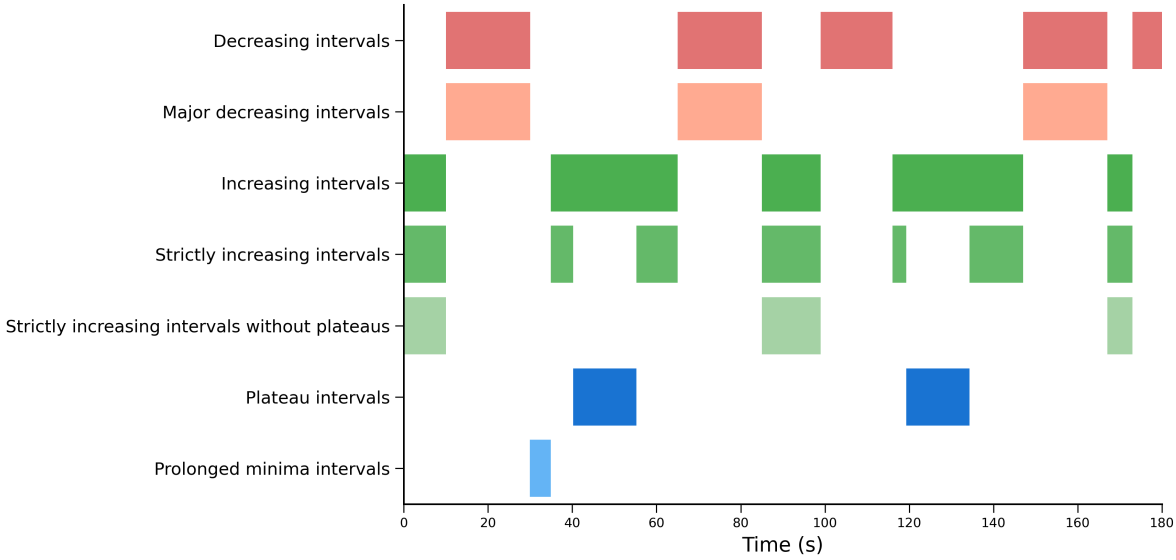

Figure S1: Stimulus interval types. Samples for the classification task were based on strictly increasing intervals without plateaus, strictly increasing intervals after plateaus, and plateau intervals for non-decreases, and major decreasing intervals for decreases. All interval labels were automatically created within the stimulus generation. The shown stimulus had the random seed 133.

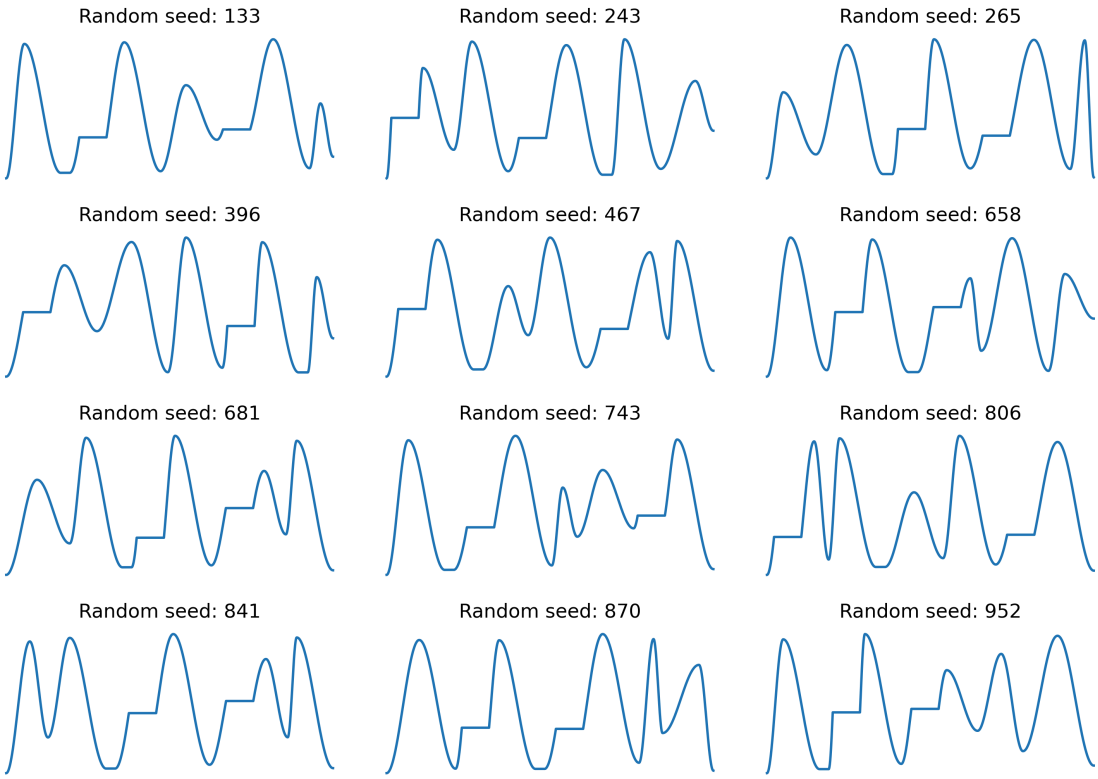

Figure S2: Temperature curves used in the pain measurement experiment

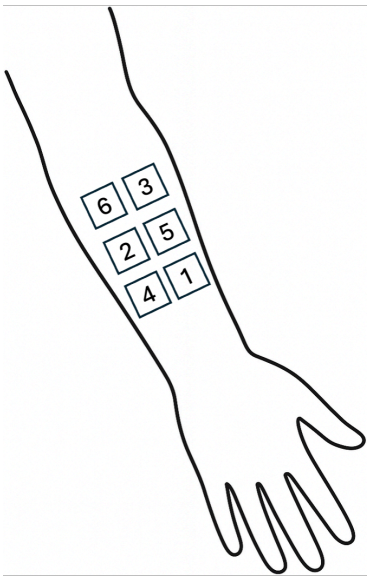

Figure S3: Distribution of skin patches. Six patches on the left inner forearm were arranged in two parallel columns (three per column). To minimize habituation and sensitization effects, we alternated between patches to maximize the distance between two consecutive stimulations. Throughout the experiment, each patch received two stimulations. Stimulation order was counterbalanced across participants: participants with odd IDs received stimulation in ascending order (1-6, repeated twice) with calibration on patch 5; even IDs received descending order (6-1, repeated twice) with calibration on patch 2.

Supplementary Material B: Participant Characteristics

Table S1: Demographic characteristics of all participants

ID	Age	Gender	BMI	Education	Employment	Physical activity	Meditation
1	22	f	22	University degree	Student	Several times a week	No
2	23	f	27	Abitur (university entrance qualification)	Student	Daily	No
3	23	f	23	University degree	Employed part-time	Several times a week	No
4	35	f	20	University degree	Unemployed	Several times a week	No
5	25	f	27	Trade/technical/vocational training	Student	Several times a week	No
6	20	f	23	Abitur (university entrance qualification)	Student	Several times a week	No
7	27	f	19	Abitur (university entrance qualification)	Student	Several times a week	No
8	23	f	20	Abitur (university entrance qualification)	Student	Several times a week	No
9	23	f	27	Abitur (university entrance qualification)	Student	Once a week	No
10	22	f	21	Abitur (university entrance qualification)	Student	Several times a week	Yes
11	18	f	20	Abitur (university entrance qualification)	Student	Several times a week	No
12	22	f	21	Abitur (university entrance qualification)	Student	Once a week	No
13	39	f	26	Secondary school leaving certificate (Mittlere Reife)	Employed full-time	Once a week	No
14	30	f	22	Abitur (university entrance qualification)	Student	A few times a month	No
15	23	f	28	Abitur (university entrance qualification)	Student	Once a week	No
16	29	f	29	Abitur (university entrance qualification)	Unemployed	A few times a month	No
17	26	f	21	University degree	Student	Several times a week	No
18	24	f	26	Abitur (university entrance qualification)	Student	Several times a week	No
19	27	f	20	Abitur (university entrance qualification)	Student	Several times a week	No
20	27	f	22	Abitur (university entrance qualification)	Student	Daily	No
21	22	f	20	University degree	Student	Several times a week	No

ID	Age	Gender	BMI	Education	Employment	Physical activity	Meditation
22	27	f	24	University degree	Employed part-time	Once a week	No
23	27	f	23	Trade/technical/vocational training	Student	A few times a month	No
24	22	f	23	Abitur (university entrance qualification)	Student	Several times a week	No
25	25	f	21	Abitur (university entrance qualification)	Student	Several times a week	No
26	27	f	21	University degree	Employed full-time	Several times a week	Yes
27	34	f	22	University degree	Student	Several times a week	No
28	22	f	26	Abitur (university entrance qualification)	Student	Several times a week	No
29	22	f	26	Abitur (university entrance qualification)	Student	Several times a week	No
30	34	f	24	Abitur (university entrance qualification)	Student	Several times a week	No
31	28	f	26	University degree	Unemployed	Several times a week	No
32	26	f	20	University degree	Student	Several times a week	No
33	22	f	21	Abitur (university entrance qualification)	Student	Once a week	No
34	25	f	25	University degree	Student	Several times a week	No
35	20	f	20	Abitur (university entrance qualification)	Student	Several times a week	No
36	39	f	27	Trade/technical/vocational training	Student	A few times a month	No
37	35	f	23	University degree	Student	A few times a month	No
38	24	f	25	Abitur (university entrance qualification)	Student	Once a week	No
39	25	f	20	University degree	Student	Several times a week	No
40	35	f	22	University degree	Unemployed	A few times a month	No
41	28	f	30	Trade/technical/vocational training	Student	Several times a week	No
42	22	f	28	Abitur (university entrance qualification)	Student	Several times a week	No

