BMC Neuroscience

RESEARCH

Open Access

Machine learning and EEG can classify passive viewing of discrete categories of visual stimuli but not the observation of pain

Tyler Mari[1*], Jessica Henderson[1], S. Hasan Ali[1], Danielle Hewitt[1], Christopher Brown[1], Andrej Stancak[1] and Nicholas Fallon[1]

Abstract

Previous studies have demonstrated the potential of machine learning (ML) in classifying physical pain from non-pain states using electroencephalographic (EEG) data. However, the application of ML to EEG data to categorise the observation of pain versus non-pain images of human facial expressions or scenes depicting pain being inflicted has not been explored. The present study aimed to address this by training Random Forest (RF) models on cortical event-related potentials (ERPs) recorded while participants passively viewed faces displaying either pain or neutral expressions, as well as action scenes depicting pain or matched non-pain (neutral) scenarios. Ninety-one participants were recruited across three samples, which included a model development group (n = 40) and a cross-subject validation group (n = 51). Additionally, 25 participants from the model development group completed a second experimental session, providing a within-subject temporal validation sample. The analysis of ERPs revealed an enhanced N170 component in response to faces compared to action scenes. Moreover, an increased late positive potential (LPP) was observed during the viewing of pain scenes compared to neutral scenes. Additionally, an enhanced P3 response was found when participants viewed faces displaying pain expressions compared to neutral expressions. Subsequently, three RF models were developed to classify images into faces and scenes, neutral and pain scenes, and neutral and pain expressions. The RF model achieved classification accuracies of 75%, 64%, and 69% for cross-validation, cross-subject, and within-subject classifications, respectively, along with reasonably calibrated predictions for the classification of face versus scene images. However, the RF model was unable to classify pain versus neutral stimuli above chance levels when presented with subsequent tasks involving images from either category. These results expand upon previous findings by externally validating the use of ML in classifying ERPs related to different categories of visual images, namely faces and scenes. The results also indicate the limitations of ML in distinguishing pain and non-pain connotations using ERP responses to the passive viewing of visually similar images.

Keywords Empathy, Electroencephalography, N170, Event-related potential, External validation

Introduction

Machine learning (ML) and EEG have demonstrated promise for predicting discrete categories of visual stimuli (e.g., objects, scenes, faces etc.) [1–7], subjective pain intensity in response to physical pain [8–10], and response to pharmaceutical intervention [11–13], to name but a few. Research from our group previously

*Correspondence:
Tyler Mari
Tyler.Mari@liverpool.ac.uk
[1] Department of Psychology, Institute of Population Health, University of Liverpool, 2.21 Eleanor Rathbone Building, Bedford Street South, Liverpool L69 7ZA, UK

demonstrated that high and low pain stimuli can be predicted with approximately 70% accuracy using time–frequency analysis of EEG features distributed across the scalp [9]. However, the effectiveness of ML and EEG for the classification of human facial expressions and scenes depicting pain and non-pain conditions has yet to be explored. This is despite a wealth of research demonstrating the importance of neurobiological empathic responses to observed pain, which has particular relevance to clinical, physiological, and societal domains [14–17]. For example, elucidating the neurobiology of empathy is important for understanding the development of empathy and for clinical conditions where empathy is reduced or absent (e.g., autism) [18–20]. Moreover, from a societal perspective, understanding the neurobiology of empathy may support areas such as medical education [21]. Therefore, this study aimed to address this gap by developing ML models using single-trial EEG responses during the passive observation of both facial expressions and action scenes depicting neutral and painful conditions.

Traditional ERP research studies exploring empathic responses to the observation of pain demonstrate differences in ERP amplitudes, which may enable accurate ML classification at the single-trial level. A meta-analysis of up to 36 studies demonstrated an enhanced P3 and late positive potential (LPP) during pain observation, with the maximal effect observed at central-parietal sites [22]. Previous research by our lab demonstrated that images depicting pain scenes elicited an enhanced LPP over central-parietal regions compared to situation-matched neutral images in both healthy people and a chronic pain population [23]. Therefore, single-trial EEG responses over central-parietal electrode sites may be an important candidate feature for the ML algorithm.

In addition to classifying EEG responses to images depicting neutral and pain conditions, we also aimed to externally validate ML for the classification of single-trial neural responses to broad categories of visual stimuli (faces versus scenes) regardless of the pain component, which to the best of our knowledge has yet to be attempted. Here, the N170 component may be the most informative feature for classification. The N170 component is an early negative waveform deflection which is maximally observed over occipitotemporal regions between 140 and 200 ms after stimulus onset, peaking at approximately 170 ms, which is enhanced during the observation of faces [24, 25]. The N170 is maximal when viewing faces and is attenuated or missing in response to other stimulus categories [25, 26]. The N170 has been reliably reproduced in stationary and mobile EEG experiments [24–30]. Additionally, the vertex positive potential (VPP), which is a large positive potential across frontal-central regions peaking between 140 and 180 ms, is observed after the presentation of a face stimulus [24, 31, 32]. Given the similarity in the characteristics of the N170 and VPP, the evidence suggests that both components originate from the same neural dipole [33, 34]. Therefore, neural responses located over occipitotemporal and frontal-central regions may enable accurate classification of face versus scene images.

Indeed, previous research has successfully combined EEG and ML to classify neural responses to visual stimuli including faces, objects, and scenes. A support vector machine (SVM) trained on EEG components over occipital electrodes has successfully classified the presence of visual objects in 7 subjects; achieving a cross-validated accuracy and AUC of 87% and 0.7, respectively [1]. Additionally, research has demonstrated that neural networks could successfully classify 40 image classes from the ImageNet database (e.g., animals, objects, food) with an average accuracy of 90.16% using EEG recorded from 6 subjects [2]. Further research exhibits comparable results in decoding neural responses to objects, scenes, human and animal bodies and faces [3–6]. Finally, an attention-based convolutional bidirectional long short-term memory network has been developed to classify EEG responses to familiar and unfamiliar faces [7]. Using time–frequency features from pre-frontal, frontal, and temporal regions, the authors classified familiar and unfamiliar faces with an accuracy of 91.34%. Therefore, the literature suggests that EEG and ML can potentially be used to successfully decode brain responses to categories of visual stimuli.

Despite promising results, the field is not without significant limitations. ML research is often insufficiently validated, with only internal validation methods used to evaluate models. This potentially leads to inflated performance estimates, overfitting and ungeneralisable models [35–37]. Therefore, ML models should be evaluated using data independent of model development [38]. One such approach is external validation, whereby ML performance is assessed using novel data obtained from other cohorts, facilities, and repositories or collected from a different location (geographical), time (temporal) or experimental paradigm [37, 39]. Research has demonstrated reduced performance on external validation datasets [9, 40, 41]. Due to the omission of external validation, it is challenging to reasonably interpret the generalisability of existing research, as the results are potentially inflated.

The present study aimed to externally validate ML and EEG for visual stimuli decoding both across and

within subjects for the first time. Firstly, we trained a Random Forest (RF) model on EEG features to classify data into either faces or scenes. Moreover, we developed two further RF models to classify EEG data into either neutral or pain classes for both scenes and faces respectively. All models were externally validated using two separate samples: cross-subject which consisted of a new cohort, and within-subject which consisted of participants from the model development sample who were recruited for a second experimental session at a later time (temporal validation). We hypothesised that the RF model would classify visual stimuli with an accuracy significantly greater than the chance level (≈ 50%) for each classification task: (1) faces—scenes, (2) scenes: neutral—pain, and (3) faces: neutral—pain for both external validation samples.

Methods

Participants

A total of three samples, consisting of 116 EEG sessions, were collected for this study. Forty participants (22 female; 7 left-handed) aged between 18 and 52 (Mean = 27.70 years, standard deviation {SD} = 7.43) years were recruited for sample one (model development sample/cross-validation). Sample two (cross-subject validation) consisted of 51 participants (34 female; 6 left-handed) aged between 19 and 60 (Mean = 27.63 years, SD = 9.65), whilst sample three consisted of 25 participants aged between 21 and 53 (14 female; 4 left-handed; Mean = 28.96 years, SD = 8.01). Twenty-five participants from sample one completed a second experimental session a minimum of 12 weeks after their first session (Mean = 108.68 days, SD = 10.92). This cohort represented a temporal within-subject validation sample (sample three) for the ML analysis. We aimed to recruit a large sample, particularly for external validation, to provide robust estimates of model generalisability, as small external validation datasets can also provide imprecise estimates of model discrimination and calibration [42]. Participants provided written informed consent before participation and all methods were conducted in compliance with the Declaration of Helsinki. The study received ethical approval from the University of Liverpool Health and Life Sciences Research Ethics Committee. Eligibility criteria included: at least 18 years old, normal, or corrected-to-normal vision, no acute pain at the time of participating, no history of chronic pain, and no neurological conditions. Participants were compensated with a total of £40 for time and travel expenses. The raw data is available on reasonable request.

Materials

Pain faces

In the present study, we employed a passive viewing paradigm where participants were required to observe a series of visual stimuli but were not required to respond. This differs from a free viewing task, as participants were requested to pay attention to the image, which imposes a task and is arguably not truly free viewing [43]. Here, a 2 × 2 factorial design was used in this study: faces (expressions) and scenes, each with two levels, namely neutral and pain. The neutral and pain faces were selected from the Delaware Pain Database [44]. The Delaware Pain Database is an image database that contains photographs of the faces of individuals who are displaying a painful expression (e.g., grimacing) and matched neutral controls. We selected a total of 56 faces (28 painful and 28 matched neutral images). The faces were selected using several criteria. Firstly, we aimed to broadly recreate the ethnicity and gender distribution of the UK to provide representative stimuli. A total of 22 white subjects (80%) consisting of 11 males and females, 3 Asian subjects (10%) including 2 males and 1 female and 3 black subjects (10%) consisting of 1 male and 2 females were selected, which broadly matched the racial distribution of the UK [45]. Within the individual categories (e.g., white males) the images with the highest pain rating were selected, providing pain was listed as the dominant emotion. The 28 neutral images were selected as the matched version (e.g., same subject) of the pain expressions. Face images were approximately 1382 × 925 in size. Figure 1A demonstrates an example of neutral and pain expressions.

