Brainsourcing: Crowdsourcing Recognition Tasks via Collaborative Brain-Computer Interfacing

Keith M. Davis III[1], **Lauri Kangassalo**[1], **Michiel Spapé**[2], **Tuukka Ruotsalo**[1]
[1]Department of Computer Science, [2]Department of Psychology and Logopedics
University of Helsinki
Helsinki, Finland
first.last@helsinki.fi

ABSTRACT

This paper introduces brainsourcing: utilizing brain responses of a group of human contributors each performing a recognition task to determine classes of stimuli. We investigate to what extent it is possible to infer reliable class labels using data collected utilizing electroencephalography (EEG) from participants given a set of common stimuli. An experiment (N=30) measuring EEG responses to visual features of faces (gender, hair color, age, smile) revealed an improved F1 score of 0.94 for a crowd of twelve participants compared to an F1 score of 0.67 derived from individual participants and a random chance of 0.50. Our results demonstrate the methodological and pragmatic feasibility of brainsourcing in labeling tasks and opens avenues for more general applications using brain-computer interfacing in a crowdsourced setting.

Author Keywords

Crowdsourcing; Brainsourcing; Brain-computer interfaces

CCS Concepts

•Human-centered computing → Human computer interaction (HCI); •Information systems → Crowdsourcing;

INTRODUCTION

Many tasks that are trivial for humans continue to challenge computer programs. For example, image annotation has turned out to be dependent on human labeling and even the most sophisticated machine learning systems need human supervision in order to recognize objects appearing in images [71]. These tasks are difficult to fully automatize, but humans and computing systems can together effectively solve these tasks by distributing the work to several individuals - a paradigm often called crowdsourcing [34]. Crowdsourcing enables human workers to perform designated tasks over networked computing systems unrestricted by time and location [75]. Crowdsourcing has become popular for a variety of tasks ranging from microtasks, such as simple labeling, to complex macrotasks for design and knowledge work. However, while crowdsourcing has allowed communities to collaborate to solve complex tasks, many succesful applications of crowdsourcing are still based on combining results of repetitive microtasks that typically require users to recognize whether a certain stimulus meets a certain task criteria. Such tasks include, for example, detecting whether an image contains a certain object or whether a piece of text is grammatically correct. The contributions of individuals are then collected to solve large-scale tasks more reliably and efficiently [13, 40].

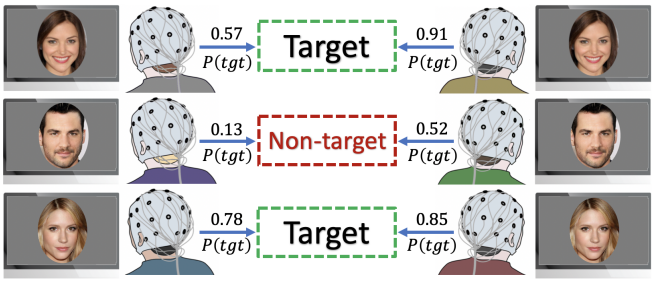

Figure 1: Brainsourcing utilizes brain responses of a group of human contributors each performing a recognition task to determine the consensus label of a stimulus.

An important factor in crowdsourcing microtasks is the economics and effort of collecting crowd input. Making explicit selections via a user interface can be tedious, physically and mentally demanding, and cause crowd workers to exhibit fatigue [60]. Implicit crowdsourcing can avoid this problem by observing natural human behavior, but it often requires very large user populations to properly divide the work, and is limited to tasks for which implicit behavior can be observed. Thus, it is often not possible to acquire user input that is low-cost, high-quality, and convenient for the users [3, 35, 50].

Here, we investigate a novel paradigm collecting implicit user input in response to either an explicit or implicit task in a crowdsourcing setting: brainsourcing. Brainsourcing, as conceptually illustrated in Figure 1, allows direct mapping of a group of users' implicit, neurally measured reactions within recognition tasks to predict when targets of interest appear.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/10.1145/3313831.3376288

Instead of relying on manual input, brainsourcing utilizes the reactions of individuals measured directly from their brain signals via EEG. Such signals enable explicit and implicit microtasks to be distributed without the users committing to any explicit interaction with the system. To put it simply, the users only need to visually perceive the content and react to it.

To this end, we study whether brainsourcing is possible and whether it has pragmatic value in real-world labeling tasks where humans identify objects or features of objects appearing in images. In detail, we ask the following research questions:

RQ1: Can human brain responses be utilized in a crowdsourcing setting to predict a class label for a given stimulus, so that the crowd performance is better than the performance of an individual user?

RQ2: How many participants are enough to reliably make labeling decisions and at what point does adding additional participants fail to significantly improve results?

In order to answer the research questions, we report an experiment (N=30) where brainsourcing is utilized for image labeling. In the experiment, the participants complete recognition tasks, in which the participants are presented with images of faces that have an obvious distinguishing feature, such as blond hair color. Subjects are instructed to mentally note whenever they see a face containing the distinguishing feature while their brain signals are measured via EEG. This technique is conceptually similar to brain-computer interfaces, in which an individual's brain activity is measured in order to control a computer interface. A classic example of brain-computer interfacing are the speller applications that enable people with locked-in syndrome to spell out words [44, 27]. These often work by repeating the same stimuli to draw decisions on an individual user's attention to a particular letter. Brainsourcing, however, takes a collaborative approach and uses signals of a *crowd of users* to reduce noise and increase the reliability of the classification output.

We focus on predicting a consensus label; to identify a commonly determined label by collecting multiple assessments from human workers. Each recognition task presents participants with stimuli representing the target class and non-target class. For example, in the 'blond' task, participants are presented with images of blond (target) and dark-haired (non-target) persons. The goal of the brainsourcing is to separate these classes using a consensus label from the signals of many participants. The inferred consensus labels are then evaluated against ground truth labels.

In summary, our contributions are the following:

1. We present brainsourcing; a novel crowdsourcing paradigm that utilizes brain responses of a group of human contributors each performing recognition tasks to determine classes of stimuli.

2. We report an experiment utilizing brainsourcing for image labeling tasks showing significant performance gains compared to a performance of individual participants.

3. Our results demonstrate the methodological and pragmatic feasibility of brainsourcing; a nearly perfect performance for simple, yet well-defined recognition tasks.

BACKGROUND

Our work is based on two distinct areas: crowdsourcing microtasks and brain-computer interfacing, which are shortly reviewed below.

Crowdsourcing

Crowdsourcing refers to a distributed problem-solving and production model in which human work perform designated tasks over networked computing systems unbounded by time and location [11, 75]. These tasks are difficult to fully automatize, but they can be solved together by humans and computer systems by dividing the work to several individuals [33, 34]. A complex problem is split into smaller parts, which can then be solved by a host of contributors. Crowdsourcing can provide complete or partial solutions to many non-trivial problems, including tasks that range from purely routine and cognitively simple to complicated and creative tasks [24]. Independent of the actual task, a generic crowdsourcing task must be divisible into lower level tasks, each one of which can be accomplished by individual members of the crowd.

Crowdsourcing can be roughly divided into macrowork tasks and microwork tasks [17]. While macrowork typically requires special skills and normally takes longer (such as open source development of software), microwork allows humans to do small tasks that are difficult for computers, but require no specialized skills from the humans (such as labeling videos). Microwork can further be divided into implicit and explicit tasks [22] based on whether the tasks are explicitly defined for the workers or the workers are performing tasks implicitly without explicit awareness that they are conducting microtasks. Most crowdsourcing tasks are explicit. Examples of explicit microtasks are image and video labeling [43] and approximating regions of images that draw visual attention [42]. Similar explicit microtasks have also been used for recognition tasks that have a specific, fine-grained categorization goal [21].

Rather than users actively participating in solving a problem or providing information, implicit crowdsourcing involves users doing a primary task from which the system can gain information for another task based on the user's actions. Implicit crowdsourcing has been used, for example, in estimating preferences in recommender systems [47] and relevance assessments from search engine usage [46, 16]. A classic example of implicit crowdsourcing is the ESP game, where users guess what images are and then these labels are used to tag Google images [71]. Another example of implicit crowdsourcing is through reCAPTCHA, which asks people to solve simple recognition tasks to prove they are human, and then provides texts from old books that cannot be deciphered by computers as recognition tasks, in order to digitize them for the web [72]. As implicit crowdsourcing allows a computing system to infer useful information from a crowd of users simply by observing their interactions with the system, the crowd is not *a priori* requested to perform a particular task, but their behavior is mined to distill useful information.

