Decoding Study-Independent Mind-Wandering from EEG using Convolutional Neural Networks

Christina Yi Jin*, Jelmer P. Borst, Marieke K. van Vugt

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Netherlands

yi.jin@rug.nl

j.p.borst@rug.nl

m.k.van.vugt@rug.nl

*Corresponding author: University of Groningen, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, Nijenborgh 9, 9747 AG Groningen, Netherlands.

E-mail address: yi.jin@rug.nl.

Abstract

Mind-wandering is a mental phenomenon where the internal thought process disengages from the external environment periodically. In the current study, we trained EEG classifiers using convolutional neural networks (CNN) to track mind-wandering. We transformed the input from raw EEG to band-frequency information (power), single-trial ERP patterns, and connectivity matrices between channels (based on inter-site phase clustering, ISPC). We used two independent EEG datasets for training and testing purposes (dataset A and B). In Experiment 1, we trained intra-subject models using the training dataset (dataset A) and compared the performance from three CNN architectures. We found that a simple CNN with two convolutional layers is sufficient. Accuracies varied across individual datasets and higher accuracies during validation were achieved with more unbalanced datasets, even though class size was balanced during training. In Experiment 2, we trained inter-subject models using dataset A and tested them on dataset B for an across-task prediction. Here, we used a stacking model to combine the outcomes of multiple input models. We achieved an accuracy of .68 during the across-task predictions, verifying the generalizability of our CNN. In both Experiment 1 and 2, the CNN performed best when trained on the raw EEG. Increased performance of the minority cases was found when training the models with a subset of more balanced datasets. Our study indicates the potential of training study-independent EEG classifiers with CNN using the raw EEG We also suggest that future EEG-machine learning studies should report the balancing status of the training dataset and the accuracies for each class.

Keywords: EEG, convolutional neural network, mind-wandering, balancing, across-study prediction

1 Introduction

Mind-wandering is a thought process that is characterized by being not directly relevant to the primary goals in the current context (Smallwood & Schooler, 2015). Mind-wandering tends to manifest itself as attentional lapses which often contribute to making errors in a task (Cheyne, Carriere, & Smilek, 2006). However, mind-wandering does not result in errors in all cases. Sometimes people can handle their primary tasks well when they start enjoying these periods of self-distraction as a temporary escape from the current situation (Schooler et al., 2011). This positive effect of mind-wandering is particularly common when the current task has a low cognitive load—in other words, performance can be achieved automatically with little executive control involved (Randall, Beier, & Villado, 2019) and therefore low performance is not a fool-proof indicator of mind-wandering. Another behavioral measure that has been proposed to characterize mind-wandering is increased response time variability. Several studies have shown that even when no obvious mistakes are observed, participants display increased variance in their response times when their mind wanders (Bastian & Sackur, 2013; Seli, Cheyne, & Smilek, 2013).

In addition to using behavior, mind-wandering can also be detected by means of physiological and neural measures. For example, researchers found that the pupil response to stimuli becomes smaller in an off-task state (Huijser, van Vugt, & Taatgen, 2018), possibly related to a vigilance decrement that tends to co-occur with mind-wandering (Unsworth & Robison, 2016). On the level of the cerebral cortex, mind-wandering appears to be associated with inhibited sensory processing to the visual stimuli, referred to as 'perceptual decoupling' (Schooler et al., 2011). This perceptual decoupling manifests itself in the EEG as a reduced P1 and increased alpha power (frequency range 8.5 ~ 12Hz) observed at the parietal-occipital regions (Compton, Gearinger, & Wild, 2019; Jin, Borst, & van Vugt, 2019; Kam & Handy, 2013). In fMRI studies, mind-wandering is associated with increased activation of the default mode network (DMN), together with changes in the connectivity between the DMN and other networks (Christoff, Gordon, Smallwood, Smith, & Schooler, 2009; Ho et al., 2019). Through indicating a memory retrieval process, the involvement of the DMN supports the functional role of mind-wandering as "spontaneous future cognition" (Cole & Kvavilashvili, 2019), potentially aiding problem-solving and creativity (Schooler et al., 2011).

Given this set of neural and physiological correlates of mind-wandering, several studies have explored the possibility of predicting mind-wandering on a single-trial level using machine learning. Mittner and colleagues used multimodal signatures of the (co)activations in the DMN, ACN (anti-correlated network), and the pupil diameter as the features for a machine learning model to predict mind wandering. They found that neural data could reliably predict mind-wandering with a median accuracy of 79.7% using leave-one-participant-out cross-validation (LOPOCV). They also found that mind-wandering was not linked to DMN, CAN or pupil diameter alone, but that instead all the features were necessary for the optimal predictive performance (Mittner et al., 2014). This seems sensible, given that the contents of mind-wandering can vary considerably. In their recent work, they achieved a 65% accuracy across participants by training another multimodality machine learning model (Groot et al., 2020).

Kawashima and Kumano (2017) found that training an EEG-based mind-wandering classifier with a non-linear support vector machine (SVM) performed better than the linear SVM. They also found

training with a selected subset of electrodes from frontal and parietal-occipital regions performs better than using all the electrodes. This suggests that the correlation between EEG and mental state fluctuations is more complex than a simple linear relationship and a subset of electrodes at key positions might contain sufficient information for discriminating mind-wandering. Jin et al. (2019) explored whether it was possible to predict mind wandering on the basis of various features of EEG data, ranging from power and inter-site phase clustering in the alpha and theta bands, to single-trial ERPs. They were able to predict mind-wandering with an average accuracy of 64% for a sustained attention to response task, and 69% for a visual search task. Additionally, they performed across-task predictions between the sustained attention to response and visual search tasks, with an accuracy of 59% and 60% (Jin et al., 2019). In a related study, the researchers used ICA-decomposed EEG in the alpha band to predict the occurrence of mind-wandering—albeit this time in a model that generalized across participants. They reported again an average accuracy of around 60% when testing the models on a left-out dataset (LOPOCV), for both within- and across-task predictions (Jin, Borst, & van Vugt, 2020).

Several conclusions can be drawn from the studies above. The accuracy with EEG seemed to hit a ceiling at 60% when making generalization predictions either during inter-subject modelling or when making across-task predictions of mind-wandering states. One notable feature of these studies was that most of their predictions were done on the basis of features that were derived from the prior literature on mind-wandering. It is possible that other spatial and frequency points could provide better performance when included in the classifier. Another limitation of these studies is that SVM was used as the algorithm to train the classifiers. SVM—when used with a nonlinear kernel—is a powerful tool for learning features following non-linear relationships, but its computational cost is relatively high. Moreover, when datasets become very large, it becomes inconvenient to train SVMs. Thus, the original time series are often reduced to features when training SVM models, as was done in the previous mind-wandering classification studies.

In the current study, we seek to address these problems by re-analyzing the data from Jin et al. (2019) and Jin et al. (2020) using convolutional neural networks (CNN, Roy et al., 2019) that can help to address the aforementioned points. CNN is one of the deep neural networks rapidly gaining in popularity. It has achieved remarkable performance in modelling large datasets, because it is able to learn from data that has only undergone minimal preprocessing. Such minimal processing requirements are helpful since our objective is to build a CNN classifier that can do reliable inter-subject classification, which would potentially be able to assist in detecting mind-wandering in an online setting in the future. When working with the EEG, CNNs have proven to be efficient in decoding the raw data, rather than relying on hand-crafted EEG features (reviewed by Roy et al., 2019). This makes CNNs an interesting candidate in the current study.

