fMRI Neurofeedback Learning Patterns are Predictive of Personal and Clinical Traits

Rotem Leibovitz ROTEMCZ1@GMAIL.COM
Jhonathan Osin YONI.OSIN@GMAIL.COM
Lior Wolf WOLF@CS.TAU.AC.IL
Guy Gurevitch GUYGU4@GMAIL.COM
Talma Hendler HENDLERT@GMAIL.COM
Tel Aviv University

Editors: Under Review for MIDL 2022

Abstract

We obtain a personal signature of a person's learning progress in a self-neuromodulation task, guided by functional MRI (fMRI). The signature is based on predicting the activity of the Amygdala in a second neurofeedback session, given a similar fMRI-derived brain state in the first session. The prediction is made by a deep neural network, which is trained on the entire training cohort of patients. This signal, which is indicative of a person's progress in performing the task of Amygdala modulation, is aggregated across multiple prototypical brain states and then classified by a linear classifier to various personal and clinical indications. The predictive power of the obtained signature is stronger than previous approaches for obtaining a personal signature from fMRI neurofeedback and provides an indication that a person's learning pattern may be used as a diagnostic tool. Our code has been made available, [1] and data would be shared, subject to ethical approvals.

1. Introduction

An individual's ability to learn to perform a specific task in a specific context is influenced by their transient task demand and context-specific mental capacity (Crump et al., 2008), as well as their motivation (Utman, 1997), all of which vary considerably between individuals (Ackerman, 1988). We hypothesize that the learning pattern is highly indicative of both personal information, such as age and previous experience in performing similar tasks, and personality and clinical traits, such as emotion expressivity, anxiety levels, and even specific psychiatric indications.

Neurofeedback (NF), a closed-loop self-neuromodulation learning procedure, provides a convenient environment for testing our hypothesis. This is because the learning task is well-defined, yet individualized, is presented in a controlled and repeatable manner, and the level of success is measured on a continuous scale of designated neural changes. NF is a reinforcement learning procedure guided by feedback presented depending on self-acquired association between mental and neural states. Mental states that happened to be associated with on-line modulation in the neural target (e.g. lower or higher activity) are rewarded, and eventually result in a desired modification of the brain signal (i.e. learning success) (Sitaram et al., 2017; Taschereau-Dumouchel et al., 2022). We hypothesize that

1. Our code is available via https://github.com/rotemcz/fmri_nf

the established association between internally-generated mental process and neural signal modulation closely signifies personal brain-mind relation and could therefore serve as an informative marker for personality and/or psychopathology.

fMRI-based neurofeedback enables precise modulation of specific brain regions in real time, leading to sustained neural and behavioral changes. However, the utilization of this method in clinical practice is limited due to its high cost and limited availability, which also hinder further research into the sustained benefit (Sulzer et al., 2013; Lubianiker et al., 2019). In the NF task we consider, one learns to reduce the activity of the Amygdala, while observing a signal that is directly correlated with it. We consider the activity of the rest of the brain as the context or states in which the learning task takes place, and discretize this space by performing clustering.

A personal signature is constructed by measuring the progress of performing the NF task in each of these clusters. Progress is obtained by comparing the Amygdala activity at a first training session with that observed at the second session for the most similar brain state. Specifically, we consider the difference between the second activity and the one predicted by a neural network that is conditioned on the brain state in the first.

The individual representation obtained by aggregating these differences across the brain-state clusters is shown to be highly predictive of multiple psychiatric traits and conditions in three datasets: (i) individuals suffering from PTSD, (ii) individuals diagnosed with Fibromyalgia, and (iii) a control dataset of healthy individuals. This predictive power is demonstrated with linear classifiers, in order to demonstrate that the personal information is encoded in an explicit way and to reduce the risk of overfitting by repeating the test with multiple hyperparameters (Asano et al., 2019).

2. Related work

fMRI is widely used in the study of psychiatric disorders (Calhoun et al., 2014; Oksuz et al., 2019). Recent applications of deep learning methods mostly focus on fully-supervised binary classification of psychopathology-diagnosed versus healthy subjects in resting state (Dvornek et al., 2017; Yan et al., 2019), i.e., when not performing a task. Contributions that perform such diagnosis while performing a task, e.g., (Bleich-Cohen et al., 2014; Jacob et al., 2019; Hendler et al., 2018; Raz et al., 2016; Lerner et al., 2018), focus on comparing entire segments that correspond to phases of the task, and have shown improved ability to predict subjects' traits, w.r.t to resting-state fMRI (Gal et al., 2022). Our analysis is based on aggregating statistics across individual time points along the acquired fMRI.

The work closest related to ours applies self-supervised learning to the same fMRI NF data in order to diagnose participants suffering from various psychopathologies and healthy controls (Osin et al., 2020). There are major differences in the approaches. First, while our method is based on a **meaningful** signature (it accumulates meaningful statistics) that indicates learning patterns, their work is based on an implicit embedding obtained by training a deep neural network. Second, while we focus on modeling the success in performing the task over a training period, their method is based on the self-supervised task of next frame prediction, which involves both the preparation ("passive") and training ("active") periods (they require more data). Third, while our method compares progress between two active NF sessions, their method is based on mapping a passive session, in

which the participant does not try to self-modulate, and the subsequent active NF session. The methods are, therefore, completely different. Finally, in a direct empirical evaluation, our method is shown to outperform (Osin et al., 2020) by a sizable gap across all datasets and prediction tasks, as further elaborated in Sec. 5.

