Journal of Neural Engineering

CrossMark

OPEN ACCESS

RECEIVED
18 May 2022

REVISED
29 July 2022

ACCEPTED FOR PUBLICATION
19 August 2022

PUBLISHED
5 September 2022

PAPER

Decoding working memory-related information from repeated psychophysiological EEG experiments using convolutional and contrastive neural networks

Jarosław Żygierewicz[2,*], Romuald A Janik[1,4], Igor T Podolak[1,3], Alan Drozd[1], Urszula Malinowska[1], Martyna Poziomska[2], Jakub Wojciechowski[1,6], Paweł Ogniewski[7], Paweł Niedbalski[7], Iwona Terczynska[5] and Jacek Rogala[6,8,*]

[1] Nencki Institute of Experimental Biology, Polish Academy of Science, Pasteura 3, 02-093 Warsaw, Poland
[2] Faculty of Physics, University of Warsaw, Pasteura 5, 02-093 Warsaw, Poland
[3] Faculty of Mathematics and Computer Science, Jagiellonian University, Łojasiewicza 6, 30-348 Kraków, Poland
[4] Institute of Theoretical Physics, Jagiellonian University, Łojasiewicza 11, 30-348 Kraków, Poland
[5] Institute of Mother and Child (IMC), Clinic of Paediatric Neurology, Kasprzaka 17A, 01-211 Warsaw, Poland
[6] Bioimaging Research Center, World Hearing Center, Institute of Physiology and Pathology of Hearing, Mokra 17, 05-830 Nadarzyn, Poland
[7] ELMIKO BIOSIGNALS LTD, Sportowa 3, 05-822 Milanowek, Poland
[8] The Center for Systemic Risk Analysis, Faculty of 'Artes Liberales', University of Warsaw, Nowy Świat 69, 00-046 Warsaw, Poland
[*] Authors to whom any correspondence should be addressed.

E-mail: j.zygierewicz@uw.edu.pl **and** j.rogala@nencki.edu.pl

Keywords: explainable EEG classification, EEG decoding, neurofeedback

Abstract

Objective. Extracting reliable information from electroencephalogram (EEG) is difficult because the low signal-to-noise ratio and significant intersubject variability seriously hinder statistical analyses. However, recent advances in explainable machine learning open a new strategy to address this problem. *Approach.* The current study evaluates this approach using results from the classification and decoding of electrical brain activity associated with information retention. We designed four neural network models differing in architecture, training strategies, and input representation to classify single experimental trials of a working memory task. *Main results.* Our best models achieved an accuracy (ACC) of 65.29 ± 0.76 and Matthews correlation coefficient of 0.288 ± 0.018, outperforming the reference model trained on the same data. The highest correlation between classification score and behavioral performance was 0.36 ($p = 0.0007$). Using analysis of input perturbation, we estimated the importance of EEG channels and frequency bands in the task at hand. The set of essential features identified for each network varies. We identified a subset of features common to all models that identified brain regions and frequency bands consistent with current neurophysiological knowledge of the processes critical to attention and working memory. Finally, we proposed sanity checks to examine further the robustness of each model's set of features. *Significance.* Our results indicate that explainable deep learning is a powerful tool for decoding information from EEG signals. It is crucial to train and analyze a range of models to identify stable and reliable features. Our results highlight the need for explainable modeling as the model with the highest ACC appeared to use residual artifactual activity.

1. Introduction

Neurofeedback is a therapy aimed at improving cognitive abilities through self-regulation of brain activity in a direction considered desirable by the therapist (Hammond 2011). The current and desired levels of electroencephalogram (EEG) activity (typically power in a chosen frequency band on selected electrodes) are presented to patients in the form of a simple game in which the patient's task is to reach the target state using mental manipulations. Despite many years of development, the training protocols used to treat different disorders and deficits are still based on arbitrary selected frequencies

and electrodes. This lack of individual diagnosis and arbitrary selection of signal features seriously limit the method's effectiveness. Moreover, the complex nature and inter and intra-subject variability of the EEG signal hinder improvements using traditional computational techniques.

Machine learning (ML) methods are promising to overcome these limitations because they can incorporate a wide range of signal features and simultaneously generalize the knowledge. Deep neural networks (DNNs) have the potential to learn effective features end-to-end and to classify raw input data. Given their effectiveness in other fields, DNNs are expected to lead to better features and classifiers and thus to a much more robust EEG classification (Lotte *et al* 2018). Schirrmeister *et al* (2017) rigorously and convincingly demonstrated that their Shallow ConvNet—a simple convolutional neural network (CNN)—could outperform a classical filter-bank-common-spatial-pattern-based[9] classifier.

However, the practical application of modern ML methods such as deep and CNNs for EEG classification is hindered by two main problems. First, small datasets carry the risk of overfitting. Second, black-box approaches risk using artifacts as features (Comstock *et al* 1992, Nathan and Contreras-Vidal 2016). While this might not be crucial for Brain-Computer Interfaces, in the case of neurofeedback or diagnostic applications, this may lead to false diagnoses and therapeutic adverse effects. The challenges of small data size and classification interpretability are not new in ML. Mitigation techniques have been developed and applied in different fields, including EEG. In the case of addressing small data sizes, a commonly used technique is transfer learning. It was applied as early as 1998 (Thrun and Pratt 1998) and today is commonly used in EEG classification (for review, see Wan *et al* 2021). Newer solutions include techniques such as self-supervised contrastive learning (Hyvarinen and Morioka 2016). This solution has already yielded promising results in brain imaging EEG investigations (Mohsenvand *et al* 2020, Banville *et al* 2021). The techniques used to overcome the interpretability problem include sensitivity analysis and backpropagation approaches. Sensitivity analysis is a more popular method that is based on the local evaluation of the output gradient with respect to input features. Sensitivity analysis results are presented in heat maps consisting of input features with the most significant impact on the output. These new developments improve classification results comparable to the level of specialists (Schweizer *et al* 2017) without the burden of tedious work. The application of these methods may have practical assistive applications in diagnostics and therapeutics.

In this work, we study the application of different classifiers trained to detect desired EEG states as potential EEG-neurofeedback protocols. The classifiers have been trained on the data collected from individual participants during a cognitive task conducted prior to the actual EEG-neurofeedback training. The feasibility of CNN to detect different mental states has already been confirmed by several studies (Bird *et al* 2018, Chakladar *et al* 2020, Han *et al* 2020).

We chose memory as the target for the neurofeedback procedure, as memory is one of the basic cognitive functions. Using datasets collected during three diagnostic sessions of each participant, we trained four neural network models to classify trials based on the EEG signal. The application of four different models to the same dataset allowed us to compare the effects of different architectures, training strategies, and input representations on classification results and features' importance. Furthermore, we performed sanity checks to verify that the extracted features important for classification were plausibly related to the memory task. This study contributes to the field by demonstrating that:

(a) perturbation analysis (section 2.5.1) allows for identifying physiologically relevant features for neurofeedback training
(b) the application of different training strategies and models to the same problem may result in classification based on different sets of features; still, it allows the identification of a robust subset of features common to all the models,
(c) the relationship between classification results and behavioral performance in cognitive tests (often significant in diagnostic applications) is moderate and varies between investigated models.

2. Materials and methods

2.1. Data

In the current study we used two different sets of data. The first one comes from the delay matched to sample experiment which we refer to as the experimental data. The experiment was designed to closely resemble a neurofeedback session. The second set was a big dataset of resting state clinical recordings which we used to train the model using a transfer learning approach.

2.1.1. Experimental procedure and data

2.1.1.1. Participants

We examined 87 healthy adults (including 42 women; age range: 23–44 years). Participants were recruited through announcements at local universities and work agencies. A neurological screening and questionnaires were administered to all potential

[9] Filter-bank-common-spatial-pattern in a supervised technique that computes spatial filters (linear combinations of EEG channels) that enhance class-discriminative band power features contained in the EEG.