Although providing advantages in handling the EEG in a data-driven way, CNNs also have their shortcomings when it comes to interpreting the results of the classification. Because a CNN architecture relies on hidden layers for its predictions, it is often not clear how the network learns to predict. There are some ways to tackle this problem (Roy et al., 2019). In the current study, we will address this by plotting the feature maps – a way to inspect the detected features for the given inputs (Brownlee, 2019).

Another potential issue is not limited to CNNs but applies to all kinds of machine learning – the generalizability. Intra-subject models are often easier to train and perform better. Nevertheless, for real applications, pre-trained inter-subject models are required. It is challenging to train inter-subject models given the substantial individual variability in brain signals. Also, scalp EEG has a relatively low signal-to-noise ratio compared to other measures like intracranial EEG or fMRI. Thus, it is common to

observe a decline in accuracy when predicting the left-out datasets using the inter-subject models compared to intra-subject models (Roy et al., 2019). However, given that we would like to develop a biomarker for mind-wandering, inter-subject classification is required.

Given that we focus on generalizability, we also have to take another concern into account: the context in which the training datasets were collected. Even with inter-subject modelling, the training and validation data samples were often collected during the same experiment, meaning the same context. EEG are sensitive physiological measures so that experiment settings may influence various features of the data such as drifts, general amplitude and variance. Models predicting task-independent mental phenomena or stable personality traits should be tested on data from different experimental contexts to guarantee their generalizability. Deep learning has been proposed to improve generalization across contexts (Gemein et al., 2020; Xu et al., 2020).

The goal of the current study is to examine whether predictions of mind-wandering state on the basis of EEG data can be improved with CNNs. Apart from possibly improving decoding accuracy, CNNs may also help us obtain a more generalizable EEG decoder that can work not only across tasks and individuals, but also across experiments. To assess this generalizability in our study, we are making use of two datasets from two independent experiments. We will train the data from Jin et al. (2019) and test the models on the data from Jin et al. (2020). In both studies, participants performed a sustained attention to response task (SART) and a visual search task, but the design, stimuli, and participants were different between the studies.

This paper is structured as follows: we will first discuss the general experimental methods and data preparation. This will be followed by two experiments: in Experiment 1 we investigate the best model and EEG input type for classifying mind-wandering from EEG by training intra-subject models, while in Experiment 2 we test the generalization across datasets by training inter-subject models. We finish with a general discussion.

2 Method

2.1 Datasets

2.1.1 Training dataset

The training dataset is derived from Jin et al. (2019). Thirty participants (13 females, ages 18–30 years, M= 23.33, SD = 2.81) took part in the study. They performed a visual search task and a SART for six blocks each in two sessions (Figure 1a). The main stimuli in the SART were English words that occurred in lowercase for 89% of the time and in uppercase for 11% of the time. Participants were required to press “m” whenever they saw a lowercase word and to withhold their response when an uppercase word appeared. In the visual search task, participants were given a target shape to search at the beginning of each block. They were required to look for the target in each trial of that block and indicated if the target was in the search panel (yes/no) by pressing the left or right arrow key corresponding to their response. There was an equal probability of the target-present and the target-

absent trials. A SART block had 135 trials and a visual search block had 140 trials. Details can be found in the original study (Jin et al., 2019).

Participants were interrupted by probe questions asking them about the content of their thinking at that moment (Figure 1a). They could respond to this question with one of six options: (1) I was entirely concentrated on the ongoing task; (2) I evaluated aspects of the task (e.g., my performance or how long it takes); (3) I thought about personal matters; (4) I was distracted by my surroundings (e.g., noise, temperature, my physical condition); (5) I was daydreaming, thinking of task-unrelated things; (6) I was not paying attention, but my thought wasn't anywhere specifically. Each task had 54 probes that were interspersed with the trials. Two consecutive probes were separated by 7–24 trials, roughly accounting for 34–144 seconds. In the analysis, response 1 and 2 were labelled as an on-task state; response 3 and 5 were labelled as a mind-wandering state, and the other responses were ignored, following the same classification rule as the original work.

2.1.2 Testing dataset

The testing dataset is derived from Jin et al. (2020). Thirty participants (16 females, age 18–31 years, M = 23.73, SD = 3.47) took part in the study (Figure 1b). In the visual search task, participants either counted the specified target in the following search panel and indicated their response by pressing the number key (counting condition), or passively viewed the search panel and pressed 'j' as a standard response (non-counting condition). In the SART, participants viewed single digits ranging between 1 to 9 drawn from a uniform distribution. They pressed 'j' whenever they saw a digit other than '3'; if '3' appeared (11%), they were to withhold their response. The visual search task had 21 trials in each block and 20 blocks in total. The SART had twelve blocks with a block length that varied between two to seven repetitions of the nine digits (18-63 trials). Further details can be found in the original study (Jin et al., 2020).

In this dataset, the probes appeared at the end of each block (which were much shorter than in the training dataset). Participants indicated their momentary attentional state on a rating scale of -5 to 5 with anchor '-5' for 'totally mind-wandering', '-2' for 'mind-wandering', '0' for 'uncertain', '2' for 'focused', and '5' for 'highly focused'. Thus, positive ratings were classified as on-task and negative ratings as mind-wandering.

[Figure 1 insert]

Figure 1 Trial sequences of the SART and the visual search (VS) task, and the probe illustrations for the (a) training and (b) testing dataset. The orange frame indicates the main stimuli in the tasks, around which the EEG epochs were extracted (each epoch comprised the period starting at stimulus onset).

2.2 EEG preparation

2.2.1 Recording hardware

The training dataset had 128 channels, while the testing dataset had 32 channels. All electrode locations were within the International 10-20 System. In the current analysis, we only considered the 32 channels that overlapped between the studies (Figure 2a). The data were recorded using the Biosemi ActiveTwo recording system. The online sampling rate was 512Hz. The Biosemi hardware does not have any high-pass filtering. An anti-aliasing filtering is performed in the ADC's decimation filter (https://www.biosemi.com/faq/adjust_filter.htm).

2.2.2 Preprocessing

Data had already been preprocessed in the original studies, and we reuse these preprocessed data here. This offline EEG preprocessing was done with the EEGLAB toolbox in MATLAB. Continuous EEG was re-referenced to the average signal of both mastoids. The band-bass filtering was set to be 0.5-40 Hz and 0.1-42 Hz for the training and testing datasets, respectively. Both datasets were down-sampled to 256 Hz. The original segmentation was [-400 1200] ms with respect to stimulus onset for the training dataset, and [-1000 3000] ms for the testing dataset. In the current study, we took the overlapping [-400 1000] ms as the time window for analysis. The 200ms before the stimulus onset was used as the baseline. Ocular artifacts were detected and removed using the infomax independent component analysis (ICA).

2.2.3 Transformation

Given that we do not limit ourselves to particular channels of interest, we will train on the EEG time series from the entire 32-channel dataset. Apart from the raw EEG data, three kinds of transformations were performed on the original EEG matrix (Figure 2b) – power (Figure 2c) and inter-site phase clustering (ISPC, Figure 2d) derived from a time-frequency decomposition using the complex Morlet wavelets (Cohen, 2014), and single-trial ERP (stERP, Figure 2e) analysis (Bostanov, 2004; Bostanov & Kotchoubey, 2006; Jin et al., 2019). These transformations were performed in MATLAB.

A complex Morlet wavelet is created using the equation

$$cmw = e^{-t^2/2s^2} e^{i2\pi ft}$$

$$s = \frac{n}{2\pi f}$$

where f denotes the frequency in Hz, and n refers to the number of wavelet cycles. Convolving the complex wavelet with the original signal gives complex dot products, from which the amplitude can be extracted through the vector length of each complex data point (power is obtained from the squared amplitudes, Figure 2c, and the phase angle is obtained from the angle with respect to the positive real axis).