Brain-Computer Interfacing

Brain-computer interfaces often employ electroencephalography (EEG) to measure the brain activity of users while they perform tasks. EEG uses electrodes to measure voltage fluctuations from the scalp in a non-invasive way. EEG provides high temporal resolution, meaning that the changes in brain activity are recorded in the millisecond range [31]. Due to this high temporal precision, it is possible to present stimuli to participants in a rapid succession, measuring the brain activity associated with each stimulus using the event-related potential technique, where slices of EEG signal time-locked to a set of stimuli, such as images or sounds, are segmented and analyzed as single waveforms. Such segments are also known as "epochs", while measured changes in electrical activity are known as event-related potentials (ERPs) [49]. The temporal precision of EEG allows dissociation between various mental processes, allowing the detection of cognitive operations. Relevance detection, such as in the context of categorical relevance [32], can be discerned in the ERP as reflected by the P300 component, a parietal positivity that appears around 300ms after a task-relevant stimulus is presented to the subject [26, 38, 25]. Psychophysiological theory suggest the process necessarily follows perceptual processing and attentional filtering, suggesting the mental operation quantified by the P300 is either part of a process relating attention to memory or is directly related to storing in memory [58, 23].

As the P300 component is particularly reliable whenever infrequent target stimuli appear within a sequence of stimuli, Brain-Computer Interfaces (BCIs) repurposed older findings from psychophysiology [68] to detect mental relevance so as to control devices using the mind [27]. This was achieved by adapting the *'oddball'* paradigm, in which users selectively focused on task-relevant, 'odd' stimuli amongst normal. The paradigm remains a widely used method for a variety of BCI applications [2, 7, 1, 37]. In BCI applications, a method called single-trial classification is often used to harness the ERPs to the use of a computer system [8]. Single-trial classification means simply the computerized classification of each of the ERPs evoked by stimuli.

Our Approach

Our approach differs from the previous crowdsourcing work as we use purely implicit user signals measured directly from the human brain. Previous research using EEG signals to classify stimuli suffers from a variety of limitations. Most of this work does not attempt to combine results from multiple individuals in any meaningful way [8, 12]. Previous attempts to integrate EEG signals from multiple users may rely upon features from the stimuli to improve performance, or use stimuli that are from distinct meta-categories (i.e. distinguishing between images of faces and images of inanimate objects) [39, 62]. We conduct an experiment where participants each complete a set of visual recognition tasks and are connected to the system via BCI. We use robust BCI methods relying on the oddball paradigm and single-trial classification of ERPs to classify human reactions toward stimuli. We then implement collaborative decision-making in the form of brainsourcing. Brainsourcing is used to determine crowd consensus labels for images by combining single-trial predictions from multiple individuals. While our tasks are explicit crowdsourcing tasks, brainsourcing does not require the participants to perform any physical interaction with the computing system; they only perceive the stimuli and are asked to make a mental note when the stimuli represents a target class.

In the following sections we describe the neurophysiological experiment to collect EEG data, BCI methodology to build predictive models of individuals, and brainsourcing experiments. We then report the results and discuss their impact, limitations and future avenues enabled by brainsourcing.

NEUROPHYSIOLOGICAL EXPERIMENT

Brainsourcing relies on the neural signals measured from participants in response to a recognition task. The corresponding neurophysiological experiment, illustrated in the upper part of Figure 2, is described in this section.

Participants

Thirty-one volunteers, 13 female and 18 male, were recruited from the University of Helsinki. They self-reported as being healthy with regard to neuropathological history. There were 29 right-handed participants and 2 left-handed participants. Prior to the study they were fully briefed as to the nature and purpose of the study, signed informed consent in accordance with the Declaration of Helsinki, and were instructed on their rights as participants, including the right to withdraw from the experiment at any time without fear of negative consequences. One participant indeed withdrew from the study early, which prevented data from being collected for a single task (old), resulting in full data from 30 participants. However, the remaining data for this participant were still valid and thus used to perform the analyses detailed in this paper. The study was approved by the Ethical Board of the University of Helsinki. The participants received ticket vouchers to the local cinema as a compensation for their participation.

Stimuli

An important aspect of the experiment was to use stimuli that would be relatively easy for the participants to classify as a target or non-target. It was also required that it would be straightforward for the participants to recognize salient and non-salient features, but not recognize the individual object so that the judgment would be based on the targets class instead of confouding individual features of the stimuli. In order to avoid these biases, we decided to use images of artificially created faces as stimuli. This allowed us to use homogeneous stimuli (human faces) that didn't represent any known individual (artificial faces) and had easily recognizable features (e.g. hair color, gender, age). Stimuli were generated using a GAN architecture trained on a large dataset of celebrity faces, sampled via a random process from 70,000 latent vectors from a 512-dimensional multivariate normal distribution [41]. Each generated face was manually screened by a human assesor and placed into a distinct category corresponding to one of the eight recognition tasks, as presented in Table 1. In total, 1961 unique images were used in the experiment.

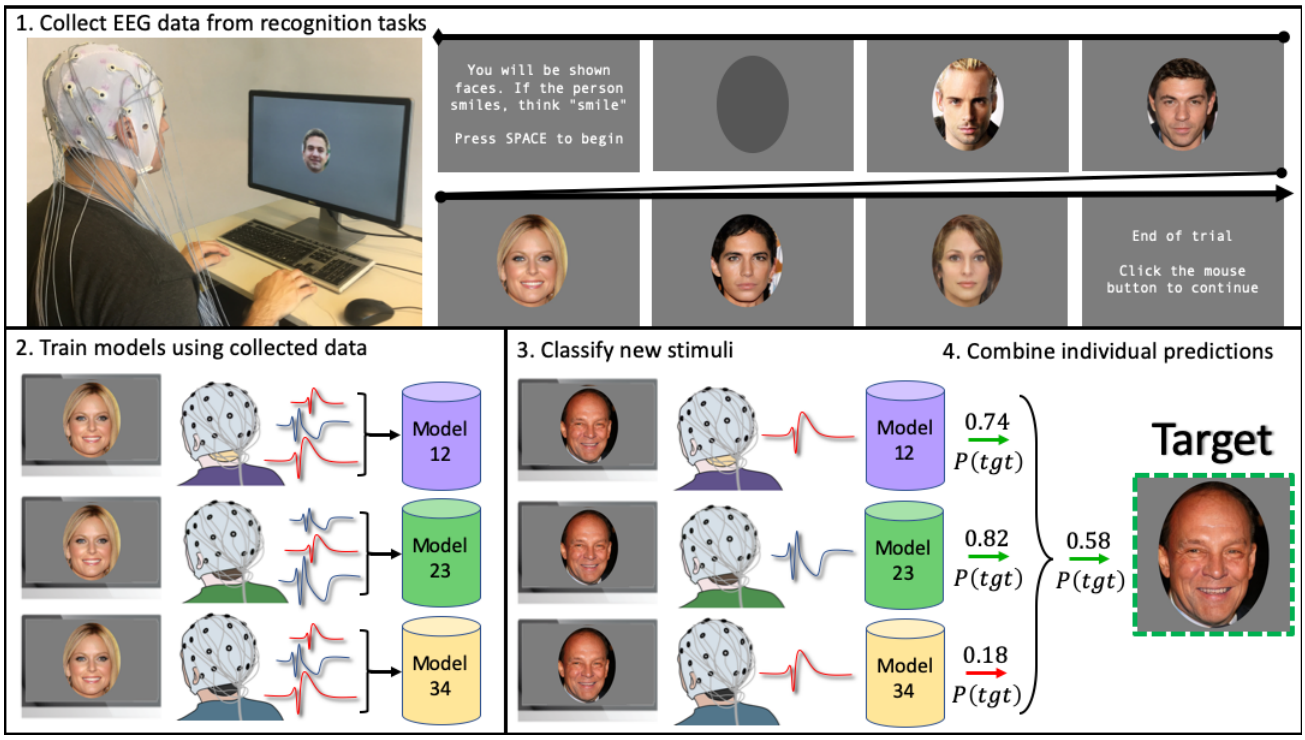

Figure 2: Diagram of brainsourcing steps. 1. An example of a recognition task (in this case, task "smile"), with preparatory prompt, masking image, sample stimuli, and ending prompt. Data is collected in one minute segments, where the subject is shown approximately 100 stimuli spaced 500ms apart. 2. Classifiers are trained individually for each subject. 3. EEG data from new stimuli are classified using these models. 4. Predictions from separate models are combined to produce a brainsourced estimation of class probability, which is used to determine the consensus label of the new stimuli.

