Received: 20 December 2022 | Revised: 28 August 2023 | Accepted: 29 August 2023

DOI: 10.1111/psyp.14446

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

Emotional arousal pattern (EMAP): A new database for modeling momentary subjective and psychophysiological responding to affective stimuli

Hedwig Eisenbarth[1] | Matt Oxner[1,2] | Harisu Abdullahi Shehu[3] | Tim Gastrell[1] | Amy Walsh[4,5] | Will N. Browne[6] | Bing Xue[3]

[1]School of Psychology, Victoria University of Wellington, Wellington, New Zealand

[2]Institute of Psychology, University of Leipzig, Leipzig, Germany

[3]School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand

[4]Department of Clinical Neuroscience, Psychology Division, Care Sciences and Society at Karolinska Institutet Solnavagen, Solna, Sweden

[5]Department of Neurobiology, Aging Research Centre, Care Sciences and Society at Karolinska Institutet Solnavagen, Solna, Sweden

[6]School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, Queensland, Australia

Correspondence

Hedwig Eisenbarth, School of Psychology, Victoria University of Wellington, Wellington 6012, New Zealand.

Email: hedwig.eisenbarth@vuw.ac.nz

Funding information

Faculty of Science, Victoria University of Wellington

Abstract

This article describes a new database (named "EMAP") of 145 individuals' reactions to emotion-provoking film clips. It includes electroencephalographic and peripheral physiological data as well as moment-by-moment ratings for emotional arousal in addition to overall and categorical ratings. The resulting variation in continuous ratings reflects inter-individual variability in emotional responding. To make use of the moment-by-moment data for ratings as well as neurophysiological activity, we used a machine learning approach. The results show that algorithms that are based on temporal information improve predictions compared to algorithms without a temporal component, both within and across participant modeling. Although predicting moment-by-moment changes in emotional experiences by analyzing neurophysiological activity was more difficult than using aggregated experience ratings, selecting a subset of predictors improved the prediction. This also showed that not only single features, for example, skin conductance, but a range of neurophysiological parameters explain variation in subjective fluctuations of subjective experience.

KEYWORDS

affective computing, data set, electroencephalogram (EEG), emotional corpora, machine learning, modeling human emotion, physiological signals

Hedwig Eisenbarth, Matt Oxner, and Harisu Abdullahi Shehu contributed equally to this work and are listed in alphabetical order.

1 | INTRODUCTION

Research on psychophysiological responses to emotional valence cues in our environment has provoked more questions for affective science than answers (Barrett & Satpute, 2017). Moreover, the two fields of computer science and affective neuroscience have been investigating emotional responding from very different viewpoints, more or less independently. By combining approaches from both areas, researchers can begin to investigate complex experiences, such as emotions, with much greater precision. The ability to record and model large data sets capturing moment-by-moment changes in subjective and physiological measures of emotional responding is critical for understanding diverse mechanisms that underlie both healthy and unhealthy emotional experiences and their development over the short term. The database presented here enables researchers to test specific hypotheses about changes in and interactions between subjective experiences of emotional affect and body/brain activity as they unfold over time. In addition, we include modeling of moment-by-moment emotional experiences based on neurophysiological activity using machine learning approaches, which allow data-intensive analysis without prior aggregation, for instance by averaging activity across a trial.

Several widely accessible databases have been used to investigate human emotional responding that include central and peripheral physiological measures (Duan et al., 2013; Joshi & Ghongade, 2022; Katsigiannis & Ramzan, 2018; Koelstra et al., 2012; Seal et al., 2020; Subramanian et al., 2018; Zheng & Bao-Liang, 2015). However, these focus on participants' retrospective, single data point ratings of their experiences and include small sample sizes (Sharma et al., 2019), nonconcurrent experience sampling (Ding et al., 2021), or low time-resolution brain activity derived from blood oxygenation that is not publicly available (Nummenmaa et al., 2012). Therefore, we created a new shared database that includes not only electroencephalographic and peripheral physiological activity measures but also more parsimonious subjective experience sampling from participants exposed to emotional stimulus material. Further, we used the new database to investigate simple baseline models based on central and peripheral physiological responses that predict moment-by-moment ratings of emotional arousal, defined henceforth as the state of heightened autonomic activity associated with strong feelings.

Human reactions to emotion-provoking situations have been studied in a vast variety of experimental settings, using distinct stimuli in the form of pictures, videos, sounds, odors, or words (Adolphs & Anderson, 2018; Brosch et al., 2010; Croy et al., 2011). By measuring electroencephalography (EEG), pervasive evidence has shown that during processing of such emotional stimuli, there are characteristic reactions indexed by event-related potentials (ERPs), which vary with the level of emotional arousal (Schupp, Flaisch, et al., 2006). Since the introduction of time-frequency analyses, this work has been further extended to understand the temporal development of emotion correlates by investigating the frequency bands related to those ERPs (Harper et al., 2014; Pourtois et al., 2008). For instance, a variation in the N2 (negative component at 150–250 ms after stimulus onset) in response to facial expressions has been linked to theta frequency (Eisenbarth et al., 2013), and feedback negativity has shown variation in related frequencies (theta vs. delta) depending on a positive or negative context (Watts & Bernat, 2018). But our understanding of the role of specific frequency bands is only evolving. While computer science researchers have been using EEG frequency bands for a while to develop models to categorize emotional valence and arousal (Ayata et al., 2017; Chanel et al., 2007; Gaeta et al., 2015; Koelstra et al., 2012; Nagya et al., 2014; Wang et al., 2022), the interpretation of their role in emotion processing seems to have come into focus only more recently (Schubring & Schupp, 2021). To be able to interpret results from those computational models of emotional states, we need to understand better how EEG-based oscillations used as predictors vary with affect and covary with each other across electrode sites. Such investigations require rich data from multiple individuals in a variety of emotional scenarios.

Reports of computational modeling of emotional states have mainly used a few shared datasets, with much of the investigation to date focusing on the Dataset for Emotion Analysis with Physiological Signals (DEAP; Koelstra et al., 2012). This data set is based on participants ($n=32$) watching a series of music videos (40 1-min trials), while a 32-channel EEG, electrodermal activity (EDA), respiration, blood pressure, electrocardiogram (ECG), blood volume (plethysmography), electromyogram (EMG) of musculus zygomaticus and musculus trapezius, and electrooculogram (EOG) were recorded. Each video received labels from participants for valence, arousal, and liking. The modeling approaches therefore focused on features extracted from psychophysiological signals to predict valence and arousal domains of personal experience using the overall ratings of emotional experience of participants.

Based on this type of data set, most existing investigations with a combination of EEG and peripheral physiological activity measures have used overall (summative) retrospective ratings of emotional valence and arousal for each trial as outcome variables for modeling and often used a classification approach, that is, high versus low arousal (Shon et al., 2018) or emotion categories (Du et al., 2022;

Thejaswini & Ravikumar, 2020). These attempts have used different classification algorithms such as K-nearest neighbors (KNN), convolutional neural networks (CNN), etc., along with various methods for selecting features, and have been able to predict high versus low reported arousal and valence with a high degree of accuracy (Ayata et al., 2017; Chen et al., 2019; Koelstra et al., 2012; Shon et al., 2018). However, considering the emotional space to be a categorical one might be restricting. Emotional states can vary in intensity and might not follow the emotion categories that have been underlying emotion research for the longest time (Keltner, 2019). Therefore, we need to improve linear outcome predictions to be able to investigate the noncategorical report of emotional states (Azari et al., 2020). A few studies investigated linear scales instead of high versus low rating categories, using a linear ridge regression algorithm, and found significantly lower error than a random analysis based on EEG and features extracted from the stimulus material (Lakhan et al., 2019; Soleymani et al., 2011). As pointed out by Soleymani et al. (2011), prediction of a continuous rating is more difficult, and it is less clear how to best evaluate the success of the prediction.

Most importantly, an aspect that has not yet been considered is the variability of emotional experiences *within* an ongoing dynamic stimulus such as a video. Existing data sets include overall ratings that participants give at the end of each video, and therefore the rating could be biased by subjects' experiences retrospectively (Fredrickson, 2000). Therefore, online ratings during the experience allow researchers to model the momentary and continuous changes in emotional experience (Poß et al., 2020). However, as only one experiential aspect can realistically be evaluated by a participant at a given time (Jolly et al., 2022), ratings are restricted to one dependent variable. While emotion categorization is a prevalent way to describe emotional states—that is, by valence—there is longstanding evidence that many cognitive processes vary depending on emotional *arousal* (e.g., Vogt et al., 2008). Therefore, continuous, moment-by-moment ratings of emotional arousal are a crucial dependent variable to be included in a new database for scientists to understand emotional experience and are central to our baseline models presented here.