A set of wavelets was created with frequencies ranging from 4 to 40 Hz with 35 frequency sampling points in the logarithmic space, covering multiple frequencies in each of the theta/alpha/beta/gamma bands. Delta oscillations (1-3Hz) were excluded because our time window of 1400ms was not long

enough to estimate delta power with sufficient accuracy. The upper boundary of the analyzed frequency was limited by the bandpass filtering during the preprocessing. The number of cycles used for the wavelet increased from 3 to 7 in a logarithmic spacing on the frequency axis.

Inter-site phase clustering (ISPC) was computed as a measure of the connectivity between channels:

$$ISPC_f = |n^{-1}\sum_{t=1}^{n} e^{i(\Phi_{xt}-\Phi_{yt})}|$$

in which Φ_x and Φ_y are the phase angles from electrode x and y. ISPC is computed as the averaged phase angle differences (which were also mapped to the complex plane) in a moving time window. The length of the averaged angle difference denotes the clustering (i.e. the more the phase difference remains constant, the more their length adds up). The window length increased from 3 to 5 cycles linearly with the frequency. As shown in Figure 2d, there is an empty region of half the wave length at both sides of each frequency with no averaging data. Those regions were set to zero during machine learning. The ISPC was computed in a 16-channel layout resulting in 120 channel pairs (Figure 2a, red font) to reduce the number of features for the classifier.

Both the power and ISPC matrices were down-sampled to 50Hz given the fact that time-frequency analysis smears out the signal over time.

The single-trial ERP analysis, which quantifies the bumps in the EEG in terms of temporal location and amplitude was performed through computing the cross-covariance between the signal and the kernel ψ(t):

$$W(s,t) = \frac{1}{\sqrt{s}}\int_{-\infty}^{\infty} f(\tau)\psi\left(\frac{\tau - t}{s}\right)d\tau$$

$$\psi(t) = (1 - 16t^2)e^{-8t^2}.$$

With two varying parameters - time lag τ and scale s (an indication of wavelength), the cross-covariances can be mapped to a contour graph (Figure 2e). In the current study, the time lag (τ) ranges from 0 ms (stimulus onset) to 1000 ms (end of the EEG epoch) following the same temporal resolution as the raw EEG. The scale (s) ranges from 1 to 2500 ms in logarithmic space with 300 sampling points. The obtained matrix is down-sampled to 65 sampling points in the time lag and 30 points in the scale (frequency dimension).

Together, this creates four types of inputs for the CNNs: raw EEG, power, ISPC, and stERP (Figure 2).

[Figure 2 insert]

Figure 2 (a) Channel locations. (b) Raw EEG of one trial where color indicates signal amplitude. C-E transformations. (a) The topoplot indicating the locations of the 32 channels. Red font highlights the subset 16 channels that were used for the ISPC. (b) One trial-normalized EEG. The y-axis denotes the channel, of which a subset are labelled. (c) Power matrix sliced by channels. Color indicates the magnitude of oscillatory power (d) Inter-site phase clustering (ISPC) matrix sliced by channel pairs. The blue region at both sides of each frequency contained no data originally and was set to be zero during the machine learning (e) The single-trial ERP (stERP) matrix sliced by channels. The z-axis in (c-e) was logarithmically scaled. The data shown were scaled, which restricted them to the interval [0,1]. (f) Summary of the input size for the networks where n refers to the trial number.

3 Experiment 1

The goal of Experiment 1 is to find the best model for training on EEG data and its transformations. This will be achieved through training intra-subject models to save the computational cost compared to inter-subject modelling. We also examine how performance is affected by different data transformations. Because there are large individual differences in brain and behavior, we will perform a clustering analysis on the achieved performance to group the datasets into participants that are similar. This then allows us to examine whether classification performance depends on individual differences.

3.1 Method

3.1.1 Neural network

In this experiment, the training and validation is performed within each dataset (intra-subject learning). There is no firm answer of how many layers or neurons should be included, thus we built three architectures for comparisons (Figure 3a). The base architecture (M2) has two convolutional layers each followed by a max pooling layer. This choice is based on the idea that shallow models can perform similarly to the deeper convolutional models and tend to be less susceptible to overfitting (Ba & Caruana, 2014). The filter number increases over the convolutional layers from 64 to 128. The activation function used in the convolutional layer is ReLU. Two fully connected layers with ReLU are used before the final output layer. The final output is produced by a softmax function on the output layer. M3 is similar to M2 except for the filter number, which is 16 in the second convolutional layer. Inspired the VGG-16 architecture (Simonyan & Zisserman, 2014), we made M1 repeat each of the convolutional layers in M2 twice and keep the rest of the structure the same. We used 1D convolutional layers when the raw EEG was the input and 2D convolutional layers when transformed EEG (power/ISPC/stERP) was the input.

Learning was achieved over 50 epochs with categorical cross-entropy as the loss function, and Adam as the optimization algorithm. The learning rate was 0.001 and the batch size was 500. Five-fold cross-validation was used to assess modelling performance. An averaged accuracy during the 5-fold cross-

validation over three randomization seeds will be reported for each model. The training data were balanced between the two classes (mind-wandering vs. on-task) using over-sampling randomization.

3.1.2 Label generation

Following our previous studies, we incorporated twenty seconds of data before each probe into the analysis; the data in this window were labelled as "on-task" or "mind-wandering" according to the probe responses. This twenty-second time window accounted for 4 trials in the visual search task and 3 trials in the SART. The trial count for each dataset ranged from 198 to 363 (M = 302, SD = 42.5) with a mind-wandering rate ranging from 0.04 to .92 (median = .32).

3.1.3 Statistics

The performance of the intra-subject models was compared through repeated measures ANOVAs with model and input type the main factors. A correlation matrix using the spearman correlation was computed between the models. The k-means (k=3) clustering analysis was performed to group the datasets according to their achieved training and validation accuracy.

3.2 Results

The training and validation accuracy are summarized in Figure 3b. A repeated measures ANOVA on validation accuracy with model and input as the main factors shows that there were main effects of both model and input (model: $F(2,58)=47.01, p<.001, \eta^2_{ges}=.03$; input: $F(3,87)=53.16, p<.001, \eta^2_{ges}=.15$). The interaction between model and input was also significant ($F(6,174)=29.31, p<.001, \eta^2_{ges}=.07$). M2 performed best compared to M1 and M3 (t(119)=*6.99 and 5.91, ps < .001*) while no difference was found between M1 and M3 (t(119)=2.12, p=.109). Training with raw EEG resulted in the highest classification accuracy (.71), followed by the stERP (.69), ISPC (.62) and power (.57). The difference between each pair was confirmed by paired t-tests (all *ts*(89)>2.85, *ps* < 0.032).

We then examined whether the different models were picking up on similar data features by computing correlations between their performance on the different participants. The spearman correlation between the performance of the various models is shown in Figure 3c. The worst model – M3 with power – exhibited no correlation with any of the other models. Most of the rest of the models were correlated with each other (*rs*>.47, *ps*<0.01).

To examine whether there were subgroups of participants for whom classification was better, we did a k-means clustering analysis with k=3 on the training and validation accuracies. Three clusters were identified (Figure 3d). Cluster 1 (center: [.98 .86], [training validation]) reflects participants achieving high training and validation accuracy. Cluster 2 reflects participants with high training and medium validation accuracy (center: [.94 .61]). Cluster 3 reflects participants with low training and validation accuracy ([.66 .51]).

The class size ratio (on-task vs. mind-wandering) in each cluster is shown in Figure 3c. Cluster 3 represents the models that were not well-trained, as can be also observed because Cluster 3 is absent

in the models with higher performance (raw EEG and stERP) while it accounts for all the datasets in the worst model (M3 with power). Cluster 1 comes from more unbalanced datasets compared to Cluster 2, as it is concentrated on either end of the class size ratio.

