Finding neural correlates of depersonalisation/derealisation disorder via explainable CNN-based analysis guided by clinical assessment scores

Abbas Salami[1], **Javier Andreu-Perez**[1,2,3], **Helge Gillmeister**[2,4]

[1]School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, U.K.
[2]Centre for Computational Intelligence, Smart Health Technologies Group, Institute of Public Health and Wellbeing, University of Essex, Colchester, CO4 3SQ, U.K.
[3]Simbad2, Department of Computer Science, University of Jaén, 23071 Jaen, Spain
[4]Department of Psychology, University of Essex, Colchester, CO4 3SQ, U.K.

Authors Email Addresses and ORCID IDs:

Abbas Salami (a.salami@essex.ac.uk) / 0000-0003-2156-8554
Javier Andreu-Perez* (j.andreu-perez@essex.ac.uk) / 0000-0002-7421-4808
Helge Gillmeister (helge@essex.ac.uk) / 0000-0001-5999-5303

*Corresponding author

Conflict of Interest Disclosure

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Keywords:

Depersonalisation/derealisation disorder, EEG, Biomarker, Convolutional neural network, Clustering

Key points:

- The first EEG-based biomarker discovery system for depersonalisation/derealisation disorder is proposed.
- The system is made up of a novel multi-input multi-output deep neural network, which requires no prior knowledge of the disorder.
- An explainable visualisation approach is proposed to investigate the neural correlates associated with depersonalisation/derealisation symptoms.

ABSTRACT

Mental health disorders are typically diagnosed based on subjective reports (e.g., through questionnaires) followed by clinical interviews to evaluate the self-reported symptoms. Therefore, considering the interconnected nature of psychiatric disorders, their accurate diagnosis is a real challenge without indicators of underlying physiological dysfunction. Depersonalisation/derealisation disorder (DPD) is an example of dissociative disorder affecting 1–2 % of the population. DPD is characterised mainly by persistent disembodiment, detachment from surroundings, and feelings of emotional numbness, which can significantly impact patients' quality of life. The underlying neural correlates of DPD have been investigated for years to understand and help with a more accurate and in-time diagnosis of the disorder. However, in terms of EEG studies, which hold great importance due to their convenient and inexpensive nature, the literature has often been based on hypotheses proposed by experts in the field, which require prior knowledge of the disorder. In addition, participants' labelling in research experiments is often derived from the outcome of the Cambridge Depersonalisation Scale (CDS), a subjective assessment to quantify the level of depersonalisation/derealisation, the threshold and reliability of which might be challenged. As a result, we aimed to propose a novel end-to-end EEG processing pipeline based on deep neural networks for DPD biomarker discovery, which requires no prior hand-crafted labelled data. Alternatively, it can assimilate knowledge from clinical outcomes like CDS as well as data-driven patterns that differentiate individual brain responses. In addition, the structure of the proposed model targets the uncertainty in CDS scores by using them as prior information only to guide the unsupervised learning task in a multi-task learning scenario. A comprehensive evaluation has been done to confirm the significance of the proposed deep structure, including new ways of network visualisation to investigate spectral, spatial, and temporal information derived in the learning process. We argued that the proposed EEG analytics could also be applied to investigate other psychological and mental disorders currently indicated on the basis of clinical assessment scores. The code to reproduce the results presented in this paper is openly accessible at https://github.com/AbbasSalami/DPD_Analysis.

1. INTRODUCTION

Depersonalisation/derealisation refers to a temporary psychological condition reported in more than 50% of college students [1] and 34-70% of people without clinical history [2] during their lifespan. In traumatic situations, or when the brain faces a high level of stress or anxiety, prefrontal inhibition of the limbic emotional response system serves to protect the organism from overwhelming sensations and emotions [3, 4]. As a result, individuals may experience temporary emotional numbing, disembodiment, out-of-body experiences, or a sense of unreality about the outside world [5]. Other factors have also been proposed as a cause of transient depersonalisation/derealisation, such as sleep deprivation [6], fatigue [7], or travelling to unfamiliar places [8]. In some cases, the symptoms can be persistent, affecting the quality of life for individuals and intervening in their daily social life. In this regard, depersonalisation/derealisation disorder (DPD) has been classed as a type of dissociative disorder within the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) [9].

DPD is believed to be the most prevalent underdiagnosed psychiatric disorder [10]. Although several neuroimaging modalities have been shown effective for the detection of neurological and mental disorders in general [11-13], we recently urged researchers interested in depersonalisation/derealisation to invest more resources into inexpensive electrophysiological techniques, such as electroencephalogram (EEG), to develop their diagnostic potential for DPD beyond what is presently known [14]. Finding potential electrophysiological biomarkers can be a huge step toward understanding DPD and supporting its accurate and timely diagnosis. It is important to note that the term "biomarker" used throughout this paper refers to a unique, abnormal neural pattern activated by appropriate stimuli that can explain DPD symptoms.

Several studies in the literature investigate DPD using electrophysiological neuroimaging techniques. However, research on the neural correlates of DPD using EEG is often formed around an expert hypothesis. Then, the significance of the potential biomarker is validated using statistical tests. In other words, finding biomarkers is based on trial and error and requires an expert's prior knowledge of the disorder. In addition, labels for participants in experimental research as either patient or control (or high vs low symptomology) are often assigned according to the Cambridge Depersonalization Scale (CDS) [15]. CDS quantifies the level of depersonalisation/derealisation by measuring the frequency and duration of symptoms over a recent period of six months. According to the ROC curve analysis and finding the best compromise between true positive and false positive rates in a sample of 77 subjects, Sierra et al. [15] proposed a cut-off point of 70 for the CDS score. However, there is sometimes disagreement between the outcome of the CDS questionnaire and clinicians' diagnosis [16]. In addition, some studies [17-19] have suggested a threshold of 50 on different non-clinical datasets, challenging the utility of the clinical cut-off point for the purpose of cognitive neuroscience research. On the other hand, we can rely on Explainable AI (XAI) to eliminate the need for cut-offs.

In this study, we employ deep learning algorithms to develop an explainable ERP signal processing pipeline and overcome the above shortcomings in the experimental literature, such as hand-crafted labels or cut-offs. Machine learning and, specifically, deep learning algorithms are reliable techniques for analysing medical imaging data with outstanding results in identifying several mental disorders [20-23]. End-to-end deep learning architectures are particularly noteworthy since they eliminate the need for handcrafted features and prior information regarding the dataset. In addition, researchers have been trying to enhance the explainability of these so-called black-box models to improve their reputation among the communities of neuroscientists and psychologists [24]. Still, the deep learning models proposed for EEG analysis are often trained for a single supervised or unsupervised task. However, in our DPD scenario or generally for analysing mental disorders based on clinical assessment scores, a multi-task learning structure is needed to handle clinical assessment scores as prior information rather than entirely relying on them. This necessitates using a novel EEG analysis pipeline, which employs clinical assessment scores as information to guide an unsupervised task.

Accordingly, we propose an EEG-based biomarker discovery protocol using a novel deep learning structure which does not rely on hand-crafted labels (for patients and controls). Instead, CDS scores could help to conduct (but not directly imply) the learning process. CDS scores are used as a regressive predictive guide in the unsupervised group segmentation performed by the simultaneous clustering, forming a multi-task learning strategy. To clarify, we take advantage of CDS scores as prior information to guide the unsupervised part of our learning process to find more reliable discriminative neural features. Since our primary objective was to find EEG biomarkers that can explain DPD symptoms, we confirmed our results and the significance of the proposed analysis method by extracting spectral, spatial, and temporal information from our network and explaining them using evidence from the cognitive neuroscientific literature.

The highlights and novelties of this work are:

a) Proposing a novel deep learning framework for EEG signal analysis to find neural patterns associated with DPD symptoms without requiring hand-crafted labels or subjective cut-off points indicating the disorder. Our approach straightforwardly employs raw clinical evaluation scores while factoring in their uncertainties during the learning phase.

b) Proposing new ways of network visualisation, investigating spectral, spatial, and temporal information derived in the deep learning process.

Thus, in the rest of the paper, we first introduce our dataset and its experimental paradigm in detail, followed by preprocessing stages needed to apply to multi-channel EEG signals before feeding them to the deep model. We also present full details of our novel deep learning structure, its corresponding loss function, and the learning process. The results and comprehensive discussion of our findings from neuroscientific and psychological perspectives are presented thereafter. We conclude our paper with some ideas for further investigations in this crucial area.

2. MATERIALS AND METHODS

In this section, we first describe our dataset and experimental design. We then explain our proposed learning process in detail, focusing on signal preprocessing, deep learning structure, loss function, and visualisation technique.

2.1. Experimental design

The dataset used in this paper was collected by [25]. The original study was approved by the Human Research Ethics Committee of the author's institution, and all participants gave informed written consent to take part. The dataset consists of 50 participants who initially took the self-rating CDS questionnaire to quantify their (trait) level of depersonalisation. With a threshold of 50 on CDS scores, 21 subjects were evaluated as participants with a low level of depersonalisation, and 29 subjects were considered individuals with a high level of depersonalisation, henceforth termed the control group and the DPD group, respectively. In addition to the CDS, participants' levels of depression and anxiety were also recorded based on Patient Health Questionnaire-9 (PHQ9) [26] and State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA) [27], respectively. The aim was to control the effect of depression and anxiety, which are highly comorbid with depersonalisation [28, 29], in the analysis. Participants also completed the Operationalised Psychodynamic Diagnosis-Structure Questionnaire (OPD-SQ) self-object differentiation subscale, which has been associated with dissociation [30], and the Multidimensional Assessment of Interoceptive Awareness (MAIA), which assesses eight different dimensions of subjective interoception [31].

Each session in the experiment consisted of two types of trials: tactile stimulation following an emotional prime and tactile stimulation following a neutral prime. The emotional prime was in the form of a happy or angry face and voice, with happy and neutral primes and angry and neutral primes forming two distinct datasets. It should be emphasised that this prime is presented before the tactile stimulation rather than simultaneously as in multisensory experiments. The prime intends to shape the participant's mood in advance rather than to gauge a brain response to it. During the experiment, subjects with normal or corrected-to-normal vision were asked to sit in front of a computer screen. Each trial started with a 400ms window of emotional or neutral prime, followed by a 600ms neutral hand stimulus (subjects observed a left or right hand and a pencil against a white background). The next 200ms time window comprised either the pencil touching the participant's hand (touch condition) or the space next to the hand (no-touch condition) so that the perceived distance of the pencil travelling would be the same. Finally, each trial ended with replaying the neutral hand stimulus for 600ms. In the 200ms time window of both touch and no-touch conditions, the participants received a 200ms tactile stimulus to their same hand, resulting in a synchronous visual-tactile stimulation for the touch condition. All trial types within each set were randomly intermixed with equal frequency. The animated schematics and timing scheme of each trial are presented in Figure 1. EEG signals were recorded during each session using a Compumedics Neuroscan SynAmps RT 64-channel amplifier and an EasyCap scalp electrode cap at a 1000-Hz sampling frequency and an online filter of 0.01-100 Hz.

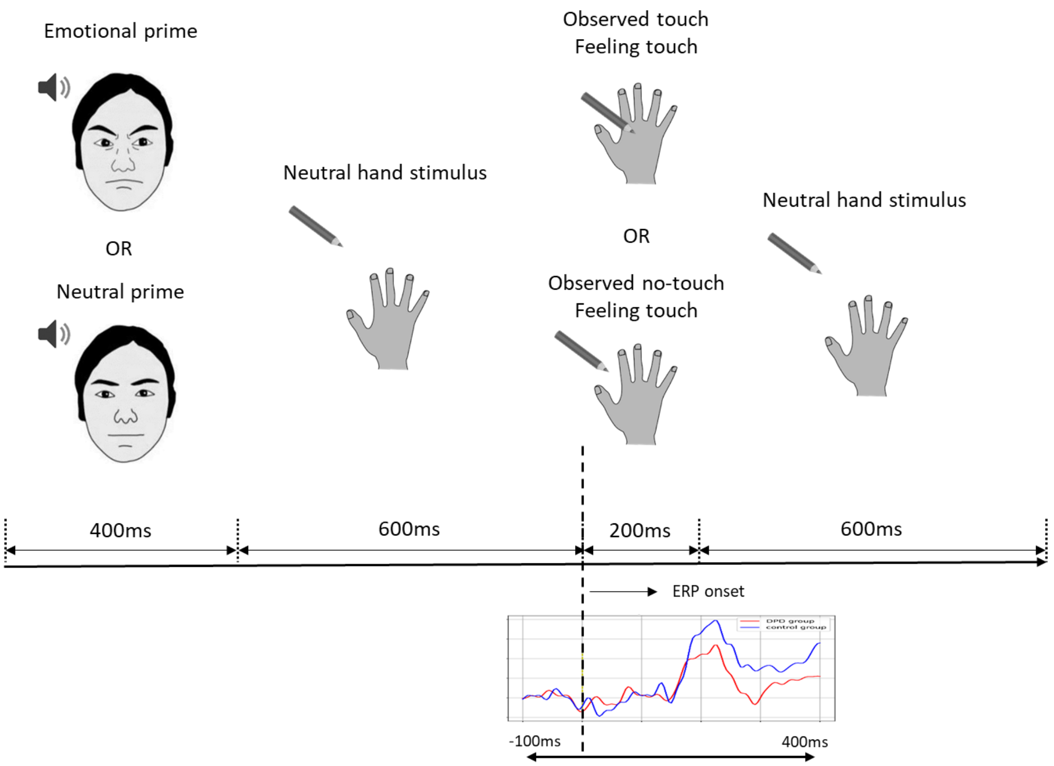


Figure 1. The schematics and timing of each trial.