,

Table S2: Individual Pain Catastrophizing Scale scores and subscales

ID	BDI-II	STAI-T-10	PCS Total	PCS Rumination	PCS Magnification	PCS Helplessness
1	1	18.57	14	7	5	2
2	8	34.29	26	10	6	10
3	3	38.57	17	7	2	8
4	0	68.57	26	13	3	10
5	4	17.14	19	9	4	6
6	1	45.71	17	6	4	7
7	0	48.57	20	8	3	9
8	3	30	22	7	6	9
9	1	32.86	20	7	5	8
10	5	35.71	29	8	9	12
11	3	38.57	9	6	0	3
12	2	62.86	12	1	3	8
13	3	30	13	4	2	7
14	0	2.86	16	8	1	7
15	6	68.57	18	9	4	5
16	4	38.57	32	13	6	13
17	3	42.86	16	4	6	6
18	5	22.86	28	13	4	11
19	0	20	30	11	5	14
20	0	15.71	7	1	0	6
21	0	5.71	1	0	1	0
22	4	45.71	17	7	6	4
23	4	41.43	23	8	8	7
24	1	21.43	0	0	0	0
25	0	55.71	33	13	8	12
26	0	14.29	14	8	3	3
27	0	41.43	28	10	5	13
28	9	48.57	20	13	0	7
29	1	27.14	21	11	4	6
30	0	7.14	2	1	0	1
31	0	12.86	0	0	0	0
32	12	57.14	16	9	6	1
33	10	35.71	16	6	4	6
34	7	38.57	24	13	4	7
35	8	51.43	25	9	6	10

ID	BDI-II	STAI-T-10	PCS Total	PCS Rumination	PCS Magnification	PCS Helplessness
36	3	25.71	19	8	4	7
37	0	31.43	0	0	0	0
38	7	47.14	22	9	6	7
39	4	25.71	21	9	4	8
40	1	48.57	9	4	2	3
41	11	35.71	20	8	3	9
42	10	48.57	15	2	5	8

Table S3: Mindful Attention Awareness Scale (MAAS) and Pain Vigilance and Awareness Questionnaire (PVAQ) scores

ID	MAAS	PVAQ Total	PVAQ Attention to pain	PVAQ Attention to changes
1	4.53	37	19	18
2	3.67	37	20	17
3	4.13	23	12	11
4	3.33	44	21	23
5	3.2	32	16	16
6	4	37	23	14
7	4.6	38	18	20
8	3.47	36	18	18
9	4.27	35	20	15
10	3	53	31	22
11	4.2	30	10	20
12	3.2	33	17	16
13	3.47	46	24	22
14	4.87	32	16	16
15	3.47	39	20	19
16	3.73	45	22	23
17	4.2	28	13	15
18	4.87	60	35	25
19	4.67	41	21	20
20	5.53	31	15	16
21	5.4	54	27	27
22	3.53	51	31	20
23	4.6	27	10	17

ID	MAAS	PVAQ Total	PVAQ Attention to pain	PVAQ Attention to changes
24	4.47	37	21	16
25	4	42	25	17
26	3.93	30	16	14
27	4.2	50	28	22
28	2.93	33	12	21
29	3.93	34	18	16
30	4.8	35	21	14
31	3.27	28	16	12
32	3.93	49	26	23
33	3.6	29	14	15
34	4.33	24	9	15
35	3.47	40	23	17
36	5.33	46	25	21
37	4	51	27	24
38	3	46	26	20
39	3.6	37	18	19
40	5.6	45	20	25
41	3.47	31	12	19
42	1.47	29	15	14

Table S4: Positive and Negative Affect Schedule (PANAS) scores before and after the experiment

ID	Positive affect (Pre)	Positive affect (Post)	Negative affect (Pre)	Negative affect (Post)
1	3.5	3.5	1.2	1
2	3.9	3.8	1.3	1.1
3	1.9	1.8	1.2	1
4	2.8	2.3	1.2	1.3
5	3.8	3	1.8	1.8
6	2.1	1.7	1	1.1
7	3.1	2.7	1.3	1.2
8	3.6	2.7	1.3	1.5
9	3.8	3.5	1.1	1.1
10	4	3.1	1.1	1.1
11	3.5	2.9	1.3	1.2

ID	Positive affect (Pre)	Positive affect (Post)	Negative affect (Pre)	Negative affect (Post)
12	2.7	2.2	1.3	1.3
13	2.9	2.6	1.2	1.1
14	4.1	3.9	1	1
15	3.3	3.1	1.4	1.2
16	3.1	2.4	1	1.2
17	3.1	2.1	1.4	1.1
18	3.4	2.9	1.3	1.6
19	3.9	3.5	1.2	1.1
20	3.5	3.1	1.2	1.2
21	3.8	4.1	1.3	1
22	2.7	2.5	1.1	1
23	3.3	2.4	1.3	1
24	4.1	4	1	1
25	3.3	2.7	1.1	1.3
26	3.4	2.2	1	1
27	3.5	3.3	1.2	1.8
28	3.3	3.1	1.2	1
29	3.4	3.1	1.3	1
30	3.9	3.9	1.1	1
31	2.3	1.7	1	1
32	2.5	2.7	1.4	1.2
33	3.4	2.9	1.8	1.1
34	2.8	2	1.7	1.3
35	3.2	2.9	1.4	1.3
36	2.9	3.1	1	1.3
37	3.1	2.6	1	1.1
38	2.9	2.1	1.2	1.3
39	3.4	1.7	1	1.1
40	3.2	3.3	1	1
41	3.4	3.1	1.2	1.1
42	3.6	4	1.1	1.2

Supplementary Material C: Deep Learning Model Architectures

Multilayer Perceptron

A multilayer perceptron (MLP) is a feedforward artificial neural network consisting of an input layer, one or more hidden layers, and an output layer. Information flows unidirectionally from the input layer through the hidden layers to the output layer. Each layer computes a weighted sum of its inputs followed by a nonlinear activation function (commonly ReLU or sigmoid). During training, the network optimizes its weights and biases through backpropagation, an algorithm that iteratively adjusts parameters to minimize a loss function by computing gradients via the chain rule. MLPs are universal function approximators [4] capable of learning complex nonlinear relationships between input features and target outputs, making them well-suited for classifications. For multivariate time series, the input is flattened into a single input vector (1D) as MLPs inherently do not understand temporal structure [7].

LightTS

LightTS is a lightweight MLP-based architecture designed for efficient multivariate time series processing [14]. It implements two complementary down-sampling strategies before applying the data to MLP layers: interval sampling and continuous sampling. Interval sampling extracts values at regular intervals to capture long-term patterns, while continuous sampling selects contiguous subsequences to preserve local temporal dependencies. These down-sampled representations are then processed through separate MLP structures, and their outputs are aggregated to produce the final output. This design exploits the observation that down-sampled time series often retain the majority of relevant information while drastically reducing computational requirements [14].

TimesNet

TimesNet is a CNN-based architecture for time series analysis that leverages the fact that many time series exhibit multiple periodic patterns [13]. Rather than processing temporal data in its native 1D format, TimesNet first identifies dominant periods in the signal and reshapes the 1D sequence into 2D tensors according to these periods. This transformation arranges variations within each cycle as columns and variations across cycles as rows, essentially converting temporal patterns into spatial patterns. The architecture then applies

special 2D convolutional kernels (inception blocks) to extract features from these reshaped representations. This approach is more parameter-efficient than modeling complex temporal dependencies directly in 1D. In the review by Wang et al. [12], this architecture achieved the best performance for classification tasks across a variety of multivariate time series compared with other state-of-the-art deep learning architectures.

PatchTST

PatchTST's architecture is based on the transformer architecture and is designed for the efficient analysis of multivariate time series [10]. Instead of treating individual time points as tokens (discrete input units processed by the model), it segments the input sequence into patches that serve as input to the transformer encoder. PatchTST employs a channel-independent strategy where each univariate series is processed separately with shared weights. This approach reduces both overfitting risk and computational requirements. The self-attention mechanism (a method that weighs the relevance of different parts of the input to each other) captures long-range dependencies between patches. This makes PatchTST particularly suitable for multivariate physiological signal processing where multiple channels need to be analyzed efficiently.

iTransformer

iTransformer is a transformer-based architecture that inverts the traditional application of attention mechanisms in multivariate time series analysis [9]. Unlike conventional transformers that apply attention across time points, iTransformer treats each variate (each individual series within the multivariate dataset) as a token and applies self-attention across the variate dimension. This inversion allows the model to capture multivariate correlations directly while reducing computational complexity. Conceptually, this resembles a learned, dynamic correlation matrix that adapts to the input data. The feed-forward network is applied to each variate token to learn nonlinear representations. The design is particularly advantageous when inter-variate relationships are more informative than temporal patterns, such as in physiological signal analysis where multiple biosignals interact simultaneously.