There are two possible explanations for this pattern of results: either the classifier learns to predict the majority class always, or it recognizes certain minority cases. If the first explanation holds, we expect to observe biased accuracies between classes; if the second explanation holds, we expect to observe high accuracies for both classes. We thus compared the harmonic mean of the sensitivity and specificity[1] between Cluster 1 and Cluster 2 and found that the harmonic mean of Cluster 1 (0.09) is significantly smaller than that of Cluster 2 (0.38, $t(195.86)=-18.56$, $p<0.001$), supporting the first explanation, that the high performance of Cluster 1 during validation is the result of a classification bias – even though the data was balanced before training.

[Figure 3 insert]

Figure 3 Experiment 1: (a) CNN architectures (b) averaged accuracies for training and validation grouped by model and input type (c) validation accuracy correlation matrix between the models (d) class size ratios (on-task vs. mind-wandering) for each cluster derived from k-means clustering on the modelling performance. The error bar in (b) denotes the 95% confidence interval.

3.3 Discussion

The goal of Experiment 1 was to explore what kind of convolutional neural structure can learn to detect mind-wandering from raw EEG and its transformations. We found that a relatively simple architecture with two convolutional layers and an neuron number that increased across layers (M2) is best suited to predicting mind-wandering. Reducing the filter (M3) or doubling the convolutional layers (M1) led to worse performance. We will therefore use the M2 architecture when training the inter-subject models in Experiment 2. In addition, we found that raw EEG and stERP worked better for classification than the time-frequency information. We will test whether the same classification architecture can generalize to a completely independent dataset with different participants and data recordings in Experiment 2.

The performance between models and input types was correlated in general, implying that overlapping information is used. The clustering results indicated that, except for model and input type, the class size ratio determined the network's performance. Surprisingly, it is easier for the classifier to achieve high performance with a more unbalanced dataset when making predictions always of the majority class even while the training dataset is balanced. This effect could potentially be due to the

[1] Sensitivity: true positive rate. Specificity: true negative rate. Harmonic mean of sensitivity and specificity = 2*sensitivity*specificity/(sensitivity+specificity).

lack of variety of the minority class since the balancing was performed through over-sampling randomization.

4 Experiment 2

The goal of Experiment 2 is to find the best spatial locations for each type of EEG input and build a stacking model using the input from the top three best locations for each input type. This will be achieved by training inter-subject models using dataset A and tested on dataset B (Figure 1). To examine the influence of the level of balance of the training dataset on the generalization of the classifiers, we built a model with a subset of relatively balanced datasets for comparison.

4.1 Method

4.1.1 Neural network

Modelling in Experiment 2 employed the architecture of M2 in Figure 3A. Two major changes were made compared to Experiment 1 – we trained inter-subject models rather than intra-subject models, and each individual channel (or channel pair) was the only input source for each model in the first step of Experiment 2. Five-fold cross-validation over datasets was used to assess performance. In each iteration, we kept 24 datasets for training and left 6 for testing. The entire five-fold cross-validation procedure was performed three times using different random seeds

In the second step of Experiment 2, we chose three top channels or channel pairs when sorting their performance in descending order for each type of input [2]. We used the outputs from those 12 models as the input for training two stacking models – a meta-learner learning the outcome from the multiple base models to improve the performance. Currently, we used two algorithms to train the stacking models for the final prediction of mind-wandering: a neural network and a logistic regression. Each base model gives a binary outcome forming an input of 24 dimensions altogether as the input for the stacking model. The neural network contained one hidden layer of 12 neurons and used ReLU as the activation function.

We performed across-task predictions using the obtained stacking models. In addition, we trained the same stacking model only with the datasets of Cluster 2 in raw_M2 as it was the best intra-subject model (Figure 3b) and Cluster 2 in this model contained the relatively balanced datasets (Figure 3d). The same five-fold cross-validation was employed to verify the modelling performance with datasets of Cluster 2. This dataset will be referred to as the 'Cluster 2 dataset'.

4.1.2 Label generation

[2] Sorting by the effect size (Cohen's d) when compared to .5 gave the same three channels/channel pairs for each type of input.

The label generation for the training dataset remained the same as in Experiment 1. The same twenty-second pre-probe time window was used to subset the data from the testing dataset. This twenty-second time window accounted for 2 trials in the visual search task and 6 trials in the SART in Study B. The trial count of each dataset ranged from 60 to 112 (M = 101, SD = 13) with mind-wandering rate ranging from 0 to .98 (median = .17).

4.2 Results

The performance of each single-channel network can be found in Supplementary Data. The best three channels and channels pairs for each type of input are shown in Figure 4a. Only channel C4 overlapped between two of the input types.

The performance of the stacking models is summarized for training, validation and across-study predictions (ASP) separately in Figure 4b. Statistical tests showed that there were no differences in classification performance between the networks using a stacked neural network and stacked logistic regression during both the validation and ASPs for all indices (all $|ts(14)| > 1.61$, $ps > 0.12$).

To investigate which input type was most important for classification, we examined the coefficients of the logistic regression model (Table 1). In correspondence with Experiment 1, higher weights were given to the outputs from the base models using raw EEG as input, followed by the stERPs (Table 1).

All models were able to predict mind-wandering significantly above chance level (.5) on the validation datasets and most importantly on the test dataset (Figure 4b, all $ts(14) > 4.85$, $ps < 0.001$), which was fully independent from the training dataset.

To investigate whether the results hold for more balanced datasets, we also trained the models only on the data from Cluster 2, the subset of participants with a more equal distribution of mind-wandering and on-task performance. As shown in Figure 4c, performance was significantly above chance for both the validation and across-study predictions (all $ts(14) > 6.31$, $ps < 0.001$).

A comparison between the modelling performance with the whole dataset and Cluster 2 alone shows the Cluster 2 models overall had worse accuracy than the whole dataset, both during cross-validation ($t(29) = -3.61$, $p = 0.001$) and testing phases ($t(29) = -5.66$, $p < 0.001$). Compared to the full dataset, the specificity dropped (all $ts(29) < -5.27$, $ps < 0.001$) but the sensitivity improved (all $ts(29) > 6.63$, $ps < 0.001$) when training on the Cluster 2 dataset.

[Figure 4 insert]

Figure 4 Experiment 2: (a) the three best-performed channel or channel pairs for each transformation that were used to train the stacking models. Performance of the stacking models using neural network (stacked_nn) and logistic regression (stacked_lr) as the hidden layer before the final output with (b) all the datasets from Study A and (c) only the datasets in Cluster 2 of raw_M2 in Figure 3. ASP = across-study predictions. Error bars denote one standard-error.

Table 1 Logistic regression coefficients of the stacking model. The outcome of each base model is binary and serves as the input for the stacking model.

Coefficient	Base model

		channel	Input type
-.81	.81	FC5	raw
-.81	.81	C4	raw
-.81	.81	FC2	raw
-.81	.81	Cz	stERP
-.77	.77	CP1	stERP
-.81	.81	C4	stERP
-.45	.45	AF3	power
-.67	.67	O2	power
-.61	.61	P3	power
-.75	.75	C3-T8	ISPC
-.81	.81	Oz-T8	ISPC
-.75	.75	O2-Fz	ISPC

4.3 Discussion

Overall, we were able to predict mind-wandering with 54-68% accuracy. The most predictive channels based on the raw EEG and stERP were more centrally distributed, and their activity may reflect either processing in working memory or involvement of the motor cortex (Smallwood, Beach, Schooler, & Handy, 2008). The most predictive channels based on power were located at both anterior and posterior sites. It matches previous studies in which mind-wandering could be detected from participants' alertness and visual processing (Kam et al., 2011). Using the synchrony between channels for predictions, the best channel pairs for ISPC included a frontal-occipital connection, an occipital-temporal connection, and a central-temporal connection.