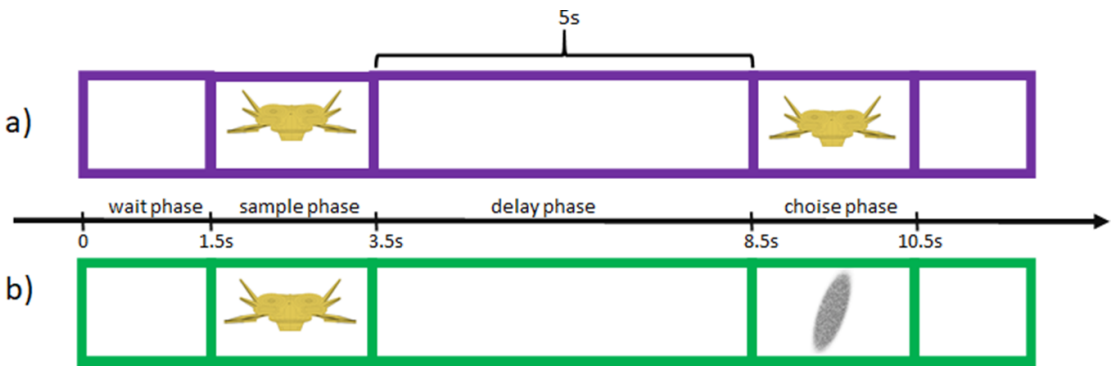

Figure 1. The scheme of the experimental procedure. Time course of (a) the retention and (b) control trials. In each case, the type of trial was signaled to the participant by the color of the background elements, purple for retention and green for control type. The 5 s delay phase indicated in the upper plot was used in the analysis described below.

participants; exclusion criteria included neurological disorders, brain injury, current use of analgesic medication, substance abuse or dependence and mental disorders. All participants were right-handed and had a normal or corrected-to-normal vision. The experimental procedures were approved by the Local Bioethics Committee at Nicolaus Copernicus University in Torun. All participants gave their written informed consent to participate in the experiment in accordance with the WMA Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects. All experiments were performed under all relevant guidelines and regulations. Neurofeedback set-up included virtual reality (VR) and standard monitor environments. Participants were randomly assigned to two experimental groups with different environment (2D or VR). Finally, the 2D group included 33 participants (including 12 women) and the VR group 54 participants (including 30 women).

2.1.1.2. Procedure

The current experimental procedure was designed for an EEG-neurofeedback experiment aimed at memory improvement in VR and standard monitor (2D) environments. The first three diagnostic sessions were aimed at decoding participants' EEG state during tasks requiring the use of working memory. During these sessions, participants used a keyboard to play a computer game implementing a working memory task while their EEG was recorded. The data collected from the three sessions were then used for decoding EEG activity. The game was identical with the one used in the proper neurofeedback training except for the participants' response which, during the proper neurofeedback training, was decoded from the modulations of their EEG. In this paper, we concentrate on identifying the working-memory-related EEG features from the three diagnostic sessions.

The game was based on a delayed match-to-sample (DMTS) task, a task which tests both attention and working memory. The original version of the DMTS has three phases which are sample, delay, and choice. During the sample phase, a participant is presented with a sample stimulus, which has to be maintained in memory during the delay period following the sample phase. When the delay phase is over, the choice phase follows. During this phase, another stimulus is presented to the participant. At this point, the participant has to decide whether or not the stimulus matches the sample.

We modified this paradigm by adding control trials that did not require retention of information (See figure 1 for the trial design). In control trials, the sample and delay phases were the same as non-control phases. However, in the choice phase, participants were instead asked to indicate the orientation (left or right) of a simple shape. The type of trial was indicated by different background colors. Therefore, participants were always aware whether or not a given trial required attention and retention of information.

Each trial lasted 10.5 s and began with a 'wait' phase lasting 1.5 s. In the 'wait' phase, no action from the participants was required. Next, in the sample phase, the silhouette of a spaceship was presented to the participants for 2 s. This was followed by a 5 s 'delay' phase (retention trial) used in further analyses for memory load detection. Finally, the 'choice' phase lasted for 2 s. Participants' responses were executed by pressing assigned buttons on the computer keyboard. They earned one point per each correct response and lost a point per each incorrect response. The three sessions of the DMTS game combined with EEG recording were separated by 2–3 d. Each EEG session lasted up to 25 min and consisted of 50 control and 50 retention trials, shown in randomized order.

2.1.2. Clinical data

The clinical data used for transfer learning in one of the approaches explored in the present paper came from University of Warsaw's database. The database contains over 100 000 healthy and abnormal anonymized clinical EEG recordings of patients of all ages collected from hospitals across Poland. The subset used

in the experiment included 12 000 randomly selected recordings. The use of this database for research purposes was approved by the Local Bioethics Committee at Nicolaus Copernicus University in Torun.

2.2. Data pre-processing

2.2.1. Experimental data preprocessing

The EEG signal was recorded using Digitrack software (Elmiko Ltd) with 19 electrodes arranged in a 10–20 system, referenced to electrodes located at the earlobes. The sampling rate of EEG signals was 500 Hz and the impedance of electrodes was kept below 10 kΩ. Data were high-pass filtered with a 0.5 Hz cutoff frequency. The experimental data were subjected to a typical EEG cleaning procedure to remove known artifacts that could bias classification results and consequently the EEG-neurofeedback therapy. To assure the same sampling rate of experimental recordings (the original sampling rate slightly differed between recordings), the data were downsampled to 400 Hz. Next, the signal was epoched into 10.5 s windows (covering a whole single trial). For automatic bad trials (heavily contaminated) detection, the ERPLAB function (Lopez-Calderon and Luck 2014) was used with adaptable threshold parameters, leading to removal of no more than 25% of threshold-exceeding trials per participant. Eye movement, electrocardiogram (ECG), and muscle artifacts components were removed using independent component analysis (EEGLAB (Delorme and Makeig 2004) plugin ICLabel with classification threshold set to 0.65 for eyes and ECG and 0.85 for muscles). The delay phase (cf figure 1) was extracted from correct trials for further analysis. Since VR and 2D groups did not differ in the behavioral results (see section 3 for details), we pooled the two groups to increase the size of the training and testing sets. The final dataset consisted of 6207 trials in session 1, 6509 trials in session 2, and 6587 trials in session 3. There were 10 937 correct control and 8366 correct retention trials in total.

2.2.2. Clinical data preprocessing

The signals were recorded using a 10–20 electrode setup, and sampled at 256 Hz. Data were bandpass filtered using 1 Hz highpass and 40 Hz lowpass filters. Next, after cropping the continuous recording into non-overlapping 5 s long samples, the ones that exceeded 500 μV were discarded. Finally, we obtained 3.5×10^6 5 s long samples.

2.3. Model architectures

In the current study, the effect of different architectures and training methods on classification results and set of important features was evaluated using the following four models:

(a) Shallow ConvNet—a reference model developed originally by Schirrmeister *et al* (2017).
(b) Parallel ConvNet and Hybrid models—both models are using channel-frequency-time input representation and share the same shallow convolutional part of the architecture. Additionally, the Hybrid model is tuned to individual participants. Both models were developed specifically for the current neurofeedback project.
(c) Contrastive model with gated multilayer perceptron (gMLP-MoCo)—also developed for the current neurofeedback project, aimed at assessment of transfer learning using contrastive training phase.

All models were trained to perform a binary classification of the input data as coming from the retention or control trials (classes) (c.f. figure 1). The details of the models are described below. The models were implemented using Pytorch 1.8.

2.3.1. Shallow ConvNet

The rationale behind choosing Shallow ConvNet (Schirrmeister *et al* 2017) as a reference model was twofold. First, the original Shallow ConvNet model was developed for classification of EEG dataset similar in size to ours (High Gamma Dataset: 14 participants consisting in total 12 320, 4 s long trials, Our dataset: 89 participants, 3 sessions per participant, in total 19 303, 5 s long trials). Second, Shallow ConvNet is a widely used reference model (over 1200 citations in Google Scholar). We implemented a variant of the original model adjusted for binary classification. In short, the first two layers of the network perform respectively: a temporal convolution and a spatial filtering analogous to Filter Bank Common Spatial Patterns (Ang *et al* 2008). The convolutional layers are followed by a squaring non-linearity, a mean pooling layer and a logarithmic activation function. The last layer is a fully connected layer with sigmoid non-linearity. In our implementation the size of the temporal filters was set to 40 to obtain the same length (in seconds) as in the original work (Schirrmeister *et al* 2017). The network's input size was $E \times T$, where E is the number of electrodes, and T is the number of time steps. The scheme of the model is presented in figure 2.

2.3.2. Parallel ConvNet and Hybrid models

Parallel ConvNet and Hybrid models were developed explicitly for the current study. For these models, we considered explicit time-frequency parametrization to study the impact of the input representation. After prepossessing described in section 2.2.1, the signal was converted into time-frequency representation using Morlet transform (wave number = 7) for central wavelet frequencies 3, 5, 8, 11, 15, 20, 25, 30, and 35 Hz. Chosen frequencies overlap with canonical EEG bands (i.e. theta 4–7 Hz, alpha 8–12 Hz, beta 1 13–20 Hz, beta 2 21–30 Hz and gamma 31–45 Hz).