2.2. Methods

The methodology used to analyse the data is detailed in the following subsections, covering every aspect of the process.

2.2.1. Preprocessing

The raw EEG data was fed into our proposed automated preprocessing pipeline, detailed in Figure 2. The pipeline starts with downsampling to 250Hz to reduce the size of the dataset and decrease the processing time without considerable loss of information. Then, a high-pass finite impulse response (FIR) filter with a 1Hz cut-off frequency was applied to the multi-channel EEG signals to remove the DC offset and low-frequency artefacts, followed by a low-pass filter with a 40Hz cut-off frequency to remove high-frequency artefacts and 50Hz line noise. One might argue that a relatively high 1Hz cut-off frequency can be detrimental in terms of affecting ERP components. However, since somatosensory processing is mainly associated with early high-frequency ERP components, using a 1Hz cut-off frequency was unlikely to have a negative impact on our results and was crucial to correct significant signal distortion in our dataset and save as many trials as possible. Nevertheless, a high-pass filter with a lower cut-off frequency is generally advisable in the case of high-quality recordings. In the next stage, the clean_rawdata() plugin in EEGLAB was used to detect and remove corrupted channels automatically, including the ones with a constant pattern,

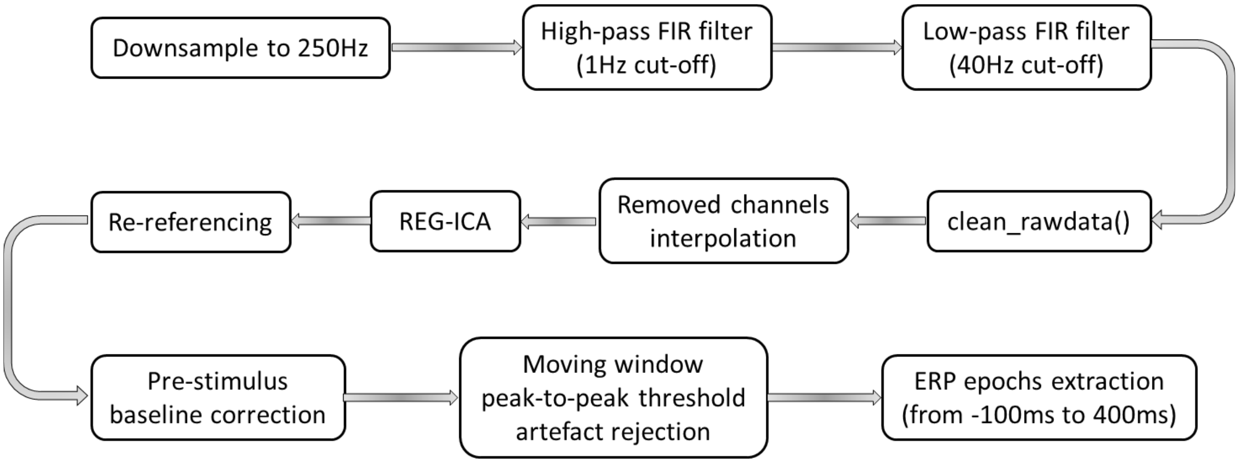

Downsample to 250Hz
High-pass FIR filter (1Hz cut-off)
Low-pass FIR filter (40Hz cut-off)
clean_rawdata()
Removed channels interpolation
REG-ICA
Re-referencing
Pre-stimulus baseline correction
Moving window peak-to-peak threshold artefact rejection
ERP epochs extraction (from -100ms to 400ms)

Figure 2. An overview of the proposed automated preprocessing pipeline.

excessive noise, or poor scalp-surface contact. clean_rawdata() is the offline version of artefact subspace reconstruction (ASR) method proposed by Christian Kothe (details can be found in [32]). Next, those rejected EEG channels were interpolated using other nearby channels. Using data from two EOG channels, we used the REG-ICA algorithm [33] to remove blinks and other EOG artefacts from our EEG signals. REG-ICA is a hybrid algorithm for EOG artefact rejection based on independent component analysis (ICA). The method applies a regression algorithm to compare independent components (ICs) with EOG channels to decontaminate them. We used preconditioned ICA (PICARD) [34, 35] as the ICA algorithm and least mean square (LMS) as the regression algorithm. After artefact rejection, we used the average of all electrodes to re-reference EEG voltages, followed by 100ms pre-tactile-stimulus baseline correction. We also applied moving window peak-to-peak threshold artefact rejection to exclude any trial that was not cleaned during the earlier steps. Finally, ERP epochs were extracted from 100ms before tactile stimulus onset to 400ms after tactile stimulus onset. Based on the quality and quantity of processed EEG signals for each participant, data from 7 participants were excluded from further analysis, resulting in a total of 19 control and 24 DPD participants.

2.2.2. Resampling-average method

In ERP analysis, the phase-locked neural components (for instance, in our dataset, locked to the tactile stimulus onset) are primarily of interest, and any other spontaneous EEG activities are stochastic signals treated as noise. This noise in the ERP experiments is assumed to be generated from a normal distribution with zero mean. Therefore, the reason behind the common approach of taking the average over multiple epochs time-locked to the stimuli of interest is to cancel out the noise and random brain activities that can affect ERP components. However, this approach results in a single grand-average ERP for each subject per condition, increasing the chance of overfitting and making it impractical to train a machine learning algorithm due to insufficient data samples. Besides, the signal-to-noise ratio (SNR) of evoked potentials is too low to extract features that can distinguish the neural activity of the control group from the patients from a single trial. To overcome this problem, we used the resampling-average method to generate multiple

ERP waves with enhanced SNR to feed into our deep model and consequently find discriminative features among the two groups. The resampling-average method used in this study is portrayed in Figure 3 and works as follows. We randomly select a subset of *r* trials for each type of trial of that condition. To enhance the output of the resampling-average method further, the trials that dropped out for averaging were randomly selected from those with high variance in the ERP pre-stimulus signal. This was based on the assumption that a high-quality ERP signal should have a low variance in the pre-stimulus signal. By repeating this process *n* times, we can generate *n* ERP waves. It is important to note that although the generated samples are not independent of each other, using this method can reduce the impact of outliers in our dataset and generate enough data to train a deep neural network.

2.2.3. Score-guided deep-learning diagnostic system

The dataset contains several types of trials, and even if we focus on a single subset (like an angry set) and a single condition (like touch) after a single prime (like emotional), we still end up with two types of trials representing tactile stimuli to the left or right hand. It is essential to consider that those trials still need to be analysed separately since somatosensory processing in the brain is only initially lateralised to the hemisphere contralateral to the touch. Accordingly, one should not combine those trials by remapping the electrodes in one of the conditions unless one is interested in only early ERP components. Thus, this paper shows how we simultaneously analyse left and right tactile stimulation trials in our deep model.

The novel deep learning architecture proposed in this study is a multi-input, multi-output deep neural network, as depicted schematically in Figure 17. The two input branches of the network consist of sequences

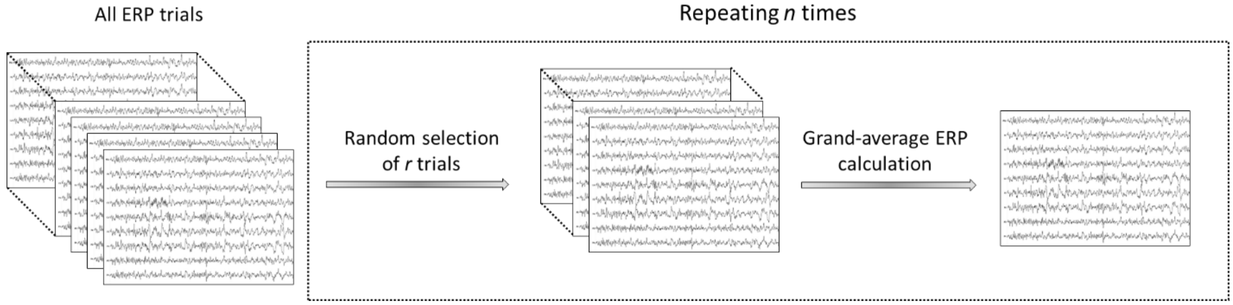


Figure 3. Resampling-average method to generate ERP samples.

of layers similar to the structure of the well-known EEGNet [36] to analyse the trials with tactile stimuli to the left and right hand separately. Generally, each branch starts with a 2D convolutional layer with a kernel size of $1 \times s$ applied along the time axis, followed by another depthwise 2D convolutional layer with a kernel size of $c \times 1$, which we later explain how they play the role of a spectral and spatial filter. The number of convolutional filters in the first layer (F_1) and the second layer (d) are the hyperparameters of the system, which determine the number of frequency bands and spatial filters in each frequency range, respectively. The goal of the later separable convolutional and flatten layers is to find a low-dimensional representation of the input from a multi-channel EEG signal. Then, low-dimensional representations derived from both types of trials are concatenated to form a more extensive feature vector, which serves as an input to the final multi-task learning structure of the network.

The output of the network consists of a supervised and an unsupervised branch. The supervised branch is a fully connected layer with one unit and a "Relu" activation function to predict the continuous CDS scores. In contrast, the unsupervised branch is a fully connected layer with two units and a "softmax" activation function to generate cluster assignments. The idea is to learn a representation that separates the dataset into two patient and control clusters (clustering branch) guided by the CDS scores (regression branch). On the one hand, we've noted that CDS scores are subjective and lack precision, so we shouldn't depend entirely on them to identify electrophysiological biomarkers. On the other hand, clustering can be accomplished by discriminative yet nonmeaningful or confounded features. Therefore, by defining an appropriate loss function, we 1) guide the network to find and extract features that represent two distinct neural patterns; 2) make sure they represent patterns of individuals with a high and low level of depersonalisation. To achieve this, our proposed loss function is as follows:

$$\begin{aligned} &\mathcal{L}_{total} = w_{regression}\mathcal{L}_{regression} + w_{clustering}\mathcal{L}_{clustering} + w_{link}\mathcal{L}_{link} \\ &\text{s.t.}\, w_{regression} + w_{clustering} + w_{link} = 1 \end{aligned} \tag{1}$$

where $\mathcal{L}_{regression}$ is the loss function associated with the regression branch, forcing the network to predict the continuous CDS scores. We used mean squared error (MSE) for this purpose, which can be formulated as follows:

$$\mathcal{L}_{regression} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

where n is the number of samples per batch, y_i is the reported CDS score, and $\hat{y}_i$ is the network predicted score from the regression branch. The $\mathcal{L}_{clustering}$ in (1) denotes the unsupervised loss function used to help the network find a low-dimensional representation that separates data points into two distinct clusters and their corresponding cluster assignments. For this purpose, we used information maximisation [37, 38], which is defined as an estimate of the mutual information between the low-dimensional representation of the input data and cluster assignments. Let $E \in (e_1, \ldots, e_n)$, where $e_i \in \mathbb{R}^d$, denotes a d-dimensional random variable representing the concatenated low-dimensional representation of left and right tactile stimulus inputs. By defining $Z \in \{0,1\}$ as a random variable expressing cluster assignments, we can estimate mutual information between E and Z as follows:

$$I(E;Z) = H(Z) - H(Z|E) \tag{3}$$

where $H(.)$ and $H(.|.)$ are entropy and conditional entropy, respectively, and can be calculated as follows on a batch:

$$H(Z) = -\sum\nolimits_{z}\left[\left(\frac{1}{n}\sum_{i=1}^{n}p(z|e_i)\right)\left(\log\left(\frac{1}{n}\sum_{i=1}^{n}p(z|e_i)\right)\right)\right] \tag{4}$$

$$H(Z|E) = -\frac{1}{n}\sum_{i=1}^{n}\left[\sum\nolimits_{z}p(z|e_i)\log\left(p(z|e_i)\right)\right] \tag{5}$$

where z is an instance of the random variable Z. Thus, the clustering loss function based on information maximisation can be finally defined as follows:

$$\mathcal{L}_{clustering} = -(H(Z) - H(Z|E)) \tag{6}$$

By this definition, a lower loss value would be subject to an increase in marginal entropy $H(Z)$, which encourages the cluster assignments toward class balance and avoids trivial solutions, and a decrease in conditional entropy $H(Z|E)$, which ensures having high confidence in each cluster assignment.