No differences of performance were found between using logistic regression and a neural network for the stacking models, indicating a simple linear relationship is already enough to assemble the base models. The weights of the logistic regression stacking models were the largest for the base models using raw EEG. This is a finding consistent with our observations in the intra-subject modelling in Experiment 1 – suggesting that the CNN learns the best from the raw EEG rather than a power or single-trial ERP transformation. Naturally, all information is present in the raw EEG, but this shows that the CNN is able to leverage that information, unlike the SVMs we used before, for which inputs were precomputed features (Jin et al., 2019, 2020). There are both pros and cons for such a raw EEG+CNN combination as discussed in the introduction. On the one hand, it allows data scientists and engineers to train efficient EEG classifiers with little neuroscientific or EEG knowledge. On the other hand, the interpretation is often not straightforward, and therefore it provides less insight into what brain areas or other data features are most relevant for the classification. As such, classifier choice should be decided by the research goal.

We found that performance of the inter-subject models was above chance level during the cross-validation. More importantly, the models were also able to predict mind-wandering on the basis of EEG from an independent study, further demonstrating the generalizability of the classifiers. The overall accuracy during the ASPs was higher than validations (Figure 2b). Considering the equivalent sensitivity and specificity during both processes, the accuracy difference was likely an outcome of the class ratio for the training and testing datasets. The mind-wandering rate was .32 and .17 for the

training and testing datasets respectively, which makes sense when the overall accuracies increased on the testing dataset when the classifiers' specificity (i.e., the accuracy of predicting on-task trials) was relatively high.

Interestingly, we found that the class balancing in the original dataset had a strong effect on the network's performance. Training with an originally more balanced dataset had an increased sensitivity when detecting the minority cases even while the overall accuracy was lower for that dataset. Real-world applications should consider this influence according to the detection goal – whether to focus on the sensitivity to the minority cases or just on a higher overall accuracy.

5 General Discussion

In the current study, we used deep learning to predict the occurrence of mind-wandering on the basis of EEG data. We did both an intra-subject and inter-subject modeling, and used the raw EEG and its transformations (power and phase clustering information in the time-frequency domain and the ERP patterns on a single-trial level). We found that training CNNs with raw EEG gives the best performance compared to the other transformations. This finding held for both the intra- and inter-subject modelling. In Experiment 1, we found that a relatively simple CNN architecture with two hidden convolutional layers is sufficient for predicting mind-wandering from EEG data. In Experiment 2, we observed that the best EEG locations for classification differed between input types, implying that different neural markers were selected. The stacking models then combined the different inputs into an overall model, which was able to classify significantly above chance across two independent studies.

For the intra-subject modelling, we found the best performance was obtained from the most biased datasets. When training with biased datasets, the models have a high likelihood of making a majority class prediction which contributes to a high accuracy. For the inter-subject modelling, using an original balanced dataset for training could improve the detection accuracy of the minority cases, although it did result in lower overall accuracy. We thus recommend future BCI-like studies that use EEG in combination with machine learning to pay attention to this and report the class size ratio and the accuracy achieved in each class separately.

Our results indicated that it is not only possible to predict mind-wandering across individuals within the same study (Jin et al., 2020), but that it is even possible to make across-study predictions. This work demonstrates the generalizability of the CNN classifiers of mind-wandering. On the other hand, it also shows that accuracy of mind-wandering classifiers is capped at around 68%, even with very powerful computational methods. Even though we did substantial experimentation with different parameter settings, it is still possible that accuracy could be improved by adjusting other hyperparameters such as learning rate, class weights, regularization and batch normalizations, etc. (see Roy et al. (2019) for a more detailed discussion about hyperparameters). Other properties with the dataset could also be adjusted, e.g., the pre-probe time window. However, when exploring these parameters, we did not find clear improvements using the current datasets.

One obvious problem influencing the classifier's performance is that the neural data were labelled by the participants. Those self-reports were inevitably subjective and might even be biased across

individuals (Weinstein 2018), and became the main source of noise in the datasets. Currently, we have no alternative to self-reports for measuring the attentional state. Other physiological measurements, such as pupillometry, are also based on self-reports during the modelling (Pelagatti et al., 2018). Behavioral performance can provide participants with some feedback about the status of their attention but it is not totally associated since at times it is possible to perform the task without mistakes even when the mind wanders. In our previous work, we tried to manipulate participants' attentional state by creating different levels of task complexity, expecting to find increased mind-wandering rate during the easier conditions. We also expected to find a temporal effect on mind-wandering based on the assumption that low vigilance leads to more mind-wandering and vigilance declines over time. While the behavioral results followed these expectations that self-reported mind-wandering occurred more during an easy task and in the latter half of the experiment, the EEG data demonstrated that vigilance, task difficulty and mind-wandering were associated with unique brain states. This indicates those experimental conditions triggering mind-wandering cannot replace self-reports for labelling mind-wandering when training classifiers, even if behavior is somewhat associated with mind-wandering status (Jin et al., 2020). One possible way to compensate for the limitations of self-reports is to improve participants' meta-awareness of their mental state, and to increase their sensitivity to their internal experience for a more accurate description of their conscious world (Seli et al., 2018), for example, through micro-phenomenology (Petitmengin, 2017).

6 Conclusion

Overall, we achieved a 68% accuracy when testing the mind-wandering classifiers on EEG collected in an independent study. We also found raw EEG as the best input for the convolutional neural network to learn. The current performance is constrained by the class balancing state in the original dataset. We suggest that a possible venue for improving classifier accuracy is to improve the quality of self-report labels, for example by training participants to be aware of their own mental state.

References

Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep?. In Advances in neural information processing systems (pp. 2654-2662).

Bastian, M., & Sackur, J. (2013). Mind wandering at the fingertips: automatic parsing of subjective states based on response time variability. *Frontiers in Psychology, 4*: 573. doi:10.3389/fpsyg.2013.00573

Bostanov, V. (2004). BCI competition 2003-data sets Ib and IIb: feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram. *IEEE Transactions on Biomedical Engineering, 51*(6), 1057-1061. doi:10.1109/TBME.2004.826702

Bostanov, V., & Kotchoubey, B. (2006). The t-CWT: a new ERP detection and quantification method based on the continuous wavelet transform and Student's t-statistics. *Clinical Neurophysiology, 117*(12), 2627-2644. doi:10.1016/j.clinph.2006.08.012

Brownlee, J. (2019). *Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python*: Machine Learning Mastery.

Cheyne, J. A., Carriere, J. S. A., & Smilek, D. (2006). Absent-mindedness: Lapses of conscious awareness and everyday cognitive failures. *Consciousness and Cognition, 15*(3), 578-592. doi:10.1016/j.concog.2005.11.009

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences of the United States of America, 106*(21), 8719-8724. doi:10.1073/pnas.0900234106

Cohen, M. X. (2014). *Analyzing neural time series data: theory and practice*. Cambridge, MA: MIT Press.

Cole, S., & Kvavilashvili, L. (2019). Spontaneous future cognition: The past, present and future of an emerging topic. *Psychological Research,* 83, 631–650. doi:10.1007/s00426-019-01193-3

Compton, R. J., Gearinger, D., & Wild, H. (2019). The wandering mind oscillates: EEG alpha power is enhanced during moments of mind-wandering. *Cognitive, Affective, & Behavioral Neuroscience, 19*(5), 1184-1191.