Here, we introduce a new shared database of EEG and peripheral physiological data of participants watching emotional video clips, including their continuous arousal ratings in addition to their overall ratings. We also provide baseline models for this database with a series of machine learning-based analyses to test whether algorithms can be trained to predict the moment-by-moment changes in continuous arousal ratings from neurophysiological activity. Machine learning is used here as it enables the exploration of various types of relationships, encompassing both linear and non-linear associations, thereby providing an adaptable approach to addressing prediction problems (Ballester & Mitchell, 2010). Given the open questions about the relationship between fluctuations of subjective arousal and fluctuations in neurophysiological activity during dynamic emotion-evoking video stimuli, these models focus on the subjective moment-by-moment ratings of emotional arousal as an outcome variable. In addition, we test for potential temporal variance between the physiological activity channels by systematically testing the effect of temporal shifts of EDA within the data set on the prediction quality. Following a computational modeling approach that uses comparisons of model effectiveness, we then compared two different but well-established and widely used machine learning algorithms, linear regression (LR) and decision tree (DT), and added a feature selection approach, which identified important physiological features using a wrapper-based feature selection method employing a Continuous Particle Swarm Optimization (CPSO) algorithm (Kennedy & Eberhart, 1995). Furthermore, we compared the impact of individual differences while considering both classification and regression.

2 | METHOD

2.1 | Participants

Participants were recruited from the student body of the university and from the surrounding community. Data from 7 out of 153 recruited participants were not usable due to technical issues. One participant withdrew from the experiment during the session. The remaining 145 participants are included in the data set (mean age: 22.7 years [*SD*: 6.7]; 93 female, 48 male, 4 other; 8 left-handed, 137 right-handed). All participants provided informed consent under the approval of the university's human ethics committee and received course credit or movie tickets in exchange for taking part.

2.2 | Stimulus material and task

The 24 videos were selected from a database of 2185 emotionally evocative short videos, rated for valence and arousal on a 9-point Likert scale (https://www.alancowen.com/; Cowen & Keltner, 2017). The selected videos were approximately 15 s long (range 13–19 s, M = 15 s, *SD* = 1 s). We selected the videos to include a range of emotional valence and arousal and to include similar content across the ranges (e.g., depictions of both animals and people).

The final video set comprised three videos in each of eight categories: low arousal negative; low arousal positive; low-mid arousal negative; low-mid arousal positive; mid-high arousal negative; mid-high arousal positive; high arousal negative; and high arousal positive. Videos varied in aspect ratio, resolution, and frame rate and were scaled such that the largest dimension subtended a 13.6° visual angle. No sound accompanied the videos.

After being fitted with the EEG cap and all peripheral sensors, each participant watched the 24 videos in a unique, random order. Each video is repeated four times in a loop, with no gap between repetitions. We used repeated presentations in order to allow for future analyses dependent on length and repetitive exposure, which reduce novelty. A prolonged exposure by repetition has been found to be resistant to habituation in the visual cortex (Schupp, Stockburger, et al., 2006). As each video played, participants continuously rated the intensity of emotion they were feeling on a slider scale from 0 (*no emotion at all*) to 10 (*very intense emotion*), using a computer mouse. They were instructed to rate the intensity of how they were truly feeling in the moment. The first repetition of each video did not begin until participants returned the mouse to the 0 point of the arousal scale. Although motor activity like using a mouse produces sensorimotor activity (Brouwer et al., 2015; Miller et al., 2007), these movements are consistent through the task and across the sample, and the advantages of moment-to-moment experience sampling outweigh their presence in the data.

Immediately after the completion of each video loop of four repetitions (i.e., each trial), participants provided a single overall rating for the video on each of nine scales. They rated how positive or negative the video made them feel, with −5 being extremely negative. A 5 being extremely positive and 0 being neither positive nor negative; the intensity of the emotion they felt while watching the video, with 0 being not intense at all and 10 being extremely intense; how much they liked the video, with 0 being not at all and 10 being very much; and how much they wanted to engage with the video, with 0 being they strongly wanted to avoid it and 10 being they strongly wanted to engage with it. They also provided separate ratings on how much they felt anger, sadness, happiness, disgust, and fear while watching the video, with 0 being not at all and 10 being very strongly, for comparability with dominant emotion research categories of emotions. The order in which these items were presented was the same for all videos and participants. There was no time limit when providing each rating, although participants were prompted to "respond faster" after 10 s. Participants practiced the task on two practice videos, during which they received feedback and could ask questions. Participants self-paced through the 24 videos of the main task by initiating each new video when ready. After every four videos, a mandatory break of at least 10 s was enforced, following guidelines for the slowest measure, EDA (Fowles et al., 1981). The mean time between the offset of one video and the onset of the next was 31.7 s (SD = 18.3 s), including questionnaire responding and breaks. The entire behavioral task took around 40 min.

2.3 | Physiological recording

During the task, electrophysiological activity at the scalp was collected using an active Ag/AgCl 64-electrode net (either actiCAP or actiCAP snap), corresponding to the international extended 10–20 system. Signals were recorded through a BrainVision actiCHamp amplifier running BrainVision Recorder (Brain Products GmbH, www.brainproducts.com). The activity was recorded at 500 Hz and was referenced online to activity at electrode FCz. Electrolyte gel was used to lower electrode impedances to below 25 kΩ wherever possible. An EOG was recorded using electrodes placed above and below one eye and near the outer canthi of both eyes.

Multiple peripheral physiological measures were collected via a PowerLab 16/35 amplifier and sampled at 1000 Hz by LabChart recording software (ADInstruments, www.adinstruments.com). These included an ECG via three electrodes placed on the shoulders and torso in Lead II configuration; respiration via a respiration belt around the ribcage; EDA through electrode plates placed on the index and ring fingers of the non-dominant hand; and blood volume through an IR plethysmograph attached to the middle finger of the inferior hand. Heart rate (BPM) was calculated online in LabChart using default parameters for human ECG. Hardware filters were applied for the ECG with a low-pass filter of 200 Hz and a high-pass filter of 0.1 Hz, as well as for the plethysmograph with a low-pass filter of 100 Hz; respiration and EDA were not online filtered.

2.4 | Database description

The emotional arousal pattern (EMAP) database is provided as two datasets (i.e., Raw Data set and Clean Data set) to researchers for the purpose of understanding human emotion (data available here: https://www.wgtn.ac.nz/psyc/research/emap-open-database). In the *Raw Data set*, we have modified the data as little as possible to provide researchers with raw data as close to the "natural state" at the time of data collection as is possible. Researchers should note that the resulting data set contains some artefactual and missing data. For the *Clean Data set*, further basic preprocessing has been performed in order to provide non-specialist researchers with data that can be used for analysis with little additional processing.

Each data set includes data from 3434 trials (video viewings) from 145 participants. Forty-six trials for which EEG data or behavioral data was not available due to technical problems are not included in the data sets. A master spreadsheet contains participant demographics, video information, and overall behavioral responses for all trials.

2.5 | Dataset description

2.5.1 | Raw Data set

All data was preprocessed in MATLAB (MathWorks) using the EEGLAB toolbox (Delorme & Makeig, 2004), the PREP Pipeline toolbox (Bigdely-Shamlo et al., 2015), and in-house scripts. The preprocessing steps described below were performed separately for each participant.

First, EEG data was loaded into MATLAB using the EEGLAB toolbox. The online reference channel, FCz, was added back in, and each EEG channel was re-referenced to an average of all channels. Event and time markers representing video onsets and offsets were co-registered across peripheral, EEG, and behavioral data.

Data from peripheral psychophysiology (including ECG, heart rate, EDA, blood volume, and respiration) and from the continuous arousal rating were then added. Peripheral measures were downsampled from 1000 to 250 Hz by a factor of four. The continuous arousal response for each trial was upsampled from 85 to 250 Hz, using the most recently provided rating for each new timepoint (i.e., via zero-order hold sampling). See Figure 2 for ratings, heart rate, and EDA for four exemplar videos across low, medium-low, medium-high, and high arousal levels according to the norms for the videos.

