Crowdsourcing Affective Annotations via fNIRS-BCI

Tuukka Ruotsalo, Kalle Mäkelä, and Michiel Spapé

Abstract—Affective annotation refers to the process of labeling media content based on the emotions they evoke. Since such experiences are inherently subjective and depend on individual differences, the central challenge is associating digital content with its affective, interindividual experience. Here, we present a first-of-its-kind methodology for affective annotation directly from brain signals by monitoring the affective experience of a crowd of individuals via functional near-infrared spectroscopy (fNIRS). An experiment is reported in which fNIRS was recorded from 31 participants to develop a brain-computer interface (BCI) for affective annotation. Brain signals evoked by images were used to draw predictions about the affective dimensions that characterize the stimuli. By combining annotations, the results show that monitoring crowd responses can draw accurate affective annotations, with performance improving significantly with increases in crowd size. Our methodology demonstrates a proof-of-concept to source affective annotations from a crowd of BCI users without requiring any auxiliary mental or physical interaction.

Index Terms—Emotion classification, Functional near-infrared spectroscopy, fNIRS, Pattern classification, Affective computing.

1 Introduction

Human-computer interaction with digital content has long focused on information value and presentation. However, in recent years, affective dimensions have become increasingly recognized as integral to the user experience.

[2], [25], [50]. A key aspect of understanding modeling, and personalizing such experiences is the ability of computing systems to infer the affective states that digital information is likely to evoke and associate that with the content via *affective annotation*. Affective annotation can then be used in downstream tasks to adjust and personalize content, avoid exposure to harmful information, and understand how people consume and react to information that provokes strong emotions [35].

A trivial solution to affective annotation is to rely on manual annotation, where users markup their affective experiences [31]. Manual annotation may be practical for limited scenarios in which users are willing to take the effort, such as marking up content in personalized social media feeds or videos in streaming services. However, the requirement for manual annotation is not likely to scale to a broader set of applications. For instance, it is unlikely that users would be willing to manually annotate their affective reactions for every video clip they watch, song they listen to, or image they view on the Web.

Another approach is to make predictions by analyzing the content itself. For example, using natural language processing to extract affective descriptions from text [68] or computer vision techniques for images and video [33]. However, these methods rely solely on features present within the content itself and do not consider the affective reactions evoked in humans experiencing that content [24]. For example, affective differences may arise from changes in how stimuli are interpreted, such as viewing a scene from a football game. The scene may evoke a variety of responses, depending on whether the person observing it is a fan of the team or not.

Here, as a viable alternative to manual and content-based annotation, we present a method for obtaining the emotional responses implicitly by monitoring human affect at the time of experience. We achieve this by directly measuring passively evoked affective states toward content via fNIRS brain-computer interfacing (fNIRS-BCI). As the brain responses can be noisy, prone to artifacts, and diverging across individuals in different contexts, we approach affective annotation as a crowdsourcing problem. This is based on a simple but powerful idea: multiple participants contribute a noisy signal that can be used to draw consensus estimates [55], [62]. Consequently, crowdsourcing allows learning affective annotations from brain responses of many individuals and can mitigate noise and artifacts.

To this end, we ask the following research questions:

> **RQ1:** *Can fNIRS-BCI monitoring be effectively employed in crowdsourcing settings to predict the affective content of stimuli?*
> **RQ2:** *To what extent does fNIRS-based affective crowdsourcing improve performance of predictive models compared to individual classification?*

To answer the research questions, we report on a neuroimaging data acquisition experiment in which 31 participants viewed visual affective stimuli while their brain responses were monitored via fNIRS.

The participants were not required to perform any artificial physical or mental activities; instead, the experiment relied solely on their natural affective reactions, as indicated

- *Tuukka Ruotsalo is with the Department of Computer Science, University of Helsinki, Finland and Department of Computer Science, University of Copenhagen, Denmark.*
E-mail: tr@di.ku.dk
- *Kalle Mäkelä is with the University of Helsinki, Finland.*
- *Michiel Spapé is with the Department of Psychology and Logopedics, University of Helsinki, Finland.*

Manuscript received February 21, 2022; revised Month DD, YYYY.

by ground truth valence and arousal labels from a well-established data source. Next, we report an affective annotation experiment in which we calibrated machine learning models for participants to distinguish between high/low valence and high/low arousal classes, using consensus labels derived from the signals of multiple participants.

In summary, our contributions are as follows:

1) We present the first-of-its-kind affective annotation from crowdsourced fNIRS-BCI to decode valence and arousal directly from natural affective reactions as they are experienced by a crowd of individuals in response to stimuli.
2) We demonstrate that affective states can be decoded with relatively high accuracy. A crowd of eight participants achieved average accuracies from 0.48 (4-class valence arousal classification) to 0.78 (two-class valence classification of high-arousal stimuli) with consistently increasing performance as a function of the crowd size.

2 Background

Our work is based on several distinct areas of study: emotion research, affective annotation, affective decoding, and crowdsourcing annotations. These are shortly reviewed below.

2.1 Models of emotion and affect

From a psychological perspective, emotion encompasses a wide range of phenomena, including the perception, experience, and expression of emotions, their neural correlates, and social contexts. Research has typically used models to reduce this complexity for empirical studies. In this manner, studies of emotional perception have investigated how stimuli with emotional content affect the body, brain, and behaviour [43], [61] Another research tradition focuses on the experience of emotion itself – the mental representation of physiological changes occurring during an emotion [17] – and the consequences thereof, for example by investigating emotional sensitivity [38], or by determining how cognition is affected by mood experience [64]. Furthermore, studies of emotional expression have explored how emotions alter facial expressions, body postures, and communication, with a long-standing debate continuing as to whether these are mostly universal [30], or primarily defined by culture and norms [57]. In reality, the boundaries between these different focuses are often blurred: seeing a gaping depth opening before you, your emotional perception will prompt fear, and a corresponding, fearful expression would probably follow. However, over a century of research on emotion has not seen a clear consensus being reached as to the exact causal relationship between perception, action, and mental states [13], [29], [49].

In addition to a model of emotion's specific focus, another critical factor for affective computing is the model's taxonomy of emotional identities. Two broad families of emotion theories are commonly found. On the one hand, discrete theories of emotions typically identify a limited number of qualitatively different emotions that give rise to the range of experiences named in most languages. For example, universal emotion theory tends to understand emotions by their evolutionary value for communication, with facial expressions signifying critical messages that can be understood even across different cultures [30]. On the other hand, dimensional theories identify a smaller number of continuous variables as latent factors that provide an internal representation of emotions. For instance, the primary dimension of *arousal* is traditionally thought to be caused by autonomic nervous activity, resulting in outward expressions of excitement [32]. The hedonic dimension of *valence*, whether affective state is experienced as pleasant or unpleasant, is often viewed as involving more cerebral cognitive processes such as attribution [58]. Dimensional theories thus account for emotions by combining the dimensions, for example explaining ”joy” as caused by high arousal and high valence.

2.2 Affective annotation

Annotation refers to adding descriptive metadata to digital content, which has traditionally been an essential component of many digital media services. By labeling media content with their evoked emotional experience, affective annotation provides particularly useful information. The methodological aim of affective annotation is to build methods to estimate how humans would experience content. For example, whether they find it pleasant, offensive, relaxing, or frightening. Traditionally, affective annotation has been approached via manual interaction [1] and content-based analysis of text or visual media content [4], [26]. The manual annotation process relies on explicit interaction enabled by interface designs that allow users to manually indicate their affective reactions when they are experiencing the content. Well-known examples of manual annotation are markup that allows expressing emotional responses or affective experiences [60].

While manual annotation can produce rich descriptions, the process is typically labor-intensive and limited by how much conscious access annotators have to affective states. For example, users might thoroughly enjoy digital media during the experience but forget the initial impact or constructively reinterpret their experience later. By not focusing on explicit, manual processes, implicit methods of affective annotation may avoid such constraints, presenting *affective decoding* techniques for detecting how content is perceived emotionally without relying on explicit interaction from users [10].