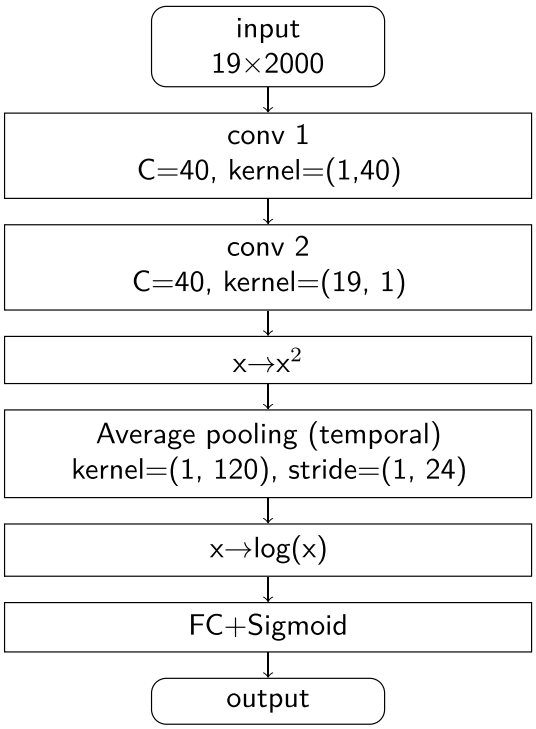

Figure 2. Outline of the Shallow ConvNet. C is the number of kernels; conv 1: temporal filters, conv 2: spatial filter mixing all electrodes, and all previous filters, FC: the fully connected layer.

We used only the real part of the complex representation for further computations and decimated it by 5 in the time domain to reduce the input size. This procedure yielded a 9×400 frequency-time map for each channel. Finally, these maps were stacked, forming a tensor of rank 3 with dimensions $19 \times 9 \times 400$ (channel $\times$ frequency $\times$ time). Consequently, the input of our network (see figure 3) had a dimension of E $\times$ F $\times$ T, where E is the number of electrodes, F the number of different frequency bands, and T the number of time steps. The number of electrodes (E) corresponds to the channel's dimension of the input tensor, i.e. to its 'depth'. Within each block, the output of convolution layer is followed by a rectified linear activation, $\mathrm{ReLU}(x) = \max(0, x)$, and then by a batch normalization layer (Ioffe and Szegedy 2015). Both models contain two convolution layers with three parallel paths of different kernel sizes (time steps) (see figure 3). The concept of parallel paths was inspired by the Inception architecture (Szegedy *et al* 2015) implemented in the context of computer vision. The Inception architecture (Szegedy *et al* 2015) introduced parallel computational pathways with different convolutional filters. We implement an analogous idea but adapted this architecture to the context of time-series multi-channel EEG data. The precise employed variant was fixed using cross-validation (CV) on a relatively small subset of the data as giving reasonable performance. We did not, however, attempt to perform a full-fledged architecture scan, as the main point of the present work was to use the neural network as a tool for neurofeedback training using interpretable EEG features. The kernel sizes in Parallel ConvNet and Hybrid models are (p, q), where p corresponds to the number of adjacent frequencies, and q to the length of the time window. Thus, the model parses the signal at three different 'speeds' depicted in kernel sizes: $(5, 10) \rightarrow (5, 5)$, $(5, 5) \rightarrow (5, 10)$, and $(5, 1) \rightarrow (5, 10$ with $2 \times$ dilation). The convolution-activation-normalization blocks are followed by an average pooling layer which aggregates features from the whole duration of the trial. Outputs of these three paths are then concatenated forming a 24-dimensional representation of the input.

In the Parallel ConvNet this representation is passed to two hidden fully connected layers. All fully connected layers receive an additional scalar input marking the type of experimental environment: 2D or VR. The model output value is computed in the third fully connected layer with a *sigmoid* function $\sigma(x) = 1/(1 + \exp(-x))$ $\sigma(x) \in (0, 1)$. This output can be interpreted as the probability of whether the trial is in the retention class.

In the Hybrid model we took advantage of multisession characteristics of our dataset to construct individual models to minimize adverse effect of large inter-subject EEG variability on classification results (Lotte *et al* 2018). To this end, the 24-dimensional representation (output of the 'concatenate' block) of the trained Parallel ConvNet was used as a fixed feature extractor input to individual logistic regression models trained on each participant's data.

2.3.3. Contrastive model

The motivation for this specific choice was twofold. First, we sought to evaluate the effectiveness of the transfer learning method using unlabeled raw clinical EEG data. The experimental EEG datasets are usually small in size, which reduces the training efficiency and carries the risk of overfitting. On the other hand, clinical EEG data are recorded according to standardized procedures, and the dataset sizes are much larger (e.g. a TUH dataset of 3000 recordings). Therefore, transfer learning from clinical EEG data may provide a solution to the problem of small size of experimental recordings, enabling more accurate classification without the risk of overfitting. Second, the application of the gMLP with self-attention mechanism to train targeted dataset using features extracted from raw clinical data could reveal importance of other features than standard analytical methods and increase the fine-tuning capabilities of the model.

The structure of the contrastive model (gMLP-MoCo) used in this study is shown in figure 4. The input is the channel-time representation of EEG signal lasting 5 s (matrix of 1280 time steps of raw EEG data in 19 channels, at 256 Hz sampling rate).

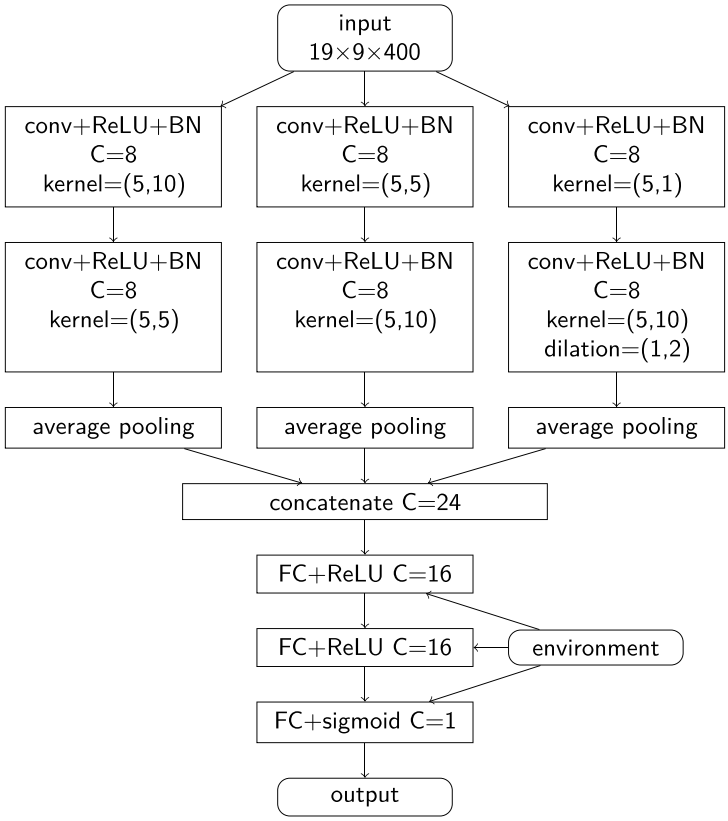

Figure 3. Parallel ConvNet architecture. Operations indicated in the blocks: conv: convolution with a kernel, ReLU: rectified linear activation function, BN: batch normalization, FC: fully connected layer, C: number of channels, i.e. kernels applied at the same location. The network returns the probability that the input trial class is retention.

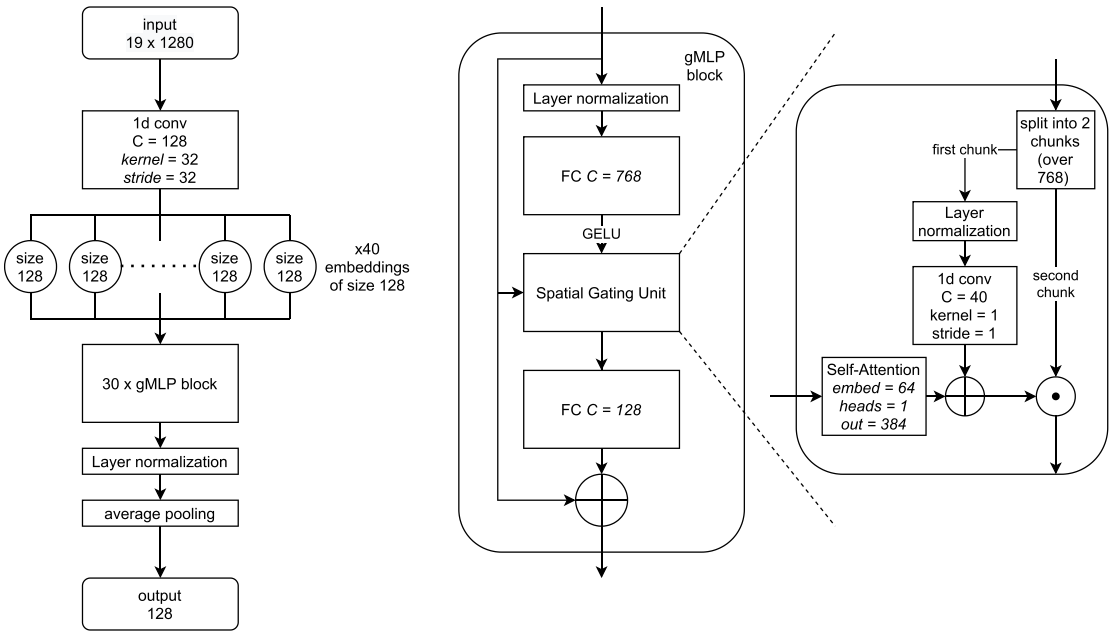

Figure 4. The gMLP network architecture. Left—overview of the network, center—structure of a gMLP block, right—zoom into the Spatial Gating Unit. See text for details. Adapted with permission from Liu *et al* (2021).