For each participant, one spreadsheet is provided containing continuous EEG, behavioral, and peripheral data from across the experimental session. Data before and after the task, and during any long break or impedance check, have been removed.

2.5.2 | Clean data set

Preprocessing of the Clean Data set started with the Raw Data set, with further processing of EEG channels to eliminate noise and artifacts. No further preprocessing was performed on peripheral channels. From the Raw Data set, a robust average reference for EEG information was computed using the PREP Pipeline toolbox (Bigdely-Shamlo et al., 2015) with default parameters. The PREP Pipeline uses a standardized, iterative, and fully automated procedure to identify electrodes with noisy or poor signal and excludes these from the average. Readers are directed to the original paper for details of the procedure (Bigdely-Shamlo et al., 2015). EEG channel data was then high- and low-pass filtered at 0.5 and 100 Hz, respectively (Kaiser window, maximum passband deviation, 0.001/0.0001; transition bandwidth, 1 Hz/25 Hz).

Independent component analysis using the extended Infomax algorithm as implemented in *binica* (Bell & Sejnowski, 1995; Lee et al., 1999) was performed separately on each participant's EEG data. Components were then labeled using an automated classification procedure, ICLabel (Pion-Tonachini et al., 2019), which classifies components as artefactual or brain-generated. Using classification probabilities, we removed from the data any components that were classified both as brain activity with <10% probability and as artifacts with >80% probability (i.e., muscle, eye, heart, line noise, or channel noise artifacts). On average, 13 components out of 64 were removed (range: 5–23). An illustration of the raw signal can be found in Figure S6. The Clean Data set includes one spreadsheet file for every trial, containing channel and behavioral data from 1 s before video onset to the end of the video.

2.6 | Modeling moment-by-moment ratings of emotional arousal

The machine learning analysis was performed with the complete data set, which comprised 3434 video viewings from 145 participants. The preprocessing of the data followed a similar path as the Clean Data set, with a few differences. EOG and ECG were not included in the analysis; EOG and ECG channels were excluded as they have many missing data. Moreover, all videos for which any EEG, peripheral, or behavioral data were missing were excluded.

To generate a set of 264 features (feature extraction) usable for machine learning, data from each EEG and peripheral channel was split into 500 ms time bins (comprising 125 data points per channel per bin) starting at the onset of each trial and ending with the last complete bin during the viewing period. Power spectra were calculated within each bin for each electrode using Welch's method (Welch, 1967) with a window size of 125. Absolute band power was calculated as area under the curve for the theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–60 Hz) frequency bands using Simpson's rule (Simpson's Rule, n.d.). Additionally, for heartrate, EDA, IR plethysmograph, and respiration data, the mean and standard deviation were calculated for each time bin. This resulted in eight peripheral features and 256 spectral EEG features for each bin. Each bin was treated as an instance for model training and used the mean self-reported moment-by-moment arousal value calculated within each bin as the dependent variable. These ready-to-use extracted features can also be downloaded from the repository (https://www.wgtn.ac.nz/psyc/research/emap-open-database) and can be used by researchers that do not wish to extract the features themselves.

2.6.1 | Arousal ratings

The primary modeling was done using the dimensional moment-by-moment ratings (rating scores ranging between 0 and 1). For the comparison of these dimensional versus categorical arousal (high vs. low) ratings, the continuous ratings were converted to high and low ratings using a threshold value of 0.5 (i.e., the middle of the scale). The EMAP database also provides overall intensity ratings for each trial, which researchers can use for classification models. However, for the analyses presented here, we chose to do the classification with dichotomous scores transformed from the continuous ratings, which provides a new perspective in comparison to previous studies that used retrospective overall categorical ratings.

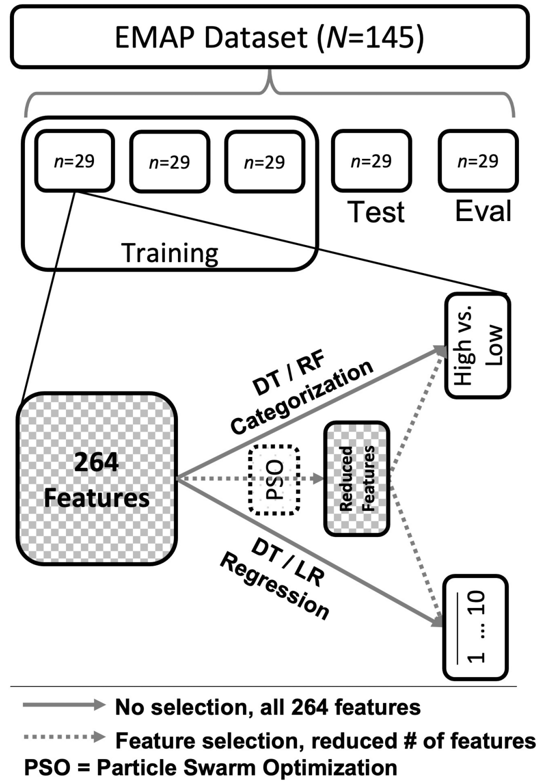

FIGURE 1 Workflow for the different modeling approaches.

2.6.2 | General analysis approach

There are different ways to model emotional arousal with machine learning algorithms. Certain researchers use classification algorithms to predict high/low arousal ratings (Becker et al., 2020; Chen et al., 2020; Ganapathy et al., 2020; Wang et al., 2020). However, as our aim was to predict the moment-by-moment-changing arousal rating from the momentary features, we used regression analysis for a more thorough prediction of the concurrent rating. Therefore, the outcome is the error rate of the predicted ratings from the actual ratings given by the participants. In addition, we also provide classification results for comparison.

While it is possible to obtain a reliable result using the complete data set, it is time-consuming and might lead to the production of a complex model that will be difficult to interpret as the EEG data has a large number of channels. One possible solution is to apply a feature selection technique to analyze the minimum number of features that can be used together to yield an optimal result, that is, reducing the training time and complexity as well as increasing the performance and interpretability of the model. This section first details how a usable set of momentary features and labels were extracted from the cleaned data for use in supervised learning. Next, the feature selection techniques used to find an optimal feature subset are described before the results of feature selection are presented. The workflow for the different modeling approaches is presented in Figure 1.

2.6.3 | Feature selection

Feature selection is an important preceding algorithm-based operation in many regression problems. It helps with removing irrelevant and redundant extracted features from a data set as well as increasing the performance by reducing the search space. Algorithms such as sequential forward selection and sequential backward elimination have been used to address the FS problem. However, these algorithms suffer from becoming stuck in local optima (Xue et al., 2013). Conversely, evolutionary computation (EC) techniques such as Particle Swarm (PSO), genetic algorithms (GA), and genetic programming (GP) are known for their ability to search for key features across the data set and have been used to solve feature selection problems (Xue et al., 2016). We used PSO for the analyses, as PSO is easy to implement, has fewer parameters, and is computationally less expensive compared with other EC algorithms. We present the details of this approach in the Supplementary Materials (1.2 Particle Swarm Optimization).

2.6.4 | Modeling

Classification and regression analysis were conducted on the extracted features of the EMAP Clean Data set. Primarily, the regression analysis was performed with arousal as the dependent variable, dependent on the remaining 264 features in the data set. Second, the classification was performed to classify high/low arousal ratings based on the given 264 features (see Section 2.6.1. for a description of how these ratings were obtained).

The data were normalized by participant such that between participant variation was conserved (see Chen et al., 2021). Moreover, the classification and regression analyses were performed both across participants (by group) and within each participant (by participant).

For the across-participants analysis, we used a feature selection using fivefold cross-validation to balance subset size with the number of participants. The data set was therefore divided into five subsets, such that each subset contained data from 29 participants (for a total of 145 participants). For each run, three subsets (i.e., data from 29×3 participants) were used for training and one subset (i.e., data from 29 participants) for testing. To avoid feature selection bias, the final subset was used for evaluating the performance of the regression method using only the selected features. The procedure was repeated five times, such that each of the five unique subsets was used for testing the performance of the model. Similarly, the classification analysis also used fivefold cross-validation to evaluate the performance of the classification method for each subset. This cross-validation ensures that the model, and thereby its internal structure, is generalizable beyond its own training data. As an additional step to reduce bias through the setup of the modeled data, in both classification and regression for the across-participants analysis, all participants were randomly selected (without replacement) and equally distributed into the groups.