2.3 Affective decoding

Affective decoding aims to estimate the affective experience of an individual by mapping the relationship between emotions and measurable signals. Neuroimaging provides information directly from the presumed origin of affective states: the brain [50]. Measures can be obtained with various non-invasive imaging techniques, such as electroencephalography (EEG) and functional magnetic neuroimaging (fMRI). In studies using EEG, alpha power asymmetry between frontal sites has been used to detect the motivational direction and valence [39]. However, the limits of localizing scalp-recorded EEG have led to controversy over the use of this biomarker [3], [18]. Previous fMRI studies have shown that valence and arousal affect both the prefrontal cortex and deeper brain structures such as the amygdala and insula [52]. Activity in the amygdala, in particular, has been associated with the highly salient emotion of fear (high arousal/negative). In contrast, prefrontal areas have been associated with affective processing of the pleasantness of images [36]. Despite their power to study the underlying structural and spatiotemporal correlates of emotions, neither electroencephalography (EEG) nor fMRI has seen strong uptake in the field of affective computing in practical human-machine interfacing settings, owing to their high cost and unwieldiness.

Functional near-infrared spectroscopy (fNIRS) presents an alternative method for quantifying cortical activity for inferring emotional processing. Since neural activity causes changes in blood oxygenation (BOLD) and since the light-absorption is affected at different wavelengths for oxygenated and deoxygenated hemoglobin [6], [12], fNIRS allows neural activity to be quantified, especially in cortical areas near the surface that are unimpeded by light-interfering tissues (e.g. hair). Thus, anterior-frontal and frontal-polar areas underneath the forehead tend to provide stronger signal-to-noise than deeper areas that reside below regions of the scalp that are typically covered by hair, such as the inferior parietal lobule.

Recent studies show fNIRS holds clear promise for affective decoding of both discrete emotions [40] and emotional dimensions [7]. In particular, fNIRS may be more successful than more ubiquitous forms of biosensing that measure activation of the autonomic nervous system, such as electrodermal activity (EDA) or heartrate, by potentially detecting valence from cortical activity in the central nervous system. Previous studies, for example, showed that viewing unpleasant (negatively valenced) images was found to particularly affect the BOLD response in the right prefrontal cortex [7]. Such findings have seen strong application within the field of human-computer interaction, in which the use of fNIRS has become increasingly common [67]. Studies in HCI have, for example, applied fNIRS during implicit interfacing between users and computing [69], enhancing real-time interfaces with additional input modality [66], evaluating visualizations [51], and determining the user experience in virtual reality [72]. Thus, although the usefulness of fNIRS as a general tool for HCI and user experience studies depends on the type of task [47], a clear consensus is forming that fNIRS can be a viable alternative to existing biological sensors and physiological measures, showing strong potential for complementing human-computer interaction studies with tools for quantifying affective experiences of users.

2.4 Crowdsourcing annotations

Crowdsourcing has emerged as a powerful approach to obtaining annotations for large media databases, such as labeling objects appearing in images, labeling text, and affective features of stimuli [14], [48], [74]. In this process, users undertake microtasks and human cognition is exploited jointly with computing systems to obtain information about stimuli. Conventionally, these tasks require simple manual input, such as selecting images that match a description [11], [73]. The majority of applications of crowdsourcing have focused on such explicit human input. However, another line of crowdsourcing research and practice relies on implicit feedback, where task-relevant information is collected implicitly as a side product of people's natural interactions. For example, search engines obtain annotations for query-document pairs by observing documents clicked in response to a query [15].

Recently, researchers have also explored physiological signals for crowdsourcing. In [20], researchers presented a methodology called brainsourcing, in which EEG responses toward facial images were decoded for relevance and consensus annotations were inferred through a crowd model. In [63], researchers approached a similar problem and presented results for predicting stimuli classes in a multi-user setting. In [28], the emotional experience of multimedia contents was detected from EEG in real-time when users were watching video clips. These responses were then used for emotion tagging. Similarly to our work, inter-brain features from a group of participants were used to find a consensus label.

EEG and fNIRS data have also been used in studying both within-subject [8] and cross-subject [9] classification scenarios. The authors have identified neural correlates of emotions using fNIRS data across subjects. However, although the models were built across subjects, which provided the capacity to generalize and predictively classify emotions in new participants, the task of predicting crowdsourced consensus estimates was not explored.

In summary, brain-computer interfacing demonstrates the potential for implicit crowdsourcing, where human opinions about stimuli are inferred from subject-independent models or collective models are trained using physiological data [22]. Our approach follows this line of research but is the first to employ fNIRS neuroimaging and adopt affective annotation that relies on natural responses to stimuli, rather than pre-assigned recognition tasks. Furthermore, we demonstrate that decoding affective states from these reactions through crowdsourcing leads to significant improvements in performance.

3 Neuroimaging data acquisition

The study was performed in compliance with the protocols laid out by the Declaration of Helsinki and was approved by the Ethical review board in humanities and social and behavioral sciences of the University of Helsinki. Participant recruitment concentrated on the undergraduate and postgraduate student population, with no requirements other than having a normal or correct-to-normal vision and having no psychiatric disorder (operationalized as having no current diagnosis and not currently taking any psychopharmaceuticals.). Thirty-one participants volunteered and took part in the study after being fully informed of the study and their rights, including the right to withdraw at any point without fear of negative consequences and signing their informed consent. Following pre-processing of data (see below), four participants were found to have fluctuations in the data recordings and were removed from the conventional statistical analysis that were conducted to study neurophysiological effects. All participants were, however, included in the machine learning experiments.

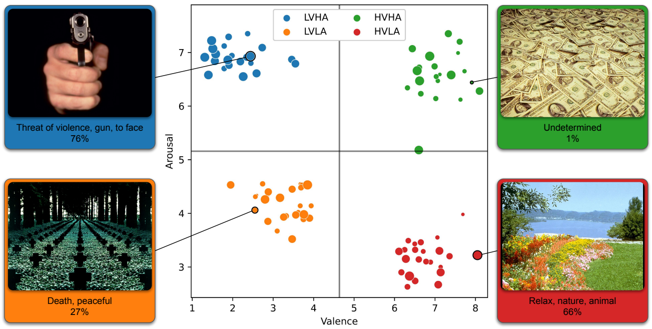


Fig. 1: Distribution and examples of stimuli samples in the four classes positioned on valence and arousal scales. Low-valence high-arousal (LVHA) in blue, high-valence low-arousal (HVLA) in green, low-valence low-arousal (LVLA) in orange, and high-valence low-arousal (LVLA) in red. Below the example images are their tags and crowdsourced image-specific classification accuracies with $N = 8$.

The average age of the participants was 31.4 (minimum 21, maximum 52, SD = 7.76) years. Regarding gender, fifteen participants reported being male, eleven female, and the rest non-binary. They were compensated for their time and efforts with local movie vouchers.

3.1 Stimuli

Stimuli were sampled from the international affective picture system (IAPS) [44] for use in the present study. The IAPS is a database of images previously rated by a large sample on their emotional reactiveness across three dimensions: arousal, valence, and dominance. Like most studies in affective computing and neuroscience, we focussed on the first two dimensions, being traditionally understood as the two main dimensions of emotion [59]. Arousal refers to the degree of nervous excitation provoked by the stimuli. The pleasantness or hedonic value of such stimulation is referred to as valence. By orthogonally crossing the dimensions, i.e. combining the classes of low and high valence with those of low and high arousal, four quadrants were defined: low valence / low arousal (LVLA), low valence / high arousal (LVHA), high valence / low arousal (HVLA), and high valence / high arousal (HVHA). Since high arousal images tend to have higher variance in valence [56], we selected the 60 images with the lowest valence (2.71 +- 01.81 on a scale of 1 to 9), and 60 with the highest valence (6.94 +- 0.53), then divided these each to form the low and high arousal samples (i.e. creating four quadrants of 30 images each). Examples and the distribution of stimuli samples are shown in Figure 1. From each quadrant, a participant viewed a random selection of 10 individual images. To increase standardization of perceptual factors, images were scaled vertically to 1024 px.