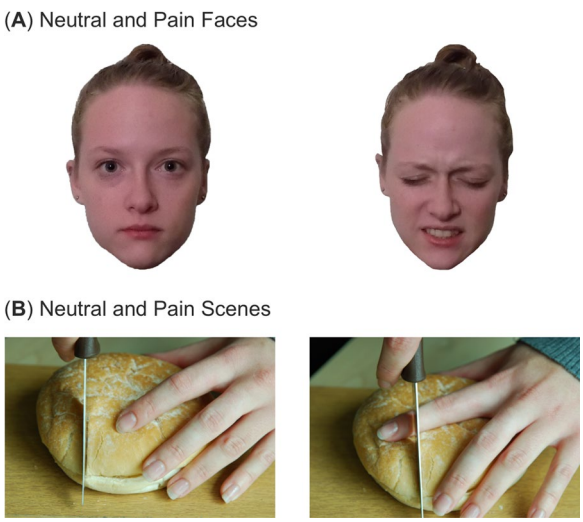

Fig. 1 **A** Example of neutral and pain face stimuli from the Delaware Pain Database [44]. **B** Example neutral and pain scene stimuli

Pain scenes

Additionally, still, photograph images of action scenes depicting pain or matched non-pain scenarios (hereinafter referred to as neutral or pain scenes) were employed in the present study. The pain scene images consisted of 28 images depicting either hands or feet in scenarios that elicit pain. For example, images of a knife cutting through bread in a way that would endanger the finger (e.g., placed under the knife). Twenty-eight matched neutral scenes, which replicate the scene but did not demonstrate pain, were also used. For example, the image depicted a knife cutting through bread without endangering the finger (e.g., the finger not placed under the knife). The same distribution of ethnicities implemented in the facial expression images was applied to the pain scene images. The images were selected from a larger internal pool of photographs depending on their pain rating. A small pilot study was conducted (n = 5) to rate each of the images in terms of pain intensity. The images that elicited the highest average pain rating in the pilot study were selected for the final experiment. The images used in this study are similar to previous research [23, 46–49]. Pain scene images were 774 × 518 in size. Figure 1B demonstrates examples of neutral and pain scene images used in this study.

Procedure

Participants attended the EEG laboratory at the University of Liverpool between June and October 2022. Following the fitting of the EEG cap, participants were seated inside a Faraday cage 1 m away from a 23-inch 1080p LCD monitor. The experimenter verbally explained the passive viewing task and the participants' questions were answered. During this time, participants were requested to pay attention to the images and minimise movement during trials. The experiment consisted of a total of 336 trials, split into three blocks of 112 stimuli. Within each block, 28 stimuli for each of the four conditions were presented. Each block lasted 6 min and was separated by approximately 15-min periods. During the block intervals, electrode impedances were checked, and additional saline solution was applied as required.

Each trial was initiated with a 2-s rest interval, where participants were shown a blank grey screen. Following the rest period, a colour photograph, that was randomly selected, was displayed for 1 s. Subsequently, the image disappeared, and the 2-s rest interval occurred before the presentation of the next image. This was repeated until all 112 images had been presented.

Following the completion of all blocks, the EEG cap was removed, and a subjective rating block was completed. Here, participants were informed that they were required to rate their perceived pain intensity of the images on a 0–100 scale with 0 reflecting no pain and 100 reflecting extreme pain. The rating scale included vertical bars denoting increments of 10. During the rating period, participants were presented with an image positioned above the rating scale and were required to rate the image by clicking the scale with the mouse in their right hand. The presentation of the images was randomised, and for each image, an infinite response time was employed. Once the participant had successfully rated the image, the screen was cleared, and the next image and scale were presented 100 ms later. Following this, participants completed the pain catastrophizing scale (PCS) [50] and were subsequently debriefed and compensated for their time and expenses.

EEG acquisition

Continuous EEG recordings were acquired using a 129-channel EGI System (Electrical Geodesic Inc., EGI, now Magstim EGI, Eugene, Oregon, USA) and a sponge-based Geodesic sensor net. The net was positioned with respect to three anatomical landmarks: two pre-auricular points and the nasion. Throughout the experiment, electrode-to-skin impedances were maintained below 50 kΩ. A recording bandpass filter was applied between 0.001 and 200 Hz and the sampling rate was set at 1000 Hz. Cz was used as the reference electrode.

EEG data analysis

The data were pre-processed using the Harvard Automated Processing Pipeline for Electroencephalography (HAPPE version 3) [51]. Firstly, low-pass and high-pass filters were applied to the data at 45 and 0.1 Hz, respectively. Secondly, the data were downsampled to 500 Hz and re-referenced using the common average approach [52]. Moreover, bad channel detection and interpolation were performed, and data contaminated by artefacts (e.g., oculographic) underwent wavelet thresholding (soft margin) to separate artefact and neural data. The data were then segmented into epochs of − 200 ms to 800 ms relative to stimulus onset (500 total time points) and baseline corrected (− 200 ms to 0 ms). Automated epoch rejection was then performed based on segment amplitude and similarity criteria. The thresholds were set at minimum and maximum segment amplitude of − 150 and 150, respectively in line with HAPPE recommendations [51]. The number of trials (mean ± SD) retained after automated trial rejection was 60.18 ± 8.44 (72% of total trials) for neutral scenes, 61.23 ± 6.19 (73%) for pain scenes, 62.93 ± 7.87 (75%) for neutral faces, and 62.15 ± 6.90 (74%) for pain faces, in sample one. In sample two, the mean number of trials remaining was 61.88 ± 5.14 (74%) for neutral scenes, 61.78 ± 6.22 (74%) for pain scenes, 62.63 ± 4.81 (75%) for neutral faces, and 62.27 ± 5.19

(74%) for pain faces. Finally, for sample three, the remaining number of trials was 62.76 ± 6.36 (75%) for neutral scenes, 60.20 ± 5.89 (72%) for pain scenes, 63.80 ± 5.97 (76%) for neutral faces, and 64.08 ± 6.49 (76%) for pain faces. Following pre-processing, the ERPs were analysed in MATLAB 2020b (The MathWorks, Inc., Natick, Massachusetts, USA) and EEGLAB 2021.1 [53]. Multiple comparisons were accounted for using the false discovery rate (FDR) method. A minimum window width of 10 ms was implemented to assess significant differences between the ERP waveforms.

Machine learning procedure

Following EEG pre-processing, the data were prepared for ML analysis. Each of the datasets (model development, cross-subject, and within-subject validation sample) were processed independently to prevent data leakage which could bias the external validation procedure [54]. Candidate features were calculated from single-trial ERP waveforms. A total of 18 candidate features, which primarily represented descriptive statistics of the ERP waveform, were calculated for each trial between 0 and 800 ms relative to stimulus onset. The features consisted of the mean, mode, median, minimum, maximum, standard deviation, root mean squared, variance, skewness, kurtosis, absolute mean, Shannon entropy, log energy entropy, range, mean squared, number of peaks, number of troughs, and the ratio between peaks and troughs. The features calculated in this study are comparable to previous research, both by our lab and external groups [9, 55–58]. The 18 features were calculated using MATLAB functions, where possible, and were computed for each of the 129 electrodes, resulting in 2322 candidate features.

Single-trial EEG is significantly impacted by noise and variability [59–61]. In line with our previous research, outlier feature values, defined as values beyond three median absolute deviations, were linearly interpolated. The interpolated values were calculated from neighbouring non-outlier data points for each condition using the MATLAB function *filloutliers* and were implemented as outliers impair the ML performance [62]. Interpolation was selected over data removal to maximise the dataset, as smaller datasets are more prone to overfitting [36]. A total of 4.77 ± 0.49%, 5.16 ± 0.31%, and 4.74 ± 0.15% of the data were interpolated for the model development sample, cross-subject validation sample, and within-subject validation sample, respectively.

After outlier interpolation in MATLAB, all ML processing and analysis were conducted using Python and Scikit-learn [63]. Here, the random seed was set to 123 for all ML analyses. The features for each dataset were scaled to between 0 and 1 and univariate feature selection was conducted. All candidate features were ranked in terms of importance using F-tests and a custom sequential feature selection was implemented. Here, a baseline RF model, with no hyperparameter tuning, was developed with one feature initially. Features were sequentially added, up to a maximum of 100 features (to limit computational complexity), to identify the optimal feature configuration. The optimal number of features for each classification task (scenes—faces; scenes: neutral—pain; and faces: neutral—pain) was defined as the baseline model that achieved the best cross-validation accuracy. Stratified k-fold validation (k = 10) was used as the cross-validation procedure.

Following the identification of the optimal features, the final ML model was developed for each task. Here, a RF model was trained on the model development dataset. Hyperparameter optimisation was achieved using random search, which searches within a range of upper and lower bounds for the optimal hyperparameter values for a user-specified number of iterations [64–66]. The external validation datasets did not inform model development as this can lead to overfitting. Therefore, hyperparameter optimisation was only performed in relation to cross-validation performance. For training and cross-validation, we evaluated model performance using stratified k-fold validation (k = 10) with accuracy as the scoring function. A maximum of 5000 iterations was specified for hyperparameter tuning. Once the optimal hyperparameters were identified, the model was refitted to the entire training dataset. This resulted in the final model that was evaluated using the external validation datasets.