EEGNet

EEGNet is an efficient convolutional neural network architecture specifically designed for EEG-based brain-computer interfaces [6]. It incorporates domain knowledge from EEG analysis by employing temporal convolution to learn frequency-specific filters, depthwise spatial convolution to capture optimal electrode combinations, and separable convolutions to extract additional features in the data. Despite its compact size, EEGNet demonstrates competitive performance across diverse EEG classification tasks and performs well even when training data is limited [6].

Ensemble Models

The ensemble models combined EEGNet for EEG data and PatchTST for peripheral physiological and facial expression data. This design grouped related signals and addressed the different sampling rates inherent in our multimodal dataset: EEG signals at 250 Hz and physiological signals (EDA, heart rate, pupil diameter, facial expressions) at 10 Hz. For feature sets combining EEG with other physiological data, non-EEG signals were upsampled to 250 Hz using linear interpolation to match the EEG sampling rate and to enable simple sample generation in PyTorch. The architecture split the input by channel, routing EEG channels through EEGNet and other channels through PatchTST. The upsampled input for PatchTST was downsampled back to 10 Hz (taking every 25th data point). Each sub-model had its final classification layers removed and instead served as feature extractors. The extracted features were concatenated and passed through a fusion network consisting of MLP layers. This feature-level fusion approach [2] allowed each sub-model to learn modality-specific representations optimally before combining them for classification, a strategy shown effective for multimodal physiological signal processing [3].

Supplementary Material D: Hyperparameter Optimization, Model Training and Evaluation

Hyperparameter Optimization

We conducted hyperparameter optimization using the Optuna library (version 4.3.0) [1], which implements Bayesian-inspired optimization to efficiently search the hyperparameter space. The hyperparameter search spaces (see Suppl. Table S8) were informed by default values from Wang et al. [12] and Lawhern et al. [6]. We further included learning rate as a tuneable hyperparameter (range: 0.00001 to 0.01, log-scaled), given its critical influence on convergence and performance [8]. For 30 trials, each model's hyperparameter space was optimized by maximizing the accuracy of the validation set. A single trial consisted of 15 epochs with a batch size of 64. We used Optuna-based median pruning (5 startup trials, 5 warmup steps) to terminate unpromising trials early. Additionally, we used a fixed learning rate decay schedule (cosine annealing without restarts and with a maximum number of iterations equaling the number of epochs) to eliminate dependence on validation monitoring during the final model training.

Model Training

We employed cross-entropy as the loss function for our binary classification task. Cross-entropy is defined as

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)],$$

where y represents the vector of true labels, $\hat{y}$ is the vector of predicted probabilities, and N is the number of samples in each batch. For optimization, we utilized the Adam optimizer [5]. To prevent exploding gradients and ensure stable training, we applied gradient clipping with a maximum norm of 4.0.

Model training was performed using PyTorch (version 2.6.0) [11] on an NVIDIA GeForce RTX 3080 Ti GPU. We set a fixed random seed (42) across all experiments for Python, NumPy, PyTorch and Optuna to ensure reproducibility of results. The complete hyperparameter optimization cycle took approximately 3 hours for all model-feature combinations.

Model Evaluation

We utilized accuracy as our primary evaluation metric during model selection. Following this, we evaluated the best-performing model for each feature set using additional classification metrics: area under the receiver operating characteristic curve (AUROC), precision, recall, F_1 -scores and Matthews correlation coefficient (MCC). We defined pain decreases as positive and non-decreases as negative cases. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Accuracy quantified the overall proportion of correct predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

AUROC quantified the model's ability to discriminate between pain decreases and non-decreases across all possible classification thresholds. This metric ranges from 0 to 1, where values near 0.5 indicate chance-level performance, and scores approaching 1 demonstrate strong discriminative capability. Precision (also called positive predictive value (PPV)) evaluated the proportion of true positives among predicted positives, indicating the reliability of positive predictions. Recall (also known as sensitivity) measured the model's ability to correctly identify positive classes (pain decreases), representing the proportion of actual positives correctly identified by the model. The F_1-score provided the harmonic mean of precision and recall, offering a balance between these two metrics particularly useful when class distributions are imbalanced:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

MCC quantified the association between observed and predicted classifications, producing a value between −1 and +1, where +1 represents perfect prediction, 0 represents random prediction, and −1 represents inverse prediction:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}.$$