3.2 Apparatus

E-Prime 3 (Psychology Software Tools, Inc., Sharpsburg PA), running on a Windows 10 PC, was used for stimulus presentation, behavioral data recording, and device synchronization. The presentation used a 22-inch LCD monitor running at 1920 x 1080 px, explicit feedback were obtained from the keyboard, and synchronization between the display and data recording was done via the DCOM interface to send triggers to the fNIRS device. Optical density data were recorded using an Artinis Brite-24 fNIRS device. The Brite uses 10 LED transmitters and 8 receiving photodiodes placed on an elastic cap to standardize localization between users. Here, a frontal configuration was used, with each receiver obtaining light from three transmitters placed at a distance of ca. 3 cm. By combining 5 transmitters and 4 receivers for each hemisphere, we were able to record optical densities from 12 left and 12 right frontal areas. These were digitized and recorded using Artinis OxySoft software at a sample rate of 50 Hz.

3.3 Procedure

The experiments took place in a designated laboratory space. After reading the instructions and signing informed consent, the participants were seated and fitted with an fNIRS device. This involved putting on the elastic cap and fitting the diodes in the holders, then adjusting hair and diode orientation so as to reduce interference and artifacts. Following this, a 1-minute resting-state measurement was obtained while participants focussed on a centrally displayed crosshair against a grey background. The recording session itself involved two blocks of 20 trials each. Each trial commenced by instructing users to carefully view the subsequently presented image and freely associate with its content. After taking the necessary time to read these instructions and pressing a key, a fixation cross was shown for 4 seconds to provide a neutral baseline for data analysis, before the experimental stimulus was presented, which was shown for 14 seconds. Finally, during a blank inter-trial interval of at least 0.1 s, trial-specific information was synchronised with the biosignal data. Note that the influence of the preceding image on the evoked response of the present was assumed to be limited for two reasons. First, the interval between two emotional images was substantial (4s + 14s + time to press, total M = 21.1s, SD = 1.9s). Second, stimuli of each quadrant were presented with their order randomised for every four trials (restricted only against emotion repetition). Thus, any carryover effect would be equal across averages. As all analysis and machine learning experiments were also averaged either by analyzing all data or through cross-validation, there should be no effect on the results. The entire experiment took about 45 minutes to complete.

4 Affective Annotation Experiment

The affective annotation experiment aimed to evaluate the predictive performance of the crowdsourcing approach to decode affective categories of stimuli from their evoked fNIRS responses. The methodology, from producing individual classifications for each epoch to combining them to create crowdsourced predictions, is described below.

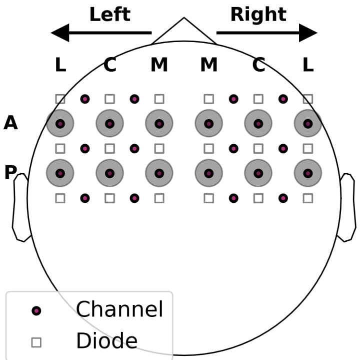


Fig. 2: fNIRS channel and diode placement. The analysis used only the channels highlighted with grey circles, with a montage separating the regions into anterior (A) and posterior (P) frontal regions and each of the hemispheres divided across lateral (L), central (C), and medial (M) channels.

4.1 Tasks

We experiment with five affective classification tasks based on the well-known dimensionality theory of affect. The dimensionality of emotion or affect is most commonly represented in a two-dimensional space spanning valence and arousal. Valence accounts for the extent to which an emotion is positive or negative, and arousal accounts for the intensity of the associated emotional state. The main task, referred to as *4-class*, aims to classify each image into one of four affective classes, *high-valence-high-arousal (HVHA)*, *high-valence-low-arousal (HVLA)*, *low-valence-high-arousal (LVHA)*, and *low-valence-low-arousal (LVLA)*. The following two tasks, *Valence* and *Arousal*, only try to predict the high or low valence (negativity or positivity) or high or low arousal (intensity level) of the stimuli, ignoring the other affective dimension. In tasks *high-arousal valence* (*HA Valence*) and *low-arousal valence* (*LA Valence*), images are also classified by valence, but the classification considers only either high-arousal or low-arousal stimuli. Studying these separately is motivated by an assumption that affective states with stronger intensity (high arousal) are more important for many downstream tasks and may be easier to decode.

4.2 Data preprocessing

The Optical Density (OD) data and stimuli are processed using MNE python [37]. We apply a 3x3 grid layout for both left and right hemispheres, closely resembling the original sensor layout. Since raw fNIRS recordings are susceptible to various noise sources, standard preprocessing is conducted. First, to detect poorly connected sensors, the scalp coupling index (SCI) [54] is applied to each channel. SCI measures whether the channels measuring activity at different wavelengths in the same location are negatively correlated at the heartbeat's frequency range (0.7 - 1.5Hz). Low SCI indicates poor coupling; hence channels with SCI below the threshold of 0.8 are interpolated by taking the average of their neighboring channels. As the final OD preprocessing step, artifacts due to, e.g., motion, are corrected with temporal derivative distribution repair [34].

After processing the OD data, it is converted to oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) concentrations with the modified Beer-Lambert law [23]. Finally, to remove physiological noise, such as the heartbeat, from the hemoglobin concentrations, a 0.1 Hz low-pass filter is used, while a 0.01 Hz high-pass filter is applied to eliminate slow drifts in the signal. After preprocessing, the data is divided into 17-second epochs, consisting of 12 seconds of recording after the stimulus and a 5-second baseline period before.

4.3 Neuroimaging analysis

To infer the effect of affect on perceiving emotional images on frontal brain activity, we performed a statistical analysis at the population level. Baseline activity was subtracted from the averaged 12 seconds of post-stimulus HbO and HbR levels. A brain-wide analysis was conducted with channels arranged along a montage using solely the transmitter/receiver diode pairs along the sagittal plane (i.e., up/down arranged on the forehead), as shown in Figure 2. For the areas, we then compared these between the left and the right *hemisphere*; between three relative levels of *lateral region* from the furthest to the side (lateral), via the central/medial, to the medial; and between the relatively anterior and the posterior *frontal region*. Thus, for every participant and each combination of low and high *valence*, and of low and high *arousal*, 12 averages were analysed for two hemispheres, three lateral regions, and two frontal regions. To determine if valence, arousal, and their interaction affected fNIRS responses across participants, two 5-way repeated measures ANOVAs were conducted, one with HbR as the measure, and the other with HbO as the measure. To reduce the chance of type-I errors, only p-values below 0.025 (i.e. with Bonferroni correction applied to the alpha criterion) were reported. To maintain brevity, we do not report non-significant effects or effects without the involvement of emotional factors.

4.4 Feature extraction

The high-dimensional epoch data was converted to lower-dimensional feature space. In fNIRS, a typical response to stimuli occurs approximately 4 to 12 seconds after stimulation, which is used here as the size of an epoch. To capture this effect with simple features, the windowed mean from three equally sized non-overlapping windows was extracted for each channel. To further reduce the dimensionality of the feature space, the HbR channels were eliminated, as HbO and HbR channel pairs are strongly dependent [16]. Finally, the features are concatenated, resulting in feature space with 72 features per epoch.

4.5 Prediction model

Linear discriminant analysis classifier with shrinkage regularization (SLDA) was used as the predictive model. SLDA offers many attributes that make it an attractive choice for fNIRS modeling, such as good performance in high-dimensional low-sample settings, fast training and inference, and output of prediction probabilities for each class,

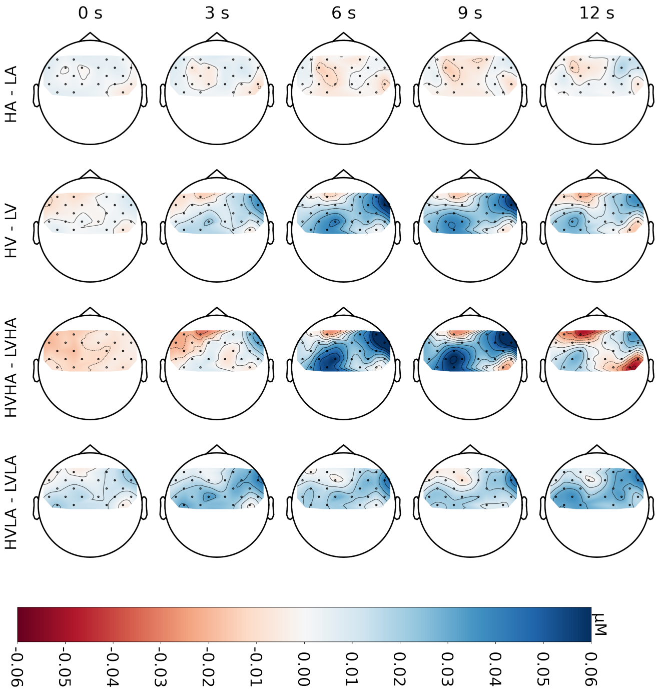


Fig. 3: Affective fNIRS oxygenated hemoglobine (HbO) response. Contrasts are shown to highlight the effect of arousal (high arousal vs. low arousal, HA-LA); valence (high valence vs. low valence, HV-LV); valence given high arousal (high valence high arousal vs. low valence high arousal, HVHA-LVHA); and valence given low arousal (high valence low arousal vs. low valence low arousal, HVLA-LVLA).

which are essential for crowdsourced predictions. The classifier does not require hyperparameters, and the regularization parameter of the SLDA model is determined by the Ledoit and Wolf lemma [46], which provides an analytical estimate for the optimal shrinkage constant.