Apparatus

The LCD display used to present stimuli was positioned at approximately 60 cm from the participants, running at 60 hz with a resolution of 1680 by 1050 pixels. Psychology Software Tools E-Prime 3.0.3.60 stimulus presentation software was used to optimise timing of display and EEG amplifier trigger control [67]. EEG was recorded from 32 Ag/AgCl electrodes, positioned on Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, TP9, CP5, CP1, CP2, CP6, TP10, P7, P3, Pz, P4, P8, PO9, PO10, O1, O2, and Iz using EasyCap elastic caps (EasyCap GmbH, Herrsching, Germany). Hardware amplification, filtering and digitization was done via a QuickAmp USB (BrainProducts GmbH, Gilching, Germany) amplifier running at 2000 Hz. Two pairs of bipolar electrodes, situated 1 cm lateral to the outer canthi of the left and right eye, and 2 cm inferior and superior to the right pupil, were used to detect eye movements.

Tasks

Participants were presented with eight recognition tasks, one task for each category in the dataset (see Table 1, left column). All stimuli presented during each task were either labeled as the target or its inverse (non-target). For example, for the task 'blond', participants were shown faces that either had blond hair or dark hair, and were instructed to make a mental note when they saw a face with blond hair. Twenty stimuli of the target class and fifty stimuli of the non-target class were shown during each iteration of the recognition task. In order to ensure that the participants understood the task, each recognition task was preceded by a demonstration task, where the participants were shown four example stimuli images and they were asked to manually select the target images. These images were not used as stimuli in the actual task. To ensure enough data was collected for each participant, the recognition task and demonstration task were repeated a total of four times for each image category, for a grand total of 32 iterations.

BCI METHODOLOGY

The EEG data of each participant was preprocessed and classified according to robust and well-known BCI methodologies [8]. Here, we detail the steps taken to increase the signal-to-noise ratio of the brain signals collected from individual participants, and define the classification model used to predict the target stimuli based on EEG data.

EEG Preprocessing

Due to the sensitivity of the measuring equipment, EEG measurements are susceptible to various kinds of signal noise, such as those caused by movements of the participant and electrical equipment. To maximise the signal-to-noise ratio of the data, standard signal cleaning procedures were employed [49]. Furthermore, the preprocessing pipeline was designed to reduce

the signal noise in an on-line application, i.e. only simple, automated operations were used for signal cleaning. First, a band-pass filter was employed at the frequency range 0.2-35 Hz, which effectively removed slow signal fluctuations (such as those caused by respiration) and high-frequency noise (such as power line noise) from the data. After filtering, the data were split to time-locked epochs ranging from -200 to 900 ms relative to stimulus onset. Each epoch was baseline-corrected with its pre-stimulus period (-200-0ms). A threshold-based heuristic was used to remove transient artefacts from the data, such as those caused by eye blinks; calibration periods consisting of the first 2000 epochs of each participant were used to determine a maximum per-participant voltage threshold, which in turn was used to identify contaminated epochs. This threshold was set at the 90th percentile in the distribution of epochs' maximum voltages in the calibration period, and capped to be at least $10\mu V$ and at most $80\mu V$. Data with epochs where the maximum voltage exceeded the threshold were removed, which led to the removal of approximately 11% of each participants' epochs. The final dataset consisted of on average 3251 epochs per participant.

Classification

A regularized Linear Discriminant Analysis (LDA) classifier was trained for each of the participants, for each task, using the procedure described in [8]. The classifier was trained with vectorized representations of the ERPs along with a binary label indicating class membership. The label indicated whether the ERP was associated with a target stimulus or not. For example, in the case of the 'blond' task, where the participant was asked to take a mental note of blond persons, the label would indicate whether an ERP was associated with a blond person or not.

The vectorized ERP representations consisted of spatio-temporal features. The vector representation utilized time points in the 50 - 800ms post-stimuli time range and all of the available channels. This produced a data tensor $X^{n\times c\times t}$ with n epochs, c = 32 channels, and t = 125 time points. To reduce the data dimensionality and speed up the training procedure, the time points were split to t' = 15 equidistant time windows for which voltage averages were computed. Finally, the spatial and temporal dimensions of the tensor were concatenated resulting in a data matrix $X^{n\times ct'}$.

Each classifier predicted the class probabilities for one task and used the other tasks as training data. Since there was overlap in the stimuli used for a task and its inverse, the task's inverse was omitted from the training dataset. For example, when training a classifier for the blond task, data from all other tasks except blond and dark-haired were used. By leaving out the inverse task, we ensure that the classifier is only predicting class probabilities for ERPs of unseen stimuli. This technique also allowed us to assign class probabilities for all stimuli presented to each subject that remained after the pre-processing step, with an average of 250 stimuli per participant, per task.

Predicted class probabilites were converted to binary class labels using a simple threshold technique. For each classifier, the mean class probability for all of its predicted outputs was computed. Class probabilities that were greater than the mean indicated a target, while those that were less than the mean indicated a non-target. These binary labels were then used to assess classifier performance, where ground truth labels for the stimuli were compared to the predicted class labels.

Classifier Evaluation

The LDA classifier performance was measured using F1 score, which was compared to F1 scores for classifiers trained with randomly permutated class labels. With a sufficient amount of permutations, this leads to permutation-based p-values [55]. $k = 100$ permutations were run per participant, which led to a minimum possible p-value of 0.01 [29].

BRAINSOURCING EXPERIMENT

The brainsourcing methodology, depicted graphically in the lower part of Figure 2, consists of taking class predictions produced by the individual classification models and averaging them to produce a crowd consensus label. The brainsourcing methodology is described in detail below.

Experimental Setup

Predictions are based on the estimated class probabilities produced by the individually trained LDA classifiers. These class probabilities are stored in a probability matrix $A^{X\times Y}$ for each task, where X denotes the participants and Y denotes the stimuli. Finally, we use crowd decisions to infer a label for a given stimulus. We select estimations for a given stimulus x from N randomly chosen participants in Y, and then take the mean of these estimations to produce the crowdsourced estimation.

Brainsourcing Model

From the probability matrix A we created 100 unique datasets, by drawing submatrices of A with all participants and a selection of stimuli. The datasets were randomly downsampled in a manner which ensured each dataset was distinct and contained predictions for an equal number of target and non-target stimuli.

For the brainsourcing step, we randomly selected one of the downsampled datasets. Next, a stimulus column was chosen at random. From this, N random datapoints were sampled, each containing a class prediction for the selected stimulus. The mean of these individual predictions was computed and stored as the consensus probability. The selected stimulus column was then dropped from the downsampled dataset, and the sampling procedure was repeated until 250 brainsourced estimations were produced. To determine the class labels, the same thresholding procedure used to determine binary labels for individual classification models was also used. The global mean for the entire iteration was computed, and consensus probabilities greater than the mean indicated a target, while those falling below the mean were indicated a non-target. After computing the binary labels, the brainsourcing procedure was repeated, until a total of 100 iterations and 250 brainsourcing estimations per iteration were produced. The results of each iteration were compared with the ground truth label to evaluate performance of brainsourcing.

The brainsourcing steps were conducted in this manner to simulate many unique cohorts of individuals collaborating to

produce crowd consensus labels for a set of stimuli. Additionally, this technique further reduces the likelihood of introducing bias in our results (such as by selecting individuals who consistently perform well on the individual classification tasks).

Selection of Performance Measures

In the experiment, the goal of brainsourcing was to produce accurate binary labels (target / non-target) for the set of stimuli as measured against ground truth labels.

We selected precision, recall, and F1 score to quantify the performance of brainsourcing. These widely used [65, 59, 6] information retrieval measures were selected for several reasons. First, we are approaching the classification problem presented in this experiment as an information retrieval task, where we prioritize good classification performance for target (positive) classes. Intuitively, it is less harmful to miss a target than to produce a false positive target. For example, given a task to recognize females, we value performance on correctly recognized females higher than missing one female target, as repeating the task on a large crowd could be used to correct for missed targets, but not necessarily to correct false positives. Second, the EEG data used to train individual LDA classifiers is unbalanced, containing approximately 30% target stimuli and 70% non-target stimuli. Given the unbalanced nature of the data, using a simple measure of accuracy to asses model performance would produce misleading results; a model that predicts everything to be non-target stimuli would yield an accuracy of around 70%. Due to how the F1 score relies upon precision and recall to assess performance (as opposite to accuracy), it is a suitable method for assessing the performance of models trained on unbalanced data [6]. While the dataset used for the brainsourcing model predictions was artifically balanced using downsampling, we still wanted to quantify the precision / recall tradeoff, which the F1 score captures.

RESULTS

Here we present briefly the neurophysiological findings and the results for the individual classification models, and detail the results of the actual brainsourcing experiment.