This input is fed to a one-dimensional convolutional layer which translates the signal into a sequence of 40 embeddings (corresponding to tokens), each representing the subsequent fragment of signal as a lower dimensional (here, 128) vector. The embeddings are then processed by the gMLP module. The gMLP (Liu *et al* 2021) is built of 30 identical blocks. In each block, embeddings are first normalized and

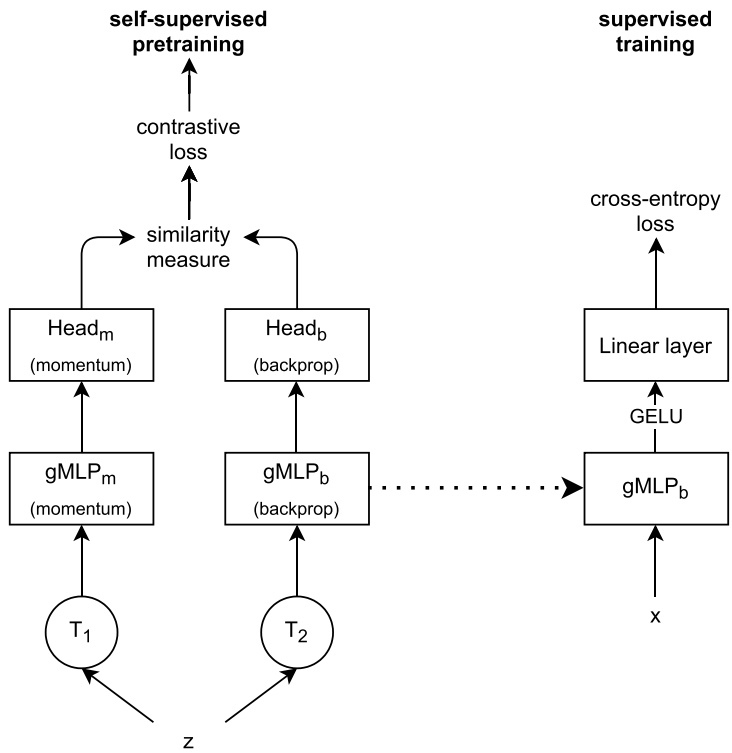

Figure 5. Training of the gMLP Momentum Contrastive gMLP-MoCo network occurs in two stages. In the self-supervised pretraining stage (left), the z is a training example from the clinical database that is distorted with T_1 and T_2 transformations, transformed with network and compared using a contrastive loss function as distance. In the supervised stage (right) an example x from the current experiment is passed through the network.

then projected by a shared linear layer into a higher dimension. These expanded tokens enter the Spatial Gating Unit module and are split into two equally sized chunks.

The first chunk after the normalization layer enters a trainable one-dimensional convolutional network and then after combining it with Self-Attention layer, it is pointwise multiplied by the second unprocessed chunk assuring interaction between tokens. The Self-Attention layer uses non-expanded tokens that enters the gMLP block. Processing finishes with normalization layer across embeddings and average pooling. The original gMLP model was adjusted to EEG signal processing: instead of a two-dimensional convolution, we used a one-dimensional along the time axis and electrodes acting as channels. Kernel size was doubled in linear size from 16 to 32, with appropriate changes to the strides.

2.4. Training and evaluation of the models

The low signal-to-noise ratio of EEG results in high instability of predictions in successive training epochs (i.e. iterations through the whole training dataset). To reduce the adverse effect of high variability, we averaged the predictions from the last three epochs.

2.4.1. Shallow, Parallel and Hybrid models

The Shallow ConvNet, Parallel ConvNet and pre-trained part of the Hybrid model were trained using the AdamW optimizer. Training was performed in batches of $N = 64$ with standard binary cross-entropy loss. The number of training epochs was chosen experimentally to mitigate overfitting and resulted in 20 epochs for Shallow ConvNet and 10 epochs for the Parallel ConvNet and Hybrid models.

2.4.2. Contrastive model

The gMLP-MoCo model was trained in two steps:

(a) self-supervised pre-training using a Momentum Contrastive Learning framework (MoCo). Our main motivation behind this choice was to extract EEG features which could bring a fresh insight into the EEG correlates of information retention. Contrastive learning methods (Wu *et al* 2018, Chen *et al* 2020, He *et al* 2020, Tian *et al* 2020) attempt to learn how to be invariant to transformations introduced to the training set while maintaining discrimination over other features;
(b) tuning of the pre-trained network on the data from the current experiment.

In both steps we used AdamW optimizer, with a learning rate of 1×10^{-4} and the batch size of 64.

In our implementation of the MoCo framework, outlined in figure 5, the original training example z (from the clinical dataset) were augmented using T_1 and T_2 transformations constructed by automated augment policy RandAugment from the list of basic

Table 1. The list of the basic transformations used in contrastive training.

Transformation	Description
Identity	Identity transformation
Noise	Add noise generated by a normal distribution
Signal cutout	Zero the signal across all electrodes on a randomly selected continuous section
Mean shift	Change the mean of an electrode by adding a randomly sampled number
Sensor dropout	Zero the signal on a randomly chosen set of electrodes
Sensor flipping	Flip upside down the signal on a randomly chosen set of electrodes
Bandstop filtering	Bandstop randomly selected range of frequency
Constant scaling	Multiply signal by a number randomly chosen for each electrode
Irregular scaling	Multiply signal by a cubic spline

transformations (table 1). Next each transformation was fed to one of two subnetworks: momentum or backprop. The subnetworks learned representations by maximizing contrastive loss InfoNCE on samples organized into similar and dissimilar pairs. The parameters of gMLP and Head modules of both sub-networks were updated with the back-propagation algorithm. However, in the case of the momentum sub-network (gMLPm and Headm), we used momentum mechanism (Chen *et al* 2020), i.e. the parameter vector, θ_m, of the momentum network was updated using past values of the θ_b parameter vector of non-momentum network according to $\theta_m = \alpha \cdot \theta_m + (1-\alpha) \cdot \theta_b$. The smoothly evolving momentum networks enabled us to reuse the old batches during the calculation of contrastive loss, which inherently requires a large batch size to work properly. We used 100 epochs for gMLP pre-training.

In the second step, the pre-trained gMLP_b module was coupled with a linear layer and tuned to the data examples x from the current experiment using cross-entropy loss. To prevent overfitting, the augment policy from the previous step was reused with the same set of basic transformations. The experimental data were downsampled to 256 Hz to match the sampling rate of clinical recordings. We used eight epochs for training.

2.4.3. Evaluation of models' performance

For estimation of models' performance we applied 3-fold CV on the experimental data from three individual experimental sessions. To further increase the number of estimates of the investigated measures, we repeated the CV for five random neural networks' initializations. Thus, we have finally results for $3 \times 5 = 15$ instantiations of each model. All models were evaluated using accuracy (ACC) and Matthews correlation coefficient (MCC). The latter was chosen for its insensitivity to class imbalance.

2.5. Feature importance

2.5.1. Evaluation of feature importance

In order to determine EEG features related to information retention from the neural network perspective, we isolate features which are relevant for the classification of the trial as retention versus control. Indeed, the only difference between the control and retention trials is the retention of information in memory during the analyzed segment of the trial.

To determine the features' importance for the classification results, we applied perturbation analysis using automatic gradient evaluation in Pytorch. We focused on the power of EEG signal in the canonical frequency bands across electrodes in order to facilitate comparison with classical EEG methods of analysis.

As part of the analysis, the input to the trained model was perturbed by multiplying the amplitude at a given electrode and a given frequency band by a factor $c_{e,f}$. Next, the derivative of the class probability p over the perturbation parameter at $c_{e,f} = 1$ was averaged over all versions of a given model (trained on 3 folds and for 5 random initializations) and evaluated on the relevant test set for the given fold to obtain the average feature importance index $\overline{\text{FI}}_{e,f}$ defined as:

$$\overline{\text{FI}}_{e,f} = \left\langle \frac{\partial p}{\partial c_{e,f}}_{|c_{e,f}=1} \right\rangle. \tag{1}$$

A positive value of $\overline{\text{FI}}_{e,f}$ means that an increase of power at electrode e in frequency band around f increases the probability of classification of a given input trial as the retention one.