4.6 Prediction setup

The prediction model's target is to predict each class's probability for each epoch using the feature representation. The data are split into training and testing sets with the stratified k-fold cross-validation scheme, where k is the number of samples in the least common class for that participant. Selecting k in this manner ensures that each test set has at least one sample from each class. For 29 out of 31 participants, the cross-validation is equivalent to stratified 10-fold, but for participants with missing epochs, a smaller k is required. Since each sample belongs to exactly one test set, this process yields one test set prediction for each epoch, which are used in the latter steps.

4.7 Crowdsourced prediction setup

The crowdsourcing experiment follows a scenario where groups of $N \in \{1, ..., 8\}$ participants produce crowdsourced predictions for images in a way that allows comparison between different group sizes.

Before producing the crowdsourced predictions, 22 images were eliminated because there was data from less than 8 participants available for them. The varying amount of predictions for different images is due to the sampling in the stimuli selection process; each subject is shown 10 randomly sampled images from each class. Eliminating images with less than eight predictions allowed the use of the same set of images for all group sizes. The remaining 98 images had 8 to 17 unique predictions, 11 on average, and the class distribution was as follows: LVLA=27, LVHA=26, HVLA=24, and HVHA=21.

The crowdsourced predictions were produced iteratively for each image individually. On each iteration, a new participant is sampled with replacement from the participants to whom the image was shown and added to the image's participant pool. Then, the predictions from the image's updated participant pool are combined via soft voting, i.e., by taking the average of class probabilities over each participant's predictions, and choosing the class with the largest mean probability, which forms the new crowdsourced prediction. Soft voting was chosen as it was found to perform the best among several voting schemes (See Appendix A). The iteration is stopped when crowdsourced predictions for $N \in \{1, ..., 8\}$ are created. Adding one participant to the previous iteration's participant pool minimizes noise factors due to, e.g., entirely different participants, and the difference in results between N can be attributed to the change in group size. This process was repeated 100 times for each of the 98 images with the aim of simulating crowdsourcing's effectiveness across different, varying groups. Each repetition produced eight predictions for different group sizes, resulting in $98 \times 100 \times 8$ crowdsourced predictions.

4.8 Control model and statistical testing

A random model was trained for a control model to find an empirical random performance. The training followed the same procedure as the model with real data, but the labels were permutated. The mean accuracy scores for each N were then evaluated with permutation tests with 100 permutations. All tasks achieved the minimum p-value, $p = 0.01$, with all N.

5 Results

5.1 Neuroimaging effects

To determine whether emotion generally affected the Oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) responses to viewing images, repeated measures ANOVAs were conducted with *valence* (low, high), *arousal* (low, high), *hemisphere* (left, right), *lateral region* (lateral, central, medial), and *frontal region* (anterior, posterior) as factors, and HbO and HbR as measures. In HbO, this showed significant effects of valence, F(1,26) = 8.88, p = 0.006, with more negative responses in low (-1.95 +- 0.36) than high (-1.14 +- 0.40) valence conditions. Valence furthermore interacted with the hemisphere and frontal region, F(1, 26) = 7.15, p = 0.01, and entered a three-way interaction with the frontal region and arousal, f(1, 26) = 16.46, p < 0.001. This effect could be characterized in reference to the general negative effect of low valence being especially large in the more anterior area in the high arousal condition (D = 1.44) compared to low arousal (0.47) or the more posterior region (0.89). With HbR, only one significant effect was

Task	N=1		N=2		N=4		N=8	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
4 class	0.40	0.39	0.40	0.40	0.45	0.45	0.48	0.48
Valence	0.59	0.58	0.62	0.62	0.64	0.64	0.67	0.66
Arousal	0.56	0.56	0.60	0.60	0.61	0.61	0.63	0.63
HA Valence	0.67	0.67	0.70	0.70	0.74	0.74	0.78	0.78
LA Valence	0.57	0.57	0.59	0.58	0.61	0.61	0.63	0.63

TABLE 1: Accuracy and F1 scores for different N for each task. The datasets are nearly balanced for all prediction tasks.

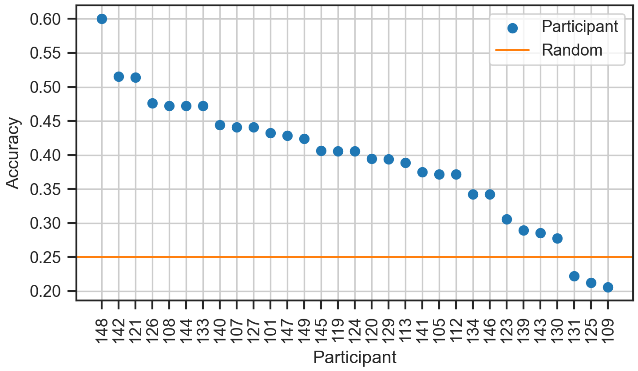


Fig. 4: Per-participant model accuracies in the 4-class prediction task.

observed, the interaction between valence, hemisphere, and frontal region. This suggested a more positive effect of low valence in left posterior areas than left frontal areas (-0.002) or right hemisphere areas (0.04). A more comprehensive, exploratory analysis is presented in Figure 3 with all diode-pairs included, showing effects for HbO, particularly in left medial-posterior and right frontolateral areas. Valence generally shows a stronger response than arousal, although the two lower rows in the figure suggest this effect occurs mainly in conditions of high arousal.

5.2 Classification performance

Participant-specific models. The participants' individual classification performance was evaluated before the crowdsourcing task. Each participant's individual classification accuracy was calculated from all predictions made by that participant. The participant-specific 4 class accuracies are shown in Figure 4. In the 4-class task, the average overall accuracy for a participant was 0.40 ± 0.02 (± standard error). For other tasks, the mean accuracies were Valence 0.59 ± 0.01, Arousal 0.56 ± 0.02, HA Valence 0.67 ± 0.02, and LA Valence 0.57 ± 0.02. All mean accuracies were significantly different from the accuracies of the random model using permutation tests with 100 permutations ($p = 0.01$).

Crowdsourced models. Table 1 and Figure 5 show the classification accuracies for different group sizes. First, 100 combination scores were calculated for each N by, for $i \in \{1, .., 100\}$, taking the prediction from the ith participant group of each image and calculating their classification accuracy. For example, the first combination score is calculated by taking the classification accuracy over the crowdsourced predictions from the first participant combination of each image. This is conducted for each participant combination, resulting in 100 combination scores per N. Figure 5 visualizes the mean and standard deviation of the accuracies for different group sizes, and Table 1 shows the numerical values of the mean accuracies and F1 scores. The classification performance consistently improves as the crowd gets larger in all tasks. This is also visible in classifier decision probabilities in Figure 6. The distribution converges as crowd size increases.