Neurophysiological Results

ERPs for target and non-target stimuli were averaged across all participants and analyzed by channel. Large differences between grand average scalp voltages between target and non-target classes were oberved over Cp1, Cp2, Cz, P3, P4, and Pz sites with the largest difference in average found for the Pz electrode. This difference, and grand average scalp voltages represented as a cranial topographical plot, are provided in Figure 3. Target images are associated with a strong scalp positivity, beginning at around 200 ms, peaking at ca. 280 ms, and continuing until approximately 600 ms. This suggests target detections in general amplified the P300 component to images, conforming to the known psychophysiology literature [4, 32].

Per-participant Classification

To ensure that per-participant classification found meaningful structure in the data discriminating ERPs between target/non-target stimuli, the per-participant classifiers were tested against a random baseline, computed using label permutation [55]. F1 scores for single participants averaged at 0.67 across all tasks, with task “old” achieving the lowest F1 score at 0.63, and task “female” achieving the highest at 0.74. The classifiers of all of the participants performed better than the random baseline ($p < 0.05$), indicating that the classifiers had learned meaningful structure from the data.

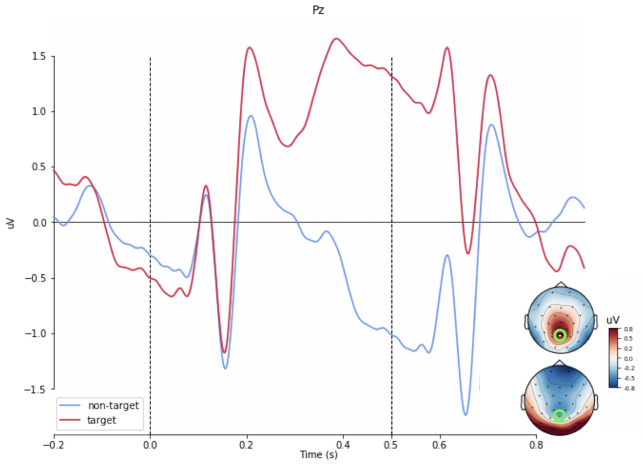

Figure 3: Grand average voltages of each ERP-component for target and non-target stimuli at the Pz channel and cranial topomap for average scalp voltages from 250 ms to 800 ms, with the Pz channel circled in green.

Brainsourcing Performance

Table 1 shows the classifier performance for the brainsourcing experiment. This performance, with random baseline computed using permuted labels, is depicted in Figure 5.

As additional participants were added, performance significantly improved, with the largest incremental improvement occurring at $N = 2$, with an average F1 score of 0.77 and improvement of 0.10 over $N = 1$. Significant performance improvements continued through $N = 12$, with an average F1 score of 0.94 across all tasks. Task “old” experienced the largest improvement in performance, with an F1 score of 0.98 at $N = 12$ and a $\Delta N = 1$ of 0.35. Conversely, task “young” performance improved the least, with an F1 score of 0.65 for $N = 1$, compared with 0.86 at $N = 12$, for a $\Delta N = 1$ of 0.21.

Throughout all numbers of participants used, precision was significantly higher than recall. These differences were most pronounced at low values of N ($N < 7$), however they become less pronounced at higher values of N ($N \geq 8$), where precision and recall begin to converge for most tasks.

Using the Wilcoxon signed-rank test, results estimated using different numbers of participants N were compared to each other to determine if the differences in estimations were statistically significant (Figure 7). Estimations calculated using $N \geq 2$ participants were all significantly better than those from a single participant $N = 1$. A representative sample of labeling

Task	N = 1			N = 2				N = 4				N = 6				N = 12			
	P	R	**F1**	P	R	**F1**	ΔN=1	P	R	**F1**	ΔN=1	P	R	**F1**	ΔN=1	P	R	**F1**	ΔN=1
Blond	0.77	0.64	**0.70**	0.84	0.75	**0.79**	0.09	0.91	0.84	**0.87**	0.17	0.95	0.89	**0.92**	0.22	0.98	0.95	**0.97**	0.27
Female	0.81	0.68	**0.74**	0.85	0.82	**0.83**	0.09	0.92	0.89	**0.91**	0.17	0.94	0.93	**0.93**	0.19	0.97	0.99	**0.98**	0.24
Young	0.72	0.59	**0.65**	0.77	0.71	**0.74**	0.09	0.82	0.76	**0.79**	0.14	0.85	0.78	**0.82**	0.17	0.85	0.87	**0.86**	0.21
Smiling	0.75	0.59	**0.65**	0.81	0.75	**0.78**	0.13	0.89	0.83	**0.86**	0.19	0.93	0.86	**0.89**	0.22	0.97	0.87	**0.92**	0.27
Dark-haired	0.76	0.60	**0.67**	0.81	0.74	**0.78**	0.11	0.90	0.82	**0.86**	0.19	0.94	0.86	**0.89**	0.22	0.98	0.92	**0.95**	0.28
Male	0.71	0.62	**0.69**	0.82	0.72	**0.77**	0.08	0.90	0.79	**0.84**	0.15	0.93	0.82	**0.87**	0.18	0.97	0.86	**0.91**	0.22
Old	0.71	0.57	**0.63**	0.76	0.69	**0.72**	0.09	0.82	0.69	**0.76**	0.13	0.89	0.75	**0.81**	0.18	0.96	0.99	**0.98**	0.35
Not smiling	0.74	0.59	**0.66**	0.79	0.74	**0.76**	0.10	0.87	0.78	**0.82**	0.16	0.91	0.87	**0.88**	0.22	0.97	0.94	**0.96**	0.30
Mean	*0.75*	*0.61*	***0.67***	*0.81*	*0.74*	***0.77***	*0.10*	*0.88*	*0.80*	***0.82***	*0.15*	*0.92*	*0.85*	***0.88***	*0.21*	*0.96*	*0.92*	***0.94***	*0.27*

Table 1: Precision, recall, F1 score, and the improvement of the F1 score with respect to $N = 1$ for target task, given N participants used in the brainsourcing estimation. All $\Delta N = 1$ are statistically significant $p \leq 0.0001$. Performance for each task improved dramatically by increasing the number of participants used to estimate class labels for stimuli.

outputs with corresponding stimuli and varying N is shown in Figure 6, where the performance can also be seen to improve significantly as the number of participants increases.

Characteristics of Brainsourcing Output

The stimuli used in this experiment were produced by sampling from a multivariate normal distribution. By grouping the stimuli into binary categories containing a target class and its inverse, we expected that this would lead to a bimodal distribution. Intuition suggests that a properly functioning brainsourcing model should output predictions that also follow a bimodal distribution. Histograms in Figure 4 produced from the results of the brainsourcing model reveal the predictions indeed follow a bimodal distribution, corresponding with the nature of the stimuli dataset. This bimodal property becomes more pronounced when the number of participants used to create the brainsourced estimation increases. The results indicate that as the size of the crowd increases, the confidence of the classifier predictions improve as they start to follow the distribution of the ground truth data more closely. This suggests that the improved empirical performance of brainsourcing can be attributed to increased confidence of the brainsourcing model.

Analysis of Sparsity and Crowd Size

After accounting for observations lost due to eye blinks and other sources of noise, an average of 11 participants viewed a given stimulus. Most stimuli for a given task were viewed by at least two participants, with the total sparsity of the cleaned data averaging 60%. Less than 3% of stimuli were viewed only by a single participant. These stimuli were equally distributed between target and non-target classes and did not significantly affect model performance.

DISCUSSION AND CONCLUSIONS

The objective of our research was to study whether crowdsourced microtasks can be solved by a group of individuals

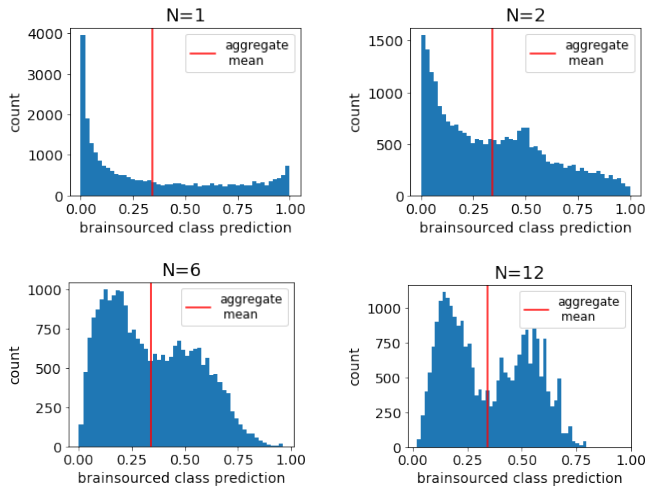

Figure 4: Distribution of mean predictions produced from the brainsourcing model. A brainsourcing prediction above the mean indicates that a given stimulus is more likely to be of the target class. Conversely, a brainsourcing prediction below the mean indicates that a given stimulus is more likely a non-target stimulus. The results correctly show the bimodal nature of the classified data, which becomes increasingly distinct as more participants are used to estimate class labels.

via Brain-Computer Interfacing (BCI) requiring no explicit feedback from the individuals.