Gemein, L. A. W., Schirrmeister, R. T., Chrabąszcz, P., Wilson, D., Boedecker, J., Schulze-Bonhage, A., Hutter, F., & Ball, T. (2020). Machine-learning-based diagnostics of EEG pathology. *NeuroImage, 220*: 117021. doi:https://doi.org/10.1016/j.neuroimage.2020.117021

Groot, J. M., Boayue, N. M., Csifcsák, G., Boekel, W., Huster, R., Forstmann, B. U., & Mittner, M. (2020). Probing the neural signature of mind wandering with simultaneous fMRI-EEG and pupillometry. *NeuroImage*, *224*: 117412. doi: 10.1016/j.neuroimage.2020.117412

Ho, N. S. P., Wang, X., Vatansever, D., Margulies, D. S., Bernhardt, B., Jefferies, E., & Smallwood, J. (2019). Individual variation in patterns of task focused, and detailed, thought are uniquely associated within the architecture of the medial temporal lobe. *NeuroImage*, *202*: 116045. doi:10.1016/j.neuroimage.2019.116045

Huijser, S., van Vugt, M. K., & Taatgen, N. A. (2018). The wandering self: Tracking distracting self-generated thought in a cognitively demanding context. *Consciousness and Cognition, 58*, 170-185. doi:10.1016/j.concog.2017.12.004

Jin, C. Y., Borst, J. P., & van Vugt, M. K. (2019). Predicting task-general mind-wandering with EEG. *Cognitive, Affective, & Behavioral Neuroscience, 19*(4), 1059-1073. doi:10.3758/s13415-019-00707-1

Jin, C. Y., Borst, J. P., & van Vugt, M. K. (2020). Distinguishing vigilance decrement and low task demands from mind-wandering: A machine learning analysis of EEG. *European Journal of Neuroscience*. doi: 10.1111/ejn.14863

Kam, J. W. Y., Dao, E., Farley, J., Fitzpatrick, K., Smallwood, J., Schooler, J. W., & Handy, T. C. (2011). Slow fluctuations in attentional control of sensory cortex. *Journal of Cognitive Neuroscience, 23*(2), 460-470. doi:10.1162/jocn.2010.21443

Kam, J. W. Y., & Handy, T. C. (2013). The neurocognitive consequences of the wandering mind: a mechanistic account of sensory-motor decoupling. *Frontiers in Psychology, 4*: 725. doi:10.3389/fpsyg.2013.00725

Kawashima, I., & Kumano, H. (2017). Prediction of Mind-Wandering with Electroencephalogram and Non-linear Regression Modeling. *Frontiers in Human Neuroscience*, *11*: 365. doi:10.3389/fnhum.2017.00365

Mittner, M., Boekel, W., Tucker, A. M., Turner, B. M., Heathcote, A., & Forstmann, B. U. (2014). When the brain takes a break: a model-based analysis of mind wandering. *Journal of Neuroscience, 34*(49), 16286-16295. doi:10.1523/jneurosci.2062-14.2014

Randall, J. G., Beier, M. E., & Villado, A. J. (2019). Multiple routes to mind wandering: Predicting mind wandering with resource theories. *Consciousness and Cognition, 67*, 26-43. doi:10.1016/j.concog.2018.11.006

Pelagatti, C., Binda, P., & Vannucci, M. (2018). Tracking the Dynamics of Mind Wandering: Insights from Pupillometry. *Journal of Cognition*, *1*(1), 38-38. doi:10.5334/joc.41

Petitmengin, C. (2017). Exploring the hidden side of lived experience through Micro-phenomenology. Paper presented at the xTalks: Digital Discourses, Massachussets Instutute Of Technology, Boston, United States. https://hal.archives-ouvertes.fr/hal-01663598

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, *16*(5), 051001. doi:10.1088/1741-2552/ab260c

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences, 15*(7), 319-326. doi:10.1016/j.tics.2011.05.006

Seli, P., Cheyne, J. A., & Smilek, D. (2013). Wandering minds and wavering rhythms: Linking mind wandering and behavioral variability. *Journal of Experimental Psychology: Human Perception and Performance, 39*(1), 1-5. doi:10.1037/a0030954

Seli, P., Smilek, D., Ralph, B. C. W., & Schacter, D. L. (2018). The awakening of the attention: Evidence for a link between the monitoring of mind wandering and prospective goals. *Journal of Experimental Psychology-General, 147*(3), 431-443. doi:10.1037/xge0000385

Sherkatghanad, Z., Akhondzadeh, M., Salari, S., Zomorodi-Moghadam, M., Abdar, M., Acharya, U. R., Khosrowabadi, R., & Salari, V. (2020). Automated detection of autism spectrum disorder using a convolutional neural network. *Frontiers in Neuroscience, 13*: 1325. doi:10.3389/fnins.2019.01325

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Smallwood, J., Beach, E., Schooler, J. W., & Handy, T. C. (2008). Going AWOL in the brain: Mind wandering reduces cortical analysis of external events. *Journal of Cognitive Neuroscience, 20*(3), 458-469. doi:10.1162/jocn.2008.20.3.458

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: empirically navigating the stream of consciousness. *Annual Review of Psychology, 66*, 487-518. doi:10.1146/annurev-psych-010814-015331

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective, & Behavioral Neuroscience, 16*(4), 601-615. doi:10.3758/s13415-016-0417-4

Weinstein, Y. (2018). Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods, 50*(2), 642-661. doi:10.3758/s13428-017-0891-9

Xu, L., Xu, M., Ke, Y., An, X., Liu, S., & Ming, D. (2020). Cross-dataset variability problem in EEG decoding with deep learning. *Frontiers in Human Neuroscience, 14*: 103. doi:10.3389/fnhum.2020.00103