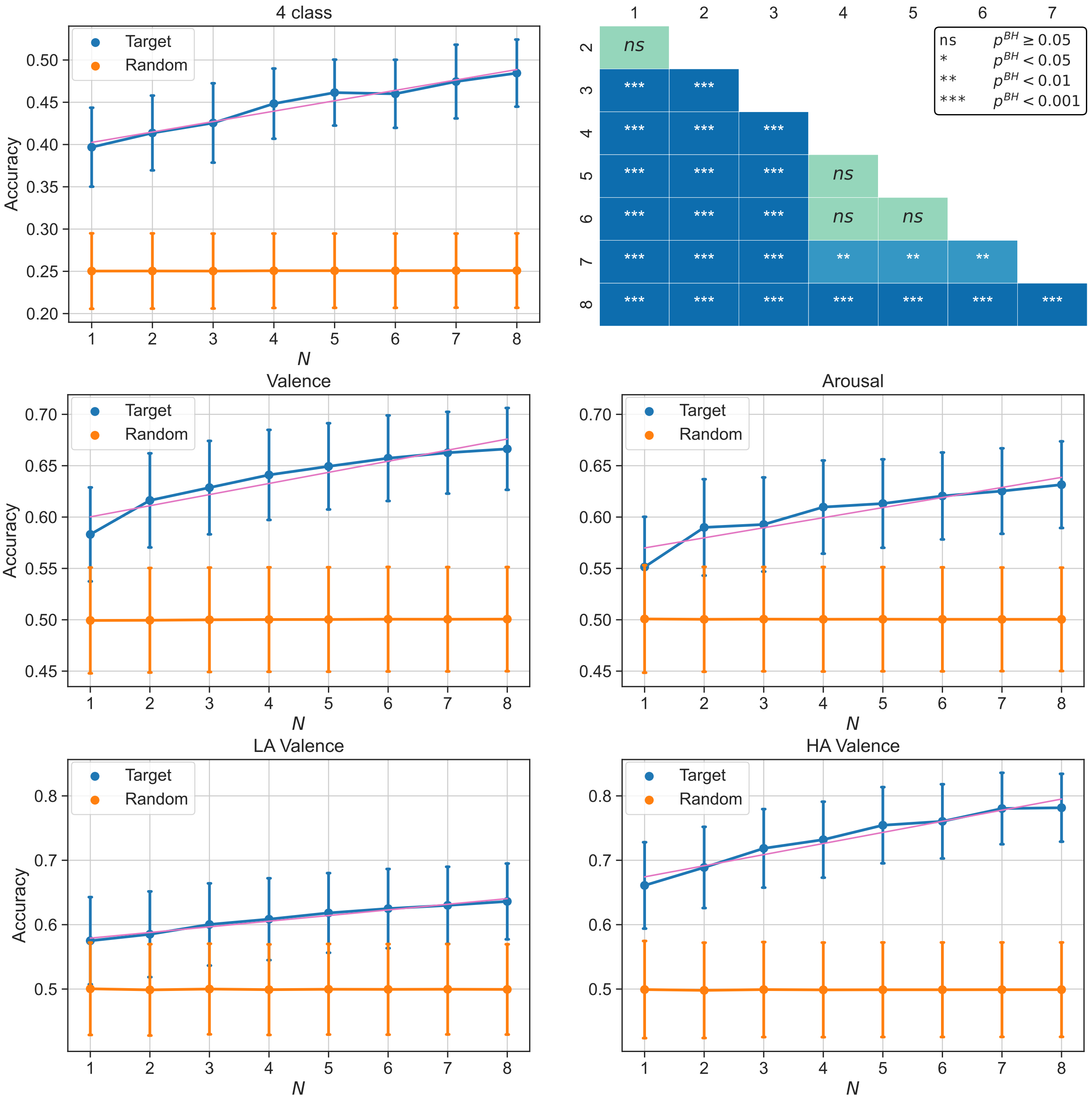
Task	β_N	p
4 class	0.012	< 0.001
Valence	0.010	< 0.001
Arousal	0.009	< 0.01
HA Valence	0.017	< 0.001
LA Valence	0.008	< 0.001

TABLE 2: The effect of group size on the accuracy, measured by coefficients β_N and their corresponding p-values.

Significance of crowd size. The improvement in performance relative to group size was evaluated by testing for linear dependence between N and mean accuracy. This test was conducted by first fitting an OLS simple linear regression model to $\{(N_i, \overline{Acc_i})\}_{i=1}^{8}$ for each task. The fits of these models are visualized in Figure 5 as purple lines. Then, testing if the coefficient of N, β_N, is significantly different from 0 with a Student's t-test. The coefficients β_N and their corresponding p-values are shown in Table 2.

Differences in the performance of crowdsourced predictions with respect to group sizes were also compared at the image level to outrule the possibility that different stimuli would account for the performance differences. The accuracies were calculated by taking the classification accuracy over all combinations for each image. This results in 98 image scores for each N. The image scores of different N were compared with each other using the Wilcoxon signed-rank test, with the alternative hypothesis that the larger group outperforms the smaller one. The image-specific accuracies of larger groups are predominantly greater than those of smaller groups, especially when the difference in size is substantial. The Benjamini-Hochberg adjusted pairwise statistically significant differences across different crowd sizes are visualized in the top-right corner of Figure 5.

Significance of affective class and stimulus content. There were substantial differences between crowdsourced classification accuracies of different images in the 4 class task with 8 participants. Figure 1 illustrates the image-specific accuracies by the relative size of the dot markers.



	1	2	3	4	5	6	7
2	ns						
3	***	***					
4	***	***	***				
5	***	***	***	ns			
6	***	***	***	ns	ns		
7	***	***	***	**	**	**	
8	***	***	***	***	***	***	***

ns	$p^{BH} \geq 0.05$
*	$p^{BH} < 0.05$
**	$p^{BH} < 0.01$
***	$p^{BH} < 0.001$

Fig. 5: **Top left:** Classification accuracy for the full 4-class (low/high valence, low/high arousal) as a function of crowd size. **Top right:** Statistical significance for differences between models with different crowd size (Benjamini-Hochberg adjusted). **Middle:** Classification accuracy for high/low valence (left) and high/low arousal (right). **Bottom:** Classification accuracy for low-arousal valence (left) and high-arousal valence (right). All results show accuracy as a function of crowd size. The orange lines show control model performances trained with randomly permutated labels. The error bars denote the standard deviation of the accuracy scores.

Noticeably, LVHA images have higher average classification accuracy (0.62) than HVHA, LVLA, and HVLA, with accuracies of 0.45, 0.45, and 0.38, respectively. It is evident that the image class, and therefore the valence and arousal, affects the classification accuracy. Most notably, high-arousal images achieved significantly higher accuracies (Mann-Whitney U = 1515.5, $p < 0.05$ two-tailed) than low-arousal images, suggesting that images that evoke more intense emotional responses are easier to recognize.

To further investigate the distinguishability of types of images, we assigned images to smaller groups with descriptive tags (e.g., Figure 1) and examined differences in prediction accuracy for each tag. Tags with less than three representative images were not considered. In line with our previous finding, the highest scoring tags were associated with the LVHA class, more specifically with grisly images

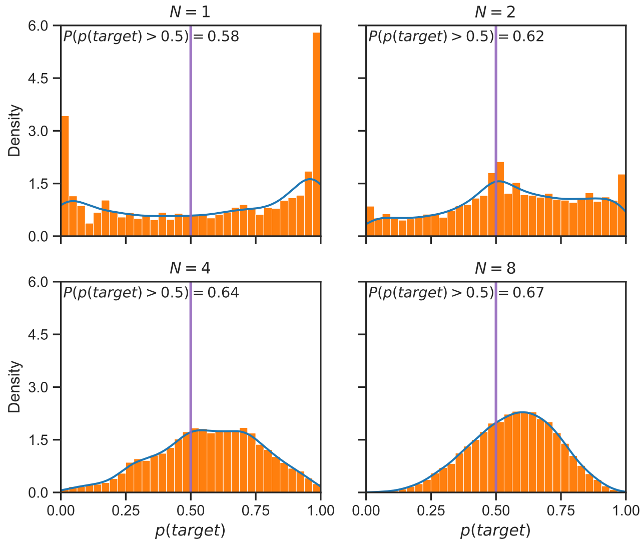


Fig. 6: Distribution of crowdsourced predictions for the target label in valence classification for increasing crowd size (upper left N=1, upper right N=2, lower left N=4, lower right N=8). The prediction probabilities converge as crowd size increases.