In the case of Parallel ConvNet and Hybrid models, the gradients were directly evaluated for the Morlet coefficient. For the models using raw signal as input (Shallow ConvNet and gMLP-MoCo), data were bandpass filtered in frequencies corresponding to the central wavelet frequencies of Morlet transforms and then signal was reconstructed by summing the bands with the weights $c = 1$, with gradient computation turned on for the weights c, c.f. figure 6. In this way we preserved information about the direction of class probability change with the change of the given feature.

2.5.2. Testing of feature importance

2.5.2.1. Statistical tests

To test if the $\overline{\text{FI}}_{e,f}$ was significantly different from zero, we used a one-sample t-test on the set of perturbation results of all versions of a given model (i.e. individual $\text{FI}_{e,f}$). To account for the multiple comparisons, we

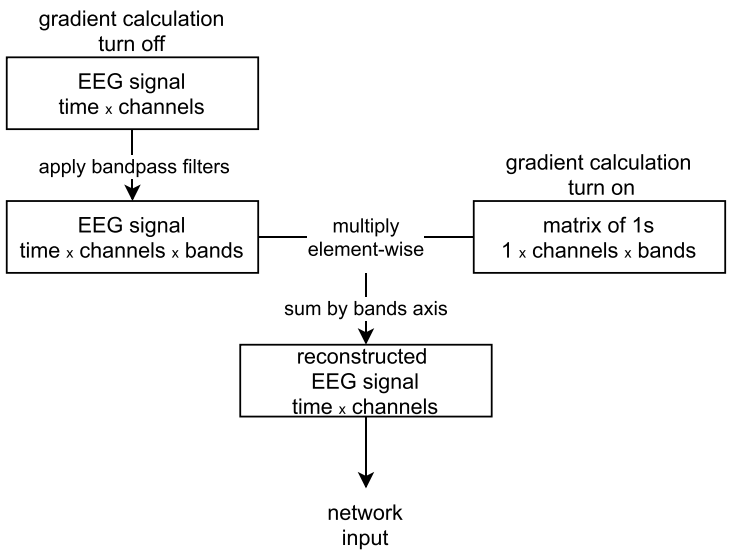

Figure 6. Methodology of perturbation analysis for time-channel signal representation.

applied a false discovery rate (FDR) correction (Benjamini and Yekutieli 2001).

2.5.2.2. Sanity checks

Comparison of perturbation analyses results obtained for different subgroups, i.e. participants with best and worst classification scores, gives opportunity to further validate the obtained feature importance with a sanity check. To this end we tested the significance of differences of values of feature importance between 30 best and 30 worst classified participants using the Mann–Whitney test. To account for the multiple comparisons, we applied FDR correction. Another simple sanity test was correlating the classification results to task performance.

2.6. Classical EEG analyses

As a reference to perturbation analyses, we used classical spectral EEG analyses. The power in each of the frequency bands was estimated by summing periodograms in ranges corresponding to the frequency bands of Morlet wavelets used in our models, namely: 1–5, 3–7, 7–9, 10–12, 13–17, 15–25, 20–30, 25–30, and 30–40 Hz. We performed a series of permutation tests, shuffling the labels 'retention' and 'control' trials for each combination of (electrode, frequency band) with FDR correction for multiple comparisons setting a 0.05 p-value threshold (we used the standard implementation from EEGLAB Matlab toolbox).

3. Results

3.1. Behavioral results

The average number of points scored by a participant in all three diagnostic sessions equaled $M = 35.7$, $SD = 5.17$. The VR and 2D group results did not differ significantly for all three sessions together (2 tailed Wilcoxon test, $p = .61$), nor for individual sessions (2 tailed Wilcoxon test $p_1 = .49$, $p_2 = .97$ and $p_3 = .69$ for sessions 1 through 3 respectively).

3.2. Model performance

ACC and MCC scores of the tested models are shown in table 2. For the sake of reference, using just the 'most frequent class' as a naive classifier yielded an ACC of 56% and 0 MCC. All the evaluated models performed better.

The statistical comparison of ACC and MCC of our models using the Kruskal–Wallis test showed that there were significant differences between the scores of different models (for ACC $\chi^2_4 = 41.92$, $p < .001$; for MCC $\chi^2_4 = 41.25$, $p < .001$).

The post-hoc tests, done using the Mann–Whitney two-sided test with FDR correction for multiple comparisons, indicated that all the pairwise differences, excluding Shallow vs Parallel ConvNets were significant, both for ACC and MCC. The details of significant post-hoc tests are reported in table 3. To summarize, the best scores were obtained for the gMLP-MoCo model, slightly lower for Hybrid, and the lowest for Parallel and Shallow ConvNets, the last two performing on the same level.

3.3. Feature importance indicated by the perturbation analysis

In diagnostic and therapeutic applications, such as EEG-neurofeedback, classification ACC and the features which were the basis of classifications are of equal significance. Therefore, we performed perturbation analysis to identify essential features. For each model, we evaluated the feature importance index according to equation (1). The results are presented in figure 7. It provides for the explainability of the model.

Table 2. ACC and MCC obtained in 3-fold CV and five random initializations training of the models.

Model	ACC	MCC	# Trainable parameters
Shallow ConvNet	61.50 ± 2.33	0.216 ± 0.030	3.5×10^4
Parallel ConvNet	62.06 ± 1.39	0.223 ± 0.025	2.1×10^4
Hybrid model	64.38 ± 0.60	0.264 ± 0.011	2.1×10^4
gMLP-MoCo	65.29 ± 0.76	0.288 ± 0.018	6.1×10^6

Table 3. Significant differences in ACC and MCC assessed with a two-sided Mann–Whitney test; p values with FDR.

	ACC		MCC	
Comparison	U	p	U	p
Shallow ConvNet vs Hybrid model	20	2×10^{-04}	20	2×10^{-04}
Shallow ConvNet vs gMLP-MoCo	4	1×10^{-05}	6	2×10^{-05}
Parallel ConvNet vs Hybrid model	2	1×10^{-05}	6	2×10^{-05}
Parallel ConvNet vs gMLP-MoCo	0	1×10^{-05}	1	2×10^{-05}
Hybrid model vs gMLP-MoCo	44	6×10^{-03}	37	2×10^{-03}