We presented brainsourcing; a novel crowdsourcing paradigm that utilizes brain responses of a group of human contributors each performing a recognition task to determine classes of stimuli. An experiment utilizing brainsourcing is reported for an image labeling task. We demonstrated significant performance gains compared to a performance of individual partici-

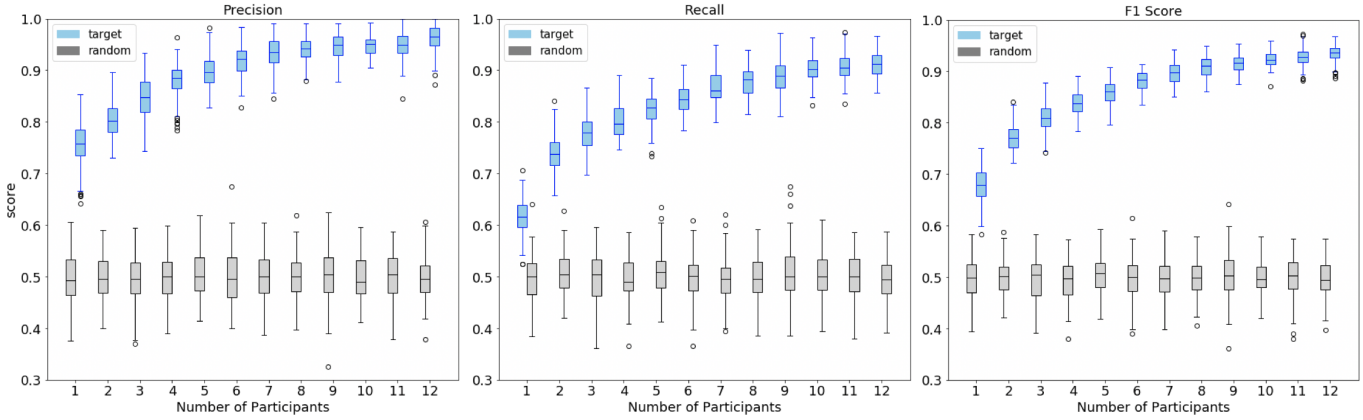

Figure 5: Performance of the brainsourcing model was significantly better than random for all values N, and the performance improves significantly as more participants N are added to the estimation. The largest per-participant increase in F1 score over $N = 1$ occurs at $N = 2$, and performance begins to converge around $N = 8$.

pants, and achieved nearly perfect crowd-level performance. A task-wide average F1 score of 0.94 shows that brainsourcing performs well enough to be considered a useful method for real-world recognition tasks.

By combining the recognition of different perspectives of individuals on the attributes in each task, we were able to increase the accuracy of the classification. Brainsourcing was also shown to help in improving the low signal-to-noise ratio present in brain imaging. By using the input of multiple users, the correct predictions of brain activity tend to override the erroneous predictions caused by signal noise.

Summary of Contributions

In order to study whether brainsourcing is possible and how it performs, we asked two research questions. Here, we discuss the results accordingly.

RQ1: *Can human brain responses be utilized in a crowdsourcing setting to predict a class label for a given stimulus, such that the performance of the crowd is better than the performance of an individual user?*

A1: Our results show that human brain responses can be utilized to collectively inference class labels for stimulus, and that the performance of a crowd with only two participants significantly outperforms the performance of an individual participant.

RQ2: *How many participants are enough to reliably draw labeling decisions and at what point does adding additional participants fail to significantly improve results?*

A2: Brainsourcing shows consistently increasing performance as a function of the crowd size; in our results significant improvements were achieved when the crowd size increased by three participants (Figure 7). The performance peaks at the maximum crowd size of 12 for all measures used. While precision seems to be relatively stable already with a smaller crowd size, recall constantly improves when data from new participants are added (Figure 5). This indicates that even small crowds can be used in a brainsourcing setting with high quality output, but ensuring high recall requires larger crowd size. This finding is also supproted by the classifier analysis showing clear convergence towards a binomial distribution as a function of the crowd size (Figure 4).

Limitations

Currently, this approach exploits the use of binary class datasets, rather than a multiclass dataset. The stimuli used were selected such that they could easily be separated as target or non-target. Edge cases, such as androgynous faces for the male/female tasks, or light brown hair for the blond/darkhaired task, were therefore underexplored. Additionally, we only investigated strict recognition tasks and relied on well-known BCI methodology [8]. More complex tasks, such as personal preference assignment, are likely to prove more challenging for this approach.

During the experiment, the stimulus dataset had to be structured in such a way so that approximately 70% of the stimuli were of the non-target class, and the remaining 30% were of the target class. ERPs generated from frequent target stimuli are more difficult to distinguish due to ERP components being selectively evoked by infrequent stimuli [68, 27, 58]. Effectively balanced data to ensure significantly more non-target than target stimuli cannot be guaranteed in real-world datasets, although this balancing can be partially achieved by randomly mixing already classified non-target stimuli into a new dataset.

Due to eye blinks, electrical intereference, and other non-preventable disturbances, approximately 10% of the data collected had to be discarded. While discarding data in undesirable, the amount of stimuli that can be shown to a participant in a single session (approxiately 1 stimulus/500 ms) compensates for the inevitable loss of data.

A significant limitation of our work in practical applications are the devices, setup time, and cost. Participants were connected to a 32-channel EEG system, requiring approximately

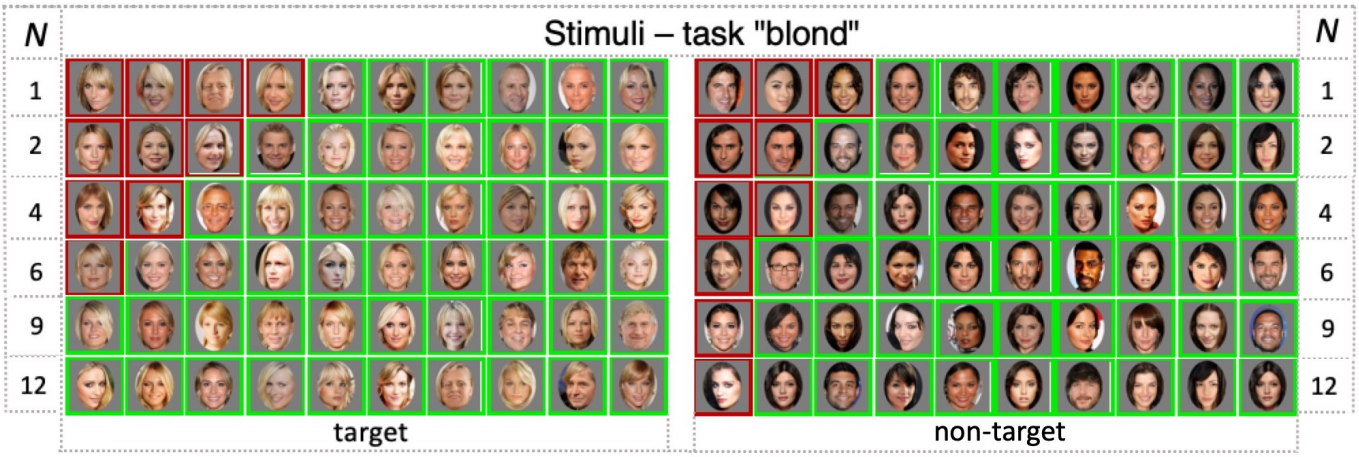

Figure 6: Representative sample of correct and incorrect labelings for target and non-target stimuli within the task "blond", by number of participants used to estimate the true label. Brainsourcing classification performance continues to improve as particpants N increases. Performance at $N > 9$ achieves an F1 score of 0.90.

n	1	2	4	6	9	12
1	-	****	****	****	****	****
2	****	-	ns	*	*	****
4	****	ns	-	ns	*	***
6	****	*	ns	-	ns	*
9	****	*	*	ns	-	ns
12	****	****	***	*	ns	-

ns : not significant * : $p < 0.05$

*** : $p < 0.001$ **** : $p < 0.0001$

Figure 7: A matrix demonstrating statistical significance between different brainsourcing population sizes. We show that for any crowd size greater than N=1, the improvements in performance are statistically significant. Similarly, a crowd size of N=6 is significantly better than N=2; N=9 is significantly better than N=4.