Model evaluation: discrimination and calibration

The predictive capability of each model was assessed using several performance metrics for each of the validation sets (cross-validation and two external validation datasets). The primary discrimination metrics in this study were the model accuracy and area under the receiver operating characteristics curve (AUC). In addition, we also assessed model performance using alternative metrics including the Brier score, F1 score, precision, and recall. Overviews of these metrics have been reported elsewhere [8, 9, 67–69]. For the external validation datasets, we calculated model performance for each subject and averaged across the entire sample to achieve both individual subject and whole sample accuracies.

In addition to model discrimination performance, we also assessed calibration for models that exceed chance discrimination performance. Prediction algorithms can be subject to bias even when the models demonstrate excellent discrimination performance [70]. Consequently, model calibration, which evaluates the agreement

between the model's predicted probability of an event compared to the reference or observed value, should be assessed [54, 69, 70]. We assessed model calibration using calibration curves for both the cross-subject and within-subject validation sets, segmenting each dataset into 20 bins (see [70]). Calibration curves display the predicted probability on the x-axis and the true probability on the y-axis. Perfect calibration is represented by a 45° line, whereby the predicted and observed probabilities are identical [9]. Calibration has been extensively reviewed elsewhere [70, 71]. Calibration assessment is only necessary when the ML models demonstrate good discrimination ability, as models with poor performance do not require additional calibration assessment [69].

Statistical thresholding

Theoretically, the chance level for a binary classification task with infinite sample size is 50%. However, sample sizes are not infinite and are often small in neuroscience, resulting in variable chance levels. To quantitatively evaluate whether the ML model significantly outperformed the chance level for each subject, we implemented a statistical thresholding approach based on a binomial cumulative distribution method proposed by Combrisson and Jerbi (2015). The statistical threshold to exceed the chance level can be calculated using the following approach that applies the *binoinv* MATLAB function:

$$Statistical\ Threshold = binoinv\left(1-\alpha, n, \frac{1}{c}\right) * \frac{100}{n}$$

where α is the significance level, n is the number of trials per participant, and c is the number of classes.

For a given participant with $n=200$ and $c=2$, the model accuracy must be above 56%, 58%, and 61% to be significant at the 0.05, 0.01, and 0.001 levels, respectively [72]. If the model accuracy exceeds the given threshold, the performance is significantly greater than the chance level. A minimum of 100 data samples is required to achieve comparable results to permutation testing [72]. For all classification attempts, all subjects had more than 100 trials meaning that the use of binomial testing is acceptable. In all classifications, we use a threshold of $p=0.05$. The average chance level for cross-subject and within-subject predictions was 55.20 ± 0.20% and 55.26 ± 0.24%, 57.34 ± 0.37% and 57.41 ± 0.39%, and 57.39 ± 0.36% and 57.24 ± 0.38%, for faces—scenes, scenes: neutral—pain, and faces: neutral—pain classifications, respectively. Finally, to test whether the average sample performance exceeded the average chance threshold for each sample and classification attempt, the individual subject accuracies and chance levels were compared using paired samples t-tests.

Results

Self-report ratings

Descriptive statistics of the average self-report pain ratings for each of the four image types across the three samples are presented in Table 1. A 2×2 repeated measures ANOVA was conducted using IBM SPSS 27 (IBM Corp., Armonk, New York, USA) to assess the differences between participant pain ratings for the different conditions. The data from samples one (model development) and two (cross-subject validation) were combined for the analysis. There was a significant main effect of image type on the participant's perceived pain intensity ratings ($F(1,90)=19.89$, $p<0.001$, $\eta_p^2=0.18$), with the action scene images being rated as more painful than faces. Moreover, there was a significant main effect of pain condition ($F(1,90)=1568.26$, $p<0.001$, $\eta_p^2=0.95$). Here, the pain condition images received significantly higher pain ratings than the neutral condition images. Additionally, there was a significant interaction between image type and pain condition ($F(1,90)=22.10$, $p<0.001$, $\eta_p^2=0.20$). Post hoc paired samples t-tests demonstrated that pain ratings were significantly higher in the pain scenes condition when compared to the pain faces condition ($t(90)=4.89$, $p<0.001$, $d=0.51$). There was no significant difference between pain ratings for the neutral faces or scenes conditions ($t(90)=0.68$, $p=0.497$, $d=0.07$). Furthermore, the pain scene images had significantly higher pain ratings when compared to the neutral scene images ($t(90)=38.72$, $p<0.001$, $d=4.06$). Finally, the pain face images received significantly higher pain ratings when compared to the neutral face images ($t(90)=31.09$, $p<0.001$, $d=3.26$).

Table 1 Mean ± SD of perceived pain intensity for each condition and sample

Sample	Neutral scenes	Neutral faces	Pain scenes	Pain faces
Development Sample	5.96 ± 8.32	4.87 ± 8.35	61.74 ± 14.04	52.63 ± 18.19
Cross-subject Validation Sample	3.80 ± 3.98	3.93 ± 5.10	63.55 ± 14.49	57.28 ± 14.80
Within-subject Validation Sample	4.87 ± 8.31	4.56 ± 8.91	61.59 ± 10.69	58.38 ± 14.84

ERP analyses

Figure 2A–C show the averaged ERP waveform from select electrodes and the scalp isopotential maps for each condition and comparison (scenes—faces, scenes: neutral—pain, faces: neutral—pain). A significantly stronger negative deflection in response to face images compared to scene images was observed over bilateral occipital-temporal electrodes during the N170 time window (142–214 ms; peak 170 ms; $p < 0.00001$). Regarding neutral and pain scene images, a significantly stronger positive deflection was observed in a cluster of central-parietal electrodes during the LPP (524–796 ms; $p < 0.05$), peaking at 578 ms. Similarly, for neutral and pain faces, a significantly enhanced P3 potential (270–348 ms; peak 318 ms; $p < 0.05$) was observed over central-parietal electrodes in the pain condition relative to the neutral condition.

Machine learning analyses

Following ERP analyses, the ML analysis was conducted for each of the three classification attempts. From the feature selection procedure, a total of 89, 94, and 90 features were deemed optimal for each classification task, respectively. The scalp locations of the optimal features for each of the different classification paradigms are presented in Fig. 3. Additionally, the number of trials/observations

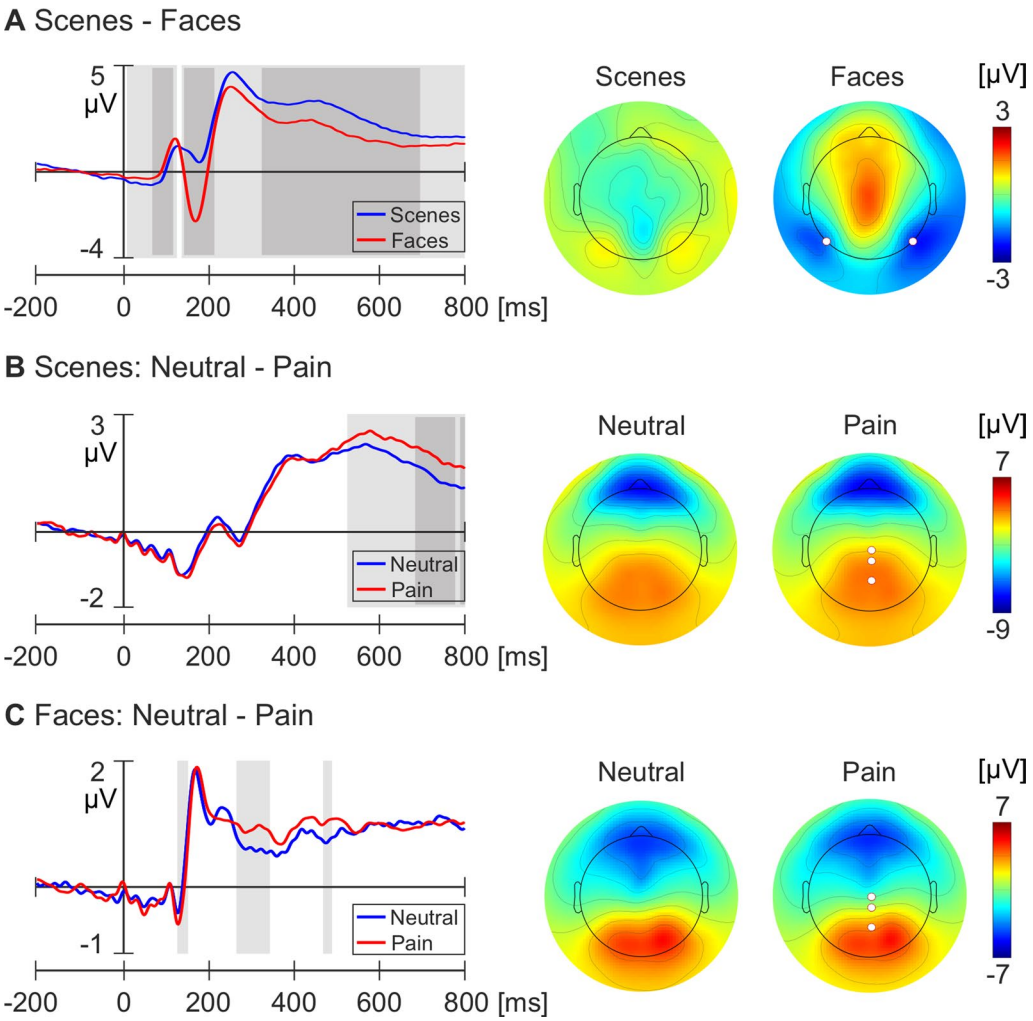

Fig. 2 Average ERP waveforms and scalp isopotential maps for each comparison from the unique 91 subjects within samples one and two. **A** Brain responses to scene and face images. Left: Average ERP waveforms from electrodes 58 (P7) and 96 (P8) for each condition. Right: Average scalp potential for each condition between 150 and 190 ms. **B** Brain responses to neutral and pain scenes. Left: Average ERP waveforms from electrodes Cz, 55, and 62 (Pz). Right: Average scalp potential between 524 and 674 ms for each condition. (**C**) Brain responses to neutral and pain face images. Left: Average ERP waveforms at electrodes Cz, 55, and 62 (Pz). Right: Average scalp potential between 270 and 348 ms for each condition. White circles indicate electrode locations of the average ERP waveforms. Light grey bars denote significant differences at $p < .05$. Dark grey bars represent significant differences at $p < .00001$