For the within-participant analysis, the analysis was performed for each participant using a threefold cross-validation to allow sufficient data in each fold across the 24 trials. Thus, at any given time, eight out of the 24 video trials watched by the participant were used for testing, whereas the remainder of the 16 videos were used for training. The procedure was repeated three times until all folds were tested. For the sake of comparison, we present both the classification and regression results of the by-participant analysis.

The PSO algorithm was run with the following parameters: $c_1 = c_2 = 0.5$, $w = 0.9$, *number of particles* = 30, dimension D = number of features, maximum iteration $T = 100$, $\alpha = .88$, and threshold value $\theta \in [0, 1] = 0.5$. These parameters were chosen based on common settings in the literature (Miranda, 2018; Shi & Eberhart, 1998). The algorithm was run with two different regression learning methods: LR and DT with a *depth* of 5, within the wrapper method. LR and DT were chosen because they are relatively fast. We named these two PSO-based feature selection methods CPSO-LR and CPSO-DT, respectively.

For the classification analysis, baseline results are provided using DT and random forests (RF). RF was chosen as it is an ensemble learning algorithm that uses many DTs and therefore represents the next higher level of model. Comparing this higher-level modeling approach should lead to less error, allowing us to test for the validity of the models. A number of different trees, such as 2, 5, 10, 15, and 20, were tested, and as using two trees resulted in the best accuracy, this setting was used.

To test the potential effects of a delay in the EDA response, which is characteristic of EDA responses (Horn et al., 2020), we further added an analysis with different temporal offsets for the EDA data of 0, 5, and 10 time bins, thus" shifting" the EDA data by 0, 2500, and 5000 ms.

The data and code are available at https://www.wgtn.ac.nz/psyc/research/emap-open-database.

3 | RESULTS

3.1 | Behavioral characteristics

Video stimuli were pre-selected to cover a range of emotions on the affective dimensions of arousal and valence. To describe the subjective response of participants to each stimulus, we present emotional ratings provided during and after viewing each video. As can be seen in Figure 2, participants varied greatly in their subjective response to the videos, providing a rich corpus of information on naturalistic responding to affective stimuli.

Participants' moment-by-moment arousal ratings generally increased quickly over the course of the first repetition of each video as they recognized the emotional content of what they watched (Figure 2). Even videos selected a priori to elicit low arousal evoked a range of ratings, which could be a result of tonic arousal driven by the experimental context. Overall arousal ratings, provided after each video, could be assumed to reflect a summation of ratings reported during the trial. Indeed, a Spearman's rank correlation on all trials (independent of participant) between the overall arousal rating and the median of the momentary arousal ratings given while viewing showed a monotonically increasing relationship between arousal reports (Spearman's

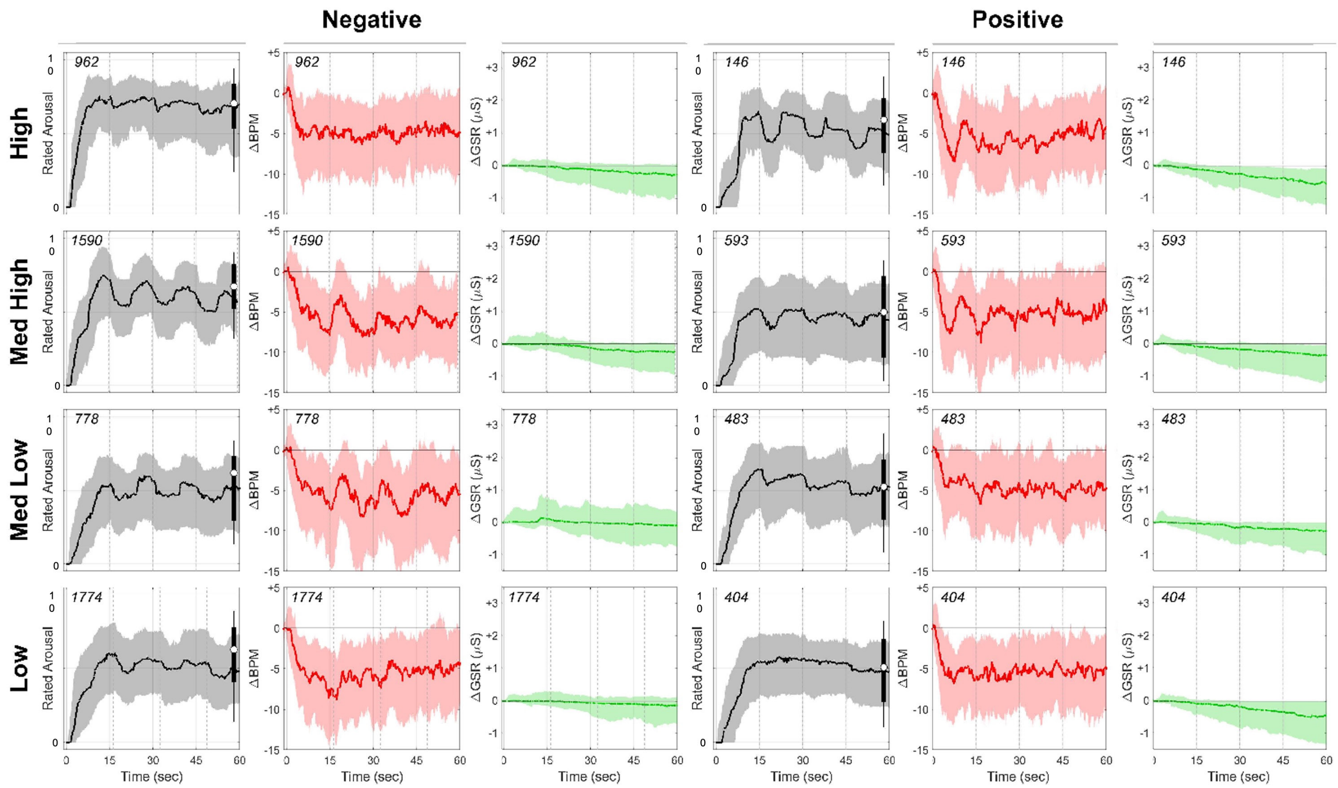

FIGURE 2 Median moment-by-moment arousal ratings (gray), heart rate (red), and electrodermal activity (EDA, green) with ribbon area for interquartile range, in response to four example videos, one for low, medium-low, medium-high, and high arousal levels each, according to their norms. Refer to Figures S3–S5 for a full moment-by-moment arousal rating, EDA, and heart rate response of each video.

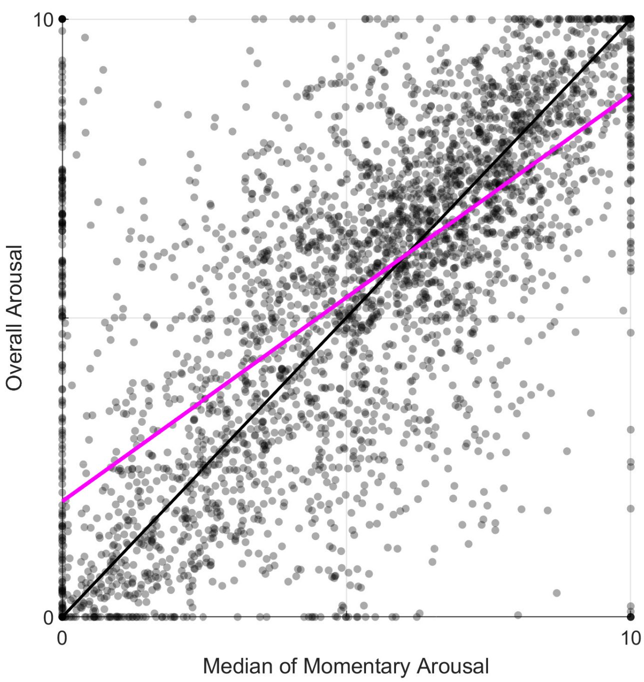

FIGURE 3 Comparing median momentary and overall arousal. For each video viewing, the median momentary arousal across the trial is plotted against the overall rating provided after the trial (gray dots). Post-trial ratings of overall arousal correlated with during-trial arousal, though appeared to be less extreme (least squares fit shown in magenta).