Using the weighted sum of $\mathcal{L}_{regression}$ and $\mathcal{L}_{clustering}$ as the loss function, our model's supervised and unsupervised tasks could be achieved independently due to their independent units, making the network highly prone to overfit on uninformative features. In other words, this did not guarantee to find clusters representing our desired groups of people with low and high levels of depersonalisation. To tackle this problem, we introduced a third term in our loss function called $\mathcal{L}_{link}$, which bridges the gap between the supervised and unsupervised branches and ensures the finding of meaningful, desirable neural patterns to distinguish our patients from the control group. To achieve that, we first used a smooth logistic function in the form of $f(x) = \frac{1}{1+e^{-(\frac{Thr}{10}+0.1x)}}$ (see Figure 4-right) to scale the predicted scores from the supervised task to numbers between 0 and 1. The *Thr* in the indicated logistic function determines the turning point of the function and is the threshold on CDS scores to evaluate subjects as control or DPD patients. As mentioned earlier, there is no globally agreed threshold on CDS scores for DPD diagnosis. While researchers often

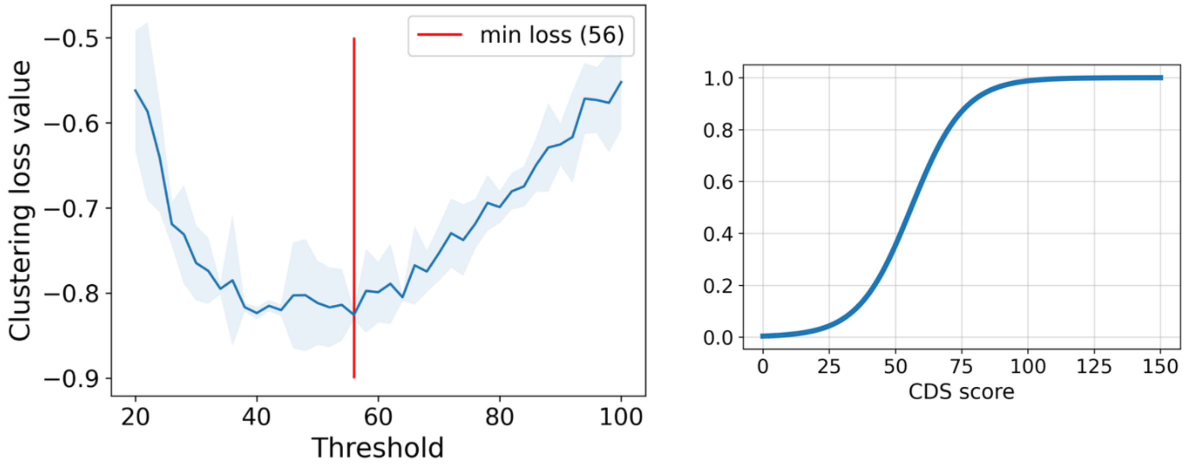


Figure 4. Minimum clustering loss value for different CDS thresholds (left) and the smooth logistic function defined to transform CDS predictions to a value between 0 and 1 with a turning point of optimum threshold (right)

choose 50 in their studies, clinicians prefer to use 70. Therefore, we performed a greedy search by sweeping all the possible values for the threshold from 20 to 100 to find the one that results in the lowest loss value, with the idea that an optimum threshold would be the one that performs subjects' separability with high confidence while making an accurate prediction of the CDS scores. Figure 4-left shows the result of our sweep, illustrating the mean and standard deviation (as shadow) of the clustering loss values over ten iterations for each threshold value. We found an optimum threshold of 56, which lies between the common threshold in the literature and clinician preference and used that value for *Thr* in the rest of our analysis.

The scaled scores following the logistic function were then compared to the cluster assignments using the cross-entropy loss function. The idea was based on the fair assumption that participants with extreme scores (too low or too high) should be assigned to their corresponding cluster with higher confidence. So assume $\hat{y}_i$ the network predicted score from the regression branch for the i-th data point, and $f\colon \hat{y}_i \to s_i$ the optimised logistic function, where s_i donates the scaled scores ($s_i \in [0,1]$). The scaled scores can form a vector $\vec{s} = \begin{bmatrix} 1-s_i \\ s_i \end{bmatrix}$ showing how likely each input data belongs to each group. Similarly, in our binary problem, the output of the clustering branch for the input i is a vector $\vec{z} = \begin{bmatrix} z_{i1} \\ z_{i2} \end{bmatrix}$ (where $z_{i1} + z_{i2} = 1$) containing the cluster assignments. As a result, the $\mathcal{L}_{link}$ can be defined as follows:

$$\mathcal{L}_{link} = -\frac{1}{n}\sum_{i=1}^{n}[(1-s_i).logz_{i1} + (s_i).\log(1-z_{i1})] \tag{7}$$

In sum, the proposed loss function (1) can be trained using gradient descent to minimise the CDS prediction error while forming two clusters and guarantee getting clusters representing our two groups of participants with low and high levels of depersonalisation.

Notice that our proposed deep learning model does not require a validation set, as our primary objective was to find EEG biomarkers that can explain DPD symptoms and serve as an adjunct to CDS in the DPD diagnostic process rather than accurately predicting CDS scores. We will validate our model by performing statistical analysis on the identified EEG biomarkers and providing neuroscientific evidence supporting our results. To address the concern regarding model overfitting, we have performed an ablation study in section 3.1 to show how $\mathcal{L}_{link}$ causes a trade-off between the regressions and the clustering task, preventing the network from overfitting or converging to trivial solutions. Therefore, we only used early stopping in our model to terminate the learning process once we no longer see improvements in the total loss, meaning the supervised and unsupervised tasks have reached an equilibrium point. Our tailored loss function enables identifying relevant features for distinguishing our participant groups while retaining an estimate of their CDS scores. This strategy was crucial in mitigating the limitations associated with subjective clinical assessment scores and advancing our goal of identifying EEG biomarkers for DPD.

2.2.4. Network visualisation

The goal of our study was to find potential electrophysiological biomarkers for DPD. For that, we need to investigate our deep model learning process by visualising the spectral, spatial, and temporal information that the model used to make a decision. Since the initial blocks in our model are similar to our recently proposed explainable EEG-ITNet [39], we have used the same techniques to visualise the learned spectral and spatial information. Simply put, both input branches of our model depicted in Figure 17 initially employ 2D convolutional layers on multi-channel EEG data along the time axis. These layers act as a filter bank to decompose EEG signals into different frequency components, and one can extract the network's targeted frequency sub-bands by taking the Fourier transforms of the convolutional kernels after training. This is due to the fact that convolution in the time domain is equivalent to multiplication in the frequency domain. In addition, a depthwise convolutional layer with "valid" padding was applied following the 2D vanilla convolution in our model. This second layer in our model functions as a spatial filter because of its specific kernel dimension. This layer combines signals across all electrodes and acts as a mapping function from the surface to the source domain. Therefore, the learned weights of this layer can be used to identify distinctive sources and visualise spatial filters. For more details and the math behind our visualisation approach, please refer to [39].

In addition, we also aim to visualise the temporal information in both the source and electrode domains in the current paper. Since the first two blocks of our model behave as frequency and spatial filters, they map the input multi-channel EEG data from the electrode domain to the source domain. Hence, after training and learning the optimal network weights, we can extract and depict the corresponding source activity for each input sample. The signal of the closest electrodes to each source can also be used to investigate the neural patterns in the electrode domain. To enhance our understanding of the temporal and spatial aspects, we plan to utilise two XAI visualisers: the Layer-Wise Relevance Propagation (LRP) method [ref] and the Deep Taylor Decomposition (DTD) technique [ref]. These methods, effectively used in diverse neural network applications within EEG analysis research [40-42], will provide further insight as supplementary tools in our deep learning visualisation.

The basic idea behind LRP is to assign relevance scores to the input features, indicating their importance in the final prediction. These relevance scores are propagated backwards through the network, layer by layer, in a way that ensures the conservation of relevance. This propagation process helps in attributing the model's decision to specific input features or neurons. Additional information and the mathematical principles underlying LRP can be found in [41, 43]. DTD is another XAI technique for interpreting decisions made by deep neural networks. It simplifies the network's complex output by tracing the influence of input features across each layer. Using Taylor decomposition, it assesses the contribution of individual neurons in one layer and redistributes their relevance to connected neurons in preceding layers. This step-by-step process reveals how specific inputs affect the network's final decision. By breaking down the network's function into simpler components, DTD offers a clearer understanding of how deep learning models arrive at their conclusions, enhancing transparency and trust in these sophisticated systems. Readers seeking further information on DTD are advised to refer to [44].

2.2.5. Statistical analysis

Discovered biomarkers were evaluated by finding the statistical significance of their difference among our two groups, individuals with low and high levels of depersonalisation. For that, an independent samples t-test was applied on the average of each biomarker on its corresponding time window. The test assumptions, including normal distribution and homogeneity of variance, were examined using the Shapiro-Wilk test for normality and the Mauchly test for sphericity. In addition to the degree of freedom for each biomarker, the following were reported: t-statistics, p-value at the .05 significance level, Cohen's d effect size, and 95% confidence interval. All statistical analysis was performed in Python 3.10 using the SciPy [45] and pingouin libraries [46].

3. RESULTS

Studies have shown abnormal responses in both the brain and the autonomic nervous system of DPD patients, which tend to be more evident for unpleasant and threatening emotional stimuli [47]. Therefore, this paper only focused on our angry set and synchronous visual-tactile stimulation (touch condition) following emotional primes (angry faces and sounds). We wanted to investigate if our two groups have distinguishable neural patterns in response to synchronous visual-tactile stimulation following negative emotional primes and identify the spectral, temporal, and spatial information of those potential biomarkers. In order to simplify our analysis, we disregarded almost half of the EEG channels and analysed only those presented in Figure 5. Note that the channel selection was in line with the nature of our experiment, which

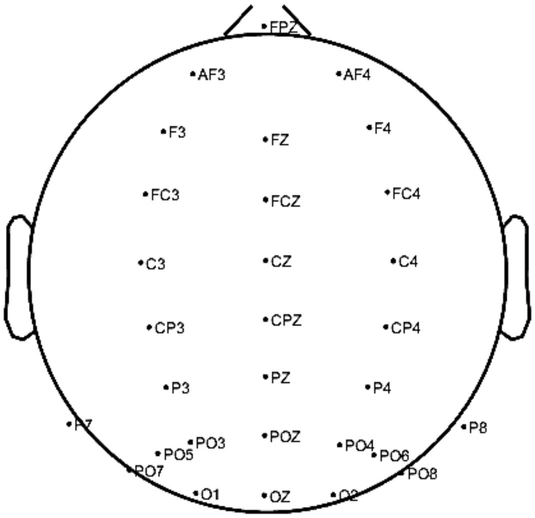


Figure 5. The names and placement of electrodes used in this study.

contained visual and tactical stimulation. The hyperparameters chosen for our model are summarised in Table 3, causing our network to have a total of 2947 trainable parameters.

3.1. Investigation of the effectiveness of multi-task learning

We proposed a third term in our loss function for our multi-task learning scenario called $\mathcal{L}_{link}$ as we argued that only using the weighted sum of $\mathcal{L}_{regression}$ and $\mathcal{L}_{clustering}$ as our loss function would make the network highly prone to overfit and result in finding uninformative and meaningless features in our case. To show the impact of $\mathcal{L}_{link}$, we investigated and compared the low-dimensional representation of the

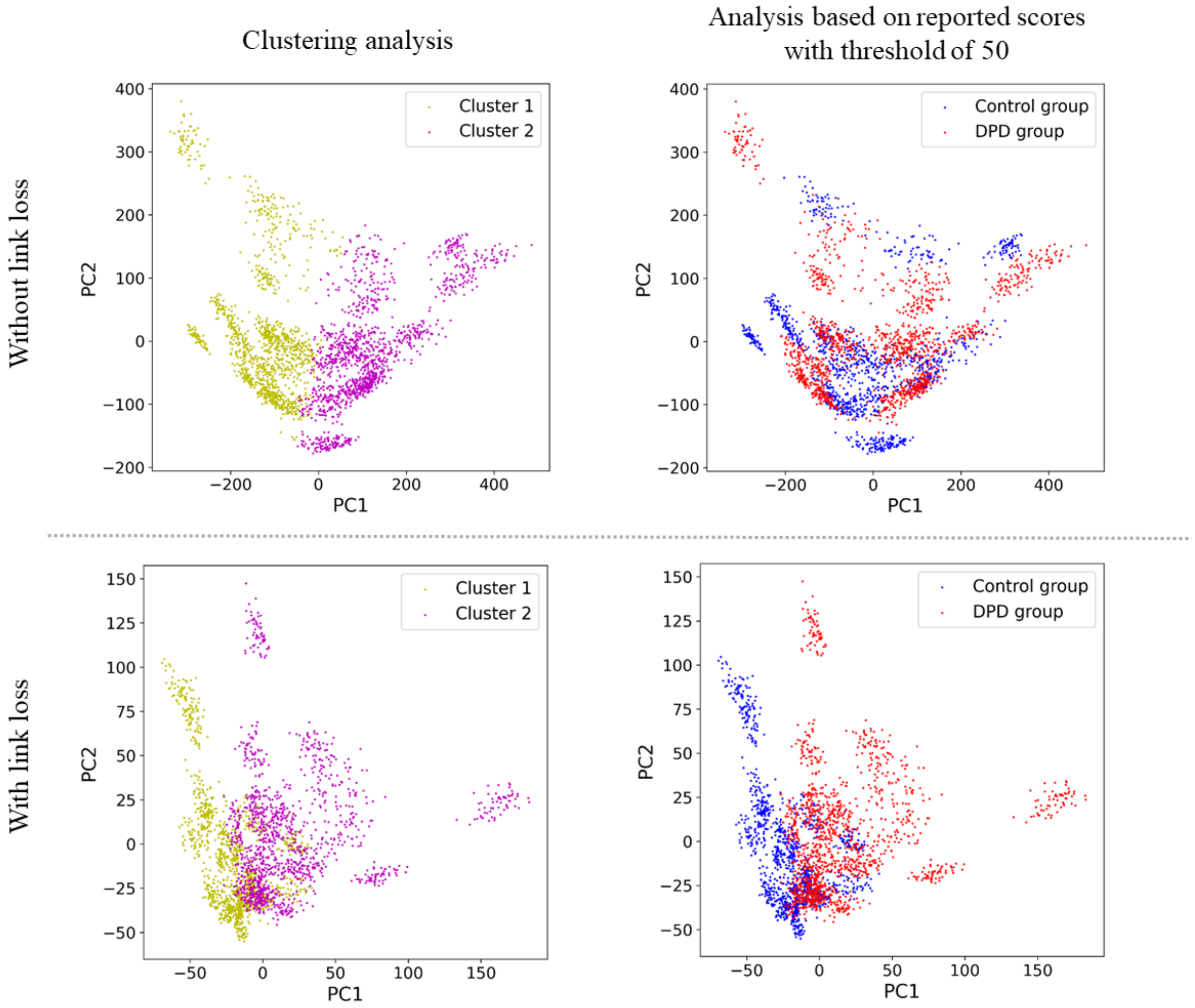


Figure 6. The scatter plots of the two strongest principal components of the learned low-dimensional representation of the input multi-channel EEG signals. The illustrations have been derived after training without (top) and with (bottom) link loss. The grouping is based on the output of the clustering layer (left) or reported CDS scores (right).

multi-channel EEG signals obtained from two networks with and without the $\mathcal{L}_{link}$ term in their loss function in Figure 6. For the sake of visualisation, we used the two strongest principal components of the learned low-dimensional representations. Figure 6 shows how $\mathcal{L}_{link}$ guides the network to find clusters representing our two groups of people with low and high levels of depersonalisation.