Supplementary Material E: Exploratory Data Analysis

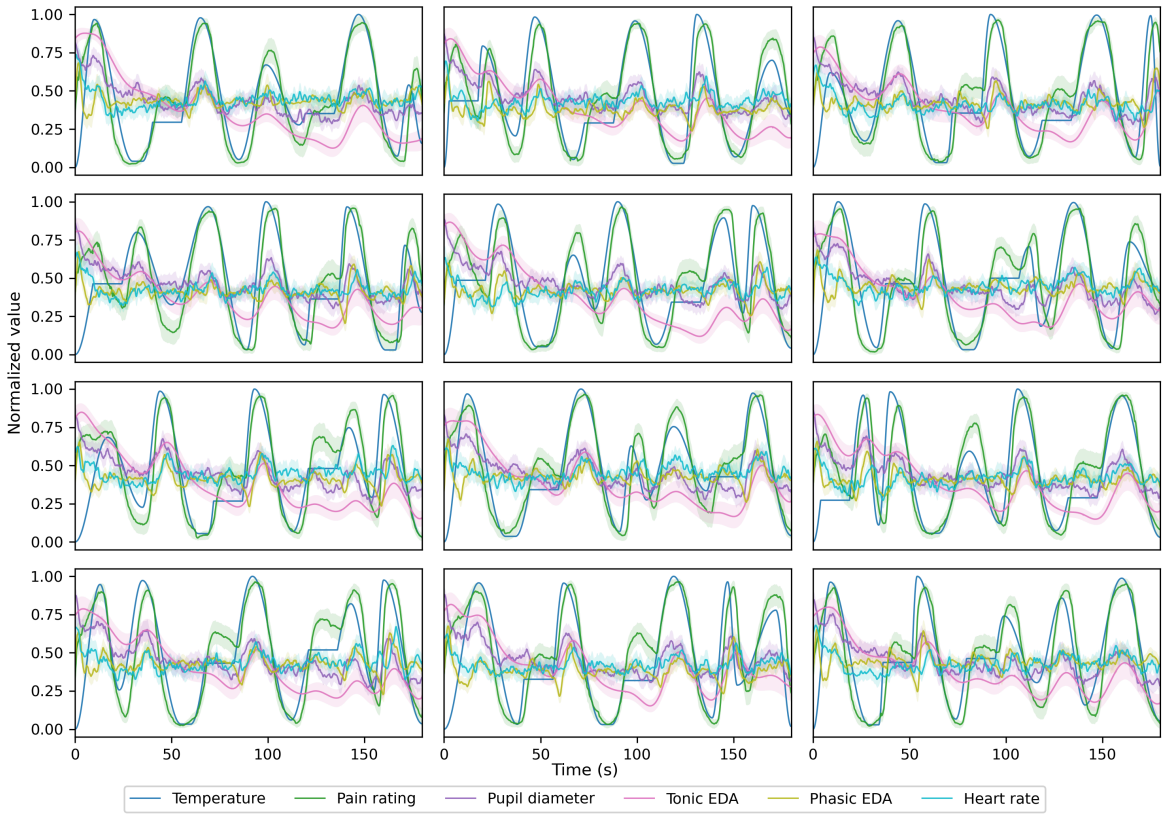

Figure S4: Grand-averaged means of physiological signals

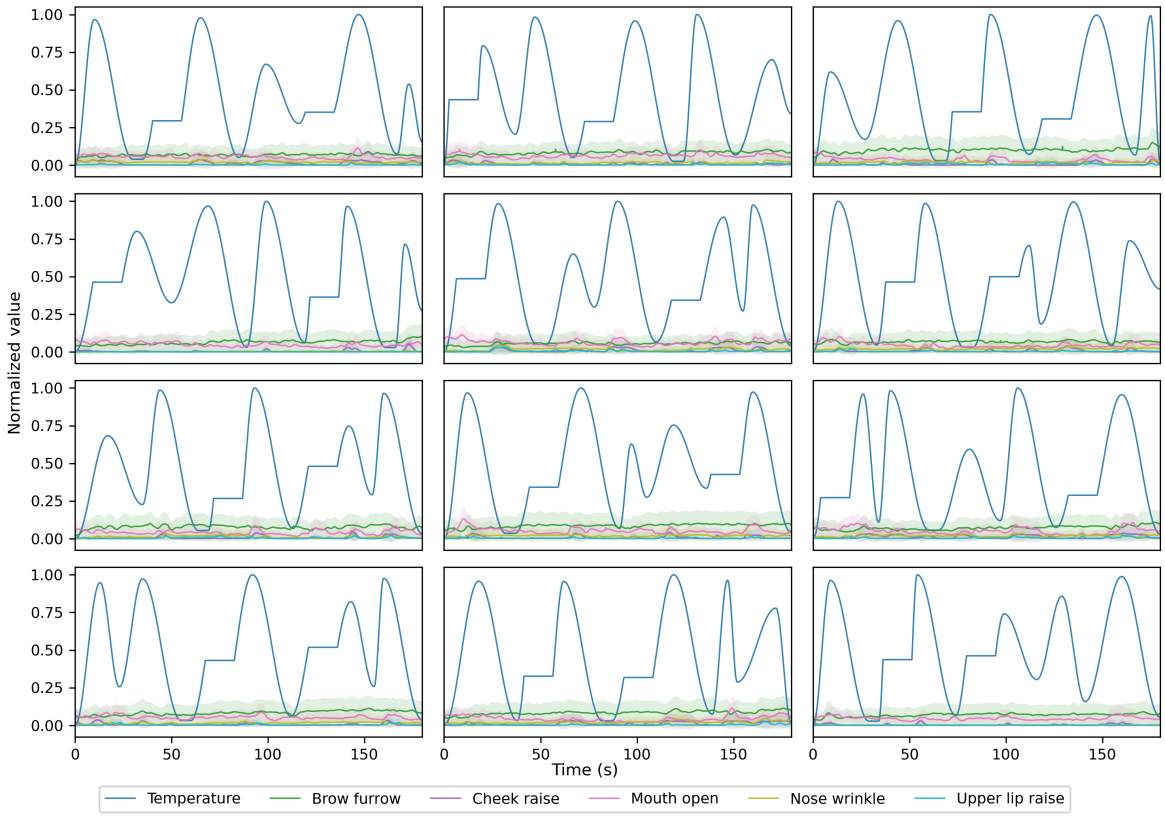

Figure S5: Grand-averaged means of facial expressions

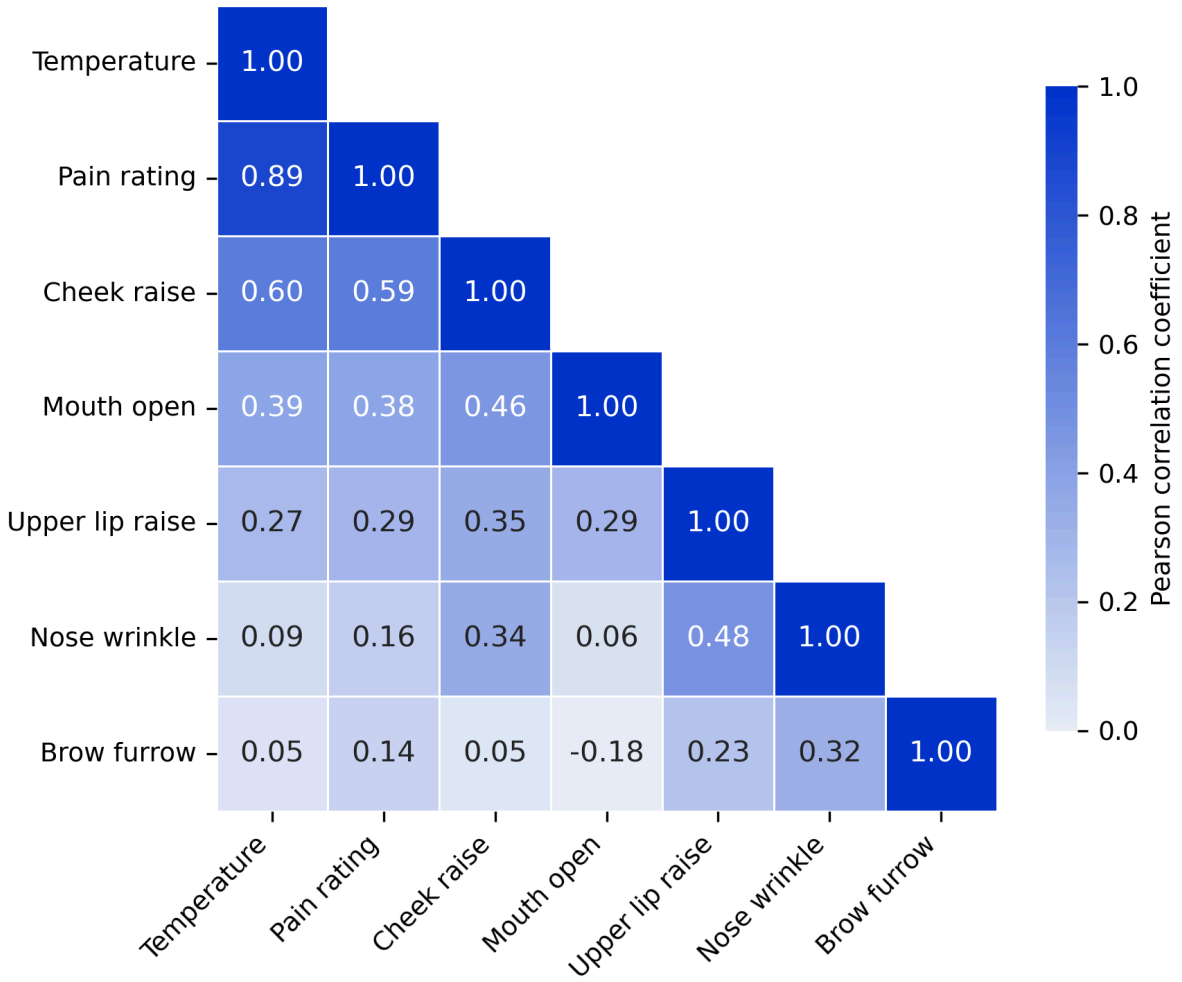

Figure S6: Correlation matrix of grand-averaged facial expressions, temperatures, and pain ratings. The heatmap shows Pearson correlation coefficients between all measured variables, with darker blue indicating stronger positive correlations.

Table S5: Temperature-referenced cross-correlation lags

Reference signal	Target signal	Mean lag (s) ± SD
Temperature	Upper Lip Raise	1.5 ± 7.5
Temperature	Nose Wrinkle	1.43 ± 4.41
Temperature	Pupil Diameter	−0.07 ± 0.22
Temperature	Mouth Open	−0.56 ± 0.71
Temperature	Cheek Raise	−0.63 ± 0.87
Temperature	EDA (tonic)	−0.95 ± 0.59
Temperature	Pain Rating	−1.01 ± 0.34
Temperature	Heart Rate	−1.76 ± 1.36
Temperature	Brow Furrow	−2.45 ± 1.91
Temperature	EDA (phasic)	−2.45 ± 0.89

Note: Positive values indicate the target signal precedes the temperature curve; negative values indicate the target signal follows the temperature curve. Values are presented as mean ± standard deviation across all grand-averaged stimulus functions.

Table S6: Pain rating-referenced cross-correlation lags

Reference signal	Target signal	Mean lag (s) ± SD
Pain Rating	Nose Wrinkle	2.58 ± 4.19
Pain Rating	Upper Lip Raise	2.58 ± 7.52
Pain Rating	Temperature	1.01 ± 0.34
Pain Rating	Cheek Raise	0.95 ± 1.07
Pain Rating	Pupil Diameter	0.61 ± 0.71
Pain Rating	Mouth Open	0.1 ± 0.32
Pain Rating	EDA (tonic)	−0.1 ± 0.18
Pain Rating	Heart Rate	−0.34 ± 0.53
Pain Rating	EDA (phasic)	−0.52 ± 0.47
Pain Rating	Brow Furrow	−1.39 ± 2.09

Note: Positive values indicate the target signal precedes the pain rating; negative values indicate the target signal follows the pain rating. Values are presented as mean ± standard deviation across all grand-averaged stimulus functions.