$\rho(3432) = .69$, $p < .001$; Figure 3). Nevertheless, the correlation is not perfect, and it is clear from Figure 3 and Figure S3 that ratings given in the heat of the moment provide a richer source of information than either ratings given retrospectively or a priori video classifications. For example, the clip of a baby playing with puppies (#2131) showed one of the greatest increases in arousal in the first 5 s (see Figure S3).

The distributions of overall ratings of arousal, valence, and seven other dimensions—provided after each trial—further illustrate the variation in individualized responses to emotional content (Figure 4). For example, a clip of a whale shark being freed from a rope (#1612) elicited mostly positive feelings but also anger, disgust, and sadness in some viewers. A dizzying video of mountain climbing (#436) is also of interest for featuring high across-participant variation in both fear and arousal responses.

3.2 | Modeling results

Here we describe results from models predicting moment-by-moment ratings from physiological channels: first, the effects of temporal shifting of EDA on the prediction error; second, a comparison of different

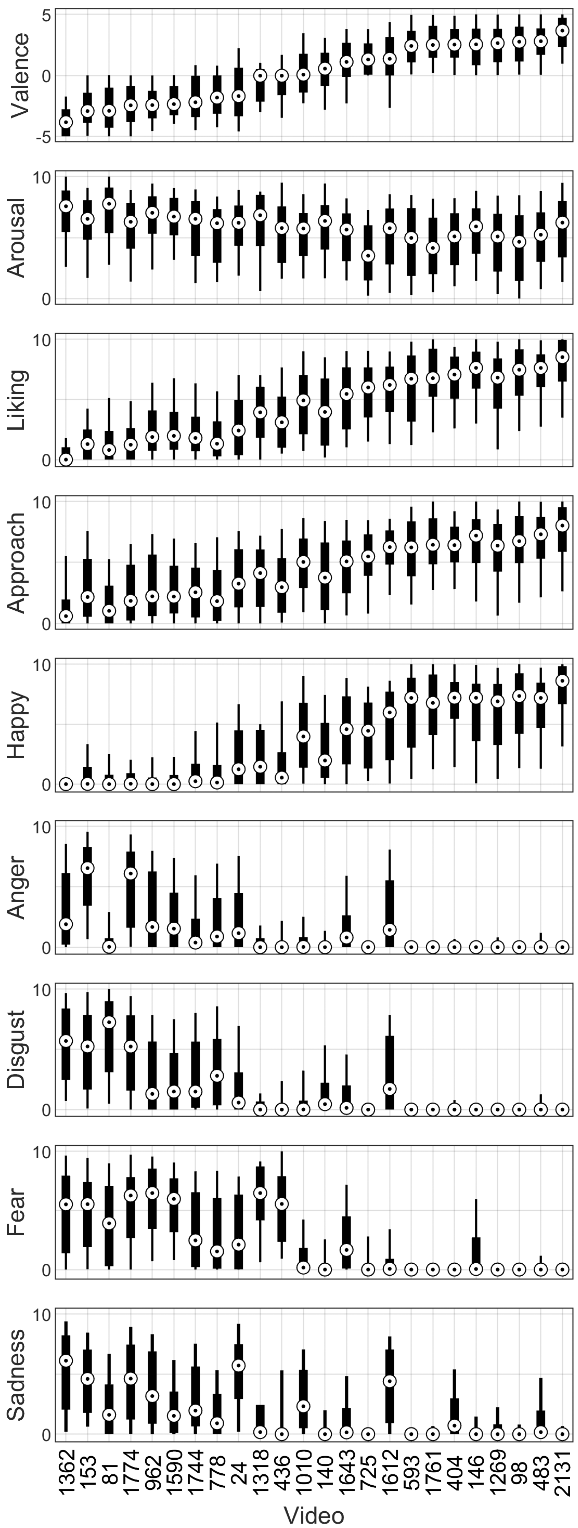

FIGURE 4 Overall rating for each video on several dimensions. Participants rated each video on nine emotional dimensions immediately after viewing. The median (circles), interquartile range (boxes), and interdecile range (whiskers) are shown for each video (video ID numbers at bottom). Videos are ordered by median valence rating.

machine learning algorithms and feature selection approaches; and finally, results from testing the impact of individual differences.

3.2.1 | The impact of temporal shifting

Table 1 presents the results obtained from analyses using four different methods (i.e., LR, CPSO-LR, DT, and CPSO-DT) before and after shifting the EDA by 0, 2500, and 5000 ms. A Wilcoxon Signed-Rank test with a Bonferroni-corrected alpha level of 0.0167 showed no significant differences (all $p > .06$) for all shift versions (i.e., 0 vs. 2500, 0 vs. 5000, 2500 vs. 5000) using the four different algorithms. These results suggest that shifting the EDA within the EMAP data set did not affect the predictive accuracy of the emotional arousal ratings. Therefore, the subsequent analyses were performed without temporal shifting of the EDA data.

3.2.2 | Impact of machine learning and feature selection algorithms

Comparing the results from the DT and LR algorithms, a Wilcoxon Signed-Rank test showed no significant difference between the results ($p = .062$). Although the mean normalized root mean squared error (M-NRMSE) obtained after feature selection is lower than the M-NRMSE obtained before feature selection for both DT and LR algorithms within the wrapper method, this difference was not significant ($p > .06$) for either DT versus CPSO-DT or LR versus CPSO-LR-based, respectively. Thus, the PSO-based feature selection methods can maintain the regression error while substantially reducing the number of features. Good and bad examples of predictions of online arousal ratings via DT are illustrated in Figures S1 and S2, respectively.

For the sake of comparison with existing work, we also report classification results by using the converted continuous ratings in their categorical form of high/low arousal using a threshold value of 0.5, the visual center of the scale (see Methods section on 2.6.1. Section). The related classification analysis using DT resulted in an accuracy slightly decreasing from 50.12% ± 0.77 before feature selection to 49.37% ± 0.42 after feature selection, whereas using RF, the accuracy improved from 49.13% ± 2.19 before feature selection to 50.89% ± 2.88 after feature selection (Table 2). However, Wilcoxon signed-rank tests showed that none of these differences were statistically significant (all $p > .31$), while the number of features has been reduced. Thus, reducing the features in the analysis leads to a comparable result for the classification-based analyses as well.

3.2.3 | Impact of individual differences

In comparison to what was reported in the section before, for the LR and DT-based regression analyses across

TABLE 1 Prediction errors obtained by different algorithms before and after application of feature selection.

Method	LR			CPSO- LR			DT			CPSO-DT		
EDA shift (ms)	**0**	**2500**	**5000**	**0**	**2500**	**5000**	**0**	**2500**	**5000**	**0**	**2500**	**5000**
M-NRMSE	3.8306	3.9155	4.9509	1.1580	0.9814	0.8200	0.3170	0.3174	0.3207	0.3141	0.3147	0.3118
Best-NRMSE	0.3618	0.3762	0.4629	0.3420	0.3195	0.3049	0.3102	0.3010	0.2962	0.2987	0.3012	0.2951
SD-NRMSE	2.7364	4.8298	5.9827	0.7521	0.6764	0.4767	0.0065	0.0111	0.0183	0.0109	0.0094	0.0124
Ave-#FS	264	264	264	127	119	126	264	264	264	157	155	152

Note: Row Method = analysis method; Shift = EDA time shift (ms) for each analysis; M-NRMSE = mean of the normalized root mean squared error (*M-NRMSE*) across five splits; Best-NRMSE = best NRMSEs; *SD*-NRMSE = standard deviation of the NRMSEs; Ave-#FS = average number of selected features.

TABLE 2 Regression and classification results for the within- and across- participants analyses before and after feature selection to predict arousal ratings.