Figure 7 shows the clustering and regression loss values during the training process for two networks with and without $\mathcal{L}_{link}$ in their loss function. The figure is included in this paper to highlight further the importance of $\mathcal{L}_{link}$ in preventing the model from overfitting (in the regression task) and converging to a trivial solution (in the clustering task). An examination of the clustering and regression loss values reveals that the model achieves an equilibrium after approximately 110 epochs, with no substantial improvement thereafter. We have implemented early stopping in our model to halt training at this point, underscoring the effectiveness of $\mathcal{L}_{link}$ in ensuring meaningful results.

One of the objectives of this study was to address the unreliability of CDS scores as a diagnostic metric. The outcome of our proposed deep multi-task learning model with the parameters in Table 3 formed two clusters representing the DPD and the control groups. After comparing the participants assigned to each cluster with their original classification based on the reported CDS scores, we noticed that the discrepancy between our clustering analysis and original grouping was two DPD participants whom our clustering

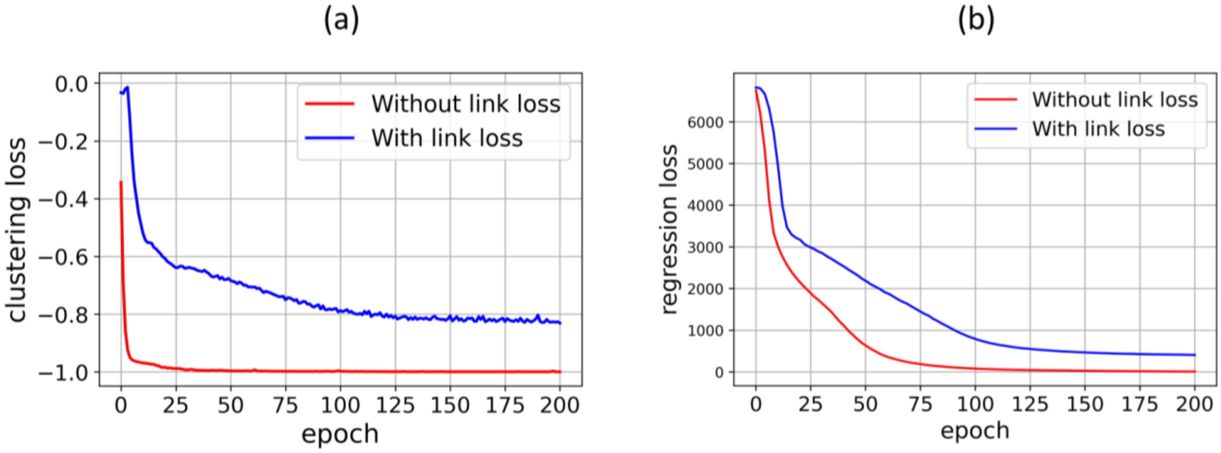


Figure 7. The clustering (a) and regression (b) loss values during the training process with and without the link loss term.

analysis assigned to the cluster representing our control group. In other words, the classification of the subjects based on our clustering analysis differed slightly from the one based on reported CDS scores. Accordingly, to investigate our method's superiority and accuracy over the conventional grouping based on CDS scores, we calculated a point-biserial correlation coefficient between the outcome of clustering and original classification with some additional data available from our participants, such as their depression, anxiety and self-object differentiation scores. Table 1 demonstrates that the outcome of clustering analysis shows a higher correlation with some psychological factors that are known to be highly comorbid with depersonalisation [14, 30] compared to the original classification based on the reported CDS scores. This suggests that classification based on our clustering analysis is at least equivalent to, if not superior, to the classification based on reported CDS scores.

3.2. Network analysis

We applied three on-system interpretability measures to investigate our trained model and identify potential electrophysiological biomarkers for DPD symptoms. We first share our results from the technique we

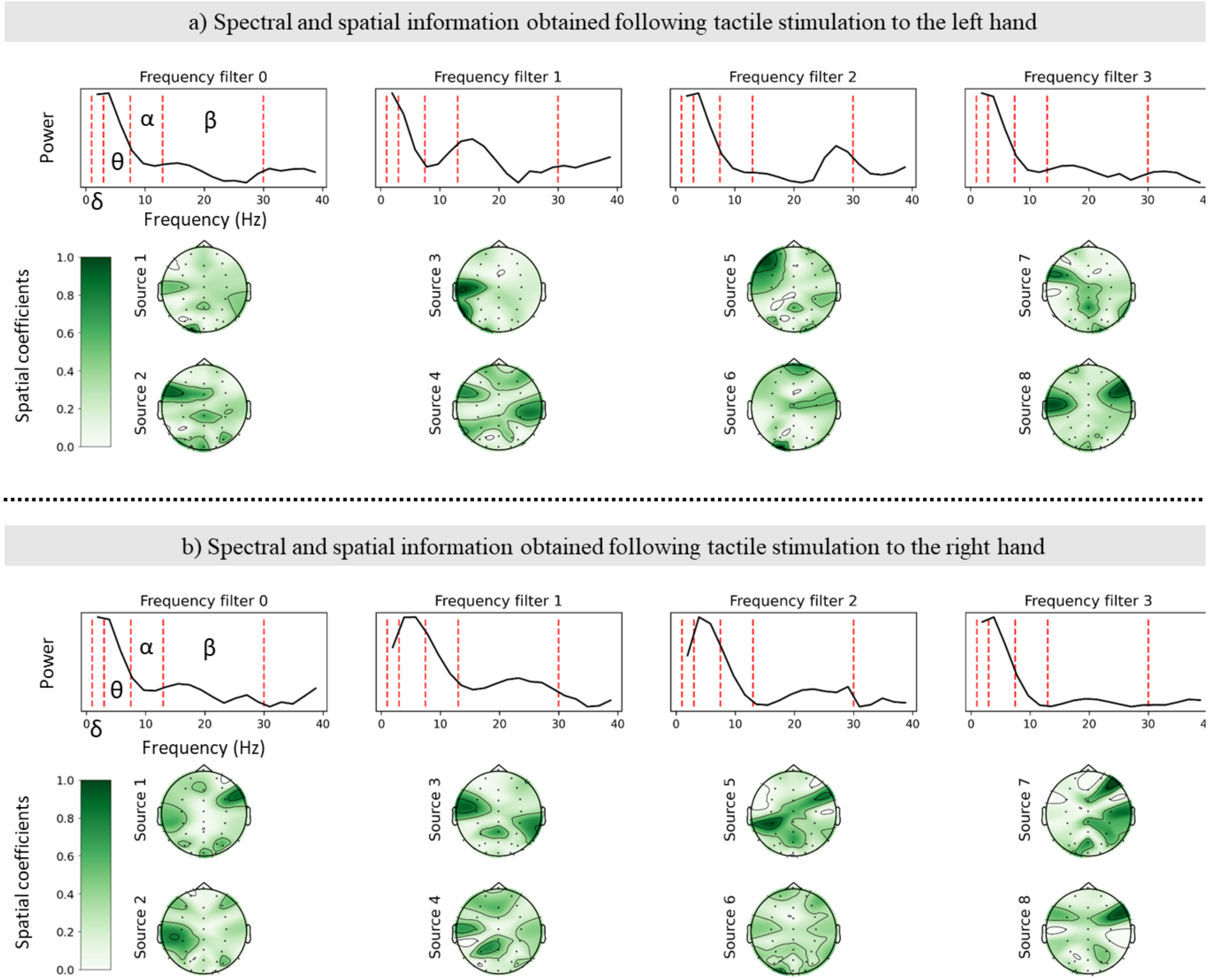
Table 1
Point-biserial correlation analysis for grouping based on reported CDS scores and our proposed learning method with participants' psychological factors.

Psychological factors	Grouping based on reported CDS scores		Grouping based on clustering analysis	
	r_{pb}	*p-value*	r_{pb}	*p-value*
Total CDS score	0.78	< 0.001	0.79	< 0.001
Anomalous body experience factor of CDS	0.66	< 0.001	0.67	< 0.001
Emotional numbing factor of CDS	0.73	< 0.001	0.72	< 0.001
PHQ-9 depression test score	0.29	0.061	0.31	0.041
Cognitive anxiety factor of STICSA	0.42	0.005	0.49	< 0.001
Somatic anxiety factor of STICSA	0.28	0.072	0.33	0.029
Self-object differentiation subscale of the Operationalized Psychodynamic Diagnosis-Structure Questionnaire (OPD-SQ) score	0.56	< 0.001	0.6	< 0.001

initially proposed in [39] and then the results from applying LRP and DTD methods. We used the iNNvestigate Python library [48] to implement LRP and DTD in our study.

3.2.1. Our visualisation technique

In order to find potential biomarkers for DPD diagnosis, we aimed to visualise spectral, spatial, and temporal information obtained after training our proposed deep model in Figure 8. Notice that our deep learning pipeline consists of two parallel branches, with the same type of layers, to simultaneously analyse trials associated with synchronous visual-tactile stimulation to the participant's left and right hand. Using the visualisation technique we proposed in [39], we illustrated the power spectrum and the spatial location


a) Spectral and spatial information obtained following tactile stimulation to the left hand

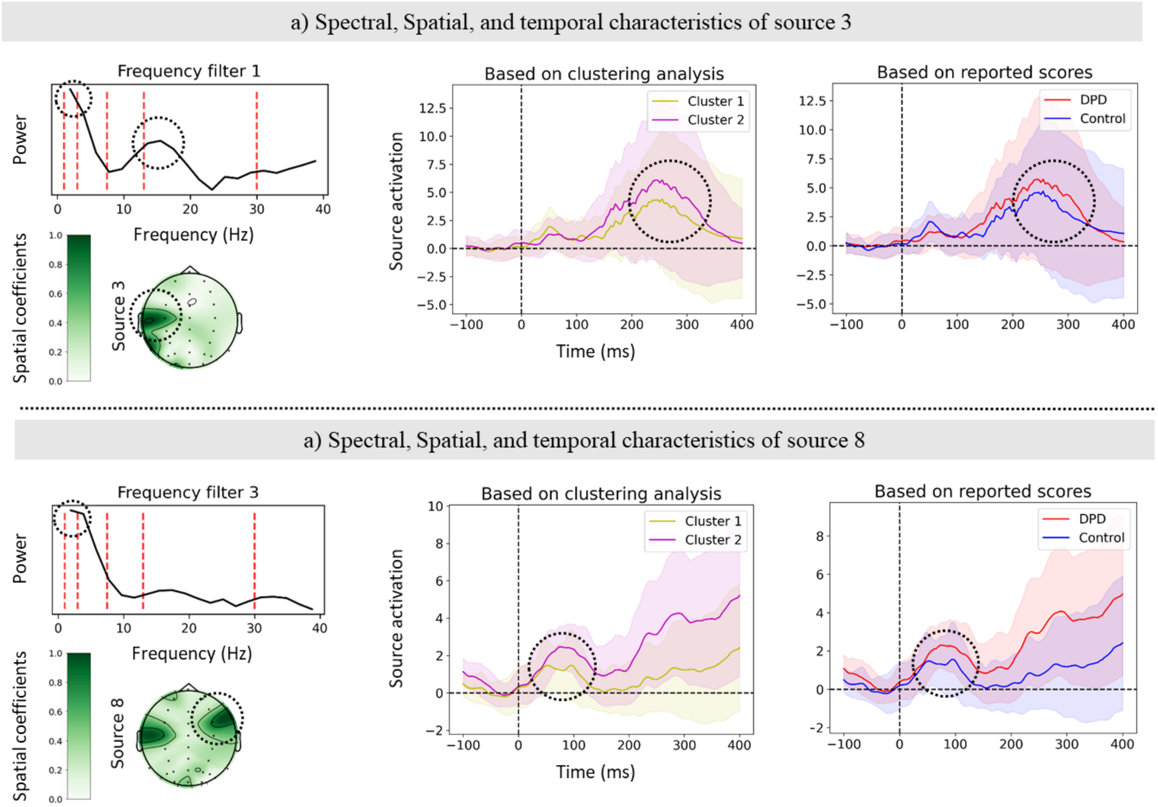
b) Spectral and spatial information obtained following tactile stimulation to the right hand

Figure 8. Spectral and spatial information of potential electrophysiological biomarkers obtained from analysing synchronous visual-tactile stimulation to the left (a) and right (b) hands.

of sources that contributed the most during our multi-task learning process to perform the clustering task and predict CDS scores simultaneously. Therefore, any sources in Figure 8 can serve as potential electrophysiological biomarker. Notice that each topoplot depicted in Figure 8 can represent the combination of more than one source of electrical activity. That is why spatial activities can be observed in multiple spatial locations for each source. The power spectrum displayed above every two sources indicates the response of the frequency filter learned to extract those sources, with the red dashed lines separating EEG waves, including delta, theta, alpha, and beta.