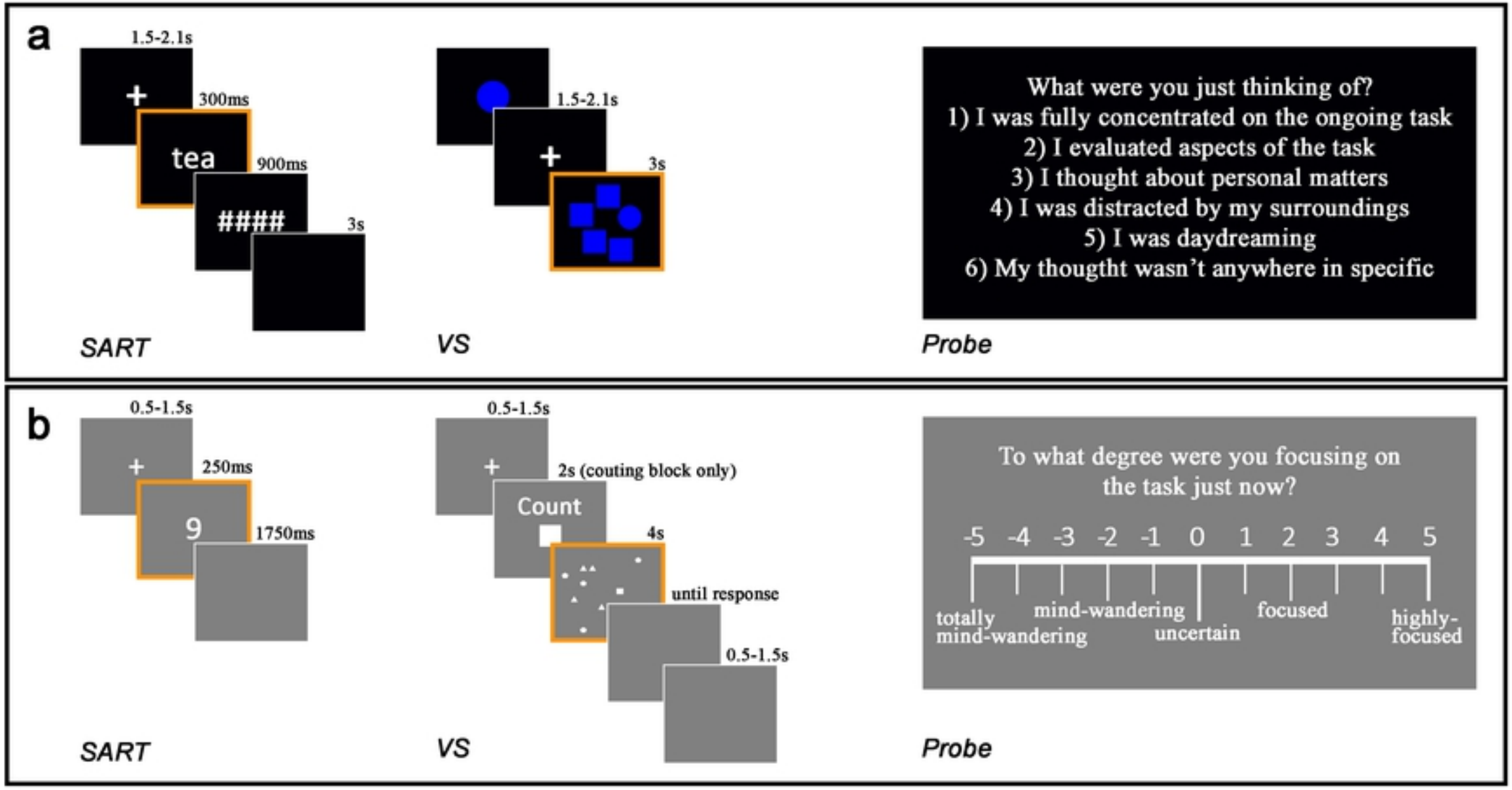
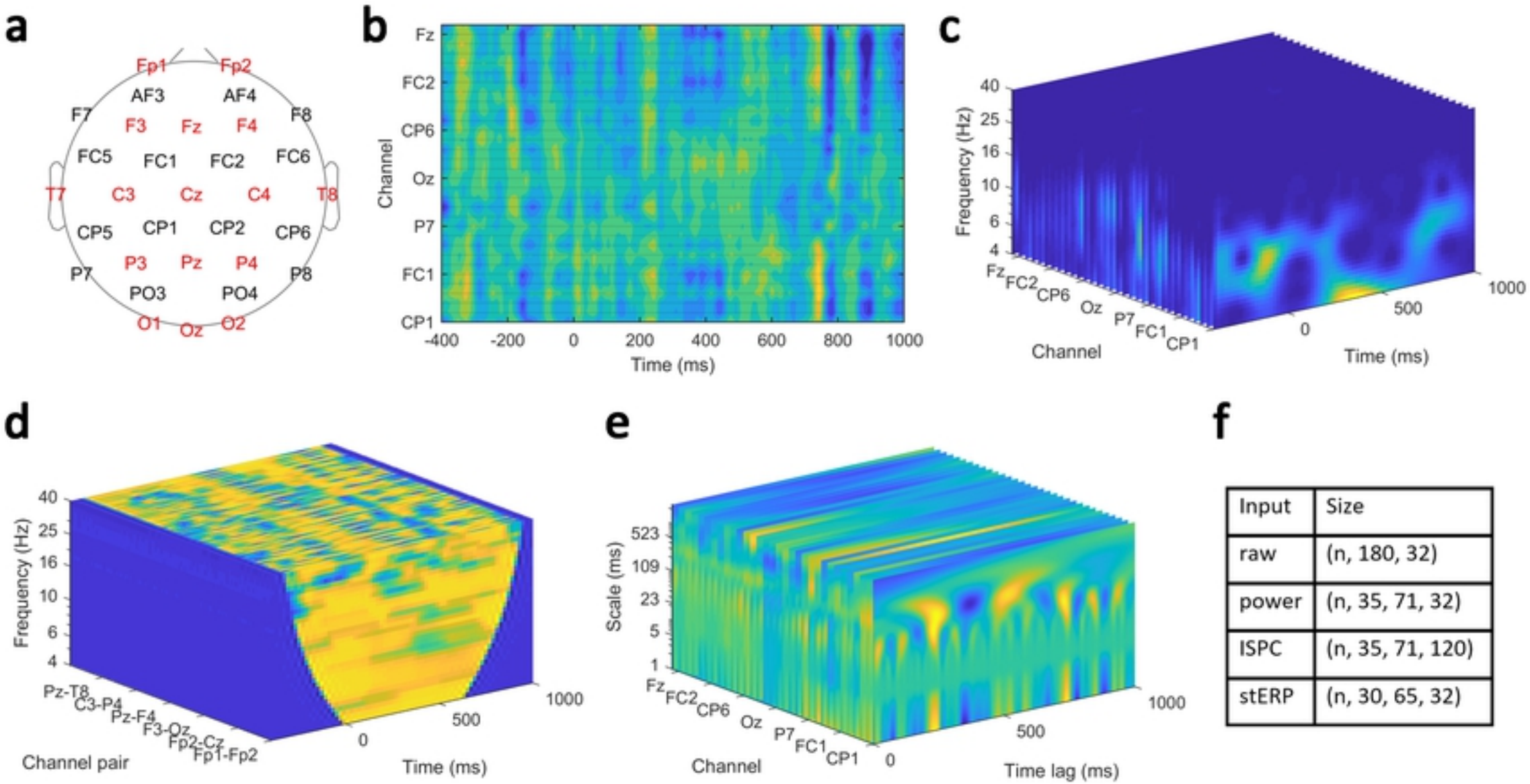

a
1.5-2.1s
+
300ms
tea
900ms
####
3s
SART
1.5-2.1s
+
3s
VS
What were you just thinking of?
1) I was fully concentrated on the ongoing task
2) I evaluated aspects of the task
3) I thought about personal matters
4) I was distracted by my surroundings
5) I was daydreaming
6) My thougtht wasn't anywhere in specific
Probe
b
0.5-1.5s
+
250ms
9
1750ms
SART
0.5-1.5s
+
2s (couting block only)
Count
4s
until response
0.5-1.5s
VS
To what degree were you focusing on the task just now?
-5 -4 -3 -2 -1 0 1 2 3 4 5
totally mind-wandering
mind-wandering
uncertain
focused
highly-focused
Probe

Figure 1



Input	Size
raw	(n, 180, 32)
power	(n, 35, 71, 32)
ISPC	(n, 35, 71, 120)
stERP	(n, 30, 65, 32)