Analysis type		Regression				Classification				
	Method	**M-NRMSE**	**Best-NRMSE**	***SD*-NRMSE**	**Ave-#FS**	**Method**	**M-Accuracy(%)**	**Best-Accuracy(%)**	***SD*-Accuracy(%)**	**Ave-#FS**
Across participants	DT	0.3170	0.3102	0.0065	264	DT	50.12	51.47	0.87	264
	CPSO-DT	0.3141	0.2987	0.0109	157	CPSO-DT	49.37	49.86	0.48	136
	LR	3.8306	0.3618	2.7364	264	RF	49.13	53.12	2.50	264
	CPSO-LR	1.1580	0.3420	0.7521	127	CPSO-RF	50.89	54.46	3.28	147
Within participants	DT	0.4116	2.55E-15	0.3198	264	DT	54.16	91.06	8.31	264
	LR	0.5642	0.1621	0.6207	264	RF	61.46	95.06	7.68	264

Note: Method = analysis method; M-NRMSE and M-Accuracy = mean of the normalized root mean squared error (M-NRMSE) and accuracy across five splits (across participants) and across 145 splits (within participants); Best-NRMSE and Best-Accuracy = best NRMSEs and accuracies; SD-NRMSE and SD-Accuracy = standard deviation of the NRMSEs and accuracies; Ave-#FS = average number of selected features.

participants (3.8306 ± 2.74 and 0.3170 ± 0.01), we found a significantly higher error rate for the DT-based regression analysis within participants (0.4116 ± 0.54) and a significantly lower error rate for the LR-based analysis across participants (0.5642 ± 0.62; Mann–Whitney U tests, all $p < .02$ at $\alpha = .05$).

A permutation test comparing the results from LR-based and DT-based regression methods to chance level by randomizing the labels for up to 1000 independent runs showed that all DT-based predictions were significantly better than chance level-based results ($p < .001$), but the LR-based predictions were not significantly better than at chance level ($p = 1.0$) for the analysis across participants. There was also no significant difference from chance for both LR- and DT-based predictions for the analyses within participants (all $ps > .99$).

We also compared the results from DT- and RF-based classification analyses across participants with DT- and RF-based classification analyses within participants. While the RF-based analysis within participants resulted in a significantly higher accuracy in comparison to the RF-based analysis across participants, there were no significant differences between the DT-based analyses across and within participants.

We re-evaluated the performance of the classification approaches in comparison to data with randomized labels. We found that both the RF-based (49.13%) and the DT-based approach (50.12%) achieved a lower accuracy for correctly labeled versus randomly labeled data (54.22%) for the analysis across participants. However, these differences were not significant. Further, while the RF (61.46%) approach with correctly labeled data resulted in significantly higher accuracy than with randomized labels (54.86%, $p < .001$), randomizing the labels resulted in a non-significantly higher accuracy for the DT-based approach (56.15%) in the analysis within participants (all $p > .06$).

We also evaluated the performance of the classification approaches in comparison to a majority voting-based method (i.e., using a zero-rule baseline classifier). We found that the majority voting-based results (63.89% and 65.75%) were higher than the results based on DT (50.12% and 54.16%) and RF (49.13% and 61.46%) methods for both analyses, across and within participants, respectively. A Wilcoxon signed-rank test showed that the majority voting-based results were significantly ($p < .001$) better than the results achieved by both DT and RF-based methods for the analyses within but not across participants (all $p > .062$).

Assuming that this result for the majority voting-based analysis could be due to the unbalanced nature of the EMAP rating data, we re-performed the analysis using a balanced class ratio of 0.5:0.5. We found that the DT-based method achieved an accuracy of 49.15% and the RF-based method achieved an accuracy of 49.96% for the analysis across participants. However, both DT and RF-based methods (51.30% and 52.83%) were significantly better than the majority voting-based method results (50.0%) for the analysis within participants ($p < .001$ $\alpha = .05$). We also tested the results of using the sole modalities of EEG and peripheral measures (see Materials S1 and related Table S1).

Overall, these findings suggest that predicting moment-by-moment emotional arousal in the EMAP dataset is notably more difficult than classifying the overall labels. However, accuracy was found to still be higher than random guessing for the majority of prediction analysis approaches.

4 | DISCUSSION

EMAP is a new and open database of emotional responses, including EEG, peripheral, and self-reported data, all in a moment-by-moment format to enable investigations of human affective reactivity in multiple emotional contexts. The database primarily facilitates the use of EEG and peripheral physiological signals to continuously model momentary ratings of emotional arousal. However, the variation in participants' overall ratings of valence, liking, interest, and emotion categories additionally permits researchers to investigate these other constructs, albeit in a less granular fashion. The descriptive results show the large variation across observers in subjective moment-by-moment ratings of arousal. While this type of rating might have some limitations in terms of temporal accuracy and memory effects (Fredrickson, 2000), other recent work has shown that a continuous rating activity does not seem to change the overall rating—at least for aesthetic stimuli (Isik & Vessel, 2019)—and can capture meaningful but fleeting emotional moments (Poß et al., 2020). Our baseline models showed that successfully modeling momentary arousal is more difficult than predicting overall and retrospective arousal ratings. The much lower accuracy of high-low arousal classification and the error rate from the regression-based analyses show that moment-by-moment changes of neurophysiological activity and potentially related subjective reports will require the development of further analytical approaches. Providing this database will allow more detailed research into the temporal dynamics of self-reported arousal, EEG, and peripheral physiological correlates and their interactions.

Several caveats have to be highlighted regarding this database. The video clips that were used to induce emotional reactions of various intensities were rather short and therefore looped, that is, repeated four times. This helped sustain the subjective affective reaction, which is reflected in the raw moment-by-moment ratings that plateau even with some variability across each trial/video (see Figure 2 and Figure S3). But the repetition also led to iteration in the visual input, which needs to be considered when using the database. Researchers who utilize the EMAP database should also note that while some responses to emotional stimuli may be immediate (perhaps reflecting specific events within a video), other signals may show a gradual drift in baseline activity (for instance, EDA). While such dynamics will certainly reflect participants' physiological activity, they may also reflect measurement artifacts (e.g., temperature of the electrode paste and its hydration of the skin increasing over time). Importantly, the database should contain no systematic variation in these artifacts across participants, given the randomized presentation of video stimuli. Still, the dynamics of gradual signal changes will not be confined to individual trials; even if the mean time interval between the trials was long enough (31.7 s) for physiological reactions to decrease, there might still be some inter-trial carryover, even if likely rather minimal. Crucially, however, continuous arousal ratings were forced to start at zero, which means that for each video, the affective intensity was baselined at the start of each trial, likely obscuring such carry-over effects from previous trials. It is most appropriate, then, that each trial is considered independently and as such, we advise researchers to baseline EEG and peripheral physiological data within trials relative to the pre-trial fixation period in a fashion similar to that described above. One final concern is the potential confounding impact of motor cortex activity resulting from hand movements involving continuously rating emotional arousal. Although these movements were constant and related to both increases and decreases in ratings, the amount of movement might be correlated with the arousal ratings. Future work in this area might include some method of analysis capable of accounting for this factor (Iriarte et al., 2003) or some novel measure of continuous subjective arousal beyond the rating system applied here. For instance, as per the suggestion of an anonymous reviewer, using the mouse control itself (e.g., acceleration or nonmovement) as an indicator of emotional engagement might be an interesting future direction.

The results of modeling the physiological data onto the moment-by-moment ratings of arousal show that the application of feature selection with Particle Swarm Optimization improved the arousal predictions (reduced NRMSE), but only to a significant extent for the LR-based analyses. They also showed that the combination of the DT algorithm and feature selection led to the lowest error in prediction of the linear outcome variable, arousal. However, these regression-based prediction results are harder to interpret in terms of their accuracy compared to classification tasks (Sutton, 2005). In addition, moment-by-moment varying labels seem to be much harder to predict due to less available data per momentary rating, similar to what was found for EDA/HR prediction in analyses in a related study (Shehu et al., 2023).

Importantly, we were able to show that including a temporal shift of the EDA data, considering the potential delays of that response in comparison to the other included measures, did not change the predictive results. This indicates that the slight temporal delays might not have a strong impact on the predictive relevance of single features. However, future work could investigate temporal shifting in more detail to investigate the impact of aligned timing across features.

When using a traditional classification task approach based on a median split for each moment-by-moment arousal rating data point, we found a moderate accuracy of 50%–56% across the various methods. Interestingly, in contrast to the regression approach, an additional feature selection step did not significantly change the accuracy of the two methods used for the classification. Also, here, using DT resulted in lower accuracy than RF. Thus, in the context of these categorical data, there is no clear advantage or disadvantage to the use of a pre-selection of features, indicating that when predicting less parsimonious ratings, a pre-selection of features would not be needed.