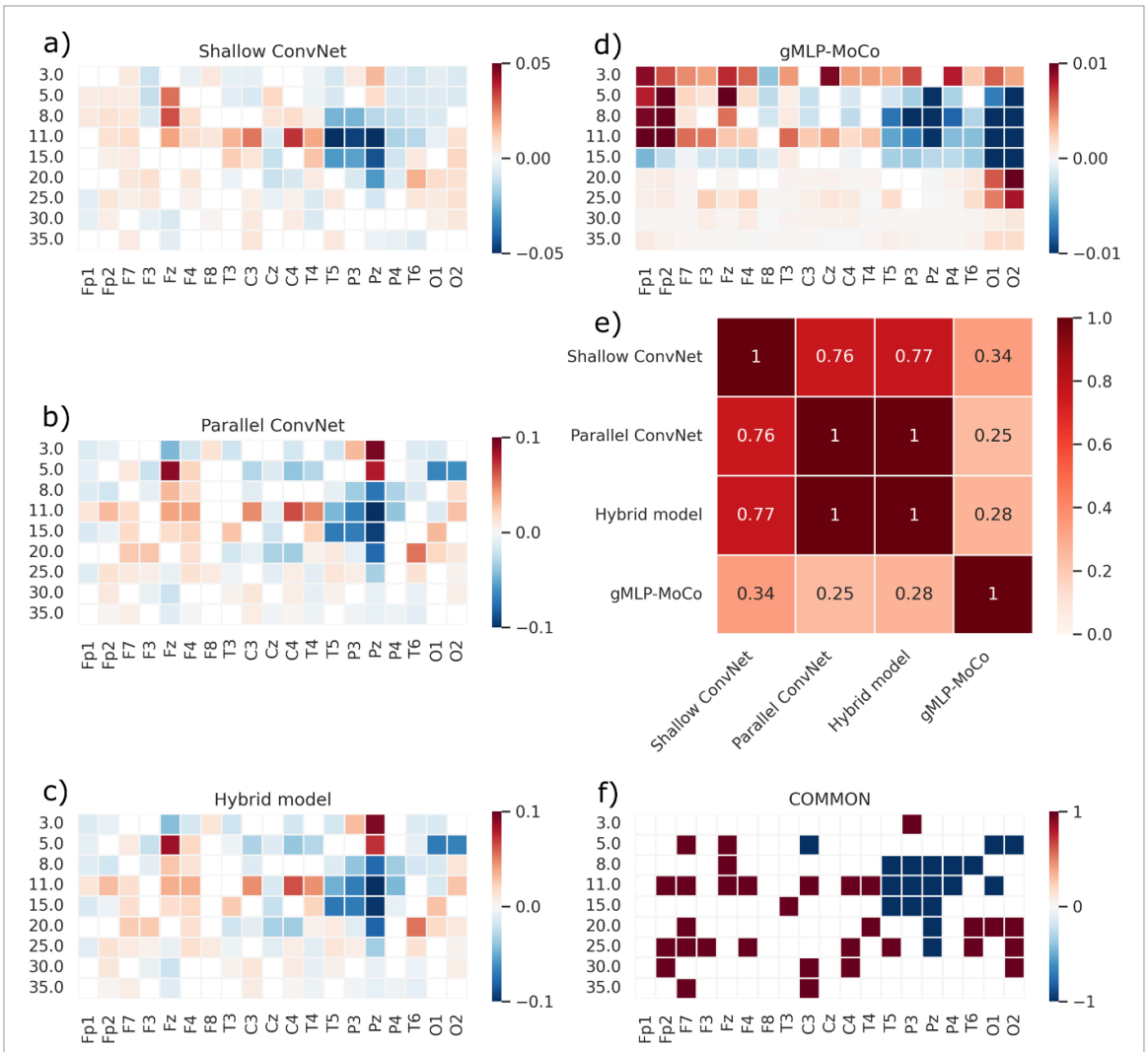

Figure 7. Results of perturbation analysis. (a)–(d) heatmaps of feature importance index for the four investigated models. The channel-frequency pairs masked white were not statistically significant. (e) Spearman correlation between the heatmaps; all correlations statistically significant ($p < .001$). (f) Elements of heatmaps common to all the models—red positive, blue negative $\overline{\mathrm{FI}}_{e,f}$ for all models.

The pattern of $\overline{\mathrm{FI}}_{e,f}$ obtained for models trained directly on the experimental task data (figures 7(a)–(c)) is very similar, which was confirmed by the strong Spearman correlation (over 0.75). Comparison with the gMLP-MoCo model revealed a weaker correlation (0.34 with Shallow,

and 0.25 and 0.28 with Parallel and Hybrid models, respectively). Nevertheless, there was an emerging pattern common to all the models (figure 7(f)). Increased power in the temporal-posterior electrodes in the alpha and beta frequencies was related to the decrease in the probability of a trial being of class retention. On the other hand, the augmented power at frontal and temporal-central electrodes in alpha and beta frequency bands was indicative of the trial having requiring more working memory. Additionally, at electrodes Fz and F3, increased theta activity corresponded to class retention trials. Finally, increased beta and decreased theta activity in the occipital were also characteristic of the retention class.

The feature importance index heatmap obtained for the gMLP-MoCo model showed the importance of the delta EEG band (3 Hz) at all electrodes and the alpha-band at occipital ones, which was not indicated by the perturbation analysis performed on other models. Another important feature identified only for the gMLP-MoCo model was the increase of the delta through alpha bands activity at the Fp1 and Fp2 electrodes indicating higher probability of the retention trials.

3.4. Sanity checks

We expect that the features which consistently indicate a higher probability of classifying a trial as a retention class should be characteristic of working memory. Thus, comparing the feature importance index calculated for best and worst classified participants may inform us whether features used by our models to detect retention trials differ between groups. Further, analysis of the correlation of the classification scores with behavioral results may clarify to what extent the classification, based on the developed set of features, corresponds to the desired behavioral aim of the neurofeedback training.

3.4.1. Differences between best and worst classified participants

We checked the differences between groups with the highest and lowest classification results. We compared the feature importance index of 30 participants with best and worst ACC and MCC for each model. Their ACC and MCC scores were significantly different (2-tailed t-test $p < .001$). The statistically significant differences in $\overline{\mathrm{FI}}_{e,f}$, as revealed by a series of Mann–Whitney tests with FDR correction, are presented in figure 8. The Shallow, Parallel ConvNets, and Hybrid models, i.e. the models trained directly on the experimental task data, showed several significant regions in frequency-electrode space. These regions can be seen as the most robust features developed during the model training. Furthermore, the signs of the gradients are the same for the highest and lowest classification scores, and, as could be expected, they have a higher absolute value for the best-classified group, which is represented as more saturated colors in figure 8 (right compared to left column).

3.4.2. Correlation between classification results and task performance

EEG features differentiating retention and control trials common to all models were highly consistent with those associated with neuronal processes engaged in memory. Furthermore, they were more pronounced in the participants with higher classification ACC. Therefore, one could expect that classification score metrics (ACC, MCC) would correspond to some extent to behavioral performance. We performed correlation analysis between game scores and classification metrics to verify whether classification results of the tested models are related to behavioral performance in the memory task. Interestingly, significant results were found only for the MCC metrics. Pearson correlation coefficients ranged from $r = .36$ for Hybrid and Shallow models to $r = .25$ for gMLP-MoCo. The correlation appeared significant for all models. Details are shown in table 4.

3.5. Classical analyses

Traditional spectral analyses (section 2.6) did not reveal any significant differences between retention and control trials. We also did not find significant group differences in EEG activity between best and worst classified participants.

4. Discussion

The application of ML methods to classify EEG signals in research and medicine is hampered by the inherent problems of low signal-to-noise ratio, high inter- and intra-subject variability, and stochasticity (Subha *et al* 2010). In the race to overcome the obstacles mentioned above and obtain the best possible classification results, explainability is often sacrificed. The need for explainability, understood as reasons a model gives to make its functioning clear or easy to understand (Barredo Arrieta *et al* 2020), has a twofold rationale. The first is to ensure that the classification is based on features related to the problems being solved (e.g. disorders) and not the accompanying artifacts (Wan *et al* 2021). The second rationale is to gain a comprehensive insight into the problem under consideration. Unfortunately, most of the previous studies dealing with the classification of cognitive functions vary widely in implemented solutions, cognitive tasks, and datasets, hindering comparison of the results and their explainability.

The present study performs classification and feature importance analyses of working memory load based on EEG as a potential method for EEG-Neurofeedback applications. To directly compare different architectures, training methods, and input representations, we implemented and investigated