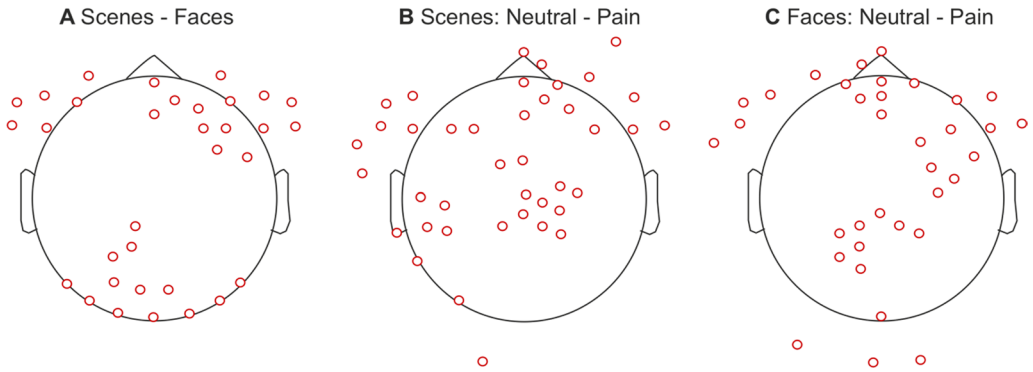

Fig. 3 Scalp locations of the important features determined during feature selection and model development for each classification task: scenes—faces (**A**), scenes: neutral—pain (**B**), and faces: neutral—pain (**C**)

Table 2 The number of observations/trials per condition and sample used in the ML analysis

Sample	Scenes		Faces		
	Neutral	Pain	Neutral	Pain	Total
Development/Cross-validation (n = 40)	2407	2449	2517	2486	9859
Cross-subject (n = 51)	3156	3151	3194	3176	12,677
Within-subject (n = 25)	1569	1505	1595	1602	6271
Total	7132	7105	7306	7264	28,807

used in the ML analysis for each condition and each sample is presented in Table 2.

Faces—scenes classification

The average of each sample's classification performance metrics and optimal hyperparameters for the classification of face versus scene photographs are reported in Table 3. Additionally, Fig. 4 shows the accuracies and chance thresholds for individual subjects in the cross-subject and within-subject validation samples. The average sample results demonstrate that the RF model achieved an accuracy (± SD) of 0.7456 (0.0459), 0.6415 (0.0634), and 0.6880 (0.0792) on the cross-validation and two external validation sets, respectively. Moreover, the model achieved an average AUC of 0.8189 (0.0406) on cross-validation, 0.7088 (0.0753) on cross-subject validation, and 0.7558 (0.0922) on within-subject validation. Paired samples t-tests demonstrated that the average sample accuracy was significantly greater than chance levels for the cross-subject sample ($t(50) = 10.08$, $p < 0.001$, $d = 1.41$) and the within-subject sample ($t(24) = 8.46$, $p < 0.001$, $d = 1.69$).

Regarding the individual subject classification performance, the results demonstrate that the model accuracy for 47 of 51 subjects was significantly greater than the chance level ($p < 0.05$) for the cross-subject validation sample. Moreover, for all participants (25/25) in the within-subject sample, the model achieved accuracies significantly greater than the chance levels.

Finally, we also assessed model calibration for the two external validation datasets. The calibration curves for both validation stages are presented in Fig. 5. To interpret the plots, if the model line falls above the reference line it is indicative of underestimating the probability of the outcome, whilst a line below the reference suggests

Table 3 Mean sample performance metrics for scenes—faces classification

Metric	Cross validation		Cross-subject validation		Within-subject validation	
	Mean	SD	Mean	SD	Mean	SD
Accuracy	0.7456	0.0459	0.6415	0.0634	0.6880	0.0792
AUC	0.8189	0.0406	0.7088	0.0753	0.7558	0.0922
Brier Score	0.1707	0.0164	0.2152	0.0253	0.1970	0.0358
F1 Score	0.7854	0.0299	0.6972	0.0460	0.7388	0.0557
Precision	0.6924	0.0495	0.6129	0.0583	0.6597	0.0959
Recall	0.9111	0.0240	0.8207	0.0890	0.8560	0.0802

Optimal hyperparameters: Number of estimators = 766, Maximum depth = 53, Minimum samples to split = 9, Minimum samples at leaf = 2, Maximum features = sqrt, Bootstrap = False

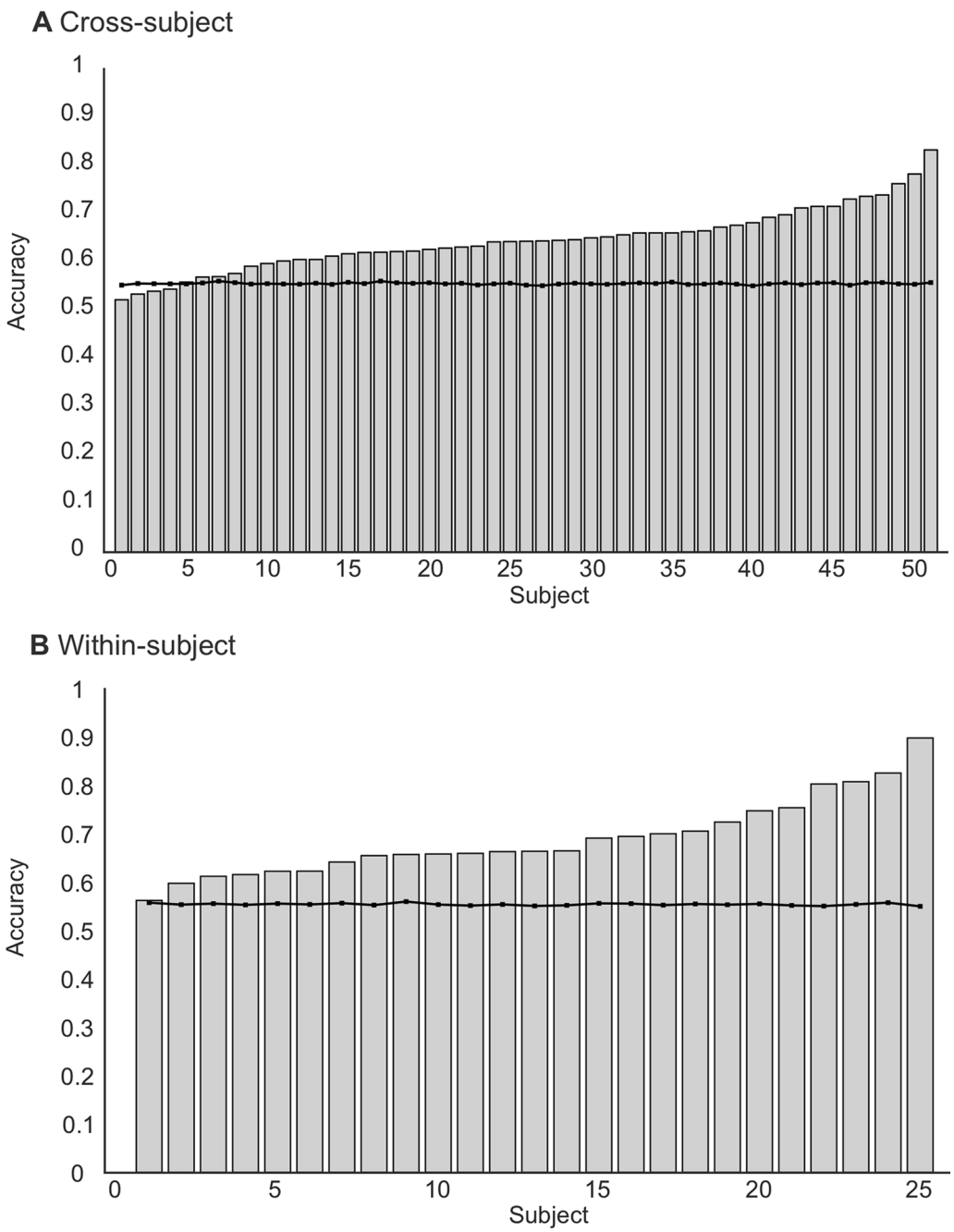

Fig. 4 Accuracies for each individual participant for the scenes—faces classification. (**A**) Cross-subject validation dataset. (**B**) Within-subject validation dataset. The black lines denote the significance threshold for chance classification performance at p = .05

the model is overestimating the probability of the event [9, 70]. The RF model for the faces versus scenes classification task generally demonstrates reasonable calibration for both cross-subject and within-subject datasets. The calibration curves follow the expected trend. Overall, the model is reasonably well-calibrated for both cross-subject and within-subject predictions.