(grisly 0.72, injury 0.68). In addition, the LVHA class had another tag type that scored high, threats of violence (knife 0.56, threat of violence 0.55). The highest scoring tags from other classes were couple (HVHA) 0.66, dirty (LVLA) 0.49, and sociability (HVLA) 0.45. Lower scoring tags were usually ambiguous, such as peaceful (LVLA) 0.22, which was most commonly predicted as HVLA, or associated with multiple classes, such as animals (HVLA, HVHA, LVHA) 0.33. This result further supports the finding that high-arousal images are easier to classify.

Prediction accuracy is dependent on the content of the stimulus image. Images that evoke strong responses are easier to classify, while it is more difficult to distinguish between milder emotional responses. This suggests that greater performance could be achieved in downstream tasks that deal with distinctive content evoking strong responses.

6 Discussion and conclusions

Existing approaches to affective annotation typically rely upon manual annotation, which is labor-intensive and necessitates explicit interactions from users. On the other hand, automatic methods that analyze *only* content to estimate users' affective responses may be unreliable and produce affective state estimations that diverge from users' actual experiences. Here, we explored an intriguing alternative to affective annotation: learning affective annotations directly from brain signals by passively monitoring the affective experiences of a crowd of participants. The present work, to the best of our knowledge, is *the first-of-its-kind to employ fNIRS brain-computer interfaces in a crowdsourcing setting for affective annotation.* Our approach is based on a simple but powerful idea: The affective states decoded from the brain responses of many participants toward stimuli can be used to infer a consensus estimate of the affective response that the stimuli are likely to evoke. Since our approach relies on implicit affective responses as they are naturally experienced by users, without requiring any artificial physical or mental activity, we envision that they could be monitored implicitly as part of everyday human-computer interaction.

6.1 Answers to research questions

To study whether crowdsourced brain-computer interfacing can be used for affective annotation we asked two research questions, which we answer below.

RQ1: Can fNIRS-BCI monitoring be effectively employed in crowdsourcing settings to predict the affective content of stimuli? Yes, we show that fNIRS measured from the frontal lobe carries information about affective states experienced by humans (Figure 3). Valence, in particular, was associated with activity in the medial left and lateral right frontal cortex. We demonstrate that from such patterns of activity, affective annotations can be decoded via machine learning with relatively high accuracy and significantly increasing performance with respect to crowd size (Figure 5). The prediction accuracy varies between 0.48 (against 0.25 random) for a four-class valence-arousal classification to 0.78 (against random 0.5) valence classification for high-arousal stimuli (see Table 2 for details). High-arousal stimuli, in general, are more likely to evoke stronger affective responses [58]. They can also be more important for downstream applications: The stronger the affective response, the higher the importance for affect detection and annotation. The accuracy of the latter result is particularly encouraging as it suggests that performance in real-world downstream tasks, such as detecting harmful content or content that evokes particularly positive responses, may perform at a similar level of quality as manual annotation. It is noteworthy that these results are achieved entirely implicitly, meaning they are based solely on perception without requiring any explicit mental or physical activity from the participants.

RQ2: To what extent does fNIRS-based affective crowdsourcing improve performance of predictive models compared to individual classification? The results show a significant increase in accuracy with respect to crowd size, exhibiting a consistently increasing performance. This suggests that relatively small crowds can be used to source affective annotations effectively, and less than 10 participants are enough to obtain high accuracy (Figure 5). The classifier analysis further supports this finding, which shows the distribution of average class probabilities stabilizing as a function of crowd size (Figure 6).

6.2 Limitations

The reported performance may overestimate or underestimate future replications or applications, depending on differences in sampling procedures and apparatus. However, the standardized acquisition setup and data processing protocols make it unlikely that the reported differences between conditions were due to confounding factors. That is, noise in the LED-diode-based fNIRS may have adversely affected accuracy compared to laser-based fNIRS, which has been shown to reduce crosstalk and improve spatial accuracy [41]. Conversely, our recruitment of healthy, relatively young participants may have improved overall accuracy due to their engagement with the task being likely stronger than would be observed in the general population. However, since the neuroimaging data acquisition employed a fully randomized experimental protocol, such effects cannot account for the observed differences between the conditions. Moreover, these effects were robust across variations in neuroimaging analysis, decoding models, and crowd-analyses. We, therefore, expect the results to generalize towards future studies and application settings.

The experimental design further places limitations on the ecological validity. For example, while the randomised order balanced interference from preceding emotional images, such that the reported averages were unlikely to have been due to carry-over effects from preceding trials, such balancing is unlikely to occur in the real world. Indeed, in common interaction, emotions may follow one another in rapid succession and repeat more frequently than alternate. Furthermore, the visual stimuli we used were selected from a standard and widely used affective image database. This allows for excluding many contextual factors that might be present in real-world content, such as news articles and associated images. It also allows for comparing and reproducing our results. On the other hand, the images are old and may not always be comparable to images that users would encounter when browsing the Web, for example. Such differences in studies of emotions within and outside the laboratory are now more frequently recognized within psychology and affective computing [45], [70], [71], and future research must determine whether the reported results will replicate towards emotions captured during real-life interaction.

Another factor in our experiment is the specific decoding model that is used to classify affective states. The model is a fairly standard classification model, and we used standard grid search to optimize pre-processing and feature extraction. All procedures were conducted in a repeated k-fold cross-validation setting, with any model tuning performed exclusively using the training data. We also experimented with other standard models and did not find performance differences that would be significant. Our consensus labeling followed a simple strategy of aggregating individual predictions that were also found successful in earlier studies with manual labels [62]. Therefore, we can be confident that the model or the learning setup does not account for the significance of the results. Nevertheless, it is possible that experimentation with a larger amount of participants, more advanced representation learning, or more sophisticated label aggregation could lead to further improvement of the results.

6.3 Ethics

Brain-computer interfacing, and physiological computing more generally, provide new opportunities for computing systems that learn directly from the human cognitive system. This is enabled by active monitoring of humans while they are interacting with their digital environments. This technology has advanced with unprecedented speed during the past decade and is transforming from laboratory experimentation in a research setting to consumer-grade devices that measure human brain activity and physiology in the wild.

These new opportunities provide novel signals from humans to be used in a variety of human-facing applications, but the technology may also raise concerns about the abuse and misuse of these susceptible signals.

For instance, fNIRS data should be considered personal medical data; protecting it becomes particularly important as it can be used as a cognitive biomarker [53], detecting cognitive load [67], detecting cognitive (dis)ability [5], and other sensitive biomarkers, such as deception [27]. On the other hand, it is clear that the current stage of technology is not such that one might unobtrusively detect emotions. That is, unlike signals such as EDA or heartrate, fNIRS is far from a ubiquitous form of biosensing, making it at present unlikely to be used without a user's explicit consent.

Data captured via BCI could also be used together with other individuals' signals. For example, combining the affective data with browsing behavior and comparing that to the data of other individuals' behavior and affective responses. Moreover, subliminal probing could be used beyond the annotation task for predicting unwanted user characteristics [21] and compared to other individuals' data to reveal even social or political views. Preventing unintended use of these signals requires future research for protecting the privacy of data.

6.4 Future work

Although ergonomics, cost, and comfort may impede the adoption of consumer-grade BCI, our methodology demonstrates a proof-of-concept approach to source affective annotations from a crowd of BCI users without requiring additional mental or physical interaction effort. Future work could experimentally investigate affective decoding with novel sensors and fewer transmitter-receiver pairs to study whether a reduced hardware setup could yield similar results.

The present machine learning models are well-suited for the scenario where a relatively small amount of data is available from each participant. Although classical machine learning methods have proven challenging to outperform in affective classification settings for various downstream tasks [19], [42], [65], conducting experiments with representation learning and contrastive learning models, along with data augmentation, should be considered. These could learn to better separate nuanced signals associated with each affective state. Furthermore, by extending the models to account for participant-independent data, a single model could be trained across participants rather than requiring per-participant models that are then fused in the crowdsourcing stage.

Our approach and study fall under implicit crowdsourcing: participants were not instructed to perform any specific tasks, and they only naturally reacted to the presented stimuli, which were successfully decoded from both individual and crowd responses. This mitigates the need for setting up specific experiments for utilizing our methodology in real-world settings. To this end, future research should explore sourcing affective annotations with accessible hardware and data outside of a pre-recorded stimuli database to capture affective annotations as they occur in our everyday interaction with digital information.