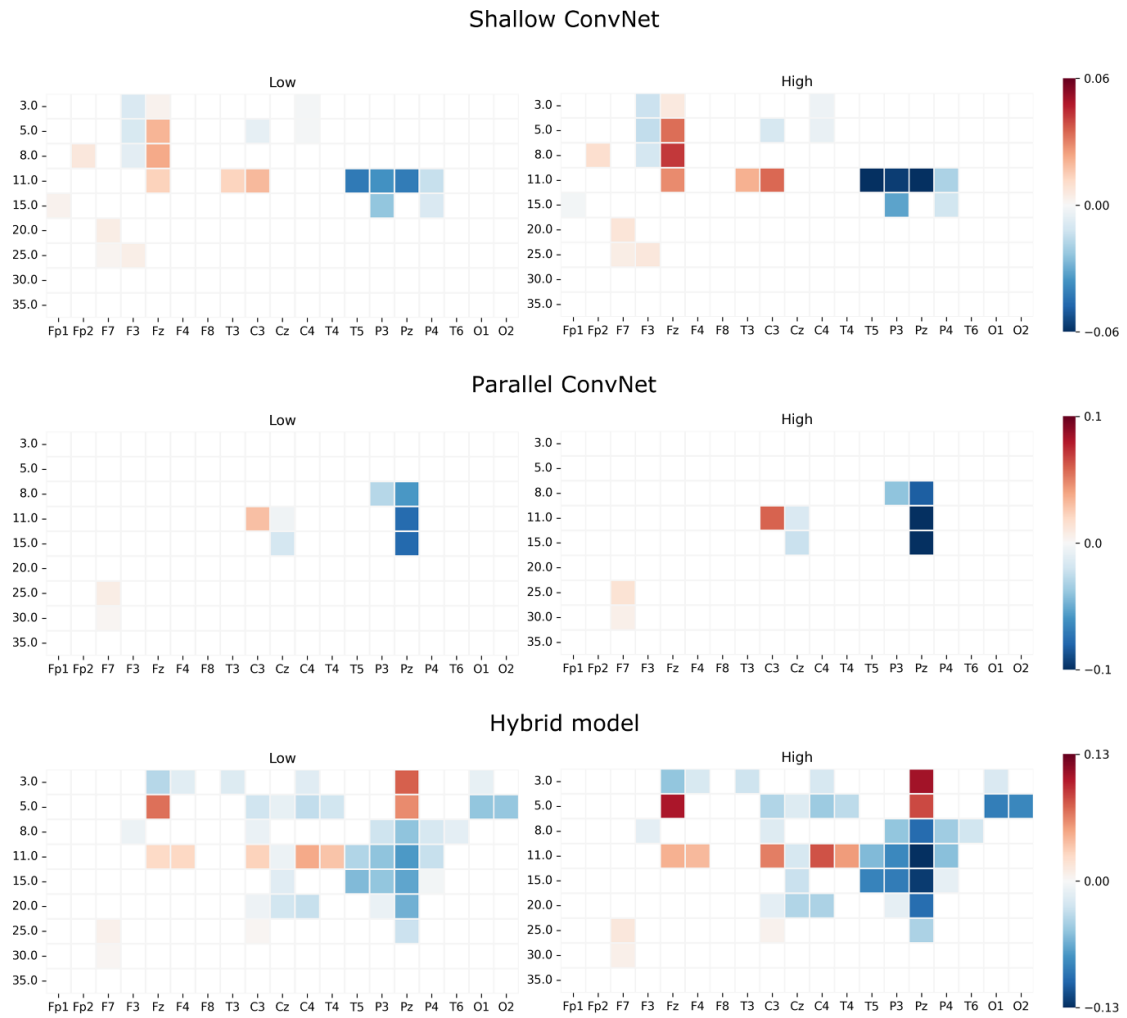

Figure 8. Sanity check. The average feature importance index for the groups of participants with low (left) or high (right) ACC in individual classification achieved by the Shallow ConvNet, Parallel ConvNet, and Hybrid model. The features within each model were compared with Mann–Whitney test. The red color indicates that the increase of the associated feature value increases the probability of the trial being of class retention; the blue color means that the increase of that feature value decreases the likelihood of the trial being of retention class. The nonsignificant differences are masked white. The vertical axis shows frequency in Hz, the horizontal axis represents EEG channels.

Table 4. Pearson correlation coefficients for correlations between MCC and game scores in the investigated models, where r is the Pearson correlation coefficient, and p is the significance of the coefficient.

	Shallow ConvNet	Hybrid	Parallel ConvNet	gMLP-MoCo
r	.36	.36	.31	.25
p	.0007	.0007	.003	.02

the properties of four neural networks using various architectures and training techniques concerning their classification results, features' importance, and correlations between the classification metrics and behavioral scores. Finally, to assure direct comparison of our results with the commonly-used reference, we implemented and trained a Shallow ConvNet model, originally designed by Schirrmeister *et al* (2017) and compared our results to those from published studies. We discuss the results in detail below.

4.1. Classification results

The best classification results in terms of ACC and MCC metrics were obtained for the gMLP-MoCo and the Hybrid models. Both models significantly outperformed the reference Shallow ConvNet. The two highest-performing networks represent different model designs and training methods, but both use some pre-training and fine-tuning.

The gMLP-MoCo network represents a transfer learning approach combining self-supervised contrastive learning on clinical EEG recordings and fine-tuning to the experimental data using standard back-propagation. This approach was motivated by the sparsity of EEG data for the current experimental task and the availability of a large dataset of the clinical EEG recordings. Our goal was to extract clinical data features that were invariant to nuisance factors such as noise, exact location, or DC shifts and use them to classify the targeted dataset. While transfer learning is a well-known approach

proposed as early as 1998 (Thrun and Pratt 1998) and widely used in EEG (for review, see Wan *et al* 2021) to overcome data scarcity (Hüebner *et al* 2018, Dutta and Nandy 2019), the self-supervised contrastive learning approach used here is relatively new. Contrastive learning was first proposed for stimulus modality classification from resting-state magnetoencephalography data (Hyvarinen and Morioka 2016). This approach was later successfully applied to other types of neuroimaging data such as MRI and fMRI (Chaitanya *et al* 2020, Li *et al* 2021) and EEG (Mohsenvand *et al* 2020, Banville *et al* 2021). Our results confirm that self-supervised contrastive learning may help extract features from unlabeled data, which can be successfully used for downstream tasks.

The other best-performing model, the Hybrid network, was tuned to individual participants. In this case, the learning relied on features formed by the convolution layers of the Parallel ConvNet, trained on task time-series data collected from all participants. Subsequently, a simple logistic classifier was fine-tuned using these features extracted from a given participant's data. This approach yielded comparable results with the much more complex gMLP-MoCo. However, a disadvantage of the Hybrid model is the need for multiple recording sessions for individual participants which are not always feasible.

Another contrasting property of the Hybrid and gMLP-MoCo models is the input data representation. The Hybrid model used an explicit time-frequency transformation of signal for all channels, while gMLP-MoCo operated on the raw channel-time data.

The results obtained suggest that neither input data representation nor model complexity was a key property of the best-performing models as measured by ACC and MCC. However, some fine-tuning appears beneficial in enabling the use of massive amounts of data or customization. Moreover, the feature importance index differed as expected between the best and worst classified participants (figure 8) only for the models trained on the task data.

4.2. Maps of feature importance index and their physiological significance

Power spectra in canonical EEG bands, their spatial distribution across electrodes, and their relationships to cognitive functions have been extensively studied since the early days of electroencephalography. Therefore, the spectral features used by the models for classification and their comparison with the classical spatial-frequency analysis and current knowledge can indicate whether the classification results reflect known physiological phenomena, which is of great importance in medicine and biology. Classic EEG analysis methods rely on statistical comparison of predefined measures at specific regions of interest between the experimental conditions. Instead, artificial neural networks perform their calculations using information about all available features and locations simultaneously, which brings additional insight into physiological mechanisms associated with features of importance for classification results. It is important to note that investigators predefine the features recovered during perturbation analyses, meaning that these features are not necessarily the ones that were most significant for classification.

Spearman's correlation between the patterns of features' importance revealed by different models showed a strong correlation of all models except gMLP-MoCo (figure 7(e)). All models showing high correlation between the patterns of feature importance were characterized by a shallow architecture, fewer trainable parameters, and a purely supervised learning strategy. The most prominent features common for these models included positive values at frontal electrodes in the theta band (5 and 8 Hz) and negative values in parietal electrodes centered around the alpha (11 Hz) and beta (15 Hz) bands. These features were also present in the gMLP-MoCo model. These features are well documented in numerous psychological and neurobiological studies on EEG and working memory. It has been shown that the retention of information in memory is associated with an increase in theta band power in the frontal electrodes (e.g. Wilson *et al* 1999, Bastiaansen *et al* 2002, Klimesch *et al* 2008, Michels *et al* 2010, Sauseng *et al* 2010). Although the interpretation of the other features, such as negative probability gradients for the alpha and beta bands found at parietal electrodes, are more challenging to interpret, they are also detected in numerous electrophysiological experiments. This shows that the proposed models are, in different ways, interpretable using the core electrophysiological literature. Our experiments show how the ML can be used as a tool in hypothesis-driven research (Yang and Wang 2020).

Probability gradients calculated for the gMLP-MoCo model with contrastive learning show different patterns of significant features. Even though the experimental data were preprocessed to exclude artifacts from the training dataset (including eye blinks and muscle activity removal), the set of important features developed by the model points to the significance of delta and theta bands on the Fp1 and Fp2 electrodes and, to a lesser extent, the importance of gamma-band in frontal and occipital sites. Notably, delta and theta activity at Fp1 and Fp2 electrodes are characteristic of eye blinks, while high gamma amplitude in the frontal and occipital regions might result from muscle activity. These findings are of particular importance for potential applications in EEG-neurofeedback training and diagnostic purposes. Namely, high sensitivity to the features associated with artifacts practically excludes the model from the application in medicine, although it may be of practical use in BCI.

The vital clue also comes from the sanity check comparing the feature importance indexes between groups of best and worst classified participants. We expect the robust features to be the same in both groups but more pronounced in the best-classified players. Indeed, comparison of the probability gradients between those two groups showed identical set of features but more pronounced in the best classified participants.

Summarizing the importance of the extracted features of the EEG signal for classification purposes, it should be noted that they correspond to physiological EEG properties observed during the biological and clinical experiments. Furthermore, it is essential to note that different training methods lead to classifications based on different sets of features that are not necessarily directly related to the task at hand. However, a subset of features is common to all models. Interestingly and importantly, this subset is highly relevant for the task solved by the participants.

4.3. Classification ACC and game performance

Finally, we investigated the relationship between classification ACC and game performance. This step was motivated by the notion that successful retention of information necessary for answering questions correctly should result in better performance in the game. This was further supported by the fact that extracted features, common for all models, overlapped with EEG activity associated with successful retention of information. Not surprisingly, the correlation was found only for MCC capable of handling unbalanced classes (Boughorbel *et al* 2017) (the number of retention trials was 25% smaller than control ones). The overall correlations for all participants and both game environments were weaker than expected, showing only a weak to moderate association between MCC and game performance.

Surprisingly, correlational analyses of the relationship between EEG classification ACC and behavioral performance in classified tasks are rare. Pang *et al* (2021), who classified mental workload on the NASA MATBII test using a stochastic configuration network, obtained a correlation between mean task ACC and classification ACC of $r = .852$ with $p < .01$ for 16 participants (15 males and one female). However, a more detailed analysis of these correlations shows two clusters, with no correlation within each cluster. The lower correlation scores obtained for our data may have several reasons: (i) the MATBII procedure used by Pang *et al* (2021) is a mixture of four tasks, (ii) the group was 94% male, and (iii) the number of participants was relatively low.