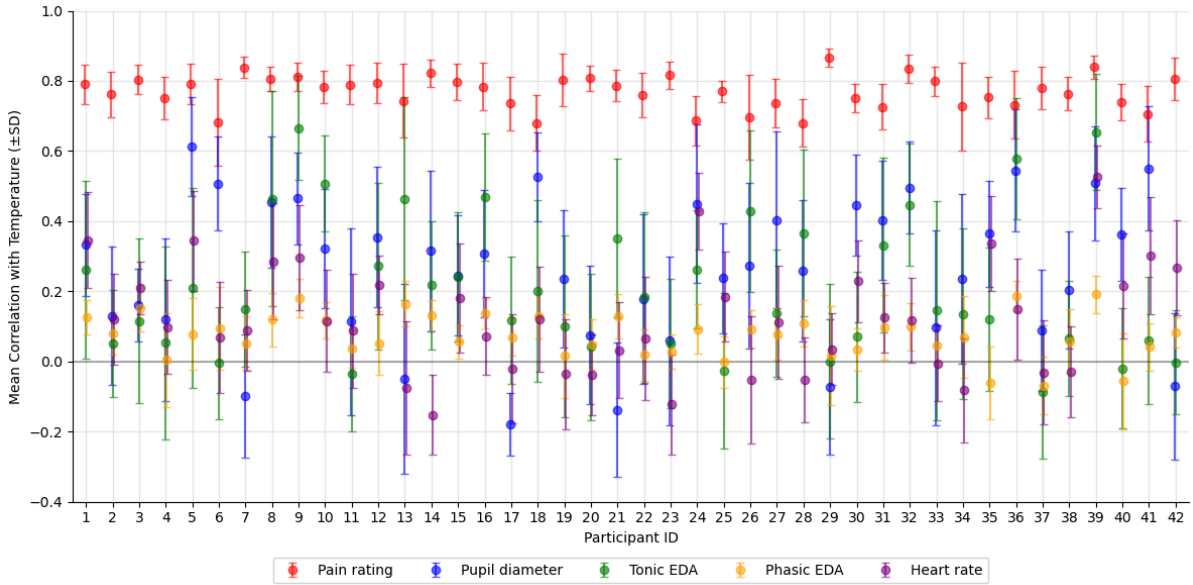

Figure S7: Correlations across participants between temperature and physiological signals

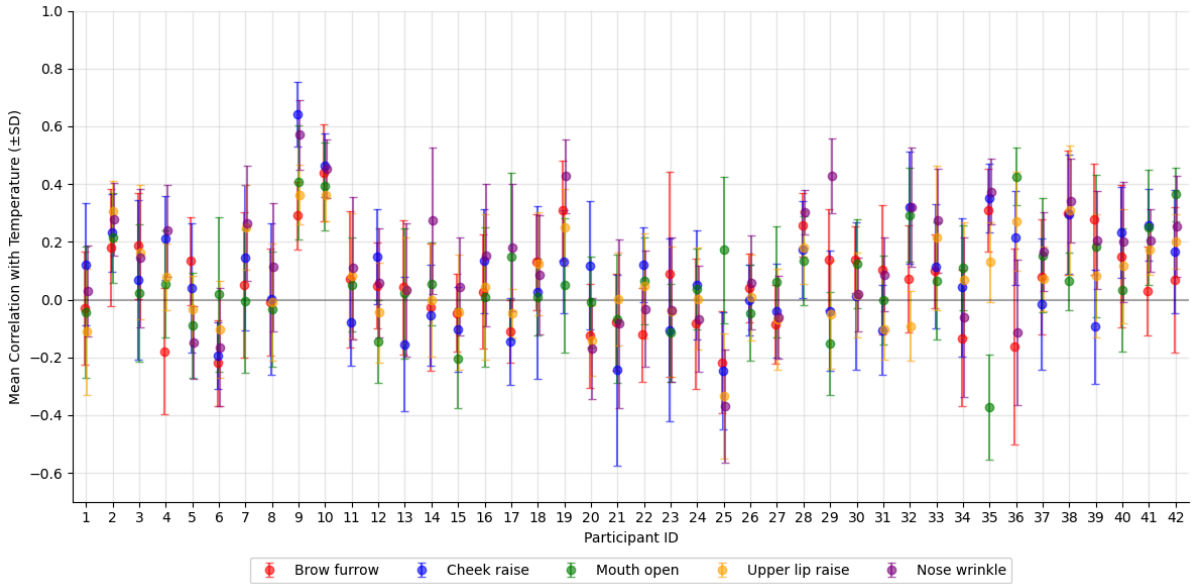

Figure S8: Correlations across participants between temperature and facial expressions

Supplementary Material F: Model Comparison and Optimization Results

Table S7: Validation set accuracies across model architectures

Feature set	MLP	LightTS	TimesNet	PatchTST	iTransf.	Ensemble	EEGNet
EDA	0.665	0.674	0.650	**0.748**	0.663	-	-
HR	**0.609**	0.596	0.572	0.575	0.585	-	-
Pupil	0.666	0.679	0.668	0.679	**0.681**	-	-
EDA, HR	0.655	0.676	0.617	**0.756**	0.628	-	-
EDA, Pupil	0.700	0.717	0.663	**0.741**	0.695	-	-
EDA, HR, Pupil	0.685	0.700	0.663	**0.751**	0.696	-	-
Facial Expressions	0.537	0.530	0.558	**0.575**	0.530	-	-
Combined (w/o EEG)	0.641	0.693	0.650	**0.741**	0.693	0.736	-
EEG	0.570	0.502	-	-	**0.591**	-	0.564
EEG + EDA	0.573	0.554	-	-	0.570	**0.746**	-
Combined	0.647	0.658	-	-	0.655	**0.748**	-

Note: Marked in bold are the highest validation accuracies. Ensemble stands for the ensemble model architecture of the specific feature set. iTransf. = iTransformer, HR = heart rate, EDA = electrodermal activity, EEG = electroencephalography.

Table S8: Hyperparameter search spaces

Model	Hyperparameter	Search space
Multilayer Perceptron	hidden_size	256 - 4096
	depth	1 - 4
	dropout	0.0 - 0.8
	lr	1e-5 - 1e-2 (log)
PatchTST	d_model	32 - 128
	e_layers	1 - 3
	d_ff	128 - 512
	patch_len	4 - 32
	stride	4 - 16
	top_k	2 - 5
	dropout	0.0 - 0.5

Model	Hyperparameter	Search space
	factor	1 - 2
	lr	1e-5 - 1e-2 (log)
iTransformer	d_model	32 - 128
	e_layers	1 - 3
	d_ff	128 - 512
	dropout	0.0 - 0.5
	factor	1 - 2
	lr	1e-5 - 1e-2 (log)
TimesNet	d_model	8 - 32
	e_layers	1 - 3
	d_ff	8 - 64
	top_k	2 - 4
	num_kernels	4 - 8
	dropout	0.0 - 0.5
	lr	1e-5 - 1e-2 (log)
LightTS	d_model	32 - 128
	dropout	0.0 - 0.5
	lr	1e-5 - 1e-2 (log)
EEGNet	F1	8 - 16
	D	2 - 5
	dropout	0.1 - 0.5
	lr	1e-5 - 1e-2 (log)

Note: Ensemble models consisting of EEGNet and PatchTST models employed the same hyperparameter search spaces as their individual components. Model parameters: hidden_size (number of neurons per layer), depth (number of hidden layers), d_model (model embedding dimension), e_layers (number of encoder layers), d_ff (feed-forward network dimension), patch_len (length of input patches), stride (patch stride), top_k (number of frequencies to keep), factor (attention scaling factor), num_kernels (number of convolutional kernels), F1

(number of temporal filters), D (depth multiplier for spatial filters), dropout (dropout probability), lr (learning rate with logarithmic sampling).

Table S9: Optimized hyperparameters from the model comparison

Feature set	Best model	Optimized hyperparameters
EDA	PatchTST	d_model=32, e_layers=3, d_ff=512, patch_len=16, stride=16, top_k=2, dropout=0.179, factor=1, lr=0.00388
Pupil	iTransformer	d_model=64, e_layers=2, d_ff=256, factor=2, dropout=0.003, lr=0.00361
HR	MLP	hidden_size=2048, depth=1, dropout_rate=0.234, lr=0.00013
EDA + Pupil	PatchTST	d_model=64, e_layers=2, d_ff=128, patch_len=16, stride=4, top_k=2, dropout=0.474, factor=2, lr=0.00266
EDA + HR	PatchTST	d_model=64, e_layers=3, d_ff=512, patch_len=16, stride=8, top_k=2, dropout=0.07, factor=1, lr=0.00063
EDA + HR + Pupil	PatchTST	d_model=32, e_layers=3, d_ff=512, patch_len=16, stride=8, top_k=2, dropout=0.145, factor=1, lr=0.00112
Facial Expressions	PatchTST	d_model=32, e_layers=3, d_ff=128, patch_len=16, stride=16, top_k=4, dropout=0.063, factor=1, lr=0.00015
Combined (w/o EEG)	PatchTST	d_model=32, e_layers=2, d_ff=256, patch_len=16, stride=16, top_k=2, dropout=0.078, factor=1, lr=0.00318
EEG	iTransformer	d_model=64, e_layers=3, d_ff=512, factor=1, dropout=0.078, lr=0.00001
EEG + EDA	EEGPhysioEnsemble	F1=16, D=3, eeg_dropout=0.427, d_model=64, e_layers=3, d_ff=256,

Feature set	Best model	Optimized hyperparameters
		patch_len=16, stride=16, physio_dropout=0.042, fusion_hidden_dim=64, fusion_dropout=0.251, lr=0.00054
Combined	EEGFacePhysioEnsemble	F1=8, D=2, eeg_dropout=0.449, face_d_model=32, face_e_layers=2, face_d_ff=128, face_patch_len=8, face_stride=4, face_dropout=0.356, physio_d_model=128, physio_e_layers=1, physio_d_ff=128, physio_patch_len=16, physio_stride=4, physio_dropout=0.167, fusion_hidden_dim=64, fusion_dropout=0.276, lr=0.00035

Note: Ensemble models consisting of EEGNet and PatchTST models employed the same hyperparameter search spaces as their individual components. Model parameters: hidden_size (number of neurons per layer), depth (number of hidden layers), d_model (model embedding dimension), e_layers (number of encoder layers), d_ff (feed-forward network dimension), patch_len (length of input patches), stride (patch stride), top_k (number of frequencies to keep), factor (attention scaling factor), num_kernels (number of convolutional kernels), F1 (number of temporal filters), D (depth multiplier for spatial filters), dropout (dropout probability), lr (learning rate with logarithmic sampling).