Finally, our comparison of training and testing across participants and within each participant revealed that including data from several participants resulted in a lower prediction error than when basing the analyses on single individuals, but only for the regression-based analysis. For the classification task, the slightly higher accuracy with the within-participants analyses was not significantly different from the training across participants. This finding was astonishing, given previous results that led to lower prediction errors when training and testing within participants when predicting EDA and HR from functional brain activity data, which was attributed to large individual differences in arousal regulation (Eisenbarth et al., 2016). Our results point more toward the advantage of pooling data from different participants and therefore to common associations between subjective evaluations of one's arousal state and physiological activity. Future work could disentangle the relevance of individual versus group-based analyses, varying across dependent measures as well as regression and classification tasks. Beyond these initial analyses, this database provides the opportunity to investigate moment-by-moment changes in emotional experience in conjunction with neurophysiological and peripheral physiological activity, which will contribute to the understanding of human emotional reactivity.

AUTHOR CONTRIBUTIONS

Hedwig Eisenbarth: Conceptualization; data curation; funding acquisition; methodology; project administration; resources; software; supervision; writing – original draft; writing – review and editing. **Matt Oxner:** Data curation; methodology; project administration; visualization; writing – review and editing. **Harisu Abdullahi Shehu:** Formal analysis; validation; writing – review and editing. **Tim Gastrell:** Formal analysis; validation; writing – review and editing. **Amy Walsh:** Data curation; methodology; project administration; writing – review and editing. **Will N. Browne:** Supervision; writing – review and editing. **Bing Xue:** Methodology; supervision; writing – review and editing.

FUNDING INFORMATION

This work was funded by Victoria University of Wellington.

DATA AVAILABILITY STATEMENT

Data are available via https://www.wgtn.ac.nz/psyc/research/emap-open-database.

ORCID

Hedwig Eisenbarth https://orcid.org/0000-0002-0521-2630
Harisu Abdullahi Shehu https://orcid.org/0000-0002-9689-3290

REFERENCES

Adolphs, R., & Anderson, D. J. (2018). *The neuroscience of emotion: A new synthesis*. Princeton University Press.

Ayata, D., Yaslan, Y., & Kamasak, M. (2017). Multi channel brain EEG signals based emotional arousal classification with unsupervised feature learning using autoencoders. 25th Signal Processing and Communications Applications Conference (SIU), 1–4 https://doi.org/10.1109/SIU.2017.7960629

Azari, B., Westlin, C., Satpute, A., Hutchinson, J. B., Kragel, P. A., Hoemann, K., Khan, Z., Wormwood, J., Quigley, K., Erdogmus, D., Dy, J., Brooks, D., & Barrett, L. F. (2020). Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body, and subjective experience. https://doi.org/10.31234/osf.io/egh2t

Ballester, P. J., & Mitchell, J. B. O. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, *26*(9), 1169–1175. https://doi.org/10.1093/bioinformatics/btq112

Barrett, L. F., & Satpute, A. B. (2017). Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience Letters*, *693*, 9–18. https://doi.org/10.1016/j.neulet.2017.07.045

Becker, H., Fleureau, J., Guillotel, P., Wendling, F., Merlet, I., & Albera, L. (2020). Emotion recognition based on high-resolution EEG recordings and reconstructed brain sources. *IEEE Transactions on Affective Computing*, *11*(2), 244–257. https://doi.org/10.1109/TAFFC.2017.2768030

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*, 16. https://doi.org/10.3389/fninf.2015.00016

Brosch, T., Pourtois, G., & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition & Emotion*, *24*(3), 377–400. https://doi.org/10.1080/02699930902975754

Brouwer, A.-M., Zander, T. O., Van Erp, J. B. F., Korteling, J. E., & Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: Six recommendations to avoid common pitfalls. *Frontiers in Neuroscience*, *9*, 136. https://doi.org/10.3389/fnins.2015.00136

Chanel, G., Ansari-Asl, K., & Pun, T. (2007). Valence-arousal evaluation using physiological signals in an emotion recall paradigm. *Conference Proceedings—IEEE International Conference on Systems, Man and Cybernetics*, *41*(22), 2662–2667. https://doi.org/10.1109/ICSMC.2007.4413638

Chen, H., Sun, S., Li, J., Yu, R., Li, N., Li, X., & Hu, B. (2021). Personal-Zscore: Eliminating individual difference for EEG-based cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, *1–1*. https://doi.org/10.1109/TAFFC.2021.3137857

Chen, J. X., Zhang, P. W., Mao, Z. J., Huang, Y. F., Jiang, D. M., & Zhang, Y. N. (2019). Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks. *IEEE Access*, *7*(c), 44317–44328. https://doi.org/10.1109/ACCESS.2019.2908285

Chen, T., Ju, S., Ren, F., Fan, M., Gu, Y., Ju, S., Ren, F., Fan, M., & Gu, Y. (2020). EEG emotion recognition model based on the LIBSVM classifier. https://doi.org/10.1016/j.measurement.2020.108047

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(38), E7909. https://doi.org/10.1073/pnas.1702247114

Croy, I., Olgun, S., & Joraschky, P. (2011). Basic emotions elicited by odors and pictures. *Emotion*, *11*(6), 1331–1335. https://doi.org/10.1037/a0024437

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Ding, Y., Hu, X., Xia, Z., Liu, Y.-J., & Zhang, D. (2021). Inter-brain EEG feature extraction and analysis for continuous implicit emotion tagging during video watching. *IEEE Transactions on Affective Computing*, *12*(1), 92–102. https://doi.org/10.1109/TAFFC.2018.2849758

Du, G., Su, J., Zhang, L., Su, K., Wang, X., Teng, S., & Liu, P. X. (2022). A multi-dimensional graph convolution network for EEG emotion recognition. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–11. https://doi.org/10.1109/TIM.2022.3204314

Duan, R.-N., Zhu, J.-Y., & Lu, B.-L. (2013). Differential entropy feature for EEG-based emotion classification. 6th international IEEE/EMBS conference on neural engineering (NER), 81–84 https://doi.org/10.1109/NER.2013.6695876

Eisenbarth, H., Angrilli, A., Calogero, A., Harper, J., Olson, L. A., & Bernat, E. (2013). Reduced negative affect response in female psychopaths. *Biological Psychology*, *94*(2), 310–318. https://doi.org/10.1016/j.biopsycho.2013.07.007

Eisenbarth, H., Chang, L. J., & Wager, T. D. (2016). Multivariate brain prediction of heart rate and skin conductance responses to social threat. *The Journal of Neuroscience*, *36*(47), 11987–11998. https://doi.org/10.1523/JNEUROSCI.3672-15.2016

Fowles, D. C., Christie, M. J., Edelberg, R., Grings, W. W., Lykken, D. T., & Venables, P. H. (1981). Committee report. Publication recommendations for electrodermal measurements. *Psychophysiology*, *18*(3), 232–239. https://doi.org/10.1111/j.1469-8986.1981.tb03024.x

Fredrickson, B. L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion*, *14*(4), 577–606. https://doi.org/10.1080/026999300402808

Gaeta, G., Susac, A., Supek, S., Babiloni, F., & Vecchiato, G. (2015). Analysis of EEG variables to measure the affective dimensions of arousal and valence related to the vision of emotional pictures. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2015-November 2518–2521 https://doi.org/10.1109/EMBC.2015.7318904

Ganapathy, N., Veeranki, Y. R., & Swaminathan, R. (2020). Convolutional neural network based emotion classification using electrodermal activity signals and time-frequency features. *Expert Systems with Applications*, *159*, 113571. https://doi.org/10.1016/j.eswa.2020.113571

Harper, J., Malone, S. M., & Bernat, E. M. (2014). Theta and delta band activity explain N2 and P3 ERP component activity in a go/no-go task. *Clinical Neurophysiology*, *125*(1), 124–132. https://doi.org/10.1016/j.clinph.2013.06.025

Horn, M., Fovet, T., Vaiva, G., Thomas, P., Amad, A., & D'Hondt, F. (2020). Emotional response in depersonalization: A systematic review of electrodermal activity studies. *Journal of Affective Disorders*, *276*, 877–882. https://doi.org/10.1016/j.jad.2020.07.064

Iriarte, J., Urrestarazu, E., Valencia, M., Alegre, M., Malanda, A., Viteri, C., & Artieda, J. (2003). Independent component analysis as a tool to eliminate artifacts in EEG: A quantitative study. *Journal of Clinical Neurophysiology*, *20*(4), 249–257.