20 minutes of physical preparatory work by a trained laboratory technician, followed by an additional 30 minutes to complete the recognition tasks. EEG systems are continually developing, with current research evolving to adopt techniques that drastically reduce setup time (e.g. as with flexible silicone technologies or active electrode amplifiers [51]). Innovations within EEG equipment [20] and other brain imaging technologies such as functional near-infrared spectroscopy [53] and magnetoencephalography [9] are likely to further to push the boundaries of what brainsourcing can achieve.

It is also important to note that while our approach is collaborative, it relies on personalized models for each individual. It is infeasible to simply take the EEG signals from all participants and train a single model using these signals. While some EEG potentials can be localized within a specific, common cortical structure, individual differences in orientations of dipole sources, skull thickness [30, 18], skin resistance, and others mean that EEG is specific to the individual and even recording setting. Therefore, the differences between two or more subjects are likely too large for such a model to perform well. However, the individual classification models enabled us to circumvent this problem.

Future Work

While the current model yields promising results, it could be improved using a variety of techniques. More advanced machine learning approaches, which could better learn the structural differences between target and non-target ERPs, could lead to better performance, particularly when the number of available crowdworkers is relatively low (N<5). Collaborative filtering techniques could be used to recover data lost due to confounding artefacts or noise by estimating missing values, as an additional step between drawing individual estimations and creating crowdsourced estimations from combined predictions. Currently the proposed approach only demonstrates the feasibility of using a supervised approach to train models for purposes of stimuli classification. Using unsupervised methods to cluster stimuli into their likely classes also warrants further exploration. The approach could also be extended to translate EEG signals from separate individuals across a common set of stimuli into a common latent space. This would enable us to combine translated EEG data or train models for all participants per task, rather than per participant, per task.

Since the participants were explicitly asked to recognize the features in the images, according to the crowdsourcing categorisation in [22], our study falls under explicit crowdsourcing; the participants are instructed to perform an explicit task. Future research could explore brainsourcing to perform implicit crowdsourcing, such as through gamification of microtasks and then recording the brain signals associated with completing a microtask. Also, the task instructions in our case were to make a mental note when a target was observed. An approach based on completely passive reactions, such that subjects are

presented with stimuli in a wide variety of classes with the only instruction to simply observe, is another potential, but unstudied extension to the present work.

The P300 evoked potential employed in our implementation of brainsourcing is produced when a stimulus matches a mental target [5]. While in the present study we harnessed the P300 evoked by images matching the target class, the P300 is actually quite insensitive regarding stimulus type or modality. For example, the P300 has been shown to be produced by target stimuli presented through the auditory [57], haptic [74], and olfactory [61] modalities. Thus, the brainsourcing paradigm could in theory be used to classify data not easily accessible to a computer system, such as the smell of a food or the feeling of a touch.

While the P300 is highly task-relevant [5], it only allows the solving of explicit microtasks; participants have to be aware of the task in order for the P300 to be robustly elicited. However, it should be possible to also leverage brain potentials that occur without explicit conscious control. The N170, for example, an early negative potential that has been associated with perceiving face-like visual stimuli and relatively unaffected by attention [15]. It therefore could be targeted in a face detection task (rather than face category detection). Another implicitly valuable potential is within the N400, which has been associated with violations of expectations in a given linguistic context, such as semantical incongruencies in text [45]. Due to this attribute, it could be possible to use brainsourcing to detect semantical incongruencies by instructing participants to simply read text presented to them while recording their brain activity. This way, the participants would not be explicitly solving a crowdsourced task, but would accomplish this as a side product of reading text.

Yet another line of research would be to extend the tasks beyond simple recognition towards tasks that could allow the crowd and the computers to detect socially cohesive opinions, and even detecting crowds with intersubjective or contradicting views. As our current results show overall performance improvement as a function of the crowd size, we also observed variance among specific tasks. The largest task-wise performance increase was in task "Old" and the smallest improvement was in task "Young". A possible explanation is that these tasks were likely the most vulnerable to subjective interpretation. What constitutes an opinion of an old or a young person may heavily depend on the participant's age, culture, and gender. This suggests that brainsourcing may have potential in unlocking subjective and implicit cohesion and diversion of opinions, emotions, and even attitudes [66, 48, 64, 56].

Ethical Concerns

In a society where BCI applications are as common as current-day smart phones, novel ways in which these technologies can be used unethically may emerge. Governments, corporations, and criminal organizations may use these systems in ways that violate individual rights to privacy or autonomy. New labor paradigms may emerge that exploit the convenience of these systems and the widespread availability of brainsourcing workers. While frameworks to facilitate labor rights among crowdworkers have been proposed, such as Turkopticon [36], these must be adapted to account for the unique challenges BCI-based labor introduces. In this section, we expand upon existing work involving crowdworker rights and design fictions within the context of ubiquitous BCI adoption [73, 14]. We discuss how the techniques presented in this paper may be used or extended in a manner such that their applications are deemed unethical or otherwise immoral.

Subliminal probing

It has been demonstrated that private information (specifically, recognition of a face) can be obtained from individuals using a BCI system without their knowledge or consent, through the use of subliminal probing techniques [28]. Such techniques may be extended or modified in a manner such that they could expose PIN codes, passwords, and private social relationships. Corporations or political campaigns could secretly probe consumer or voter attitudes to better market their products [70, 69]. Governments and corporations may be tempted to conduct BCI-driven surveillance to monitor their citizens or employees, and could enhance their findings using brainsourcing techniques.

Medical fingerprinting and privacy

EEG data should be considered personal medical data; protecting it becomes particularly important as EEG data can be used to diagnose neurological conditions [10, 63]. EEG data can also be used as a biometric identifier [52, 19], and thus could be used to identify an otherwise anonymous user if similar data is made publicly available (e.g., via a security breach in a BCI application that collects EEG data). Privacy-preserving techniques for BCI applications need to be designed to protect the raw user data collected such that it cannot be obtained by third parties without the individual's explicit knowledge and consent. Furthermore, models produced using this sensitive information should be safeguarded so that the data used to build them cannot be reverse-engineered [54].

Abuse of microtask labor

There is risk of abuse of this system through the establishment of businesses and organizations that deploy it in an unethical manner. Brainsourcing can be conducted using healthy individuals with minimal training - a typical worker can be trained in under 15 minutes. Companies that wish to perform brainsourcing for financial gain may have reduced incentives to maintain the mental and physical welfare of their employees, who require little training and are thus more likely to be treated as expendable commodities. Widespread adoption of consumer-level BCI systems could result in brainsourcing tasks being completed by the average computer user; under such circumstances, brainsourcing tasks could serve as an alternative to existing online monetization methods. Such an approach might have negative consequences when combined with subliminal probing, where users are at risk of unwittingly consenting to their sensitive data being transmitted and utilized.

ACKNOWLEDGEMENTS

The research was supported by the Academy of Finland (grant numbers 322653, 328875, 313610). We would like to thank Zania Sovijärvi-Spapé for conducting laboratory experiments.

REFERENCES

[1] Laura Acqualagna and Benjamin Blankertz. 2013. Gaze-independent BCI-spelling using rapid serial visual presentation (RSVP). *Clinical Neurophysiology* 124, 5 (2013), 901 – 908. DOI:http://dx.doi.org/https://doi.org/10.1016/j.clinph.2012.12.050

[2] Laura Acqualagna, Matthias Sebastian Treder, Martijn Schreuder, and Benjamin Blankertz. 2010. A novel brain-computer interface based on the rapid serial visual presentation paradigm. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. 2686–2689. DOI: http://dx.doi.org/10.1109/IEMBS.2010.5626548

[3] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality Control in Crowdsourcing Systems: Issues and Directions. *IEEE Internet Computing* 17, 2 (March 2013), 76–81. DOI: http://dx.doi.org/10.1109/MIC.2013.20

[4] A. Azizian, A.L. Freitas, T.D. Watson, and N.K. Squires. 2006. Electrophysiological correlates of categorization: P300 amplitude as index of target similarity. *Biological Psychology* 71, 3 (2006), 278 – 288. DOI: http://dx.doi.org/https://doi.org/10.1016/j.biopsycho.2005.05.002

[5] Theodore R. Bashore and Maurits W. van der Molen. 1991. Discovery of the P300: A tribute. *Biological Psychology* 32, 2 (Oct. 1991), 155–171. DOI: http://dx.doi.org/10.1016/0301-0511(91)90007-4

[6] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. 2013. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications* 3, 10 (2013).