Table S10: Test set sample sizes for each participant

Participant ID	Samples	Non-decreases	Decreases
5	72	36	36
9	78	42	36
14	76	40	36
20	69	33	36
26	66	30	36
27	64	28	36
30	34	16	18
31	51	24	27
40	75	39	36
total	**585**	**288**	**297**

Table S11: Test set accuracies box plot values

Feature set	Mean	Median	Q1	Q3	IQR	Min	Max
EDA	0.760	0.763	0.720	0.824	0.104	0.652	0.846
HR	0.534	0.493	0.474	0.549	0.075	0.424	0.778
Pupil	0.669	0.694	0.606	0.733	0.127	0.536	0.765
EDA + HR	0.751	0.784	0.676	0.816	0.139	0.623	0.844
EDA + Pupil	0.777	0.812	0.706	0.872	0.166	0.623	0.922
EDA + HR + Pupil	0.764	0.766	0.676	0.846	0.170	0.636	0.902
Facial Expressions	0.570	0.547	0.528	0.618	0.090	0.464	0.693
Combined (w/o EEG)	0.745	0.797	0.706	0.806	0.100	0.609	0.824
EEG	0.529	0.522	0.480	0.590	0.110	0.409	0.625
EEG + EDA	0.734	0.750	0.706	0.792	0.086	0.560	0.824
Combined	0.748	0.750	0.712	0.784	0.072	0.647	0.842

Supplementary Material G: Model Inference Visualizations

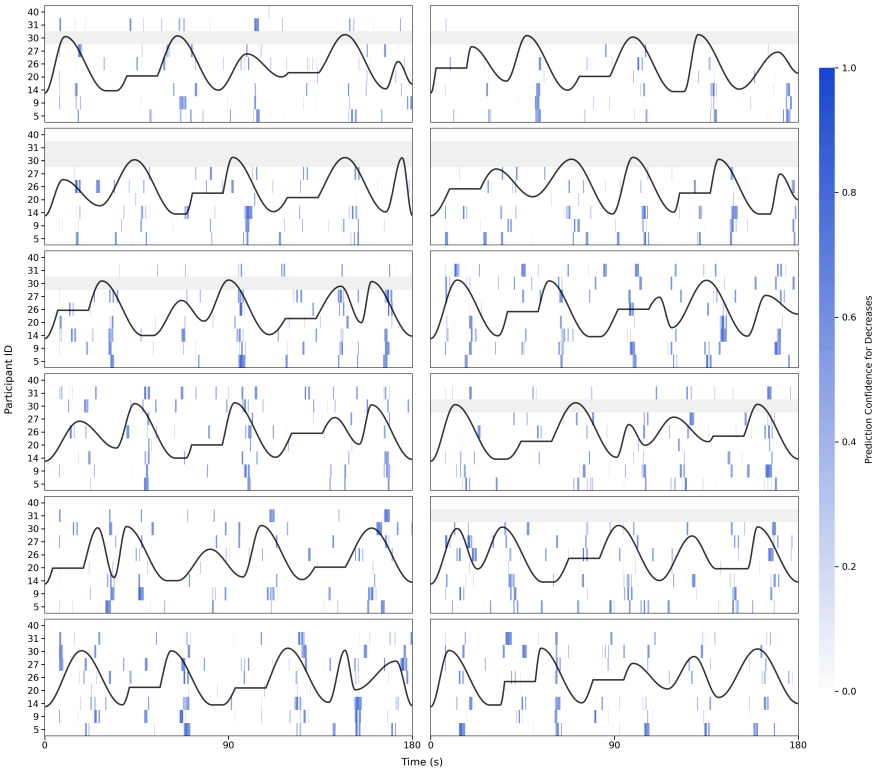

Figure S9: Continuous pain decrease prediction heatmap using EDA. Threshold = 0.85.

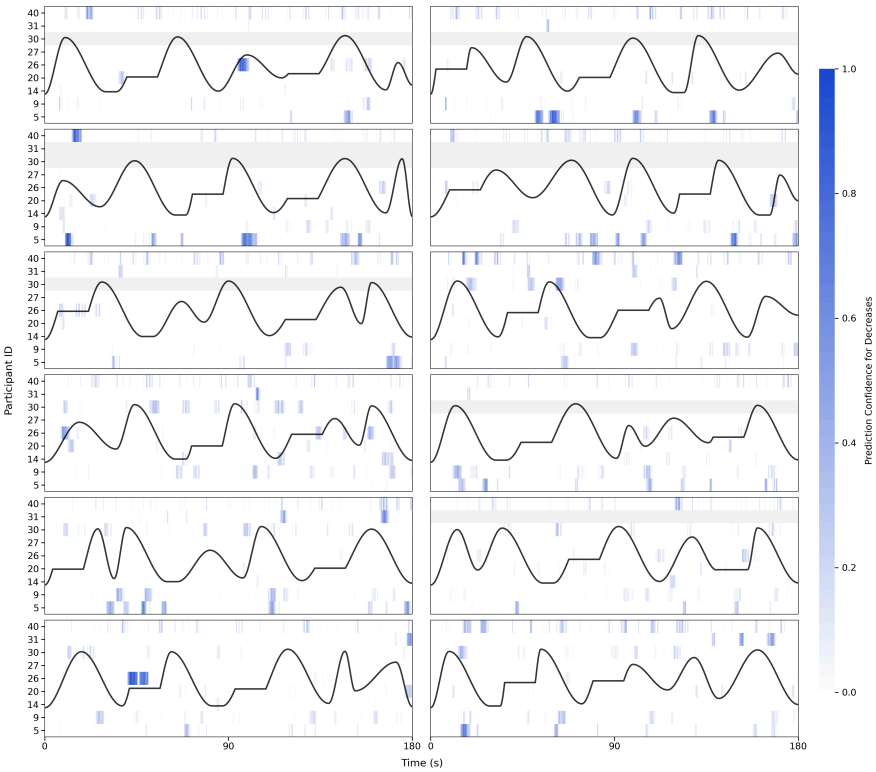

Figure S10: Continuous pain decrease prediction heatmap using HR. Threshold = 0.60.

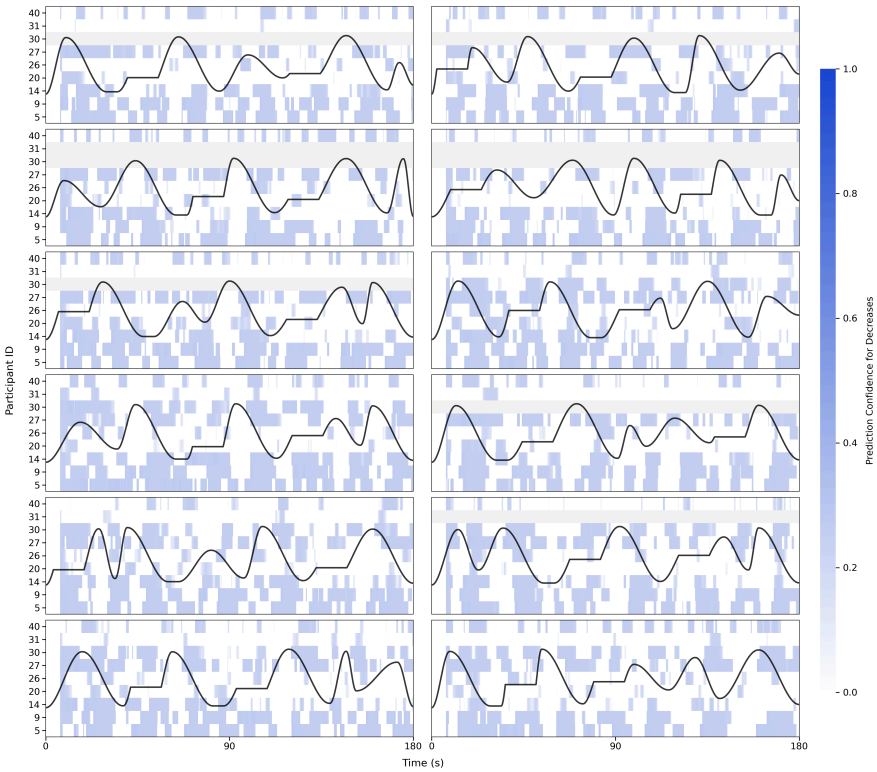

Figure S11: Continuous pain decrease prediction heatmap using Pupil. Threshold = 0.50.