Summarizing, in the context of results obtained by Pang *et al* (2021), moderate correlations between MCC and game performance may indicate one of the following: (i) information stored in the memory is not necessarily related to the task at hand, (ii) EEG features being classified may not be entirely or exclusively related to information retention, and (iii) there might be other factors interfering with memory retention/retrieval or task execution.

5. Conclusion

The input perturbation analysis showed that different training methods lead to classifications based on various features. However, the common subset of features used by all investigated models is highly relevant for the task being solved by participants. This observation suggests that if a study aims to gain some insight into the properties of the EEG signal, one should construct and train various models and identify the stable subset of features. Furthermore, our results suggest that individual tuning based on the features developed by a convolutional model may be beneficial, as in the case of the Hybrid model. The personal tuning effectively increased the ACC and MCC average values.

Interestingly, comparable classification scores were obtained for the contrastive model pretrained on the clinical data from the participants not participating in the current experiment. Still, the correlations between score metrics and behavioral performance were low for this model. Therefore, the results of our experiment identify several essential considerations when employing ML for EEG signal classification:

(a) In applications related to diagnosis and therapy, it is vital to explain the classification in terms of features used to support it. The choice of architecture and training method affects features' importance and should be made carefully.
(b) The contrastive learning method, which has shown excellent results, appeared to be sensitive to deep-rooted features (by deep-rooted we mean features which cannot be completely removed from signal using methods other than removing blocks of recording with artifacts) resistant to data manipulations such as artifact removal by ICA. This may suggest that this transfer-learning method carries the risk that classification may be based on residual artifacts correlated with task performance and should be used with caution.
(c) Depending on the architecture and training method, some of the EEG features used by the models for classification may not be directly related to the task at hand, leading to a moderate relationship between ACC and behavioral performance.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The project was partially financed by the Regional Operational Program of the Masovian Voivodeship for 2014-2020, RPMA Agreement Number 01.02.00-14-b459/18 (Project: 'EEGDigiTrack Biofeedback AI—an innovative device for personalized neurotherapy with scientifically proven effectiveness').

ORCID iD

Jarosław Żygierewicz https://orcid.org/0000-0002-7536-0735

References

Ang K K, Chin Z Y, Zhang H and Guan C 2008 Filter bank common spatial pattern (FBCSP) in brain-computer interface *2008 IEEE Int. Joint Conf. on Neural Networks (IEEE World Congress on Computational Intelligence)* (IEEE) pp 2390–7

Banville H, Chehab O, Hyvärinen A, Engemann D-A and Gramfort A 2021 Uncovering the structure of clinical EEG signals with self-supervised learning *J. Neural Eng.* **18** 046020

Barredo Arrieta A *et al* 2020 Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI *Inf. Fusion* **58** 82–115

Bastiaansen M C, Posthuma D, Groot P F and De Geus E J 2002 Event-related alpha and theta responses in a visuo-spatial working memory task *Clin. Neurophysiol.* **113** 1882–93

Benjamini Y and Yekutieli D 2001 The control of the false discovery rate in multiple testing under dependency *Ann. Stat.* **29** 1165–88

Bird J J, Manso L J, Ribeiro E P, Ekárt A and Faria D R 2018 A study on mental state classification using EEG-based brain-machine interface *Int. Conf. on Intelligent Systems (IS)* (IEEE) pp 795–800

Boughorbel S, Jarray F and El-Anbari M 2017 Optimal classifier for imbalanced data using MCC metric *PLoS One* **12** e0177678

Chaitanya K, Erdil E, Karani N and Konukoglu E 2020 Contrastive learning of global and local features for medical image segmentation with limited annotations (arXiv:2006.10511)

Chakladar D D, Dey S, Roy P P and Dogra D P 2020 EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm *Biomed. Signal Process. Control* **60** 101989

Chen T, Kornblith S, Norouzi M and Hinton G 2020 A simple framework for contrastive learning of visual representations *Int. Conf. on Machine Learning* (PMLR) pp 1597–607

Chen X, Fan H, Girshick R and He K 2020 Improved baselines with momentum contrastive learning (arXiv:2003.04297)

Comstock J R *et al* 1992 The multi-attribute task battery for human operator workload and strategic behavior research *NASA Technical Memorandum* NASA-TM-104174

Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Methods* **134** 9–21

Dutta S and Nandy A 2019 Data augmentation for ambulatory EEG based cognitive state taxonomy system with RNN-LSTM *Int. Conf. on Innovative Techniques and Applications of Artificial Intelligence* (Springer) pp 468–73

Hammond D C 2011 What is neurofeedback: an update *J. Neurotherapy* **15** 305–36

Han S-Y, Kwak N-S, Oh T and Lee S-W 2020 Classification of pilots' mental states using a multimodal deep learning network *Biocybern. Biomed. Eng.* **40** 324–36

He T, Kong R, Holmes A J, Nguyen M, Sabuncu M R, Eickhoff S B, Bzdok D, Feng J and Yeo B T 2020 Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics *NeuroImage* **206** 116276

Hüebner D, Verhoeven T, Müeller K-R, Kindermans P-J and Tangermann M 2018 Unsupervised learning for brain-computer interfaces based on event-related potentials: review and online comparison *IEEE Comput. Intell. Mag.* **13** 66–77

Hyvarinen A and Morioka H 2016 Unsupervised feature extraction by time-contrastive learning and nonlinear ICA *Advances in Neural Information Processing Systems* vol 29 pp 3765–73

Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift *Proc. 32nd Int. Conf. on Machine Learning (Proc. Machine Learning Research) (Lille, France)* vol 37, ed F Bach and D Blei (PMLR) pp 448–56

Klimesch W, Freunberger R, Sauseng P and Gruber W 2008 A short review of slow phase synchronization and memory: evidence for control processes in different memory systems? *Brain Res.* **1235** 31–44

Li J, Zhang C, Wang L, Ding P, Hu L, Yan B and Tong L 2021 A visual encoding model based on contrastive self-supervised learning for human brain activity along the ventral visual stream *Brain Sci.* **11** 1004

Liu H, Dai Z, So D R and Le Q V 2021 Pay attention to MLPs *NeurIPS'2021*

Lopez-Calderon J and Luck S J 2014 ERPLAB: an open-source toolbox for the analysis of event-related potentials *Front. Hum. Neurosci.* **8** 213

Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A and Yger F 2018 A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update *J. Neural Eng.* **15** 031005

Michels L, Bucher K, Lüchinger R, Klaver P, Martin E, Jeanmonod D and Brandeis D 2010 Simultaneous EEG-fMRI during a working memory task: modulations in low and high frequency bands *PLoS One* **5** e10298

Mohsenvand M N, Izadi M R and Maes P 2020 Contrastive representation learning for electroencephalogram classification *Machine Learning for Health* (PMLR) pp 238–53

Nathan K and Contreras-Vidal J L 2016 Negligible motion artifacts in scalp electroencephalography (EEG) during treadmill walking *Front. Hum. Neurosci.* **9** 708

Pang L, Guo L, Zhang J, Wanyan X, Qu H and Wang X 2021 Subject-specific mental workload classification using EEG and stochastic configuration network (SCN) *Biomed. Signal Process. Control* **68** 102711

Sauseng P, Griesmayr B, Freunberger R and Klimesch W 2010 Control mechanisms in working memory: a possible function of EEG theta oscillations *Neurosci. Biobehav. Rev.* **34** 1015–22

Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420

Schweizer S, Samimi Z, Hasani J, Moradi A, Mirdoraghi F and Khaleghi M 2017 Improving cognitive control in adolescents with post-traumatic stress disorder (PTSD) *Behav. Res. Therapy* **93** 88–94

Subha D P, Joseph P K, Acharya R and Lim C M 2010 EEG signal analysis: a survey *J. Med. Syst.* **34** 195–212

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 Going deeper with convolutions *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 1–9

Thrun S and Pratt L 1998 Learning to learn: introduction and overview *Learning to Learn* (Amsterdam: Kluwer Academic Publishers) pp 3–17

Tian Y, Krishnan D and Isola P 2020 Contrastive multiview coding *European Conf. on Computer Vision* (Cham: Springer) pp 776–94

Wan Z, Yang R, Huang M, Zeng N and Liu X 2021 A review on transfer learning in EEG signal analysis *Neurocomputing* **421** 1–14

Wilson G F, Swain C R and Ullsperger P 1999 EEG power changes during a multiple level memory retention task *Int. J. Psychophysiol.* **32** 107–18

Wu Z, Xiong Y, Yu S X and Lin D 2018 Unsupervised feature learning via non-parametric instance discrimination *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 3733–42

Yang G R and Wang X-J 2020 Artificial neural networks for neuroscientists: a primer *Neuron* **107** 1048–70