Each of the sources in Figure 8 can serve as a potential electrophysiological biomarker. Since each of those topoplots might represent the spatial activities of a group of sources, we may wish to choose sources that are sparse and show stronger spatial activity for further investigation. Therefore, for the left-hand touch trials (Figure 8-a), we focused on sources 3 and 8. By extracting the output of the depthwise 2D convolutional layer (see Figure 17) from our network after the training process, we can also display the temporal activity of those sources for their corresponding inputs.

Accordingly, Figure 9 shows the spectral and spatial features of sources 3 and 8 derived by analysing synchronous visual-tactile stimulation to the participant's left hand along with their average temporal activities taken on all the trials for each group of people with a low and high level of depersonalisation. As participants' grouping differed based on our clustering analysis and reported CDS scores, we present the average temporal activities for both. The shaded region shows the standard deviation of activities. Any time window with non-zero temporal activity indicates the temporal characteristic of the potential biomarker. In summary, by looking at the illustrations in Figure 9-a for source 3, one may propose a hypothesis that there is an EEG biomarker for depersonalisation (higher activation for DPD patients based on CDS score and for Cluster 1) during synchronous visual-tactile stimulation to the participant's left hand with a **delta** (and **high alpha/low beta**) component over the left centro-temporal lobe (around **channel C3**) in the time window encompassing the **P300** ERP component.



a) Spectral, Spatial, and temporal characteristics of source 3

a) Spectral, Spatial, and temporal characteristics of source 8

Figure 9. Spectral, Spatial, and temporal characteristics of sources 3 and 8 obtained from analysing synchronous visual-tactile stimulation to the participant's left hand. The temporal response slightly differs based on clustering and reported CDS scores grouping. Dashed circles indicate characteristics of potential biomarkers.

To statistically investigate the proposed hypothesis, we first visualised the average ERP responses to synchronous visual-tactile stimulation to the participant's left hand over channel C3, which is the closest electrode to the spatial characteristics of source 3 in Figure 9-a. Figure 10 shows the average ERPs based on our clustering analysis and reported CDS scores. Both plots in Figure 10 clearly show a difference around the P300 component. After checking the test assumptions, such as normal distribution and homogeneity of variance, we performed an independent samples t-test on the P300 average amplitude (average ERP in the time window of 260-360ms post-stimulus) to investigate the significance level. The result was significant, with 95% confidence for both groupings, revealing a stronger effect for grouping based on clustering analysis ($t(41)=2.57$, $p=0.014$, *Cohen's d*=0.80, 95% CI [0.14, 1.13]) than based on reported CDS scores ($t(41)=2.23$, $p=0.032$, *Cohen's d*=0.70, 95% CI [0.05, 1.08]).

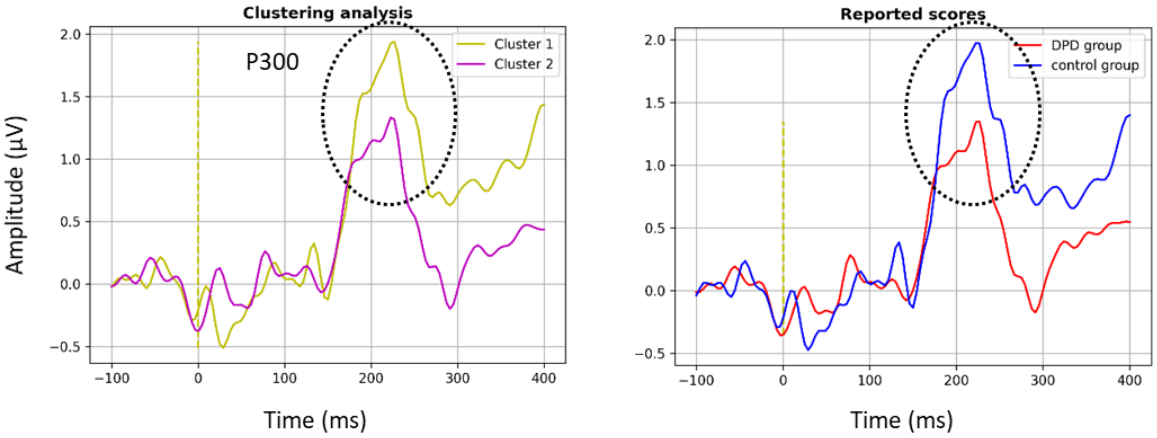


Figure 10. Average ERP responses to synchronous visual-tactile stimulation to the participant's left hand over channel C3. Clusters 1 and 2 represent DPD and control groups, respectively.

Similarly, we can propose other hypotheses by looking at spectral, spatial, and temporal characteristics of sources obtained in our learning process. As another example, looking at source 8 in Figure 9-b, we can assume that there is a potential biomarker in response to synchronous visual-tactile stimulation to the participant's left hand with delta component over channel C3 or C4/FC4 in the time window around early P45, P100, and later P300 ERP components. Notice that source 8 in Figure 9-b represents the combination of two sources; one of them (P300 component over C3) overlaps with source 3 investigated earlier. As a result, we can form our second hypothesis as an **early P45/P100** component cluster in the **delta** range activating over channel **C4/FC4** contralateral to the left-hand stimulation. In order to explore our second hypothesis further, we again aimed to visualise the time responses in the electrode domain over C4/FC4, and the results can be seen in Figure 11. We also statistically evaluated our second hypothesis and confirmed a significant difference in P45 average amplitude (time window of 40-70ms) with 95% confidence between DPD and control groups over channel C4/FC4 (see Table 2).

We should note that the group differences in the proposed biomarkers might be statistically more significant if we perform the t-test on the filtered ERP components based on the spectral characteristics of biomarkers. We did not investigate this as the length of each trial was only 500ms, which would result in high signal distortion following a bandpass filter on each trial. However, since our deep processing pipeline

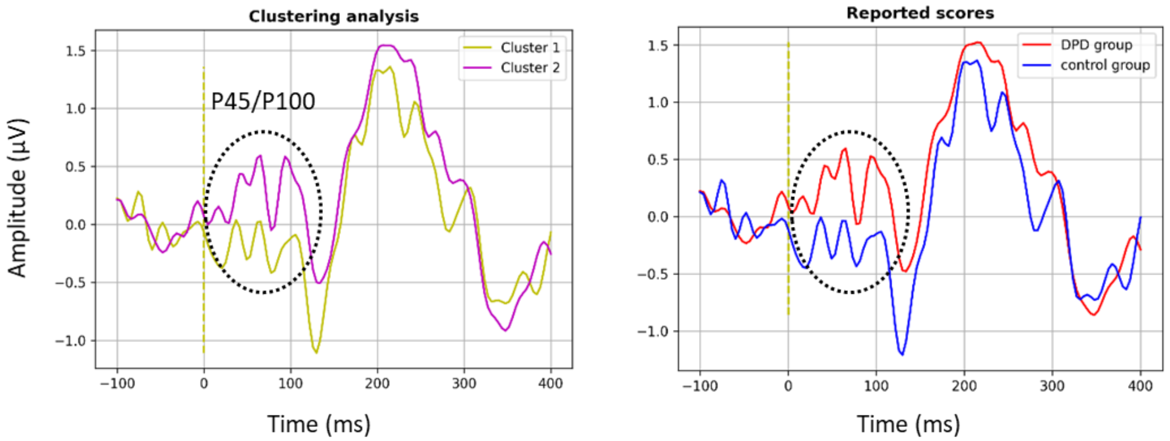


Figure 11. Average ERP responses to synchronous visual-tactile stimulation to the participant's left hand over channel C4/FC4. Clusters 1 and 2 represent DPD and control groups, respectively.

can be used for any similar psychological disorder assessed by clinical assessment scores, a bandpass filter based on the spectral map of learned features is highly recommended if a more extended time window is available.

Table 2
The summary of all the biomarkers identified using the proposed deep multi-task learning model for our DPD dataset and their statistical evaluation between DPD and control clusters.

Stimulated hand	Electrode	ERP component (Time window)	t-statistic	p-value	Effect size (Cohen's d)	95% confidence interval
left	C3	P300 (260-360ms)	2.57	**0.014***	0.80	[0.14, 1.13]
left	C4/FC4	P45 (40-70ms)	-2.78	**0.008***	0.87	[-0.99, -0.16]
right	FC4	P200 (180-280ms)	2.29	**0.027***	0.71	[0.09, 1.46]
right	C3/CP3	P45 (40-70ms)	-2.40	**0.021***	0.75	[-0.83, -0.07]
right	PO8	P200 (180-280ms)	-3.00	**0.005***	0.93	[-2.75, -0.53]

* Show significance at 0.05 level
Degree of freedom = 41

The activity of P45 over the contralateral somatosensory cortex concluded from our second hypothesis (using source 8 in Figure 8-a) was also captured by source 3 in Figure 8-b for synchronous visual-tactile stimulation to the participant's right hand, confirming the above component as a promising biomarker for DPD study and diagnosis. Therefore, using the same approach, more hypotheses can be inferred and investigated as potential electrophysiological biomarkers for DPD. The summary of discovered biomarkers and their evaluation can be found in Table 2. The statistical results reported in Table 2 are based on the clustering analysis grouping. A summarised comparison of all discovered biomarkers between the control group and DPD patients is also illustrated using boxplots in Figure 12.

3.2.2. LRP saliency-map

To gain deeper insights into the learned parameters within our multi-task learning framework, we employed the LRP technique on our model to extract saliency-maps. Note that our model in Figure 17 comprises two input branches that contain an identical sequence of layers. One branch focuses on analysing input EEG data following tactile stimulation to the left hand, while the other concentrates on the right-hand stimulation response. Accordingly, we independently applied the LRP method to each branch of the model to offer a more comprehensive understanding of the learned spatial and temporal information and potential

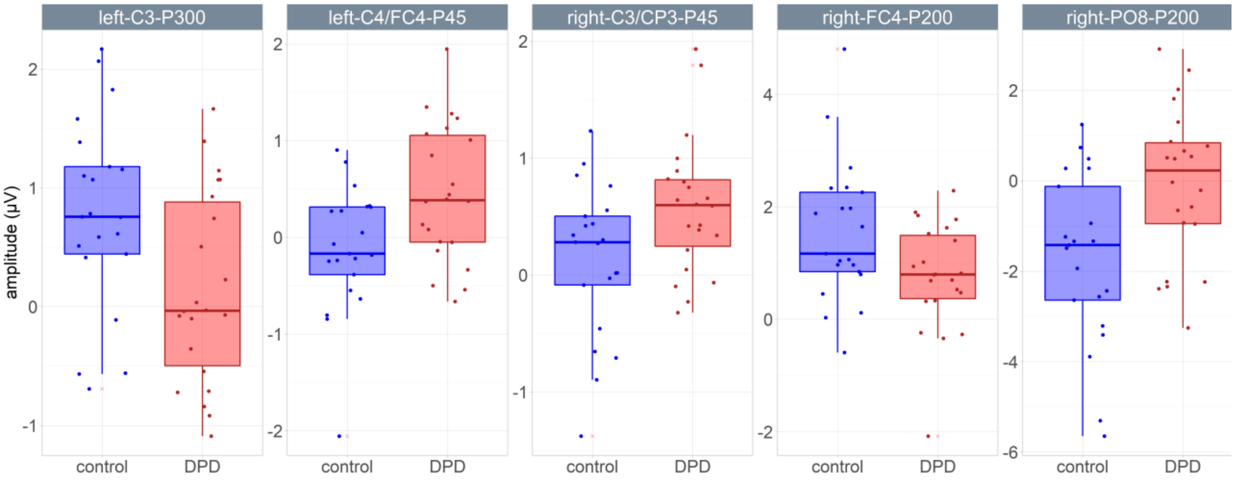


Figure 12. The component amplitude comparison between the control and DPD groups using boxplots for each biomarker discovered in Table 2.

discriminative features between individuals with low and high levels of DPD. The relevance-maps depicted in Figure 13 represent the average across all the saliency-maps derived from input signals from each group. The relevance maps in Figure 13 are normalised between -1 and 1, denoting the lowest and highest relevance, respectively.