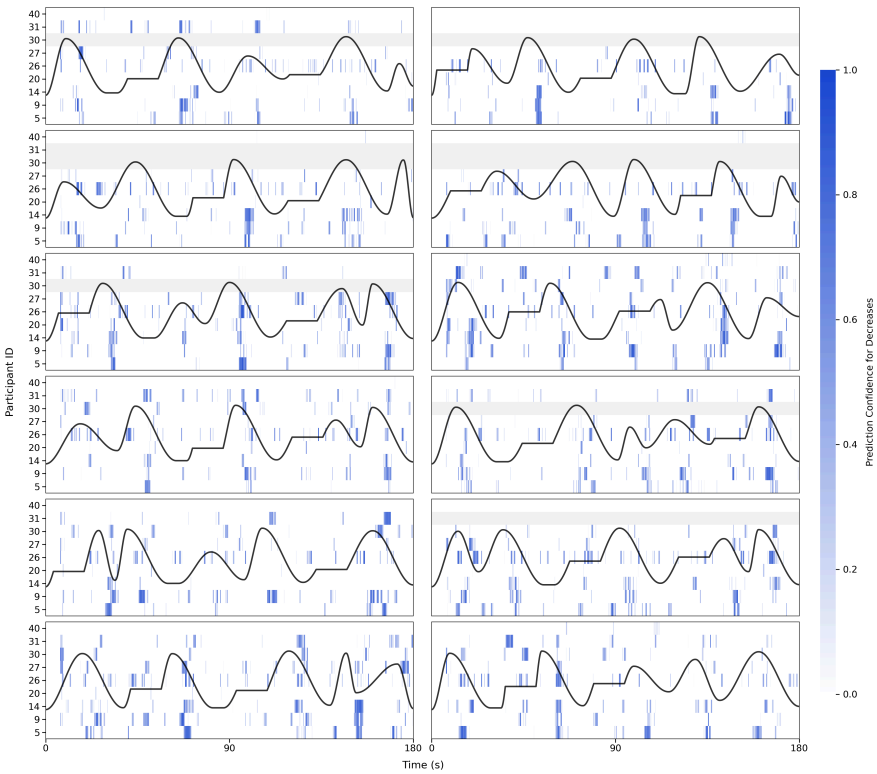

Figure S12: Continuous pain decrease prediction heatmap using EDA + HR. Threshold = 0.85.

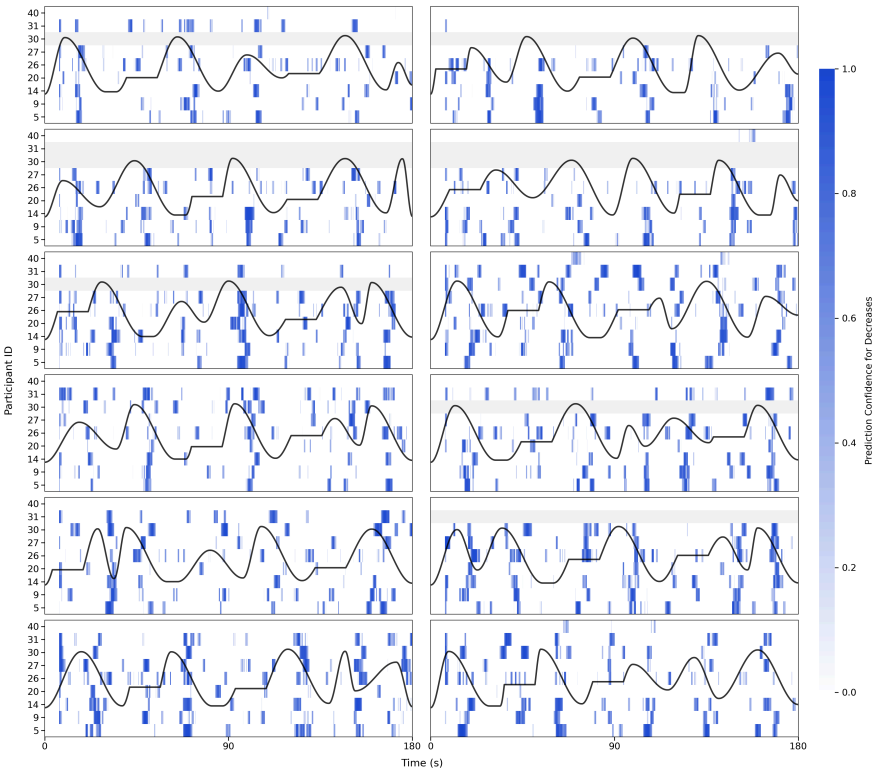

Figure S13: Continuous pain decrease prediction heatmap using EDA + Pupil. Threshold = 0.90.

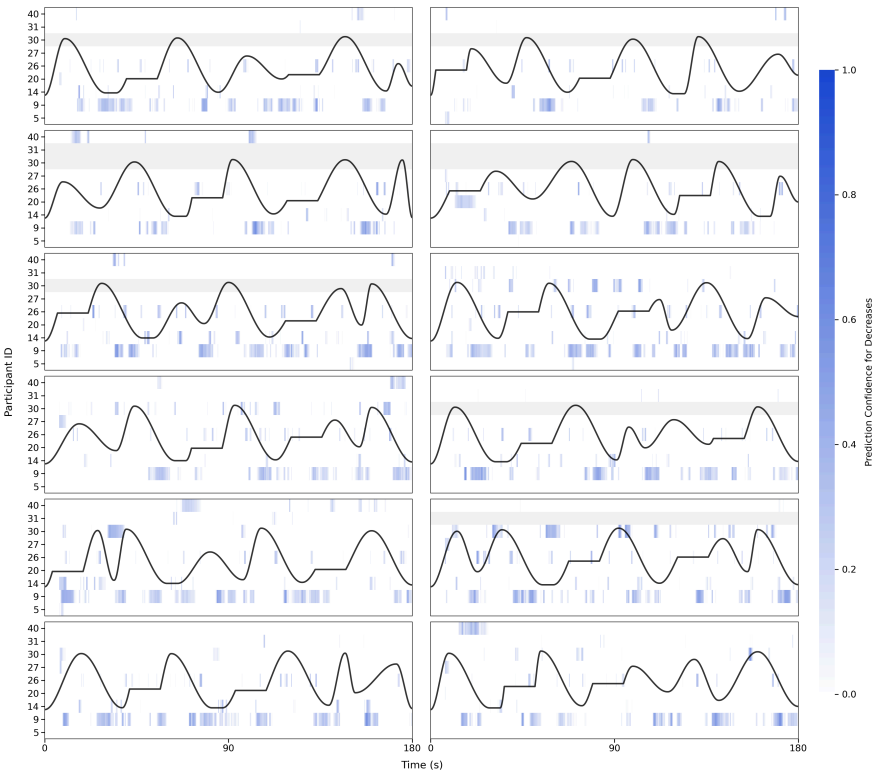

Figure S14: Continuous pain decrease prediction heatmap using facial expressions. Threshold = 0.70.

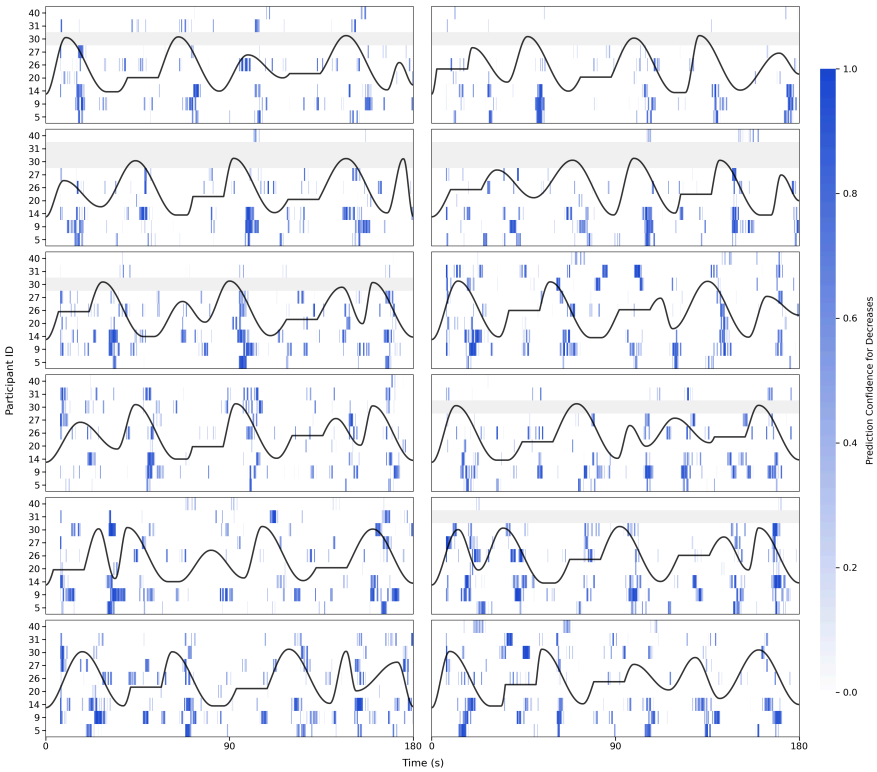

Figure S15: Continuous pain decrease prediction heatmap using combined signals (without EEG). Threshold = 0.90.