[7] Nima Bigdely-Shamlo, Andrey Vankov, Rey R. Ramirez, and Scott Makeig. 2008. Brain Activity-Based Image Classification From Rapid Serial Visual Presentation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 16, 5 (Oct 2008), 432–441. DOI:http://dx.doi.org/10.1109/TNSRE.2008.2003381

[8] Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. 2011. Single-trial analysis and classification of ERP components — A tutorial. *NeuroImage* 56, 2 (2011), 814 – 825. DOI:http://dx.doi.org/https://doi.org/10.1016/j.neuroimage.2010.06.048 Multivariate Decoding and Brain Reading.

[9] Elena Boto, Niall Holmes, James Leggett, Gillian Roberts, Vishal Shah, Sofie S. Meyer, Leonardo Duque Muñoz, Karen J. Mullinger, Tim M. Tierney, Sven Bestmann, Gareth R. Barnes, Richard Bowtell, and Matthew J. Brookes. 2018. Moving magnetoencephalography towards real-world applications with a wearable system. *Nature* 555, 7698 (2018), 657–661. DOI: http://dx.doi.org/10.1038/nature26147

[10] S. G. Boyd, A. Harden, and M. A. Patton. 1988. The EEG in early diagnosis of the Angelman (Happy Puppet) syndrome. *European Journal of Pediatrics* 147, 5 (01 Jun 1988), 508–513. DOI: http://dx.doi.org/10.1007/BF00441976

[11] Daren C. Brabham. 2008. Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies* 14, 1 (2008), 75–90. DOI: http://dx.doi.org/10.1177/1354856507084420

[12] Anne-Marie Brouwer, Boris Reuderink, Joris Vincent, Marcel A. J. van Gerven, and Jan B. F. van Erp. 2013. Distinguishing between target and nontarget fixations in a visual search task using fixation-related potentials. *Journal of Vision* 13, 3 (07 2013), 1–10. DOI: http://dx.doi.org/10.1167/13.3.17

[13] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. 2016. Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment. *IEEE Transactions on Affective Computing* 7, 4 (Oct 2016), 374–388. DOI: http://dx.doi.org/10.1109/TAFFC.2015.2493525

[14] Sasha Burwell, Matthew Sample, and Eric Racine. 2017. Ethical aspects of brain computer interfaces: a scoping review. *BMC Medical Ethics* 18, 1 (2017), 60. DOI: http://dx.doi.org/10.1186/s12910-017-0220-y

[15] Alexandra Séverac Cauquil, Gillian E. Edmonds, and Margot J. Taylor. 2000. Is the face-sensitive N170 the only ERP not affected by selective attention? *NeuroReport* 11, 10 (2000). https://journals.lww.com/neuroreport/Fulltext/2000/07140/Is_the_face_sensitive_N170_the_only_ERP_not.21.aspx

[16] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise Ranking Aggregation in a Crowdsourced Setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)*. ACM, New York, NY, USA, 193–202. DOI: http://dx.doi.org/10.1145/2433396.2433420

[17] Justin Cheng, Jaime Teevan, Shamsi T. Iqbal, and Michael S. Bernstein. 2015. Break It Down: A Comparison of Macro- and Microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 4061–4064. DOI: http://dx.doi.org/10.1145/2702123.2702146

[18] Moritz Dannhauer, Benjamin Lanfer, Carsten H. Wolters, and Thomas R. Knösche. 2011. Modeling of the human skull in EEG source analysis. *Human Brain Mapping* 32, 9 (2011), 1383–1399. DOI: http://dx.doi.org/10.1002/hbm.21114

[19] Luigi De Gennaro, Cristina Marzano, Fabiana Fratello, Fabio Moroni, Maria Concetta Pellicciari, Fabio Ferlazzo, Stefania Costa, Alessandro Couyoumdjian, Giuseppe Curcio, Emilia Sforza, Alain Malafosse, Luca A. Finelli, Patrizio Pasqualetti, Michele Ferrara, Mario Bertini, and Paolo Maria Rossini. 2008. The electroencephalographic fingerprint of sleep is genetically determined: A twin study. *Annals of Neurology* 64, 4 (2008), 455–460. DOI: http://dx.doi.org/10.1002/ana.21434

[20] Stefan Debener, Reiner Emkes, Maarten De Vos, and Martin Bleichner. 2015. Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear. *Scientific Reports* 5, 1 (2015), 16743. DOI:http://dx.doi.org/10.1038/srep16743

[21] Jia Deng, Jonathan Krause, and Li Fei-Fei. 2013. Fine-Grained Crowdsourcing for Fine-Grained Recognition. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*. IEEE Computer Society, Washington, DC, USA, 580–587. DOI: http://dx.doi.org/10.1109/CVPR.2013.81

[22] Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. 2011. Crowdsourcing Systems on the World-Wide Web. *Commun. ACM* 54, 4 (April 2011), 86–96. DOI: http://dx.doi.org/10.1145/1924421.1924442

[23] Emanuel Donchin and Michael GH Coles. 1998. Context updating and the P300. *Behavioral and brain sciences* 21, 1 (1998), 152–154. DOI: http://dx.doi.org/10.1017/S0140525X98230950

[24] Enrique Estellés-Arolas and Fernando González-Ladrón-De-Guevara. 2012. Towards an Integrated Crowdsourcing Definition. *J. Inf. Sci.* 38, 2 (April 2012), 189–200. DOI: http://dx.doi.org/10.1177/0165551512437638

[25] Manuel J.A. Eugster, Tuukka Ruotsalo, Michiel M. Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2014. Predicting Term-relevance from Brain Signals. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 425–434. DOI: http://dx.doi.org/10.1145/2600428.2609594

[26] Manuel J. A. Eugster, Tuukka Ruotsalo, Michiel M. A. Spapé, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2016. Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. *Scientific Reports* 6 (2016). DOI:http://dx.doi.org/10.1038/srep38580

[27] L.A. Farwell and E. Donchin. 1988. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology* 70, 6 (1988), 510 – 523. DOI:http://dx.doi.org/https://doi.org/10.1016/0013-4694(88)90149-6

[28] Mario Frank, Tiffany Hwu, Sakshi Jain, Robert T. Knight, Ivan Martinovic, Prateek Mittal, Daniele Perito, Ivo Sluganovic, and Dawn Song. 2017. Using EEG-Based BCI Devices to Subliminally Probe for Private Information. In *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society (WPES '17)*. ACM, New York, NY, USA, 133–136. DOI: http://dx.doi.org/10.1145/3139550.3139559

[29] Phillip.I. Good. 2000. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses* (2nd ed.). Springer. DOI:http://dx.doi.org/https://doi.org/10.1007/978-1-4757-3235-1

[30] Dirk Hagemann, Johannes Hewig, Christof Walter, and Ewald Naumann. 2008. Skull thickness and magnitude of EEG alpha activity. *Clinical Neurophysiology* 119, 6 (2008), 1271 – 1280. DOI:http://dx.doi.org/https://doi.org/10.1016/j.clinph.2008.02.010

[31] J. Craig Henry. 2006. Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, Fifth Edition. *Neurology* 67, 11 (2006), 2092–2092–a. DOI: http://dx.doi.org/10.1212/01.wnl.0000243257.85592.9a

[32] James E. Hoffman, Robert F. Simons, and Michael R. Houck. 1983. Event-Related Potentials During Controlled and Automatic Target Detection. *Psychophysiology* 20, 6 (1983), 625–632. DOI: http://dx.doi.org/10.1111/j.1469-8986.1983.tb00929.x

[33] Jeff Howe. 2006. The Rise of Crowdsourcing. *Wired Magazine* 14, 6 (06 2006). http://www.wired.com/wired/archive/14.06/crowds.html

[34] Jeff Howe. 2008. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business* (1 ed.). Crown Publishing Group, New York, NY, USA.