Isik, A. I., & Vessel, E. A. (2019). Continuous ratings of movie watching reveal idiosyncratic dynamics of aesthetic enjoyment. *PLoS One*, *14*(10), 1–22. https://doi.org/10.1371/journal.pone.0223896

Jolly, E., Farrens, M., Greenstein, N., Eisenbarth, H., Reddan, M. C., Andrews, E., Wager, T. D., & Chang, L. J. (2022). Recovering individual emotional states from sparse ratings using collaborative filtering. *Affective Science*, *3*, 799–817. https://doi.org/10.1007/s42761-022-00161-2

Joshi, V. M., & Ghongade, R. B. (2022). IDEA: Intellect database for emotion analysis using EEG signal. *Journal of King Saud University – Computer and Information Sciences*, *34*(7), 4433–4447. https://doi.org/10.1016/j.jksuci.2020.10.007

Katsigiannis, S., & Ramzan, N. (2018). DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, *22*(1), 98–107. https://doi.org/10.1109/JBHI.2017.2688239

Keltner, D. (2019). Toward a consensual taxonomy of emotions. *Cognition and Emotion*, *33*(1), 14–19. https://doi.org/10.1080/02699931.2019.1574397

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Proceedings of ICNN'95—International Conference on Neural Networks, 4, 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Koelstra, S., Mühl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I., Mühl, C., Soleymani, M., … Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, *3*(1), 18–31. https://doi.org/10.1109/T-Affc.2011.15

Lakhan, P., Banluesombatkul, N., Changniam, V., Dhithijaiyratn, R., Leelaarporn, P., Boonchieng, E., Hompoonsup, S., & Wilaiprasitporn, T. (2019). Consumer grade brain sensing for emotion recognition. *IEEE Sensors Journal*, *19*(21), 9896–9907. https://doi.org/10.1109/JSEN.2019.2928781

Lee, T.-W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, *11*(2), 417–441. https://doi.org/10.1162/089976699300016719

Miller, K. J., Leuthardt, E. C., Schalk, G., Rao, R. P. N., Anderson, N. R., Moran, D. W., Miller, J. W., & Ojemann, J. G. (2007). Spectral changes in cortical surface potentials during motor movement. *The Journal of Neuroscience*, *27*(9), 2424–2432. https://doi.org/10.1523/JNEUROSCI.3886-06.2007

Miranda, L. J. V. (2018). PySwarms: A research toolkit for particle swarm optimization in python. *The Journal of Open Source Software*, *3*(21), 433. https://doi.org/10.21105/joss.00433

Nagya, T., Telleza, D., Divák, Á., Lógób, E., Kólesb, M., & Hámornikb, B. (2014). Predicting arousal with machine learning of EEG signals. 5th IEEE international conference on cognitive Infocommunications, CogInfoCom 2014—Proceedings, 137–40 https://doi.org/10.1109/CogInfoCom.2014.7020434

Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I. P., Hari, R., & Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(24), 9599–9604. https://doi.org/10.1073/pnas.1206095109

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Poß, D., Liebl, D., Kneip, A., Eisenbarth, H., Wager, T. D., & Barrett, L. F. (2020). Superconsistent estimation of points of impact in non-parametric regression with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*, 1115–1140. https://doi.org/10.1111/rssb.12386

Pourtois, G., Delplanque, S., Michel, C., & Vuilleumier, P. (2008). Beyond conventional event-related brain potential (ERP): Exploring the time-course of visual emotion processing using topographic and principal component analyses. *Brain Topography*, *20*(4), 265–277. https://doi.org/10.1007/s10548-008-0053-6

Schubring, D., & Schupp, H. T. (2021). Emotion and brain oscillations: High arousal is associated with decreases in alpha- and lower beta-band power. *Cerebral Cortex*, *31*(3), 1597–1608. https://doi.org/10.1093/cercor/bhaa312

Schupp, H. T., Flaisch, T., Stockburger, J., & Junghofer, M. (2006). Emotion and attention: Event-related brain potential studies.

Progress in Brain Research, *156*, 31–51. https://doi.org/10.1016/S0079-6123(06)56002-9

Schupp, H. T., Stockburger, J., Codispoti, M., Junghöfer, M., Weike, A. I., & Hamm, A. O. (2006). Stimulus novelty and emotion perception: The near absence of habituation in the visual cortex. *Neuroreport*, *17*(4), 365–369. https://doi.org/10.1097/01.wnr.0000203355.88061.c6

Seal, A., Reddy, P. P. N., Chaithanya, P., Meghana, A., Jahnavi, K., Krejcar, O., & Hudak, R. (2020). An EEG database and its initial benchmark emotion classification performance. *Computational and Mathematical Methods in Medicine*, *2020*, 8303465. https://doi.org/10.1155/2020/8303465

Sharma, K., Castellini, C., van den Broek, E. L., Albu-Schaeffer, A., & Schwenker, F. (2019). A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific Data*, *6*(1), 196. https://doi.org/10.1038/s41597-019-0209-0

Shehu, H. A., Oxner, M., Browne, W. N., & Eisenbarth, H. (2023). Prediction of moment-by-moment heart rate and skin conductance changes in the context of varying emotional arousal. *Psychophysiology*, *60*(9), e14303. https://doi.org/10.1111/psyp.14303

Shi, Y., & Eberhart, R. (1998). A modified particle swarm optimizer. In *1998 IEEE International Conference on Evolutionary Computation Proceedings* (pp. 69–73). IEEE world congress on computational intelligence (Cat. No. 98TH8360). IEEE.

Shon, D., Im, K., Park, J.-H. H., Lim, D.-S. S., Jang, B., & Kim, J.-M. M. (2018). Emotional stress state detection using genetic algorithm-based feature selection on EEG signals. *International Journal of Environmental Research and Public Health*, *15*(11), 2461. https://doi.org/10.3390/ijerph15112461

Simpson's Rule. (n.d.). https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.simpson.html#scipy.integrate.simpson

Soleymani, M., Koelstra, S., Patras, I., & Pun, T. (2011). Continuous emotion detection in response to music videos. 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011, 803–808. https://doi.org/10.1109/FG.2011.5771352

Subramanian, R., Wache, J., Abadi, M. K., Vieriu, R. L., Winkler, S., & Sebe, N. (2018). ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, *9*(2), 147–160. https://doi.org/10.1109/TAFFC.2016.2625250

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. In *Handbook of statistics* (Vol. *24*, pp. 303–329) Elsevier.

Thejaswini, S., & Ravikumar, K. M. (2020). Electroencephalogram based emotion detection using hybrid longshort term memory. *European Journal of Molecular and Clinical Medicine*, *7*(8), 2786–2792.

Vogt, J., De Houwer, J., Koster, E. H., Van Damme, S., & Crombez, G. (2008). Allocation of spatial attention to emotional stimuli depends upon arousal and not valence. *Emotion*, *8*(6), 880–885. https://doi.org/10.1037/a0013981

Wang, F., Wu, S., Zhang, W., Xu, Z., Zhang, Y., Wu, C., & Coleman, S. (2020). Emotion recognition with convolutional neural network and EEG-based EFDMs. *Neuropsychologia*, *146*, 107506. https://doi.org/10.1016/j.neuropsychologia.2020.107506

Wang, Z., Wang, Y., Zhang, J., Hu, C., Yin, Z., & Song, Y. (2022). Spatial-temporal feature fusion neural network for EEG-based emotion recognition. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–12. https://doi.org/10.1109/TIM.2022.3165280

Watts, A. T. M., & Bernat, E. M. (2018). Effects of reward context on feedback processing as indexed by time-frequency analysis. *Psychophysiology*, *55*(9), e13195. https://doi.org/10.1111/psyp.13195

Welch, P. (1967). The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, *15*(2), 70–73. https://doi.org/10.1109/TAU.1967.1161901

Xue, B., Zhang, M., & Browne, W. N. (2013). Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE Transactions on Cybernetics*, *43*(6), 1656–1671. https://doi.org/10.1109/TSMCB.2012.2227469

Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*, *20*(4), 606–626. https://doi.org/10.1109/TEVC.2015.2504420

Zheng, W.-L., & Bao-Liang, L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, *7*(3), 162–175. https://doi.org/10.1109/TAMD.2015.2431497

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Data S1

How to cite this article: Eisenbarth, H., Oxner, M., Shehu, H. A., Gastrell, T., Walsh, A., Browne, W. N., & Xue, B. (2024). Emotional arousal pattern (EMAP): A new database for modeling momentary subjective and psychophysiological responding to affective stimuli. *Psychophysiology*, *61*, e14446. https://doi.org/10.1111/psyp.14446