To present a more explicit comparison of the relevance-maps between the two groups in our experiment, we additionally computed the differential LRP relevance-maps in Figure 14. Upon scrutinising the relevance-maps in Figure 14, certain resemblances with the results of our visualisation approach can be

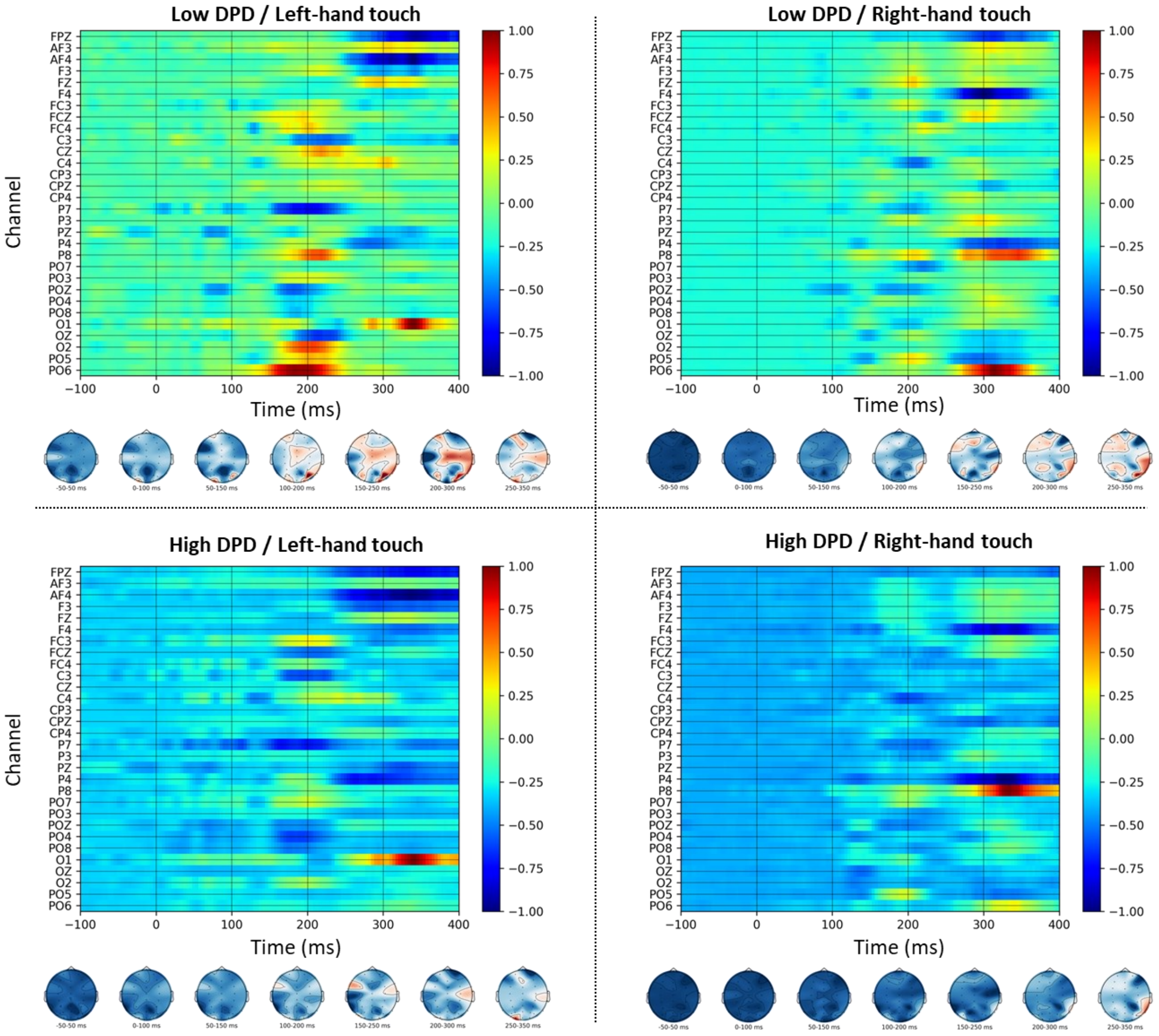


Figure 13. Average relevance-maps for Layer-Wise Relevance Propagation, normalised between -1 and 1. The topoplots of the average relevance in several time windows are depicted below each relevance-map.

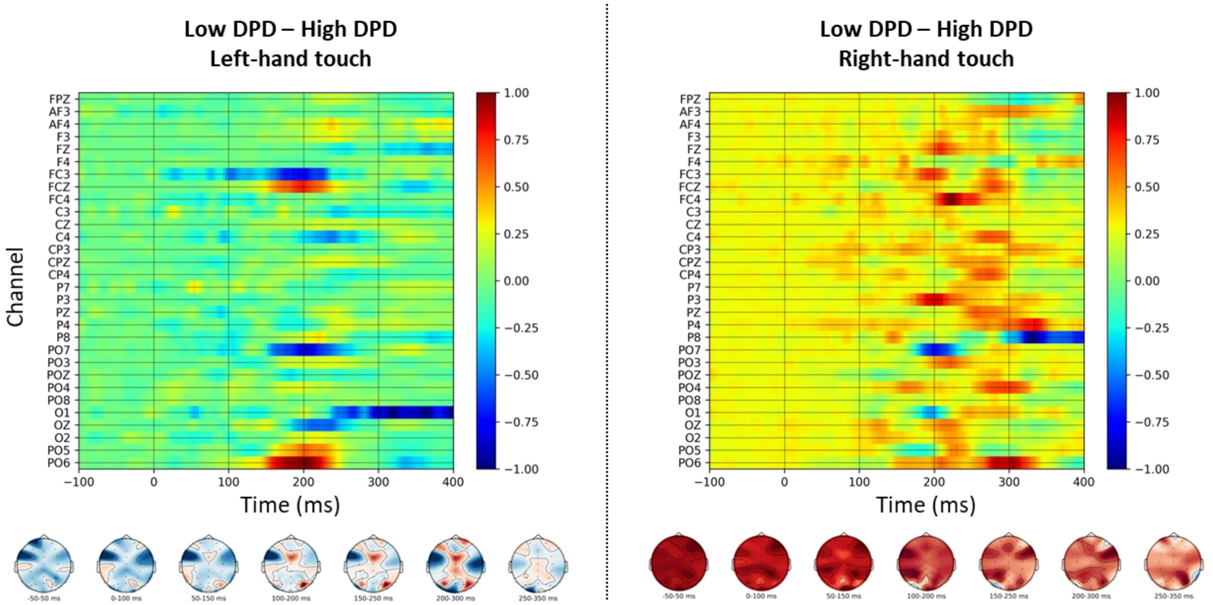


Figure 14. Differential LRP relevance-maps, derived by subtracting relevance-map for the high DPD from the low DPD group.

observed, such as the EEG component over channel FC4 within the time window spanning 200 to 300 ms post-stimulus for tactile stimulation to the participant's right hand. Although additional potential biomarkers can be suggested through the LRP relevance-maps in Figure 14, such as activity over the FCZ channel within the 150 to 250 ms timeframe following tactile stimulation to the left hand, we have not found them statistically significant between individuals with low and high DPD symptoms. Despite this, leveraging LRP and other XAI methods can offer deeper insights into the learned parameters of deep learning models, thereby aiding in the identification of electrophysiological biomarkers within EEG datasets.

3.2.3. DTD heatmap

In addition to LRP, we employed DTD to provide more understanding of the temporal and spatial characteristics learned in the training process of our model. DTD heatmaps were generated for trials following tactile stimulation to both left and right hands independently, as depicted in Figure 15. Similarly, differential DTD heatmaps were computed and illustrated in Figure 16 to facilitate a more effective comparison of these heatmaps between the two groups under study. Notably, in trials following tactile stimulation to the left hand, there is pronounced activity in the occipital-temporal area, especially in the

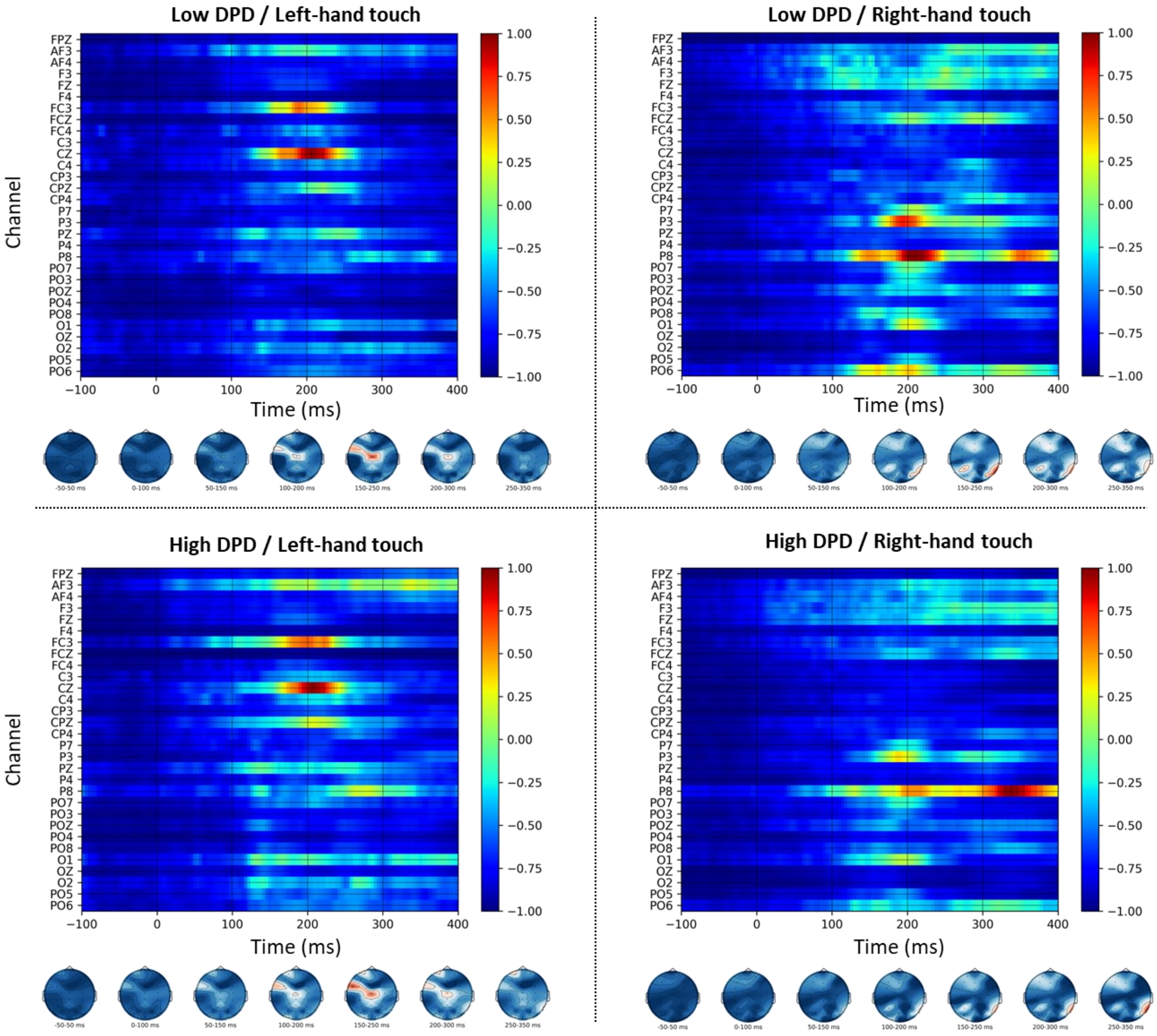


Figure 15. Average heatmaps for Deep Taylor Decomposition, normalised between -1 and 1. Spatial distribution within various time windows are illustrated beneath each corresponding heatmap.

150-250 ms time window. This observation is particularly interesting as it corroborates our findings regarding the P200 biomarker in the occipital-temporal cortex, which is relevant for understanding DPD symptoms.

4. DISCUSSION

It is important to recall that the DPD diagnostic process is not solely based on CDS, and the outcome of the questionnaire only helps the clinician in their final diagnosis of DPD as a primary condition [49]. The

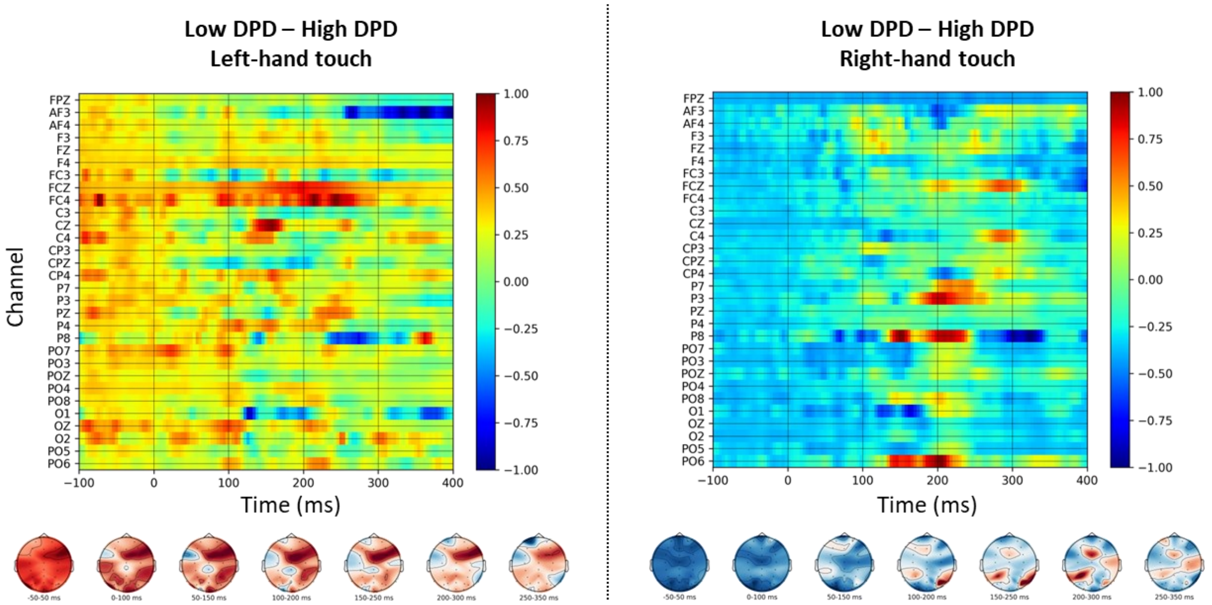


Figure 16. Differential DTD heatmaps, derived by subtracting heatmap for the high DPD from the low DPD group.

diagnostic process also involves several examinations, including but not limited to physical exams, lab tests, and psychiatric evaluations. However, since there is not always access to clinically diagnosed DPD patients in experimental research, CDS is often used as a primary metric to label participants as those with low and high levels of depersonalisation. This was a critical limitation in our dataset, indicating that our study aimed to present another reliable and less subjective metric to act as a supporting tool in the DPD diagnostic process rather than an alternative to CDS. Our deep learning driving approach fuses clinical scores with functional brain information to help clinicians make their final decisions.