Acknowledgments

The research was supported by the Academy of Finland (352915,350323,336085,322653) and the Horizon 2020 FET program of the European Union through the ERA-NET Cofund funding grant CHIST-ERA-20-BCI-001. We thank Keith Davis for his comments and suggestions on the manuscript.

References

[1] Sarkis Abrilian, Laurence Devillers, S Buisine, and Jean-Claude Martin. Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *HCI International*, volume 401, pages 407–408, 2005.

[2] Anshu Agarwal and Andrew Meyer. Beyond usability: evaluating emotional response as an integral part of the user experience. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 2919–2930. 2009.

[3] John JB Allen, James A Coan, and Maria Nazarian. Issues and assumptions on the road from raw signals to metrics of frontal eeg asymmetry in emotion. *Biological psychology*, 67(1-2):183–218, 2004.

[4] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer, 2007.

[5] Patricia M Arenth, Joseph H Ricker, and Maria T Schultheis. Applications of functional near-infrared spectroscopy (fnirs) to neurorehabilitation of cognitive disabilities. *The Clinical Neuropsychologist*, 21(1):38–57, 2007.

[6] Hasan Ayaz, Meltem Izzetoglu, Kurtulus Izzetoglu, and Banu Onaral. The use of functional near-infrared spectroscopy in neuroergonomics. In *Neuroergonomics*, pages 17–25. Elsevier, 2019.

[7] Michela Balconi, Elisabetta Grippa, and Maria Elide Vanutelli. What hemodynamic (fnirs), electrophysiological (eeg) and autonomic integrated measures can tell us about emotional processing. *Brain and cognition*, 95:67–76, 2015.

[8] Danushka Bandara, Leanne Hirshfield, and Senem Velipasalar. Classification of affect using deep learning on brain blood flow data. *Journal of Near Infrared Spectroscopy*, 27(3):206–219, 2019.

[9] Danushka Bandara, Senem Velipasalar, Sarah Bratt, and Leanne Hirshfield. Building predictive models of emotion with functional near-infrared spectroscopy. *International Journal of Human-Computer Studies*, 110:75–85, 2018.

[10] Oswald Barral, Ilkka Kosunen, Tuukka Ruotsalo, Michiel M Spapé, Manuel JA Eugster, Niklas Ravaja, Samuel Kaski, and Giulio Jacucci. Extracting relevance and affect information from physiological text annotation. *User Modeling and User-Adapted Interaction*, 26(5):493–520, 2016.

[11] Daren C. Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1):75–90, 2008.

[12] Scott C Bunce, Meltem Izzetoglu, Kurtulus Izzetoglu, Banu Onaral, and Kambiz Pourrezaei. Functional near-infrared spectroscopy. *IEEE engineering in medicine and biology magazine*, 25(4):54–62, 2006.

[13] Walter B Cannon. The james-lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology*, 39(1/4):106–124, 1927.

[14] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2334–2346, 2017.

[15] Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, and Amy Warner. Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing*, 17(4):32–38, 2012.

[16] Xu Cui, Signe Bray, and Allan L Reiss. Functional near infrared spectroscopy (nirs) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *Neuroimage*, 49(4):3039–3046, 2010.

[17] Antonio R Damasio. Emotions and feelings. In *Feelings and emotions: The Amsterdam symposium*, volume 5, pages 49–57. Cambridge University Press Cambridge, 2004.

[18] Richard J Davidson. What does the prefrontal cortex "do" in affect: perspectives on frontal eeg asymmetry research. *Biological psychology*, 67(1-2):219–234, 2004.

[19] Keith M Davis, Carlos de la Torre-Ortiz, and Tuukka Ruotsalo. Brain-supervised image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18480–18489, 2022.

[20] Keith M. Davis, Lauri Kangassalo, Michiel Spapé, and Tuukka Ruotsalo. Brainsourcing: Crowdsourcing recognition tasks via collaborative brain-computer interfacing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[21] Keith M. Davis, Michiel Spapé, and Tuukka Ruotsalo. Contradicted by the brain: Predicting individual and group preferences via brain-computer interfacing. *IEEE Transactions on Affective Computing*, pages 1–12, 2022.

[22] Keith M Davis III, Michiel Spapé, and Tuukka Ruotsalo. Collaborative filtering with preferences inferred from brain signals. In *Proceedings of the Web Conference 2021*, pages 602–611, 2021.

[23] David T Delpy, Mark Cope, Pieter van der Zee, Simon Arridge, Susan Wray, and JS Wyatt. Estimation of optical pathlength through tissue from direct time of flight measurement. *Physics in Medicine & Biology*, 33(12):1433, 1988.

[24] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

[25] Daantje Derks, Arjan ER Bos, and Jasper Von Grumbkow. Emoticons in computer-mediated communication: Social motives and social context. *Cyberpsychology & behavior*, 11(1):99–101, 2008.

[26] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005.

[27] Xiao Pan Ding, Xiaoqing Gao, Genyue Fu, and Kang Lee. Neural correlates of spontaneous deception: A functional near-infrared spectroscopy (fnirs) study. *Neuropsychologia*, 51(4):704–712, 2013.

[28] Yue Ding, Xin Hu, Zhenyi Xia, Yong-Jin Liu, and Dan Zhang. Inter-brain eeg feature extraction and analysis for continuous implicit emotion tagging during video watching. *IEEE Transactions on Affective Computing*, 12(1):92–102, 2018.

[29] Otniel E Dror. The cannon–bard thalamic theory of emotions: A brief genealogy and reappraisal. *Emotion Review*, 6(1):13–20, 2014.

[30] P Ekman. Are there basic emotions? *Psychological Review*, 99:550–553, 1992.

[31] Vyvyan Evans. *The emoji code: How smiley faces, love hearts and thumbs up are changing the way we communicate*. Michael O'Mara Books, 2017.

[32] Michael W Eysenck. Arousal, learning, and memory. *Psychological bulletin*, 83(3):389, 1976.

[33] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.

[34] Frank A Fishburn, Ruth S Ludlum, Chandan J Vaidya, and Andrei V Medvedev. Temporal derivative distribution repair (tddr): a motion correction method for fnirs. *Neuroimage*, 184:171–179, 2019.

[35] Nickolaos Fragopanagos and John G Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405, 2005.

[36] Antje Gerdes, Matthias J Wieser, Andreas Mühlberger, Peter Weyers, Georg W Alpers, Michael M Plichta, Felix Breuer, and Paul Pauli. Brain activations to emotional pictures are differentially associated with valence and arousal ratings. *Frontiers in human neuroscience*, 4:175, 2010.

[37] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, page 267, 2013.

[38] Beate M Herbert, Olga Pollatos, and Rainer Schandry. Interoceptive sensitivity and emotion processing: an eeg study. *International Journal of Psychophysiology*, 65(3):214–227, 2007.

[39] Jack van Honk and Dennis JLG Schutter. From affective valence to motivational direction: the frontal asymmetry of emotion revised. *Psychological Science*, 17(11):963–965, 2006.

[40] Xin Hu, Chu Zhuang, Fei Wang, Yong-Jin Liu, Chang-Hwan Im, and Dan Zhang. fnirs evidence for recognizably different positive emotions. *Frontiers in human neuroscience*, 13:120, 2019.

[41] Takayuki Iwano and Shinji Umeyama. Estimation of crosstalk in led fnirs by photon propagation monte carlo simulation. In *Biophotonics Japan 2015*, volume 9792, pages 110–115. SPIE, 2015.

[42] Lauri Kangassalo, Michiel Spapé, and Tuukka Ruotsalo. Neuroadaptive modelling for generating images matching perceptual categories. *Scientific reports*, 10(1):14719, 2020.

[43] Martin Klasen, Yu-Han Chen, and Klaus Mathiak. Multisensory emotions: perception, combination and underlying neural processes. *Reviews in the Neurosciences*, 23(4):381–392, 2012.

[44] Peter Lang and Margaret M Bradley. The international affective picture system (iaps) in the study of emotion and attention. *Handbook of emotion elicitation and assessment*, 29:70–73, 2007.