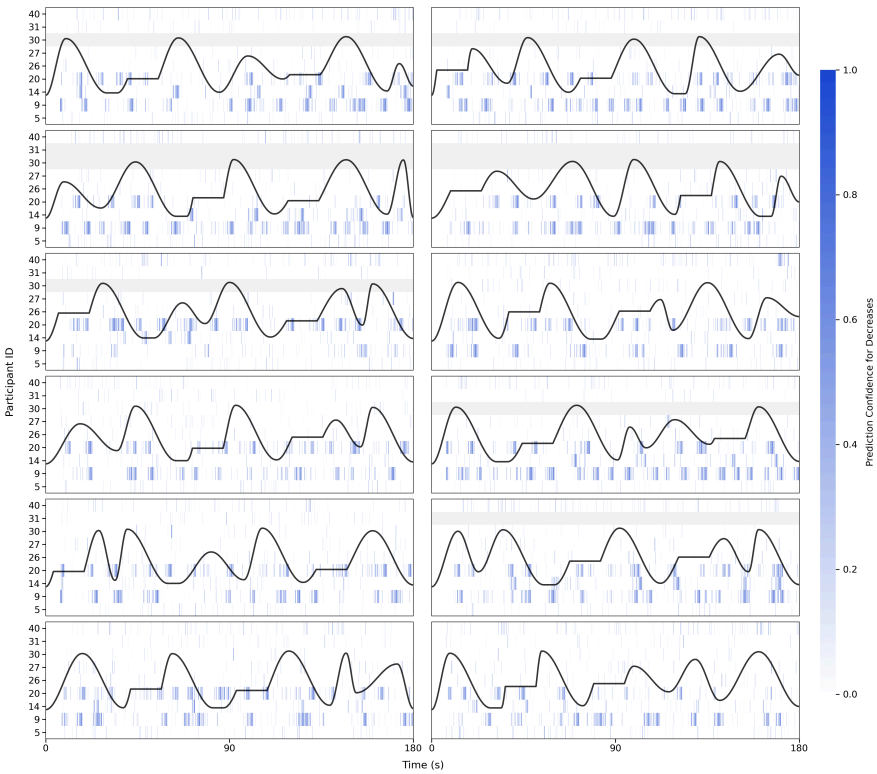

Figure S16: Continuous pain decrease prediction heatmap using EEG. Threshold = 0.65.

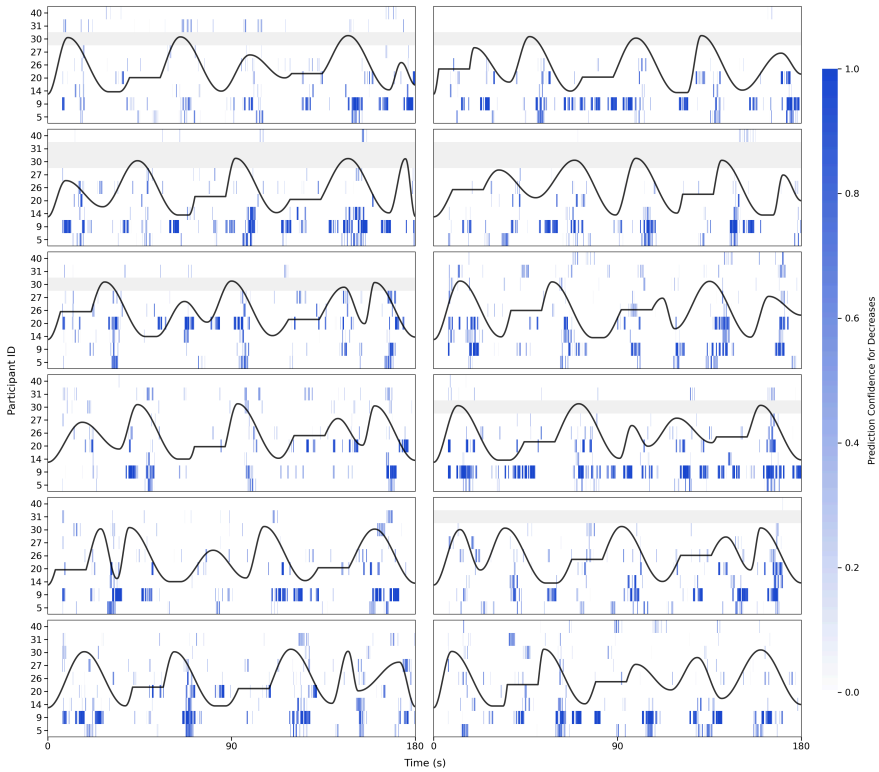

Figure S17: Continuous pain decrease prediction heatmap using EEG + EDA. Threshold = 0.90.

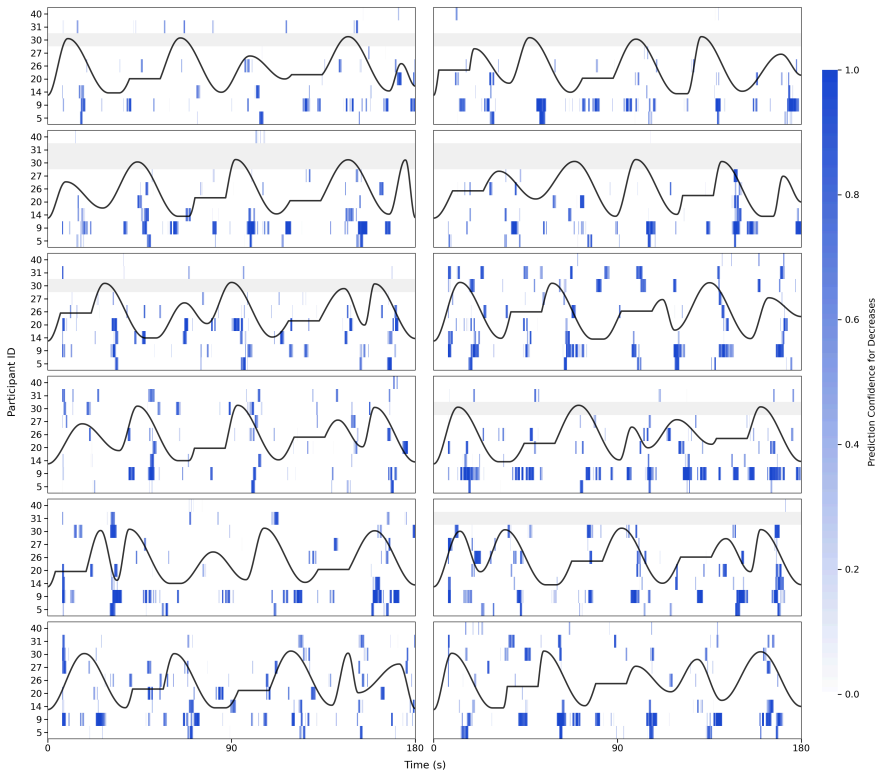

Figure S18: Continuous pain decrease prediction heatmap using combined signals. Threshold = 0.90.

Supplementary References

[1] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. arXiv preprint arXiv:1907.10902. 2019. https://doi.org/10.48550/arXiv.1907.10902

[2] Baltrušaitis T, Ahuja C, Morency L-P. Multimodal machine learning: A survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 2019;41:423–43. doi:10.1109/TPAMI.2018.2798607

[3] Bota P, Wang C, Fred A, Silva H. Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet? Sensors 2020;20:4723. doi:10.3390/s20174723

[4] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw 1989;2:359–66. doi:10.1016/0893-6080(89)90020-8

[5] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2017. https://doi.org/10.48550/arXiv.1412.6980

[6] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. J Neural Eng 2018;15:056013. doi:10.1088/1741-2552/aace8c

[7] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. doi:10.1038/nature14539

[8] LeCun YA, Bottou L, Orr GB, Müller KR. Efficient backprop. In: Neural Networks. Springer Verlag, 2012. pp. 9–48. doi:10.1007/978-3-642-35289-8_3

[9] Liu Y, Hu T, Zhang H, Wu H, Wang S, Ma L, Long M. iTransformer: Inverted transformers are effective for time series forecasting. arXiv preprint arXiv:2310.06625. 2024. https://doi.org/10.48550/arXiv.2310.06625

[10] Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A time series is worth 64 words: Long-term forecasting with transformers. arXiv preprint arXiv:2211.14730. 2023. https://doi.org/10.48550/arXiv.2211.14730

[11] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. PyTorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703. 2019. https://doi.org/10.48550/arXiv.1912.01703

[12] Wang Y, Wu H, Dong J, Liu Y, Long M, Wang J. Deep time series models: A comprehensive survey and benchmark. arXiv preprint arXiv:2407.13278. 2024. https://doi.org/10.48550/arXiv.2407.13278

[13] Wu H, Xu J, Wang J, Long M. TimesNet: Temporal 2D-variation modeling for general time series analysis. Proceedings of the 40th International Conference on Machine Learning, 2023. pp. 24539–54. https://doi.org/10.48550/arXiv.2210.02186

[14] Zhang T, Zhang Y, Cao W, Bian J, Yi X, Zheng S, Li J. Less is more: Fast multivariate time series forecasting with light sampling-oriented MLP structures. arXiv preprint arXiv:2207.01186. 2022. https://arxiv.org/abs/2207.01186v1