4.1. Neuroscientific evidence to support biomarkers

To further establish our biomarker discovery system, we aim to provide neuroscientific evidence supporting the reliability of extracted electrophysiological biomarkers in Table 2.

Contralateral P45

P45 findings are straightforward to interpret. They reflect the activity in the somatosensory cortex contralateral to tactile stimulation. This activity is heightened in DPD patients relative to controls (see Figure 12), suggesting enhanced visual-tactile processing following negative emotional primes as a biomarker for DPD. Why might tactile processing be enhanced in such circumstances? Adler et al. [17] showed that P45 visual-tactile processing is suppressed rather than enhanced for self-related stimulation conditions (self-face observation) in those with high levels of DPD compared to those with low levels. Our findings of DPD-based P45 enhancements following negative emotional primes may thus be seen as contradictory. However, the literature on DPD suggests that the stimulation conditions may have tapped into a different mechanism (other-related simulation); therefore, in our protocol, we did not show participants' own body parts but photographs of other people's hands only during tactile stimulation. Farmer et al. [19] reported enhanced visual-tactile processing for other faces in those with high vs. low levels of DPD. It may be that when the visual stimulus is more indicative of "other" than "self", visual-tactile integration is enhanced in DPD relative to control groups, which has been extracted by our current analyses. The underlying reasons for this remain yet unexplored in research. Still, it is not unlikely that they reflect a strategy of emotional over-activation or over-attunement in those with frequent depersonalization symptoms, given the well-established links between DPD and childhood trauma (typically from emotional abuse or neglect [50, 51]).

Ipsilateral P200/P300 over sensory-motor processing regions

The differences between the groups in P200 (right hand stimulation) and later P300 (left hand stimulation), with sources in ipsilateral sensory-motor processing regions, are likely related to those reported by Adler et al. [17] at frontocentral P200. In [17], brain response related to tactile stimuli was diminished in individuals with high levels of DPD compared to those with low levels. In the current study, DPD patients also showed less activation than the control group. Although our results are derived from an experiment that was not designed to directly manipulate and measure self-other distinction, the P200/P300 findings may still be speculated as a biomarker of DPD that reflects reduced self-other differentiation, as argued by Adler et al. [17]. It is frequently reported that in DPD, the mirror image feels like a stranger to the observer despite the full realisation that they are looking at themselves. This phenomenology may be underlain by less distinct self and other processing mechanisms operating in this time range. It is feasible to propose that this may be a consequence of the earlier over-attunement with the other that was seen in the time range of P45. Interestingly, the relevant right-hemispheric biomarker emerges earlier in processing (P200) than its left-hemispheric equivalent (P300), but the left-hemispheric group differences are markedly stronger. It may be speculated that this may relate to the potential left-hemispheric abnormalities that have been documented in DPD [52, 53], possibly reflecting an aberration of the typical pattern of hemispheric differences in emotional processing [54], whereby the left hemisphere predominantly processes positive emotions, and the right hemisphere predominantly processes negative emotions [55].

P200 over occipital-temporal cortex

P200 findings over the right occipital-temporal cortex (PO8) may be related to aberrant visual processing in DPD relative to controls. The identified time range of the effect and its spatial source is in line with ERP components related to the recognition of familiar faces and bodies in occipitotemporal regions, where typically a smaller P200 is obtained for familiar relative to unfamiliar shapes (e.g., [56]). Enhanced processing for the DPD group relative to controls in this time range may thus reflect greater unfamiliarity during synchronous visual-tactile stimulation. DPD is typically marked by feelings of disembodiment, where one's own hands and face may not feel like they belong to oneself, and derealisation, where one's surroundings and reality in general may appear dreamlike, intangible, and unfamiliar. It is conceivable that the identified relatively heightened P200 processing for the DPD group is a biomarker of this, specifically visual, phenomenological experience.

Since we were performing multiple hypothesis testing on the same samples, one might argue the need to apply False Discovery Rate (FDR) correction in our analysis. We acknowledge its significance in studies involving many hypothesis tests, such as in genomic analyses. However, it is important to note that our study is focused on a limited set of five EEG biomarkers, which results in a relatively small sample size. Therefore, applying FDR correction methods in such cases could lead to an overly conservative outcome. For instance, even with a relatively high 20% FDR threshold, only 2 out of the five tests would be found statistically significant using Benjamini-Hochberg method [57], potentially obscuring meaningful associations. As a result, given the exploratory nature of our research and the inherent limitations of our sample size, we have chosen not to implement FDR correction. We emphasise that our objective is to identify initial trends and associations that can guide future research directions.

To show the importance of finding a reliable electrophysiological biomarker in the DPD diagnostic process, we have previously investigated the contralateral P45 components of SEP over the somatosensory cortex as a feature to perform a classification task [58]. Using this ERP component, which was also identified by the proposed model in this paper, we managed to achieve 85% accuracy (Kappa value of 0.7) in a classification task between individuals with low and high levels of depersonalisation on a different dataset. The dataset and the experiment provide valuable insights, yet conducting large-scale studies would be optimal for additional validation of these electrophysiological biomarkers in diagnosing DPD.

4.2. Limitations and complementarity of other analyses for full diagnosis

In this study, we have directed our focus towards trials that involve tactile stimulation following an angry emotional prime within a dataset that includes a variety of stimulus types. While this enriches our analysis, it also presents challenges in distinctly isolating each condition or stimulus type. This complexity potentially activates multiple brain areas, adding intricacy to our interpretability and statistical analysis. Moreover, while the relevance maps and heatmaps generated provide valuable insights into the temporal and spatial neural responses to tactile stimulation in individuals with low and high levels of depersonalisation, they may reflect the influence of combined stimuli, including visual and tactile elements.

Moreover, EEG cannot investigate the structural brain abnormalities that are physiologically linked to the condition, which might be deemed risk factors. Such insights can be gleaned, for example, by analysing structural MRI data to detect changes, like variations in the grey matter within cortical and intra-cortical brain areas [59]. Hence, a potential approach for future research could involve a more streamlined experimental setup, such as focusing on resting-state EEG recordings to evaluate single stimulus effects. In addition, for a complete DPD diagnosis, a combination of multiple neuroimaging analyses providing different effects of the condition is a recommended option.

5. CONCLUSION

DPD affects 1-2% of the population, comparable to schizophrenia and obsessive-compulsive disorder (OCD). Yet, it takes seven to 12 years on average to be accurately diagnosed. Therefore, a correct and early diagnosis of DPD is an urgent matter in clinical research, and there is a need to find diagnostic markers highly specific to DPD to distinguish it from other alternative diagnoses. This research focused on employing deep learning techniques to design a more powerful biomarker discovery system for DPD. We aimed to develop an explainable end-to-end deep learning model that can extract rich, informative neural patterns from EEG signals and exploit neural patterns specific to DPD symptoms to help the community better understand the disorder. We discussed why our DPD scenario, or any mental disorder assessed on the basis of clinical assessment scores, should be seen as a multi-task learning problem to reduce the impact of uncertainty in clinical assessment scores. Besides, it was argued that the literature often relies on experts' knowledge of the disorder and is based on hypothesis testing to find DPD biomarkers. As a result, we proposed a multi-input multi-output deep learning structure, which is designed to find the best separability in a dataset, guided by clinical assessment score. The presented method is relevant for the analysis of clinically relevant event-related potentials.

Furthermore, a method was presented to visualise and explain the learning and decision-making process in deep neural networks designed for EEG analysis, along with a description of how it could be applied to exploit multiple reliable EEG biomarkers for DPD. Finally, we summarised and interpreted the obtained biomarkers, including P45 contralateral to tactile stimulation, ipsilateral P200/P300 over sensory-motor processing regions, and P200 over the occipital-temporal cortex, from a cognitive neuroscientific point of view. Nevertheless, the limitation in our results on the potential neural signatures of depersonalisation symptoms was the need for more discussion on the spectral information derived during the learning process

[60]. Although we visualised the frequency responses of the filters trained in the proposed model, we did not investigate them further for the ERP dataset. Since each trial was only 500ms long, there needed to be more samples to calculate the Fourier transform accurately. Besides, applying a band-pass filter on such a short window to focus on a specific frequency band would result in severe signal distortion and, therefore, the unreliability of results. That is why we encourage future studies to consider a relatively prolonged time window to allow the investigation of spectral information and some low-frequency components of EEG (such as theta), which were flagged as potential biomarkers in the literature [14].

The parameter settings in our proposed approach were primarily informed by our previous study [39], where we applied a model with similar initial layers to a dataset with the same sampling frequency. However, it is also crucial to investigate how to obtain the optimal parameter setting in our model [61]. This will be a challenge, given the absence of clinically diagnosed patients in experimental research and their reliance on potentially inaccurate CDS scores. In addition, it must be remembered that while transient depersonalisation is a common phenomenon during the lifespan, it could be an early sign of risk for developing DPD. Hence, developing a system to track the depersonalisation state and its severity could be of great importance to help prevent the symptoms from becoming chronic or overwhelming. The potential biomarkers and analytics presented in this work could help to find a solution for online tracking of depersonalisation states.

In the proposed EEG analytics, the deep learning model is no longer recognised as a black box, and its learning process can be explained. In addition, it can be modified and applied to any psychological and mental disorders currently indicated based on clinical assessment scores to exploit electrophysiological biomarkers that can help clinicians with a more accurate diagnosis. The input to the network can be ERPs, EEG recording during a mental task, or resting state EEG. Future research could employ it to extract and interpret neural patterns similarly with just a few modifications in the network parameters and layers. In sum, our deep neural network-based EEG processing pipeline is a novel contribution to explainable biomarker discovery, and we strongly encourage others to use its potential for clinical and related research.

REFERENCES

[1] Dixon J. Depersonalization phenomena in a sample population of college students. The British journal of psychiatry. 1963;109:371-5.

[2] Aderibigbe Y, Bloch R, Walker W. Prevalence of depersonalization and derealization experiences in a rural population. Soc Psychiatry Psychiatr Epidemiol. 2001;36:63-9.

[3] Hunter E, Phillips ML, Chalder T, Sierra M, David A. Depersonalisation disorder: a cognitive–behavioural conceptualisation. Behav Res Ther. 2003;41:1451-67.

[4] Stein DJ, Simeon D. Cognitive-affective neuroscience of depersonalization. CNS spectrums. 2009;14:467-71.

[5] Hunter EC, Sierra M, David AS. The epidemiology of depersonalisation and derealisation. Soc Psychiatry Psychiatr Epidemiol. 2004;39:9-18.

[6] van Heugten–van der Kloet D, Giesbrecht T, Merckelbach H. Sleep loss increases dissociation and affects memory for emotional stimuli. J Behav Ther Exp Psychiatry. 2015;47:9-17.

[7] Tibubos AN, Grammes J, Beutel ME, Michal M, Schmutzer G, Brähler E. Emotion regulation strategies moderate the relationship of fatigue with depersonalization and derealization symptoms. J Affect Disord. 2018;227:571-9.

[8] Kaplan H, Sadock B, Grebb J. Substance related disorders. Kaplan HI, Sadock BJ Kaplan and Sadock's synopsis of psychiatry: behavioral sciences, clinical psychiatry 8th ed Baltimore: Williams & Wilkins. 1998:419-26.

[9] American Psychiatric Association. Diagnostic and statistical manual of mental disorders2013.

[10] Michal M, Beutel ME, Grobe TG. How often is the Depersonalization-Derealization Disorder (ICD-10: F48. 1) diagnosed in the outpatient health-care service? Zeitschrift fur Psychosomatische Medizin und Psychotherapie. 2010;56:74-83.

[11] Strambo D, Rey V, Rossetti A, Maeder P, Dunet V, Browaeys P, et al. Perfusion-CT imaging in epileptic seizures. J Neurol. 2018;265:2972-9.

[12] Ke H, Wang F, Ma H, He Z. ADHD identification and its interpretation of functional connectivity using deep self-attention factorization. Knowledge-Based Systems. 2022;250:109082.

[13] Ke H, Chen D, Yao Q, Tang Y, Wu J, Monaghan J, et al. Deep Factor Learning for Accurate Brain Neuroimaging Data Analysis on Discrimination for Structural MRI and Functional MRI. IEEE/ACM Trans Comput Biol Bioinform. 2023.