Figure 2

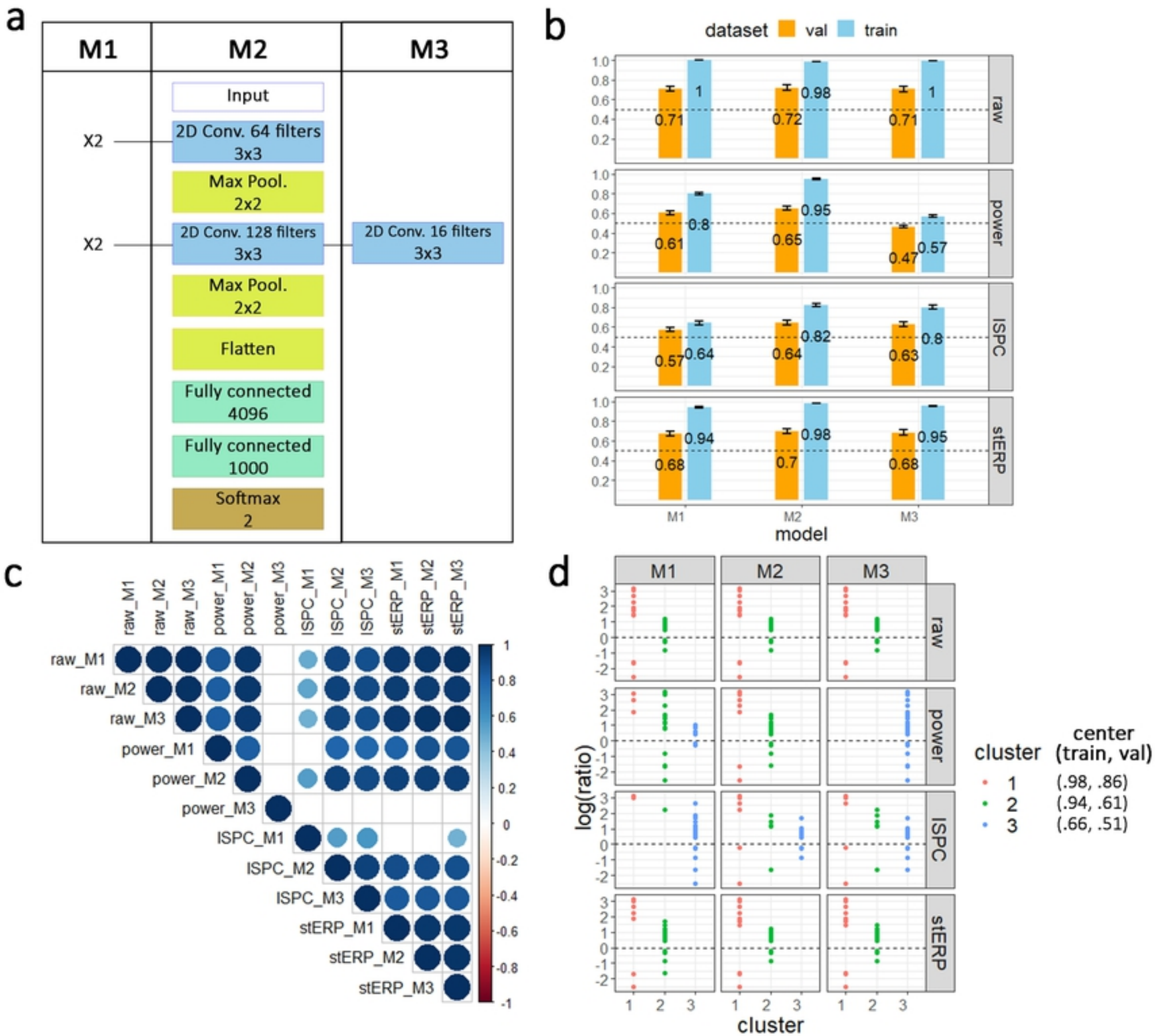

a
M1
M2
M3
Input
X2
2D Conv. 64 filters 3x3
Max Pool. 2x2
X2
2D Conv. 128 filters 3x3
2D Conv. 16 filters 3x3
Max Pool. 2x2
Flatten
Fully connected 4096
Fully connected 1000
Softmax 2
b
dataset
val
train
raw
power
ISPC
stERP
M1
M2
M3
model
c
raw_M1
raw_M2
raw_M3
power_M1
power_M2
power_M3
ISPC_M1
ISPC_M2
ISPC_M3
stERP_M1
stERP_M2
stERP_M3
d
log(ratio)
cluster
center (train, val)
1 (.98, .86)
2 (.94, .61)
3 (.66, .51)

Figure 3

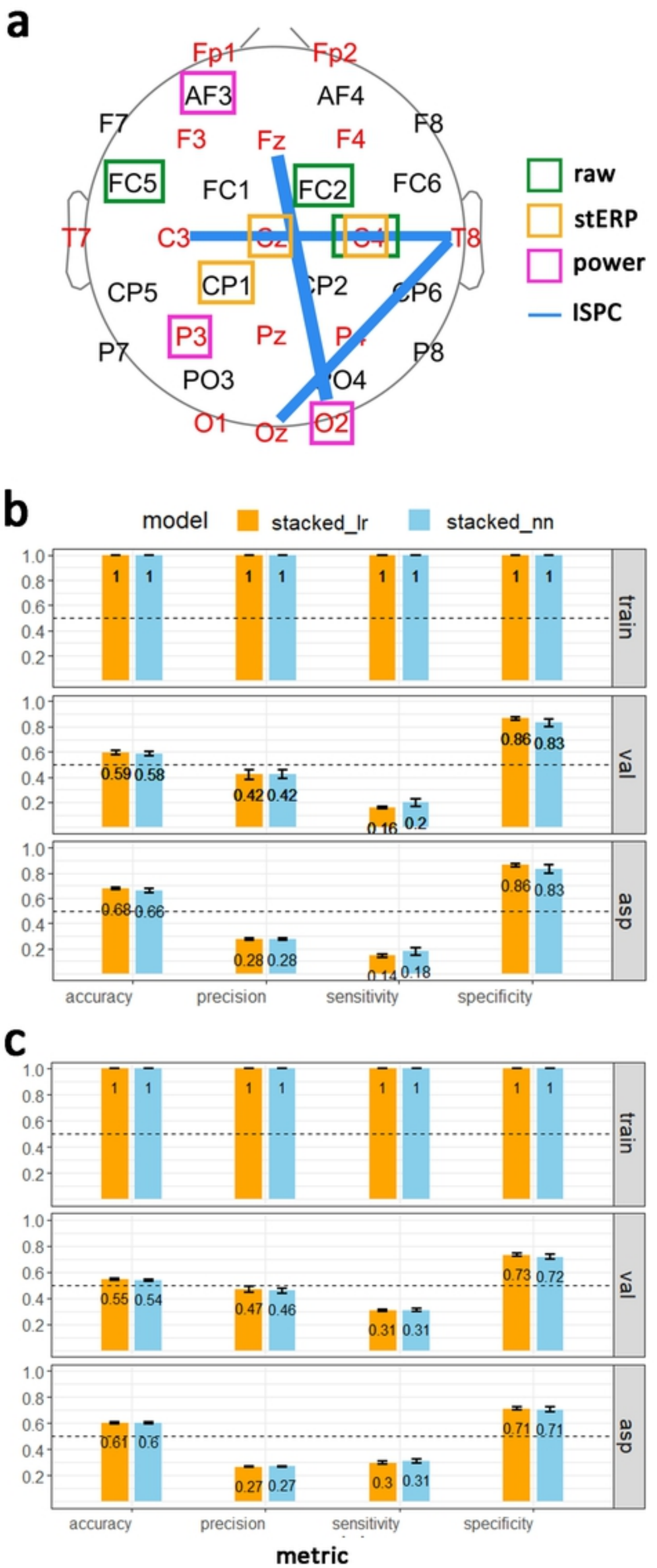

a
Fp1
Fp2
AF3
AF4
F7
F3
Fz
F4
F8
FC5
FC1
FC2
FC6
T7
C3
Cz
C4
T8
CP5
CP1
CP2
CP6
P3
Pz
P7
P8
PO3
PO4
O1
Oz
O2
raw
stERP
power
ISPC
b
model
stacked_lr
stacked_nn
train
val
asp
accuracy
precision
sensitivity
specificity
0.59 0.58
0.42 0.42
0.16 0.2
0.86 0.83
0.68 0.66
0.28 0.28
0.14 0.18
c
0.55 0.54
0.47 0.46
0.31 0.31
0.73 0.72
0.61 0.6
0.27 0.27
0.3 0.31
0.71 0.71
metric

Figure 4