Scenes: neutral—pain classification

The average classification performance and optimal hyperparameters for the neutral versus pain scenes classification are reported in Table 4. The average accuracy (SD) was 0.8038 (0.0208), 0.2837 (0.0358), and 0.5065 (0.0504) for cross-validation, cross-subject validation, and within-subject validation, respectively. The AUCs

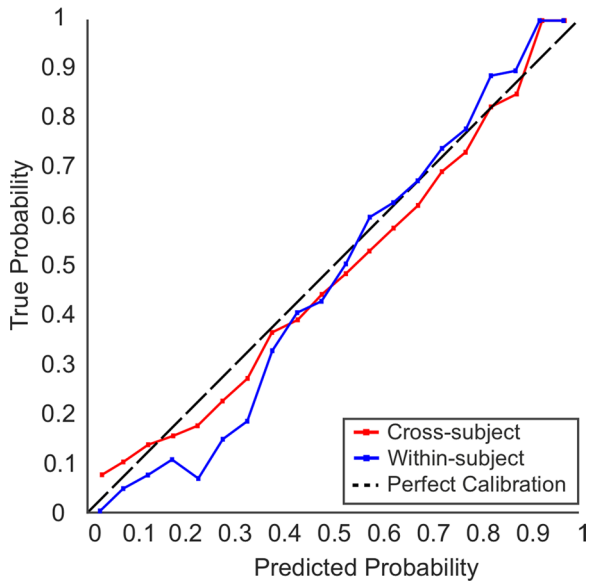

Fig. 5 Calibration curves for both cross-subject and within-subject validation datasets for the scenes—faces classification task. The black dotted line (45°) represents perfect calibration

produced a similar trend, with the evaluation procedure demonstrating an AUC of 0.8348 (0.0234), 0.2747 (0.0361), and 0.5123 (0.0518) for the three validation stages. Paired samples t-tests demonstrate that both the cross-subject ($t(50) = 57.15$, $p < 0.001$, $d = 8.00$) and within-subject ($t(24) = 6.67$, $p < 0.001$, $d = 1.33$) performance is significantly lower than the chance threshold. Regarding individual subject performance, the classification accuracy was less than the chance level for all 51 participants of the cross-subject sample. For the within-subject sample, only 2 of the 25 subjects recorded an accuracy significantly greater than the chance level. The results for individual subjects are reported in Fig. 6. Finally, as the models do not outperform chance levels for discrimination, we do not assess calibration.

Faces: neutral—pain classification

Finally, the average classification metrics and hyperparameters for the neural and pain faces classification are reported in Table 5. The results demonstrated that the RF model achieved an average accuracy (SD) of 0.6132 (0.0300), 0.5473 (0.0501), and 0.5076 (0.0383) for the cross-validation, cross-subject, and within-subject validation samples, respectively. In terms of AUC, the cross-validation AUC was 0.6717 (0.0396), the cross-subject AUC was 0.5629 (0.0667), and the within-subject AUC was 0.5241 (0.0557). Paired samples t-test indicated that the average sample accuracy was significantly lower than the chance threshold for the cross-subject validation sample ($t(50) = 3.82$, $p < 0.001$, $d = 0.53$) and the within-subject sample ($t(24) = 8.57$, $p < 0.001$, $d = 1.71$). The individual subject accuracies for both the cross and within-subject samples are reported in Fig. 6. Sixteen participants from the cross-subject sample and 2 participants from the within-subject sample achieved classification accuracies significantly greater than chance. As the model performance did not significantly exceed the chance threshold, we do not assess model calibration.

Exploratory analysis

As the RF model was unable to significantly exceed the chance thresholds for both neutral and pain scenes and faces classification, we performed exploratory analyses to assess whether a different number of features could improve the classification performance on the external validation datasets. To assess this, we developed and evaluated 100 RF models for each classification attempt, sequentially adding features on each iteration. We initially trained the model with 1 feature and progressed to a maximum of 100 features. The model was then assessed on both validation datasets. The RF was trained using the same procedure as the other models developed in this study, but the number of iterations of hyperparameter optimisation was capped at 500 to reduce computation complexity. The mean, standard deviation, minimum,

Table 4 Mean sample performance metrics for neutral—pain scenes classification

Metric	Cross validation		Cross-subject validation		Within-subject validation	
	Mean	SD	Mean	SD	Mean	SD
Accuracy	0.8038	0.0208	0.2837	0.0358	0.5065	0.0504
AUC	0.8348	0.0234	0.2747	0.0361	0.5123	0.0518
Brier Score	0.1480	0.0093	0.3966	0.0232	0.3044	0.0257
F1 Score	0.8344	0.0151	0.3866	0.0423	0.4798	0.0554
Precision	0.7277	0.0231	0.3379	0.0340	0.4960	0.0473
Recall	0.9788	0.0204	0.4553	0.0635	0.4682	0.0758

Optimal hyperparameters: Number of estimators = 735, Maximum depth = 46, Minimum samples to split = 28, Minimum samples at leaf = 17, Maximum features = sqrt, Bootstrap = False

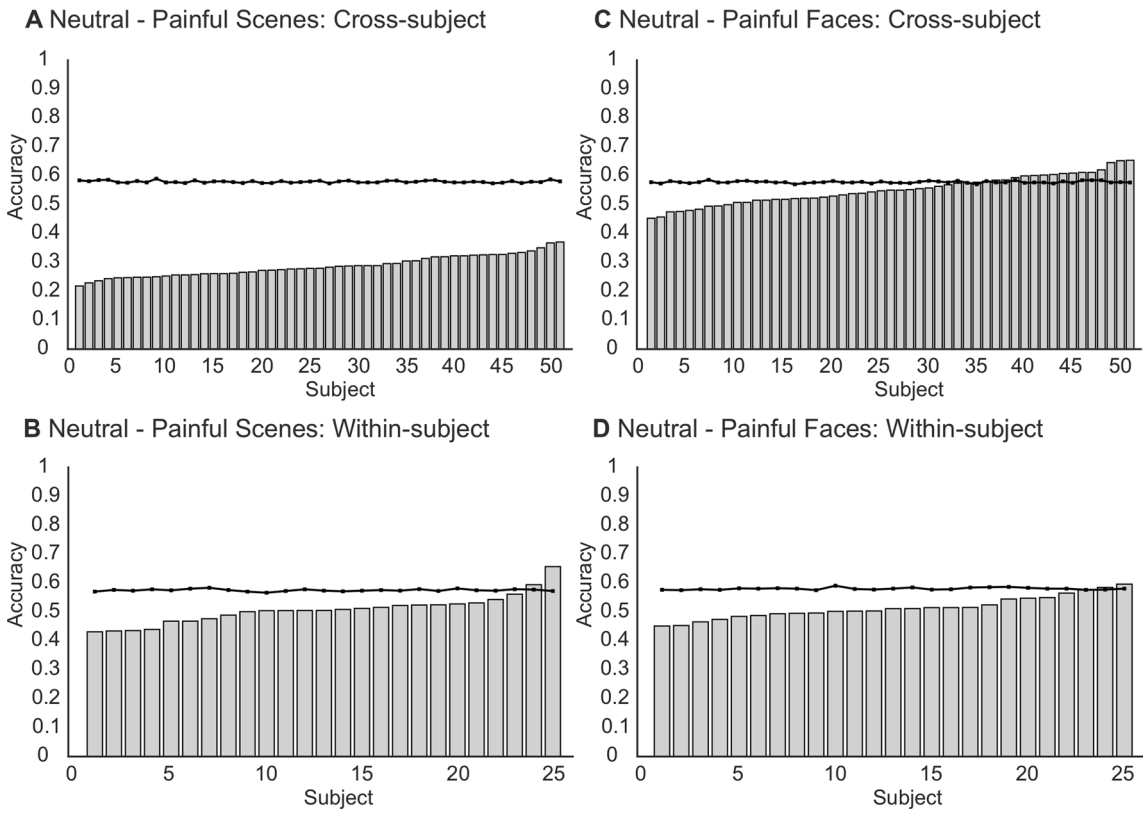

Fig. 6 Individual subject accuracies for both cross-subject (top panels) and within-subject (bottom panels) for both scenes: neutral—pain (left panels) and faces: neutral—pain (right panels). The black lines denote the significance threshold for above chance classification performance at p = .05

Table 5 Mean sample performance metrics for neutral—pain faces classification

Metric	Cross validation		Cross-subject validation		Within-subject validation	
	Mean	SD	Mean	SD	Mean	SD
Accuracy	0.6132	0.0300	0.5473	0.0501	0.5076	0.0383
AUC	0.6717	0.0396	0.5629	0.0667	0.5241	0.0557
Brier Score	0.2268	0.0073	0.2523	0.0155	0.2594	0.0108
F1 Score	0.5944	0.0505	0.5046	0.1053	0.3942	0.1003
Precision	0.6216	0.0353	0.5585	0.0720	0.5182	0.0834
Recall	0.5788	0.0930	0.4932	0.1804	0.3355	0.1200

Optimal hyperparameters: Number of estimators = 161, Maximum depth = 27, Minimum samples to split = 2, Minimum samples at leaf = 4, Maximum features = log2, Bootstrap = False

and maximum values for each of the classification tasks that did not exceed chance performance (scenes: neutral—pain and faces: neutral—pain) are reported in Table 6. The results of the exploratory analysis demonstrated comparable results to the original models developed. Minor performance improvements were observed, however, the model accuracy for both external validation sets remain around the chance classification level.

Discussion

We aimed to externally validate and classify single-trial EEG data elicited in response to visual stimuli using ML. Our results demonstrated that the RF model could classify images of scenes and faces with above-chance classification performance for all samples. However, the ML model could not discriminate between neutral and pain depictions of faces or scenes, achieving accuracies comparable to the chance classification rate, or lower. The

Table 6 Exploratory analysis results (accuracy) for feature combinations (1–100)

Classification	Sample	Mean	SD	Minimum	Maximum
Scenes	Cross-validation	0.7048	0.1155	0.5282	0.8186
	Cross-subject	0.3968	0.1226	0.2689	0.5362
	Within-subject	0.5063	0.0058	0.4889	0.5218
Faces	Cross-validation	0.5978	0.0014	0.5359	0.6128
	Cross-subject	0.5435	0.0063	0.5148	0.5540
	Within-subject	0.5166	0.0068	0.4952	0.5364

results support our first hypothesis that the RF model would outperform the chance level for the scenes versus faces classification task. However, the remaining two hypotheses that the RF model would outperform chance for both cross-subject and within-subject samples on both the neutral and pain conditions for face and scene images were not supported as the model performance was significantly lower than chance on all classification attempts. Consequently, the results suggest that large broad category differences (e.g., faces—scenes) are sufficient to achieve above-chance classification performance using external single-trial EEG data. However, more nuanced differences, such as those observed in the neutral–pain classifications, cannot be used to accurately discriminate classes with novel data using the current paradigm.