[45] Fanny Larradet, Radoslaw Niewiadomski, Giacinto Barresi, Darwin G Caldwell, and Leonardo S Mattos. Toward emotion recognition from physiological signals in the wild: approaching the methodological issues in real-life data collection. *Frontiers in psychology*, 11:1111, 2020.

[46] Olivier Ledoit and Michael Wolf. Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.

[47] Horia A. Maior, Matthew Pike, Sarah Sharples, and Max L. Wilson. Examining the reliability of using fnirs in realistic hci settings for spatial and verbal tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3039–3042, New York, NY, USA, 2015. Association for Computing Machinery.

[48] R Morris, Daniel McDuff, and R Calvo. Crowdsourcing techniques for affective computing. In *The Oxford handbook of affective computing*, pages 384–394. Oxford Univ. Press Oxford, UK, 2014.

[49] Georg Northoff. Are our emotional feelings relational? a neurophilosophical investigation of the james–lange theory. *Phenomenology and the Cognitive Sciences*, 7(4):501–527, 2008.

[50] Jaak Panksepp. At the interface of the affective, behavioral, and cognitive neurosciences: Decoding the emotional feelings of the brain. *Brain and cognition*, 52(1):4–14, 2003.

[51] Evan M M. Peck, Beste F. Yuksel, Alvitta Ottley, Robert J.K. Jacob, and Remco Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 473–482, New York, NY, USA, 2013. Association for Computing Machinery.

[52] K Luan Phan, Tor Wager, Stephan F Taylor, and Israel Liberzon. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in pet and fmri. *Neuroimage*, 16(2):331–348, 2002.

[53] Paola Pinti, Ilias Tachtsidis, Antonia Hamilton, Joy Hirsch, Clarisse Aichelburg, Sam Gilbert, and Paul W Burgess. The present and future use of functional near-infrared spectroscopy (fnirs) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1464(1):5–29, 2020.

[54] Luca Pollonini, Cristen Olds, Homer Abaya, Heather Bortfeld, Michael S Beauchamp, and John S Oghalai. Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy. *Hearing research*, 309:84–93, 2014.

[55] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010.

[56] Rainer Reisenzein. The schachter theory of emotion: two decades later. *Psychological bulletin*, 94(2):239, 1983.

[57] James A Russell. Culture and the categorization of emotions. *Psychological bulletin*, 110(3):426, 1991.

[58] Stanley Schachter and Jerome Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5):379, 1962.

[59] Harold Schlosberg. Three dimensions of emotion. *Psychological review*, 61(2):81, 1954.

[60] Marc Schröder, Laurence Devillers, Kostas Karpouzis, Jean-Claude Martin, Catherine Pelachaud, Christian Peter, Hannes Pirker, Björn Schuller, Jianhua Tao, and Ian Wilson. What should a generic emotion markup language be able to represent? In *International Conference on Affective Computing and Intelligent Interaction*, pages 440–451. Springer, 2007.

[61] Harald Schupp, Bruce Cuthbert, Margaret Bradley, Charles Hillman, Alfons Hamm, and Peter Lang. Brain processes in emotional perception: Motivated attention. *Cognition and emotion*, 18(5):593–611, 2004.

[62] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622, New York, NY, USA, 2008. Association for Computing Machinery.

[63] Pradeep Shenoy and Desney S. Tan. Human-aided computing: Utilizing implicit human processing to classify images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 845–854, New York, NY, USA, 2008. Association for Computing Machinery.

[64] Matthias Siemer. Mood-congruent cognitions constitute mood experience. *Emotion*, 5(3):296, 2005.

[65] Surjo R Soekadar, Mareike Vermehren, Annalisa Colucci, David Haslacher, Christoph Bublitz, Marcello Ienca, Jennifer A Chandler, and Benjamin Blankertz. Future developments in brain/neural–computer interface technology. In *Policy, Identity, and Neurotechnology: The Neuroethics of Brain-Computer Interfaces*, pages 65–85. Springer, 2023.

[66] Erin Solovey, Paul Schermerhorn, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, and Robert Jacob. Brainput: Enhancing interactive systems with streaming fnirs brain input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 2193–2202, New York, NY, USA, 2012. Association for Computing Machinery.

[67] Erin Treacy Solovey, Audrey Girouard, Krysta Chauncey, Leanne M. Hirshfield, Angelo Sassaroli, Feng Zheng, Sergio Fantini, and Robert J.K. Jacob. Using fnirs brain sensing in realistic hci settings: Experiments and guidelines. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, page 157–166, New York, NY, USA, 2009. Association for Computing Machinery.

[68] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, 2008.

[69] Erin Treacy Solovey, Daniel Afergan, Evan M. Peck, Samuel W. Hincks, and Robert J. K. Jacob. Designing implicit interfaces for physiological computing: Guidelines and lessons learned using fnirs. *ACM Trans. Comput.-Hum. Interact.*, 21(6), jan 2015.

[70] Hamidan Z Wijasena, Ridi Ferdiana, and Sunu Wibirama. A survey of emotion recognition using physiological signal in wearable devices. In *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, pages 1–6. IEEE, 2021.

[71] Frank H Wilhelm and Paul Grossman. Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. *Biological psychology*, 84(3):552–569, 2010.

[72] Hiroo Yamamura, Holger Baldauf, and Kai Kunze. Pleasant locomotion – towards reducing cybersickness using fnirs during walking events in vr. In *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20 Adjunct, page 56–58, New York, NY, USA, 2020. Association for Computing Machinery.

[73] Man-Ching Yuen, Irwin King, and Kwong Leung. A survey of crowdsourcing systems. pages 766–773, 10 2011.

[74] Jing Zhang, Xindong Wu, and Victor S Sheng. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46(4):543–576, 2016.

Tuukka Ruotsalo is an Associate Professor at the University of Copenhagen, Denmark and holds an Academy Research Fellowship at the University of Helsinki, Finland. He works on human-computer interaction, machine learning, and cognitive computing.

Kalle Mäkelä is a research assistant at the University of Helsinki, Finland. He works on machine learning for physiological signal analysis.

Michiel Spapé is a docent in cognitive neuroscience at the University of Helsinki, Finland. He is focusing on emotion, perception/action, and EEG.

APPENDIX A
CROWDSOURCING DECISION STRATEGIES

In addition to combining subjects' predictions via soft voting, the crowdsourced models were evaluated with two other decision strategies, hard voting and average response classification.

In hard voting, each individual classifier votes for one class, and the class assigned with the majority of the votes is considered the crowdsourced prediction. In case of a tie, the first class is selected (the classes are ordered: LANV, HANV, LAPV, HAPV). The mean accuracies and F1 scores with hard voting are shown in Table 3.

Average response classification works by averaging the fNIRS responses from N subjects for each image before feature extraction and classification. Then, each image's average response is classified with the same cross-validation scheme as in Section 4.2. The mean accuracies and F1 scores of average response classification are shown in Table 4.

Task	**N=1**		**N=2**		**N=4**		**N=8**	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
4 class	0.40	0.39	0.41	0.39	0.45	0.44	0.48	0.48
Valence	0.59	0.58	0.60	0.57	0.62	0.60	0.64	0.63
Arousal	0.56	0.56	0.57	0.55	0.59	0.58	0.59	0.59
HA Valence	0.67	0.67	0.69	0.67	0.74	0.73	0.78	0.78
LA Valence	0.57	0.57	0.59	0.57	0.61	0.59	0.64	0.63

TABLE 3: Accuracy and F1 scores for different N for each task with hard voting. The datasets are nearly balanced for all prediction tasks.

Task	**N=1**		**N=2**		**N=4**		**N=8**	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
4 class	0.26	0.26	0.27	0.27	0.28	0.28	0.30	0.29
Valence	0.52	0.52	0.55	0.55	0.57	0.57	0.59	0.58
Arousal	0.50	0.50	0.50	0.49	0.50	0.50	0.51	0.51
HA Valence	0.54	0.54	0.57	0.56	0.60	0.60	0.63	0.63
LA Valence	0.53	0.53	0.53	0.53	0.53	0.53	0.55	0.55

TABLE 4: Accuracy and F1 scores for different N for each task with average response classification. The datasets are nearly balanced for all prediction tasks.