[35] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing (HLT '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 27–35. http://dl.acm.org/citation.cfm?id=1564131.1564137

[36] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 611–620. DOI: http://dx.doi.org/10.1145/2470654.2470742

[37] Giulio Jacucci, Oswald Barral, Pedram Daee, Markus Wenzel, Baris Serim, Tuukka Ruotsalo, Patrik Pluchino, Jonathan Freeman, Luciano Gamberini, Samuel Kaski, and Benjamin Blankertz. 2019. Integrating neurophysiologic relevance feedback in intent modeling for information retrieval. *Journal of the Association for Information Science and Technology* 70, 9 (2019), 917–930. DOI:http://dx.doi.org/10.1002/asi.24161

[38] Lauri Kangassalo, Michiel Spapé, Giulio Jacucci, and Tuukka Ruotsalo. 2019. Why Do Users Issue Good Queries?: Neural Correlates of Term Specificity. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. ACM, New York, NY, USA, 375–384. DOI: http://dx.doi.org/10.1145/3331184.3331243

[39] Ashish Kapoor, Pradeep Shenoy, and Desney Tan. 2008. Combining brain computer interfaces with vision for object categorization. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8. DOI: http://dx.doi.org/10.1109/CVPR.2008.4587618

[40] David R. Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative Learning for Reliable Crowdsourcing Systems. In *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., 1953–1961. http://papers.nips.cc/paper/4396-iterative-learning-for-reliable-crowdsourcing-systems.pdf

[41] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

[42] Nam Wook Kim, Zoya Bylinskii, Michelle A. Borkin, Krzysztof Z. Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. 2017. BubbleView: An Interface for Crowdsourcing Image Importance Maps and Tracking Visual Attention. *ACM Trans. Comput.-Hum. Interact.* 24, 5, Article 36 (Nov. 2017), 40 pages. DOI: http://dx.doi.org/10.1145/3131275

[43] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. 2016. Crowdsourcing in Computer Vision. *Foundations and Trends in Computer Graphics and Vision* 10, 3 (2016), 177–243. DOI: http://dx.doi.org/10.1561/0600000071

[44] A. Kubler, V. K. Mushahwar, L. R. Hochberg, and J. P. Donoghue. 2006. BCI meeting 2005-workshop on clinical issues and applications. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14, 2 (June 2006), 131–134. DOI: http://dx.doi.org/10.1109/TNSRE.2006.875585

[45] Marta Kutas and Kara D. Federmeier. 2010. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology* 62, 1 (Dec. 2010), 621–647. DOI: http://dx.doi.org/10.1146/annurev.psych.093008.131123

[46] Matthew Lease and Emine Yilmaz. 2012. Crowdsourcing for Information Retrieval. *SIGIR Forum* 45, 2 (Jan. 2012), 66–75. DOI: http://dx.doi.org/10.1145/2093346.2093356

[47] Christopher H. Lin, Ece Kamar, and Eric Horvitz. 2014. Signals in the Silence: Models of Implicit Feedback in a Recommendation System for Crowdsourcing. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14)*. AAAI Press, 908–914. http://dl.acm.org/citation.cfm?id=2893873.2894015

[48] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. 2010. EEG-Based Emotion Recognition in Music Listening. *IEEE Transactions on Biomedical Engineering* 57, 7 (July 2010), 1798–1806. DOI:http://dx.doi.org/10.1109/TBME.2010.2048568

[49] Steven J Luck. 2014. *An introduction to the event-related potential technique*. MIT press.

[50] Lena Mamykina, Thomas N. Smyth, Jill P. Dimond, and Krzysztof Z. Gajos. 2016. Learning From the Crowd: Observational Learning in Crowdsourcing Communities. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2635–2644. DOI: http://dx.doi.org/10.1145/2858036.2858560

[51] A. C. Metting van Rijn, A. Peper, and C. A. Grimbergen. 1990. High-quality recording of bioelectric events. *Medical and Biological Engineering and Computing* 28, 5 (1990), 389–397. DOI: http://dx.doi.org/10.1007/BF02441961

[52] Chisei Miyamoto, Sadanao Baba, and Isao Nakanishi. 2009. Biometric person authentication using new spectral features of electroencephalogram (EEG). In *2008 international symposium on intelligent signal processing and communications systems*. IEEE, 1–4. DOI:http://dx.doi.org/10.1109/ISPACS.2009.4806762

[53] Noman Naseer, Melissa Jiyoun Hong, and Keum-Shik Hong. 2014. Online binary decision decoding using functional near-infrared spectroscopy for the development of brain–computer interface. *Experimental Brain Research* 232, 2 (01 Feb 2014), 555–564. DOI: http://dx.doi.org/10.1007/s00221-013-3764-1

[54] Seong Joon Oh, Bernt Schiele, and Mario Fritz. 2019. *Towards Reverse-Engineering Black-Box Neural Networks*. Springer International Publishing, Cham, 121–144. DOI: http://dx.doi.org/10.1007/978-3-030-28954-6_7

[55] Markus Ojala and Gemma C Garriga. 2010. Permutation tests for studying classifier performance. *Journal of Machine Learning Research* 11, Jun (2010), 1833–1863.

[56] Danny Plass-Oude Bos. 2006. EEG-based Emotion Recognition. *The Influence of Visual and Auditory Stimuli* (01 2006).

[57] John Polich. 1986. Attention, probability, and task demands as determinants of P300 latency from auditory stimuli. *Electroencephalography and Clinical Neurophysiology* 63, 3 (1986), 251 – 259. DOI: http://dx.doi.org/https://doi.org/10.1016/0013-4694(86)90093-3

[58] John Polich. 2007. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology* 118, 10 (2007), 2128 – 2148. DOI:http://dx.doi.org/https://doi.org/10.1016/j.clinph.2007.04.019

[59] David Powers and Ailab. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Jounral of Machine Learning Technologies* 2 (01 2011), 2229–3981. DOI: http://dx.doi.org/10.9735/2229-3981

[60] Jeffrey M Rzeszotarski, Ed Chi, Praveen Paritosh, and Peng Dai. 2013. Inserting micro-breaks into crowdsourcing workflows. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[61] Kota Sano, Yasutami Tsuda, Hisanobu Sugano, Shuji Aou, and Akikazu Hatanaka. 2002. Concentration Effects of Green Odor on Event-related Potential (P300) and Pleasantness. *Chemical Senses* 27, 3 (03 2002), 225–230. DOI: http://dx.doi.org/10.1093/chemse/27.3.225

[62] Pradeep Shenoy and Desney S. Tan. 2008. Human-aided Computing: Utilizing Implicit Human Processing to Classify Images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 845–854. DOI: http://dx.doi.org/10.1145/1357054.1357188

[63] SJM Smith. 2005. EEG in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry* 76, suppl 2 (2005), ii2–ii7. DOI: http://dx.doi.org/10.1136/jnnp.2005.069245

[64] Ahmad Tauseef Sohaib, Shahnawaz Qureshi, Johan Hagelbäck, Olle Hilborn, and Petar Jerčić. 2013. Evaluating Classifiers for Emotion Recognition Using EEG. In *Foundations of Augmented Cognition*, Dylan D. Schmorrow and Cali M. Fidopiastis (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 492–501.

[65] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427 – 437. DOI:http://dx.doi.org/https://doi.org/10.1016/j.ipm.2009.03.002

[66] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic. 2016. Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *IEEE Transactions on Affective Computing* 7, 1 (Jan 2016), 17–28. DOI: http://dx.doi.org/10.1109/TAFFC.2015.2436926

[67] Michiel Spape, Rinus Verdonschot, and Henk van Steenbergen. 2019. *The E-Primer: An introduction to creating psychological experiments in E-Prime®: Second edition updated for E-Prime 3.* Leiden Univerisity Press.

[68] Samuel Sutton, Margery Braren, Joseph Zubin, and E. R. John. 1965. Evoked-Potential Correlates of Stimulus Uncertainty. *Science* 150, 3700 (1965), 1187–1188. DOI: http://dx.doi.org/10.1126/science.150.3700.1187

[69] Ariel Telpaz, Ryan Webb, and Dino J. Levy. 2015. Using EEG to Predict Consumers' Future Choices. *Journal of Marketing Research* 52, 4 (2015), 511–529. DOI:http://dx.doi.org/10.1509/jmr.13.0564

[70] G Vecchiato, J Toppi, F Cincotti, L Astolfi, F De Vico Fallani, F Aloise, D Mattia, S Bocale, F Vernucci, and F Babiloni. 2010. Neuropolitics: EEG spectral maps related to a political vote based on the first impression of the candidate's face. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2902–2905. DOI: http://dx.doi.org/10.1109/IEMBS.2010.5626324

[71] Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 319–326. DOI: http://dx.doi.org/10.1145/985692.985733

[72] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 321, 5895 (2008), 1465–1468. DOI: http://dx.doi.org/10.1126/science.1160379

[73] Richmond Y. Wong, Nick Merrill, and John Chuang. 2018. When BCIs Have APIs: Design Fictions of Everyday Brain-Computer Interface Adoption. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 1359–1371. DOI: http://dx.doi.org/10.1145/3196709.3196746

[74] Shuhei Yamaguchi and Robert T. Knight. 1991. P300 generation by novel somatosensory stimuli. *Electroencephalography and Clinical Neurophysiology* 78, 1 (1991), 50 – 55. DOI:http://dx.doi.org/https://doi.org/10.1016/0013-4694(91)90018-Y

[75] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2011. A Survey of Crowdsourcing Systems. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 766–773. DOI: http://dx.doi.org/10.1109/PASSAT/SocialCom.2011.203