Our ERP analysis demonstrated an enhanced N170 over bilateral occipital-temporal electrodes in response to face images when compared to scenes, which has been reliably demonstrated previously [24–30]. Moreover, an increased LPP over a cluster of central-parietal electrodes was identified in the pain scene images compared to the neutral condition. Finally, an increased P3 over central-parietal electrodes was observed in response to pain faces compared to neutral expressions. The ERPs elicited in response to the empathic pain processing are also consistent with previous research [22, 23]. Meta-analyses of the ERP components observed during the empathic processing of painful stimuli demonstrated a positive shift in both the P3 and LPP components during the observation of painful stimuli, with the effect maximally observed over the central-parietal region [22]. Therefore, our ERP analysis validates the data quality and experimental paradigm and replicates the effects previously reported in a comparatively large sample of healthy participants.

The findings from this study are comparable and build upon the findings of previous research which demonstrated that discrete categories of visual stimuli could be accurately classified by ML and EEG. We successfully classified images into either faces or scenes, using features predominately located across frontal-central and occipitotemporal regions, which are active during the observation of faces (e.g., N170 and VPP) [24, 25, 31, 32]. Previous research has successfully classified neural responses to visual stimuli including faces, objects, and scenes [1, 3–7]. The present study extends the previous research by externally validating ML and EEG for image classification for both cross and within-subject prediction tasks using a large sample size. Much of the existing literature consisted of small samples (e.g., $\leq$10 subjects) [1–6], which are at higher risk of overfitting, resulting in potentially biased results [36, 73]. Furthermore, previous research did not rigorously assess model performance using external validation, which further increases the risk of poor generalisability [74]. Therefore, the performance and utility of previous models should be interpreted with caution. In addition to generalising to external data, our classification of scenes and faces demonstrated well-calibrated estimates, which provides further evidence of an effective prediction model [70, 71]. Calibration is often omitted in prediction modelling research, but it is essential to evaluating model performance [8, 75]. Consequently, our research provides methodologically superior estimates of the effectiveness of ML and EEG for classifying visual stimuli during passive viewing. To our knowledge, we are the first to externally validate ML models for EEG visual task decoding, providing robust estimates of model discrimination and calibration, and allowing for the interpretation of model generalisability.

The current study demonstrated that ML and EEG were unable to accurately classify neutral or pain faces or scenes. We believe that the low signal-to-noise ratio of EEG and the use of a passive task may have contributed to poor classification performance. Firstly, EEG has a low signal-to-noise ratio which may have resulted in poor discriminative ability for the neutral and pain stimuli classifications [76]. The N170 component offers a distinguishing characteristic between images of face and non-face classes. However, the ERP waveforms for neutral and pain images in either face or scene conditions are similar in their spatio-temporal profile, with differences mainly implicated as enhanced or augmented component fluctuations [22, 23]. Therefore, we can speculate that the differences at the single-trial level may be attenuated by noise and not detectable. Indeed, ML-EEG research often implements spatial filters to improve the signal-to-noise ratio and classification performance [77, 78]. However,

we opted against spatial filtering as it has a high risk of overfitting [77, 79]. Alternatively, the improved signal-to-noise ratio of magnetencephalography may allow for improved classification performance [80]. Moreover, the use of a passive viewing paradigm may have contributed to the classification performance. Research has demonstrated that passive viewing tasks result in reduced P300 amplitudes when compared to active viewing [81], whilst other component amplitudes (e.g., LPP) are associated with, and altered by, attention and engagement [82–84]. Therefore, any further attenuation of ERPs arising from passive viewing may have hindered the ML algorithm's ability to detect patterns. Consequently, nuanced differences (such as those elicited due to empathic responses to pain) may not enable accurate classification on the single-trial level during passive viewing. It is possible that active viewing tasks (e.g., requiring image classification performed by the viewer) may improve EEG signal and consequently ML performance. However, requiring input from the subject raises questions about the usefulness of such brain decoding tools, which should preferably allow inferences on behaviour without specific behavioural requirements. Additionally, active viewing may introduce additional confounds, leading to spurious results. Research has demonstrated that stimulus properties could be decoded solely using eye movements in an active viewing task, which was not possible during passive viewing within the same sample [85]. Whilst the impact of active viewing on EEG-ML classification systems should be investigated, it is important to note that, for the method to be genuinely useful and offer novel insight, it should preferably be able to accurately classify responses during passive viewing. Overall, the inability of the ML algorithm to classify neutral and pain images likely stems from poor signal-to-noise ratio and attenuated ERP responses.

Our results highlight the importance of external validation in ML research. Without performing robust, external validation, the generalisability of the ML model cannot be effectively assessed as the results may stem from overfitting [35–37]. Our cross-validation analysis of the pain scenes classification appears promising, with the model achieving an accuracy of approximately 80%. However, by implementing external validation, it was evident that the model was overfitting, achieving an accuracy below the chance level (28%) for the cross-subject dataset and comparable to chance (51%) for the within-subject validation. Therefore, through the external validation protocol, we were able to identify a model with poor generalisability, which may have otherwise been reported as an important finding. Indeed, we are not the first to demonstrate reduced performance when using an external validation [9, 40, 41], which is a significant, but often overlooked consideration when designing applied ML projects. Much of the prediction modelling research (regardless of research domain) does not assess model performance using external validation (e.g., only 5% of prediction modelling articles on PubMed report external validation in the title or abstract) [86]. Caution is advised when reporting or interpreting past ML-EEG results which have only been assessed using internal methods such as cross-validation, as the models are prone to overfitting, resulting in inflated, un-generalisable performance metrics [35, 37, 41]. Overall, our study highlights the importance of robust evaluation procedures when using ML, to minimise the risk of a new replication crisis [87].

The present study has several limitations. Firstly, we used a passive viewing experimental paradigm, which may have resulted in attenuated ERP responses [81]. Whilst we observed significant differences in both the P3 and LPP components in response to neutral and pain images, the differences between the conditions on a single trial level may have not been preserved due to the reduced neural responses associated with passive viewing, the low signal-to-noise ratio, and single-trial variability which may have contributed to poor ML performance [88]. Additionally, informal feedback from participants indicated that the passive viewing task was perceived as 'boring', which may have reduced attention, further impacting the neural responses [82–84]. Therefore, passive viewing may not be appropriate to elicit adequate responses that are detectable using ML at the single trial level using the approach outlined in the present study. Future research should implement active viewing paradigms and assess ML performance to build on our findings. For example, a two-alternative forced choice paradigm whereby participants are required to determine the presence or absence of pain may be more suitable for ML classification than passive viewing tasks. Similar forced choice tasks within pain empathy research have been widely reported [22]. Secondly, whilst the images in the study were similar to previous research [23, 46, 48, 49, 89], they may not be extreme enough to be detectable at the single trial level. Future research may wish to explore more intense pain imagery, such as those depicting injury [90], which may elicit larger ERP and behavioural responses. Additionally, the two stimuli categories used in this study (faces and scenes) were not matched for all physical properties (e.g., luminance), which may have confounded the EEG and impacted the classification. Research has demonstrated that properties such as brightness can alter EEG responses [91]. Therefore, we cannot entirely rule out the notion that confounds such as the physical properties of the image contributed to the classification performance. Moreover, we did not record the racial background of the participants in this study.

Research has shown that neural responses during pain observation are attenuated when viewing individuals of a different race [92]. Therefore, collecting and reporting the racial background of the subjects in this study could have provided important additional insight. Finally, the current study only recorded neural responses. Future research should aim to record composite measures (e.g., galvanic skin response) to supplement the EEG, which may improve classification performance.

The current study has important significance in the research field. Specifically, we provide the most robust estimates of EEG-ML visual stimuli decoding due to the extensive external validation procedure. We identified a potential limit of ML-EEG techniques, as ML models were unable to accurately classify pain observation above chance levels. However, assuming model performance can be improved, developing an empathy classification tool has important applications in healthcare, such as a supplementary tool for empathy training for healthcare workers [93]. However, performance improvements are imperative before such applications are considered. Currently, we can reasonably predict whether an individual was observing a face or a scene on external data, which represents an important knowledge contribution. However, the criteria typically applied to clinical contexts suggest that models that demonstrate an AUC less than or equal to 0.75 are not deemed practically useful [94]. Given that most of the AUCs in this study do not exceed this threshold, we recommend that improved model performance is pursued to increase the practical significance of the results, with a particular focus on empathic response prediction.

To the best of our knowledge, this is the first study to externally validate ML and EEG for the classification of various classes of visual stimuli including pain or neutral facial expressions and scenes with pain being inflicted on another person, or without pain. Our results demonstrate that ML and EEG can be used to decode neural responses and successfully classify face versus scene images with better-than-chance accuracy. However, the ML models were unable to discriminate between neutral and painful depictions of either face or scene images. Additionally, the ML result questions the suitability of passive viewing tasks for brain-based decoding algorithms. Overall, the study demonstrates promising results for decoding discrete categories of visual stimuli but is unable to identify the observation of pain using single-trial ERP responses. Finally, our results reiterate the importance of robust, external validation procedures to sufficiently evaluate ML-EEG performance; without which may lead to a new wave of impressive, but not replicable, findings.