[14] Salami A, Andreu-Perez J, Gillmeister H. Symptoms of depersonalisation/derealisation disorder as measured by brain electrical activity: A systematic review. Neurosci Biobehav Rev. 2020;118:524-37.

[15] Sierra M, Berrios GE. The Cambridge Depersonalisation Scale: A new instrument for the measurement of depersonalisation. Psychiatry Res. 2000;93:153-64.

[16] Merckelbach H, Giesbrecht T, van Heugten-van der Kloet D, Jong Jd, Meyer T, Rietman K. The overlap between dissociative symptoms and symptom over-reporting. The European Journal of Psychiatry. 2015;29:165-72.

[17] Adler J, Schabinger N, Michal M, Beutel ME, Gillmeister H. Is that me in the mirror? Depersonalisation modulates tactile mirroring mechanisms. Neuropsychologia. 2016;85:148-58.

[18] Kanayama N, Sato A, Ohira H. The role of gamma band oscillations and synchrony on rubber hand illusion and crossmodal integration. Brain Cogn. 2009;69:19-29.

[19] Farmer H, Cataldo A, Adel N, Wignall E, Gallese V, Deroy O, et al. The Detached Self: Investigating the Effect of Depersonalisation on Self-Bias in the Visual Remapping of Touch. 2019.

[20] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.

[21] de Bardeci M, Ip CT, Olbrich S. Deep learning applied to electroencephalogram data in mental disorders: A systematic review. Biol Psychol. 2021;162:108117.

[22] Reinders AA, Marquand AF, Schlumpf YR, Chalavi S, Vissia EM, Nijenhuis ER, et al. Aiding the diagnosis of dissociative identity disorder: pattern recognition study of brain biomarkers. The British Journal of Psychiatry. 2019;215:536-44.

[23] Ke H, Chen D, Shah T, Liu X, Zhang X, Zhang L, et al. Cloud-aided online EEG classification system for brain healthcare: A case study of depression evaluation with a lightweight CNN. Software: Practice and Experience. 2020;50:596-610.

[24] Huff DT, Weisman AJ, Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. Phys Med Biol. 2021;66:04TR1.

[25] Gillmeister H, Adler J, Savva D, Li H, Parapadakis C. Atypical multisensory and emotional body perception in adults with symptoms of Depersonalisation-Derealisation Disorder. (in preparation).

[26] Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16:606-13.

[27] Grös DF, Antony MM, Simms LJ, McCabe RE. Psychometric properties of the state-trait inventory for cognitive and somatic anxiety (STICSA): comparison to the state-trait anxiety inventory (STAI). Psychol Assess. 2007;19:369.

[28] Kessler RC, Gruber M, Hettema JM, Hwang I, Sampson N, Yonkers KA. Co-morbid major depression and generalized anxiety disorders in the National Comorbidity Survey follow-up. Psychol Med. 2008;38:365-74.

[29] Simeon D. Depersonalisation disorder: a contemporary overview. CNS drugs. 2004;18:343-54.

[30] Sole S. Dissociative symptoms and the quality of structural integration in borderline personality disorder: UCL (University College London); 2014.

[31] Mehling WE, Price C, Daubenmier JJ, Acree M, Bartmess E, Stewart A. The multidimensional assessment of interoceptive awareness (MAIA). PLoS One. 2012;7:e48230.

[32] Kothe CAE, Jung T-P. Artifact removal techniques with signal reconstruction. Google Patents; 2016.

[33] Klados MA, Papadelis C, Braun C, Bamidis PD. REG-ICA: a hybrid methodology combining blind source separation and regression techniques for the rejection of ocular artifacts. Biomedical Signal Processing and Control. 2011;6:291-300.

[34] Ablin P, Cardoso J-F, Gramfort A. Faster independent component analysis by preconditioning with Hessian approximations. IEEE Transactions on Signal Processing. 2018;66:4040-9.

[35] Ablin P, Cardoso J-F, Gramfort A. Faster ICA under orthogonal constraint. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): IEEE; 2018. p. 4464-8.

[36] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. Journal of neural engineering. 2018;15:056013.

[37] Bridle J, Heading A, MacKay D. Unsupervised classifiers, mutual information and'phantom targets. Adv Neural Inf Process Syst. 1991;4.

[38] Krause A, Perona P, Gomes R. Discriminative clustering by regularized information maximization. Adv Neural Inf Process Syst. 2010;23.

[39] Salami A, Andreu-Perez J, Gillmeister H. EEG-ITNet: An Explainable Inception Temporal Convolutional Network for Motor Imagery Classification. IEEE Access. 2022;10:36672-85.

[40] Binder A, Bach S, Montavon G, Müller K-R, Samek W. Layer-wise relevance propagation for deep neural network architectures. Information science and applications (ICISA) 2016: Springer; 2016. p. 913-22.

[41] Torres JMM, Medina-DeVilliers S, Clarkson T, Lerner MD, Riccardi G. Evaluation of interpretability for deep learning algorithms in EEG emotion recognition: A case study in autism. Artif Intell Med. 2023:102545.

[42] Jemal I, Mezghani N, Abou-Abbas L, Mitiche A. An interpretable deep learning classifier for epileptic seizure prediction using EEG data. IEEE Access. 2022;10:60141-50.

[43] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015;10:e0130140.

[44] Montavon G, Lapuschkin S, Binder A, Samek W, Müller K-R. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern recognition. 2017;65:211-22.

[45] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods. 2020;17:261-72.

[46] Vallat R. Pingouin: statistics in Python. J Open Source Softw. 2018;3:1026.

[47] Michal M, Koechel A, Canterino M, Adler J, Reiner I, Vossel G, et al. Depersonalization disorder: disconnection of cognitive evaluation from autonomic responses to emotional stimuli. PLoS One. 2013;8:e74331.

[48] Alber M, Lapuschkin S, Seegerer P, Hägele M, Schütt KT, Montavon G, et al. iNNvestigate neural networks! J Mach Learn Res. 2019;20:1-8.

[49] Hunter EC, Charlton J, David AS. Depersonalisation and derealisation: assessment and management. BMJ. 2017;356:j745.

[50] Simeon D, Guralnik O, Schmeidler J, Sirof B, Knutelska M. The role of childhood interpersonal trauma in depersonalization disorder. Am J Psychiatry. 2001;158:1027-33.

[51] Michal M, Beutel ME, Jordan J, Zimmermann M, Wolters S, Heidenreich T. Depersonalization, mindfulness, and childhood trauma. The Journal of nervous and mental disease. 2007;195:693-6.

[52] Hollander E, Carrasco JL, Mullen LS, Trungold S, DeCaria CM, Towey J. Left hemispheric activation in depersonalization disorder: a case report. Biol Psychiatry. 1992;31:1157-62.

[53] Jiménez-Genchi AM. Repetitive transcranial magnetic stimulation improves depersonalization: a case report. CNS spectrums. 2004;9:375-6.

[54] Gainotti G. The role of the right hemisphere in emotional and behavioral disorders of patients with frontotemporal lobar degeneration: an updated review. Front Aging Neurosci. 2019;11:55.

[55] Silberman EK, Weingartner H. Hemispheric lateralization of functions related to emotion. Brain Cogn. 1986;5:322-53.

[56] Itz ML, Schweinberger SR, Kaufmann JM. Effects of caricaturing in shape or color on familiarity decisions for familiar and unfamiliar faces. PLoS One. 2016;11:e0149796.

[57] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57:289-300.

[58] Salami A, Andreu-Perez J, Gillmeister H. Towards decoding of depersonalisation disorder using EEG: A time series analysis using CDTW. 2020 IEEE Symposium Series on Computational Intelligence (SSCI): IEEE; 2020. p. 548-53.

[59] Daniels JK, Gaebler M, Lamke J-P, Walter H. Grey matter alterations in patients with depersonalization disorder: a voxel-based morphometry study. J Psychiatry Neurosci. 2015;40:19-27.

[60] Ke H, Cai C, Wang F, Hu F, Tang J, Shi Y. Interpretation of frequency channel-based CNN on depression identification. Front Comput Neurosci. 2021;15:773147.

[61] Ke H, Chen D, Shi B, Zhang J, Liu X, Zhang X, et al. Improving brain E-health services via high-performance EEG classification with grouping Bayesian optimization. IEEE Transactions on Services Computing. 2019;13:696-708.

APPENDIX

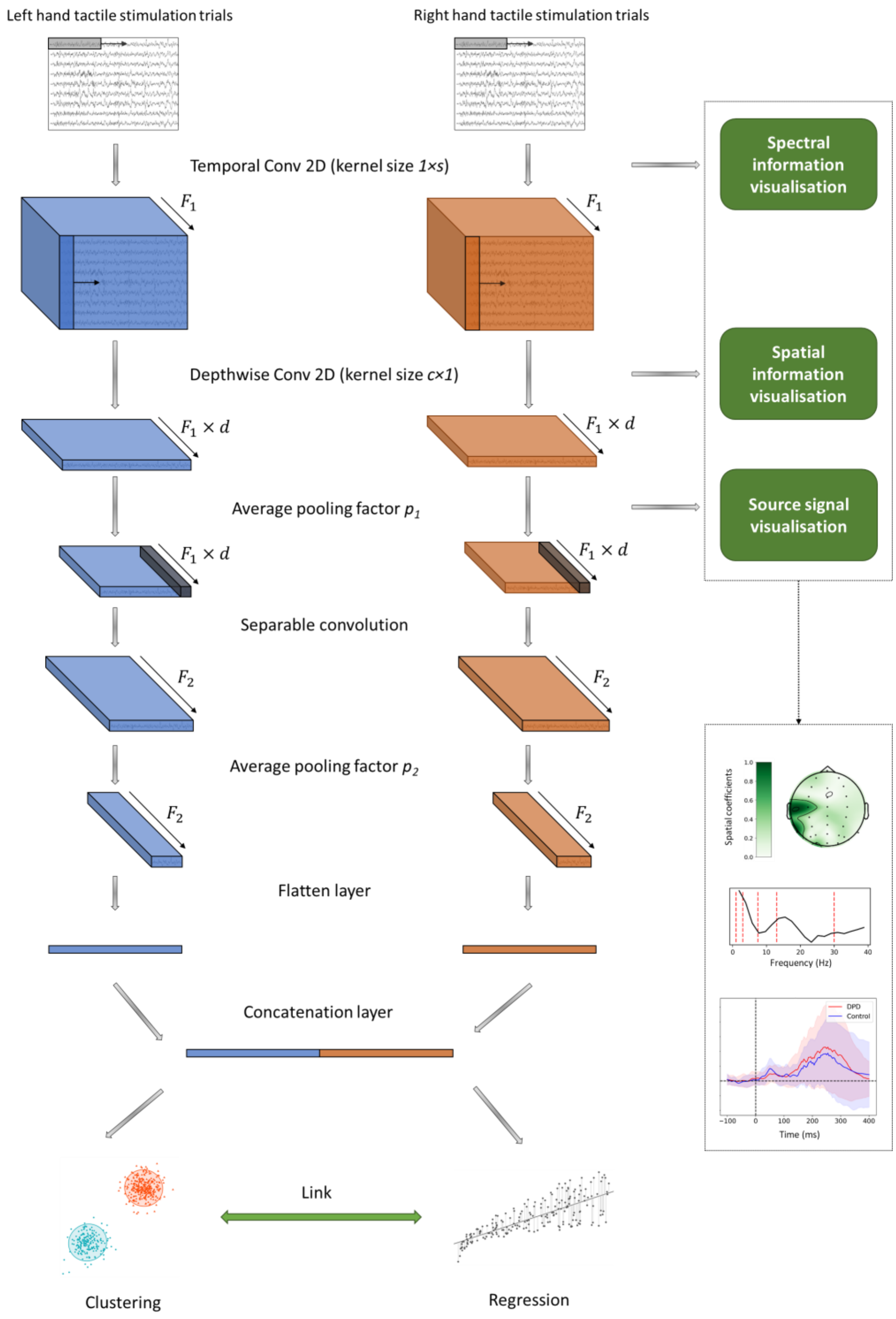

Left hand tactile stimulation trials
Right hand tactile stimulation trials
Temporal Conv 2D (kernel size *1×s*)
F_1
F_1
Spectral information visualisation
Depthwise Conv 2D (kernel size *c×1*)
$F_1 \times d$
$F_1 \times d$
Spatial information visualisation
Average pooling factor p_1
$F_1 \times d$
$F_1 \times d$
Source signal visualisation
Separable convolution
F_2
F_2
Spatial coefficients
1.0
0.8
0.6
0.4
0.2
0.0
Average pooling factor p_2
F_2
F_2
Flatten layer
0
10
20
30
40
Frequency (Hz)
Concatenation layer
DPD
Control
−100
0
100
200
300
400
Time (ms)
Link
Clustering
Regression

Figure 17. Overview of the proposed deep learning algorithm

Table 3
Selected hyperparameters for the proposed deep model

Hyperparameter name	Hyperparameter value
Number of filters in temporal convolution (F_1)	4
Number of filters in depthwise convolution (d)	2
Number of filters in separable convolution (F_2)	8
Temporal convolution kernel size ($1 \times s$)	1×128
Separable convolution kernel size	1×32
First average pooling factor (p_1)	2
Second average pooling factor (p_2)	4
Dropout rate	0.2
Batch size	32
Learning rate	10^{-4}
Early stopping patience	20
Number of generated trials using resampling-average method	60
$w_{regression}$	1.058×10^{-4}
$w_{clustering}$	3.333×10^{-4}
w_{link}	6.666×10^{-4}