Author contributions
NF, TM, CB, AS conceptualised the study. TM, JH, S.HA, DH collected the data. TM, NF, CB, AS were responsible for the methodology. TM, NF conducted the formal analysis. NF, CB, AS provided supervision. TM, NF wrote the original draft of the manuscript. All authors reviewed and edited the manuscript.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
All participants provided written informed consent before participation and all methods were conducted in compliance with the Declaration of Helsinki. The study received ethical approval from the University of Liverpool Health and Life Sciences Research Ethics Committee.

Consent for publication
Not applicable.

Competing Interests
The authors declare that they have no competing interests.

Received: 6 July 2023 Accepted: 6 September 2023
Published online: 15 September 2023

References

1. Stewart AX, Nuthmann A, Sanguinetti G. Single-trial classification of EEG in a visual object task using ICA and machine learning. J Neurosci Methods. 2014;228:1–14. https://doi.org/10.1016/j.jneumeth.2014.02.014.
2. Zheng X, Chen W, You Y, Jiang Y, Li M, Zhang T. Ensemble deep learning for automated visual classification using EEG signals. Pattern Recognit. 2020;102:107147. https://doi.org/10.1016/j.patcog.2019.107147.
3. Cudlenco N, Popescu N, Leordeanu M. Reading into the mind's eye: Boosting automatic visual recognition with EEG signals. Neurocomputing. 2020;386:281–92. https://doi.org/10.1016/j.neucom.2019.12.076.
4. Bagchi S, Bathula DR. EEG-ConvTransformer for single-trial EEG-based visual stimulus classification. Pattern Recognit. 2022;129:108757. https://doi.org/10.1016/j.patcog.2022.108757.
5. Yavandhasani M, Ghaderi F. Visual object recognition from single-trial EEG signals using machine learning wrapper techniques. IEEE Trans Biomed Eng. 2022;69(7):2176–83. https://doi.org/10.1109/TBME.2021.3138157.
6. Kaneshiro B, Perreau Guimaraes M, Kim H-S, Norcia AM, Suppes P. A Representational similarity analysis of the dynamics of object processing using single-trial EEG classification. Najbauer J, ed. PLoS ONE. 2015;10(8):e0135697. https://doi.org/10.1371/journal.pone.0135697.
7. Ghosh L, Dewan D, Chowdhury A, Konar A. Exploration of face-perceptual ability by EEG induced deep learning algorithm. Biomed Signal Process Control. 2021;66:102368. https://doi.org/10.1016/j.bspc.2020.102368.
8. Mari T, Henderson J, Maden M, Nevitt S, Duarte R, Fallon N. Systematic review of the effectiveness of machine learning algorithms for classifying pain intensity, phenotype or treatment outcomes using electroencephalogram data. J Pain. 2022;23(3):349–69. https://doi.org/10.1016/j.jpain.2021.07.011.
9. Mari T, Asgard O, Henderson J, et al. External validation of binary machine learning models for pain intensity perception classification from EEG in healthy individuals. Sci Rep. 2023;13(1):242. https://doi.org/10.1038/s41598-022-27298-1.
10. van der Miesen MM, Lindquist MA, Wager TD. Neuroimaging-based biomarkers for pain. PAIN Reports. 2019;4(4):e751. https://doi.org/10.1097/PR9.0000000000000751.

11. Jaworska N, de la Salle S, Ibrahim M-H, Blier P, Knott V. Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (EEG) and clinical data. Front Psychiatry. 2019. https://doi.org/10.3389/fpsyt.2018.00768.
12. Gram M, Erlenwein J, Petzke F, et al. Prediction of postoperative opioid analgesia using clinical-experimental parameters and electroencephalography. Eur J Pain (United Kingdom). 2017;21(2):264–77. https://doi.org/10.1002/ejp.921.
13. Graversen C, Olesen SS, Olesen AE, et al. The analgesic effect of pregabalin in patients with chronic pain is reflected by changes in pharmaco-EEG spectral indices. Br J Clin Pharmacol. 2012;73(3):363–72. https://doi.org/10.1111/j.1365-2125.2011.04104.x.
14. Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, Frith CD. Empathy for pain involves the affective but not sensory components of pain. Science (80-). 2004;303(5661):1157–62. https://doi.org/10.1126/science.1093535.
15. Decety J, Jackson PL. The Functional Architecture of Human Empathy. Behav Cogn Neurosci Rev. 2004;3(2):71–100. https://doi.org/10.1177/1534582304267187.
16. Lamm C, Decety J, Singer T. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. Neuroimage. 2011;54(3):2492–502. https://doi.org/10.1016/j.neuroimage.2010.10.014.
17. Singer T, Lamm C. The social neuroscience of empathy. Ann N Y Acad Sci. 2009;1156(1):81–96. https://doi.org/10.1111/j.1749-6632.2009.04418.x.
18. Fan Y-T, Chen C, Chen S-C, Decety J, Cheng Y. Empathic arousal and social understanding in individuals with autism: evidence from fMRI and ERP measurements. Soc Cogn Affect Neurosci. 2014;9(8):1203–13. https://doi.org/10.1093/scan/nst101.
19. Oberman LM, Hubbard EM, McCleery JP, Altschuler EL, Ramachandran VS, Pineda JA. EEG evidence for mirror neuron dysfunction in autism spectrum disorders. Cogn Brain Res. 2005;24(2):190–8. https://doi.org/10.1016/j.cogbrainres.2005.01.014.
20. Decety J, Holvoet C. The emergence of empathy: a developmental neuroscience perspective. Dev Rev. 2021;62:100999. https://doi.org/10.1016/j.dr.2021.100999.
21. Preusche I, Lamm C. Reflections on empathy in medical education: what can we learn from social neurosciences? Adv Heal Sci Educ. 2016;21(1):235–49. https://doi.org/10.1007/s10459-015-9581-5.
22. Coll M-P. Meta-analysis of ERP investigations of pain empathy underlines methodological issues in ERP research. Soc Cogn Affect Neurosci. 2018;13(10):1003–17. https://doi.org/10.1093/scan/nsy072.
23. Fallon N, Li X, Chiu Y, Nurmikko T, Stancak A. Altered cortical processing of observed pain in patients with fibromyalgia syndrome. J Pain. 2015;16(8):717–26. https://doi.org/10.1016/j.jpain.2015.04.008.
24. Bötzel K, Schulze S, Stodieck SRG. Scalp topography and analysis of intracranial sources of face-evoked potentials. Exp Brain Res. 1995. https://doi.org/10.1007/BF00229863.
25. Bentin S, Allison T, Puce A, Perez E, McCarthy G. Electrophysiological studies of face perception in humans. J Cogn Neurosci. 1996;8(6):551–65. https://doi.org/10.1162/jocn.1996.8.6.551.
26. Itier RJ. N170 or N1? spatiotemporal differences between object and face processing using ERPs. Cereb Cortex. 2004;14(2):132–42. https://doi.org/10.1093/cercor/bhg111.
27. Eimer M. Effects of face inversion on the structural encoding and recognition of faces. Cogn Brain Res. 2000;10(1–2):145–58. https://doi.org/10.1016/S0926-6410(00)00038-0.
28. Johnston P, Molyneux R, Young AW. The N170 observed 'in the wild': robust event-related potentials to faces in cluttered dynamic visual scenes. Soc Cogn Affect Neurosci. 2015;10(7):938–44. https://doi.org/10.1093/scan/nsu136.
29. Itier RJ, Taylor MJ. Source analysis of the N170 to faces and objects. NeuroReport. 2004;15(8):1261–5. https://doi.org/10.1097/01.wnr.0000127827.73576.d8.
30. Soto V, Tyson-Carr J, Kokmotou K, et al. Brain responses to emotional faces in natural settings: a wireless mobile EEG recording study. Front Psychol. 2018. https://doi.org/10.3389/fpsyg.2018.02003.
31. Jeffreys DA. Evoked potential studies of face and object processing. Vis cogn. 1996;3(1):1–38. https://doi.org/10.1080/713756729.
32. Jeffreys DA. A face-responsive potential recorded from the human scalp. Exp Brain Res. 1989. https://doi.org/10.1007/BF00230699.
33. Joyce C, Rossion B. The face-sensitive N170 and VPP components manifest the same brain processes: the effect of reference electrode site. Clin Neurophysiol. 2005;116(11):2613–31. https://doi.org/10.1016/j.clinph.2005.07.005.
34. Itier RJ, Taylor MJ. Inversion and contrast polarity reversal affect both encoding and recognition processes of unfamiliar faces: a repetition study using ERPs. Neuroimage. 2002;15(2):353–72. https://doi.org/10.1006/nimg.2001.0982.
35. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7(1):91. https://doi.org/10.1186/1471-2105-7-91.
36. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size, Hernandez-Lemus E ed. PLoS ONE. 2019;14(11):e0224365. https://doi.org/10.1371/journal.pone.0224365.
37. Cabitza F, Campagner A, Soares F, et al. The importance of being external methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed. 2021;208:106288. https://doi.org/10.1016/j.cmpb.2021.106288.
38. Lever J, Krzywinski M, Altman N. Model selection and overfitting. Nat Methods. 2016;13(9):703–4. https://doi.org/10.1038/nmeth.3968.
39. Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 2015;13(1):1. https://doi.org/10.1186/s12916-014-0241-z.
40. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol. 2019;20(2):193–201. https://doi.org/10.1016/S1470-2045(18)30762-9.
41. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68(1):25–34. https://doi.org/10.1016/j.jclinepi.2014.09.007.
42. Snell KIE, Archer L, Ensor J, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. J Clin Epidemiol. 2021;135:79–89. https://doi.org/10.1016/j.jclinepi.2021.02.011.
43. Li A, Wolfe JM, Chen Z. Implicitly and explicitly encoded features can guide attention in free viewing. J Vis. 2020;20(6):8. https://doi.org/10.1167/jov.20.6.8.
44. Mende-Siedlecki P, Qu-Lee J, Lin J, Drain A, Goharzad A. The Delaware Pain Database: a set of painful expressions and corresponding norming data. PAIN Reports. 2020;5(6):e853. https://doi.org/10.1097/PR9.0000000000000853.
45. Office for National Statistics (ONS). Ethnic Group, England and Wales: Census 2021.
46. Fan Y, Han S. Temporal dynamic of neural mechanisms involved in empathy for pain: an event-related brain potential study. Neuropsychologia. 2008;46(1):160–73. https://doi.org/10.1016/j.neuropsychologia.2007.07.023.
47. Akitsuki Y, Decety J. Social context and perceived agency affects empathy for pain: an event-related fMRI investigation. Neuroimage. 2009;47(2):722–34. https://doi.org/10.1016/j.neuroimage.2009.04.091.
48. Han S, Fan Y, Mao L. Gender difference in empathy for pain: an electrophysiological investigation. Brain Res. 2008;1196:85–93. https://doi.org/10.1016/j.brainres.2007.12.062.
49. Fallon N, Li X, Stancak A. Pain catastrophising affects cortical responses to viewing pain in others, Ptito M, ed. PLoS ONE. 2015;10(7):e0133504. https://doi.org/10.1371/journal.pone.0133504.
50. Sullivan MJL, Bishop SR, Pivik J. The pain catastrophizing scale: development and validation. Psychol Assess. 1995;7(4):524–32. https://doi.org/10.1037/1040-3590.7.4.524.
51. Gabard-Durnam LJ, Mendez Leal AS, Wilkinson CL, Levin AR. The Harvard automated processing pipeline for electroencephalography (HAPPE): standardized processing software for developmental and high-artifact data. Front Neurosci. 2018. https://doi.org/10.3389/fnins.2018.00097.
52. Lehmann D. Principles of spatial analysis. In: Gevins AS, Remond A, editors. Handbook of electroencephalography and clinical neurophysiology: methods of analysis of brain electrical and magnetic signals. Amsterdam: Elsevier; 1987. p. 309–54.
53. Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods. 2004;134(1):9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009.

54. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12):e323. https://doi.org/10.2196/jmir.5870.
55. Anuragi A, Sisodia DS. Empirical wavelet transform based automated alcoholism detecting using EEG signal features. Biomed Signal Process Control. 2020;57:101777. https://doi.org/10.1016/j.bspc.2019.101777.
56. Vargas-Lopez O, Perez-Ramirez CA, Valtierra-Rodriguez M, Yanez-Borjas JJ, Amezquita-Sanchez JP. An explainable machine learning approach based on statistical indexes and SVM for stress detection in automobile drivers using electromyographic signals. Sensors. 2021;21(9):3155. https://doi.org/10.3390/s21093155.
57. Vimala V, Ramar K, Ettappan M. An intelligent sleep Apnea classification system based on EEG signals. J Med Syst. 2019;43(2):36. https://doi.org/10.1007/s10916-018-1146-8.
58. Sai CY, Mokhtar N, Yip HW, et al. Objective identification of pain due to uterine contraction during the first stage of labour using continuous EEG signals and SVM. Sādhanā. 2019;44(4):87. https://doi.org/10.1007/s12046-019-1058-4.
59. Kaplan AY, Fingelkurts AA, Fingelkurts AA, Borisov SV, Darkhovsky BS. Nonstationary nature of the brain activity as revealed by EEG/MEG: methodological, practical and conceptual challenges. Signal Process. 2005;85(11):2190–212. https://doi.org/10.1016/j.sigpro.2005.07.010.
60. Faisal AA, Selen LPJ, Wolpert DM. Noise in the nervous system. Nat Rev Neurosci. 2008;9(4):292–303. https://doi.org/10.1038/nrn2258.
61. Marathe AR, Ries AJ, McDowell K. Sliding HDCA: single-trial EEG classification to overcome and quantify temporal variability. IEEE Trans Neural Syst Rehabil Eng. 2014;22(2):201–11. https://doi.org/10.1109/TNSRE.2014.2304884.
62. Maniruzzaman M, Rahman MJ, Al-MehediHasan M, et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. J Med Syst. 2018;42(5):92. https://doi.org/10.1007/s10916-018-0940-7.
63. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
64. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13(2):281–305.
65. Géron A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent, 2nd ed. O'Reilly; 2019.
66. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing. 2020;415:295–316. https://doi.org/10.1016/j.neucom.2020.07.061.
67. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427–37. https://doi.org/10.1016/j.ipm.2009.03.002.
68. Assel M, Sjoberg DD, Vickers AJ. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagnostic Progn Res. 2017;1(1):19. https://doi.org/10.1186/s41512-017-0020-3.
69. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models. JAMA. 2017;318(14):1377. https://doi.org/10.1001/jama.2017.12126.
70. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. https://doi.org/10.1186/s12916-019-1466-7.
71. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Informatics Assoc. 2020;27(4):621–33. https://doi.org/10.1093/jamia/ocz228.
72. Combrisson E, Jerbi K. Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci Methods. 2015;250:126–36. https://doi.org/10.1016/j.jneumeth.2015.01.010.
73. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage. 2017;145:137–65. https://doi.org/10.1016/j.neuroimage.2016.02.079.
74. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14(1):40. https://doi.org/10.1186/1471-2288-14-40.
75. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.
76. Tivadar RI, Murray MM. A primer on electroencephalography and event-related potentials for organizational neuroscience. Organ Res Methods. 2019;22(1):69–94. https://doi.org/10.1177/1094428118804657.
77. Blankertz B, Tomioka R, Lemm S, Kawanabe M, Muller K. Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Process Mag. 2008;25(1):41–56. https://doi.org/10.1109/MSP.2008.4408441.
78. Rivet B, Souloumiac A, Attina V, Gibert G. xDAWN algorithm to enhance evoked potentials: application to brain-computer interface. IEEE Trans Biomed Eng. 2009;56(8):2035–43. https://doi.org/10.1109/TBME.2009.2012869.
79. Grosse-Wentrup M, Liefhold C, Gramann K, Buss M. Beamforming in noninvasive brain-computer interfaces. IEEE Trans Biomed Eng. 2009;56(4):1209–19. https://doi.org/10.1109/TBME.2008.2009768.
80. Magnetoencephalography SS. Basic principles. Ann Indian Acad Neurol. 2014;17(5):107. https://doi.org/10.4103/0972-2327.128676.
81. Bennington JY, Polich J. Comparison of P300 from passive and active tasks for auditory and visual stimuli. Int J Psychophysiol. 1999;34(2):171–7. https://doi.org/10.1016/S0167-8760(99)00070-7.
82. Dunning JP, Hajcak G. See no evil: directing visual attention within unpleasant images modulates the electrocortical response. Psychophysiology. 2009;46(1):28–33. https://doi.org/10.1111/j.1469-8986.2008.00723.x.
83. Hajcak G, MacNamara A, Foti D, Ferri J, Keil A. The dynamic allocation of attention to emotion: simultaneous and independent evidence from the late positive potential and steady state visual evoked potentials. Biol Psychol. 2013;92(3):447–55. https://doi.org/10.1016/j.biopsycho.2011.11.012.
84. Kam JWY, Xu J, Handy TC. I don't feel your pain (as much): the desensitizing effect of mind wandering on the perception of others' discomfort. Cogn Affect Behav Neurosci. 2014;14(1):286–96. https://doi.org/10.3758/s13415-013-0197-z.
85. Thielen J, Bosch SE, van Leeuwen TM, van Gerven MAJ, van Lier R. Evidence for confounding eye movements under attempted fixation and active viewing in cognitive neuroscience. Sci Rep. 2019;9(1):17456. https://doi.org/10.1038/s41598-019-54018-z.
86. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14(1):49–58. https://doi.org/10.1093/ckj/sfaa188.
87. Hutson M. Artificial intelligence faces reproducibility crisis. Science (80-). 2018;359(6377):725–6. https://doi.org/10.1126/science.359.6377.725.
88. Blankertz B, Lemm S, Treder M, Haufe S, Müller K-R. Single-trial analysis and classification of ERP components—a tutorial. Neuroimage. 2011;56(2):814–25. https://doi.org/10.1016/j.neuroimage.2010.06.048.
89. Mende-Siedlecki P, Lin J, Ferron S, Gibbons C, Drain A, Goharzad A. Seeing no pain: assessing the generalizability of racial bias in pain perception. Emotion. 2021;21(5):932–50. https://doi.org/10.1037/emo0000953.
90. Osborn J, Derbyshire SWG. Pain sensation evoked by observing injury in others. Pain. 2010;148(2):268–74. https://doi.org/10.1016/j.pain.2009.11.007.
91. Eroğlu K, Kayıkçıoğlu T, Osman O. Effect of brightness of visual stimuli on EEG signals. Behav Brain Res. 2020;382:112486. https://doi.org/10.1016/j.bbr.2020.112486.
92. Cao Y, Contreras-Huerta LS, McFadyen J, Cunnington R. Racial bias in neural response to others' pain is reduced with other-race contact. Cortex. 2015;70:68–78. https://doi.org/10.1016/j.cortex.2015.02.010.
93. Bas-Sarmiento P, Fernández-Gutiérrez M, Baena-Baños M, Correro-Bermejo A, Soler-Martins PS, de la Torre-Moyano S. Empathy training in health sciences: A systematic review. Nurse Educ Pract. 2020;44:102739. https://doi.org/10.1016/j.nepr.2020.102739.
94. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. CJEM. 2006;8(01):19–20. https://doi.org/10.1017/S1481803500013336.

Publisher's Note