As mentioned, the success in learning is related to the internal motivation to learn. This is dependent on the brain state, and our method learns a signature that is directly linked to the internal value function. In this sense, our method solves an inverse reinforcement learning (learn a value function from an agent behavior) problem (Arora and Doshi, 2018).

3. Data

Real-time blood-oxygen-level-dependent (BOLD) signal was measured from the right Amygdala region during an interactive neurofeedback session performed inside the fMRI. The data fed into the model went through preprocessing steps using the CONN toolbox. The full preprocessing steps are detailed in App. C. In specific parts of the task, subjects were instructed to control the speed of an avatar riding a skateboard using only mental strategies (active phase), while in other parts, subjects passively watched the avatar on the screen (passive phase). During the active phase, local changes in the signal were translated into changing speed, displayed to the subjects via a speedometer and updated every three seconds.

The neurofeedback datasets used in this experiment were part of larger intervention experiments applying multiple training sessions outside the fMRI by using an EEG statistical model of the right Amygdala. The subjects went through pre/post fMRI scans with the model region as target in order to test for changes in the ability to self-regulate this area (Fruchtman-Steinbok et al., 2021; Keynan et al., 2019).

fMRI data Each subject performed several cycles of the paradigm in a single session, lasting up to one hour, in a similar fashion to common studies in the field (Paret et al., 2019). Following each active phase, a bar indicating the average speed during the current cycle was presented for six seconds. Each subject performed $M = 2$ cycles of Passive/Active phases, where each passive phase lasted one minute. Each active phase lasted one minute (for healthy controls and PTSD patients) or two minutes (Fibromyalgia patients). Following previous findings, instructions given to the subjects were not specific to the Amygdala, to allow efficient adoption of individual strategies (Marxen et al., 2016).

The active sessions were comprised of T temporal samples of the BOLD signal, each a 3D volumed box with dimensions $\mathcal{H}$[voxels] $\times$ $\mathcal{W}$[voxels] $\times$ $\mathcal{D}$[voxels]. We distinguish BOLD signals of the Amygdala signals from signals of other brain regions, each with a different spatial resolution, $\mathcal{H}_\mathcal{A} \times \mathcal{W}_\mathcal{A} \times \mathcal{D}_\mathcal{A}$, and $\mathcal{H}_\mathcal{R} \times \mathcal{W}_\mathcal{R} \times \mathcal{D}_\mathcal{R}$, respectively. Partitioning was done using a pre-calculated binary Region-of-Interest matrix.

ROI for Rest-of-Brain covers the entire gray matter of the right hemisphere, excluding the right amygdala. This mask was generated with an SPM based segmentation of the MNI brain template. This region was used for providing the feedback during the real-time fMRI experiments. See App. C for more details about the construction of the Rest-of-Brain ROI matrix. A visualisation of the selected brain regions, is shown in Fig. 5. The matrix contains a positive value for voxels that are part of the Amygdala, and a negative value for all others. Our data is, therefore, comprised of per-subject tuples of tensors: $(\mathbb{R}^{T\times\mathcal{H}_\mathcal{A}\times\mathcal{W}_\mathcal{A}\times\mathcal{D}_\mathcal{A}}, \mathbb{R}^{T\times\mathcal{H}_\mathcal{R}\times\mathcal{W}_\mathcal{R}\times\mathcal{D}_\mathcal{R}})$.

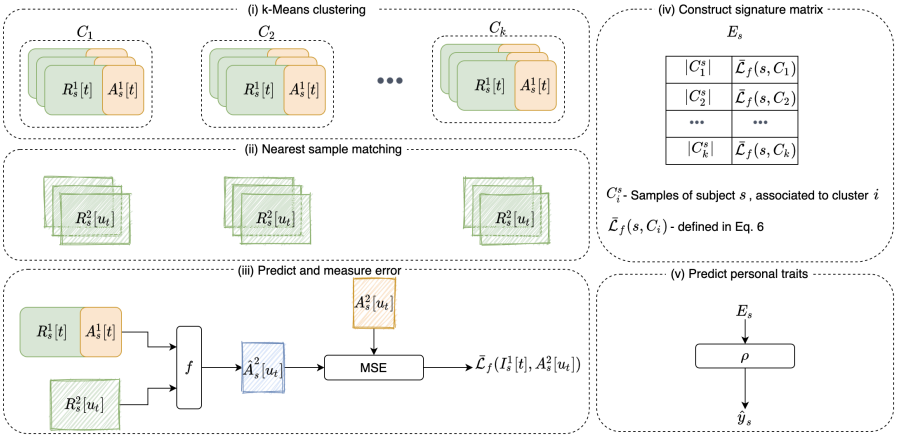


Figure 1: The training steps: (i) Clustering the frames from the first session, based on the regions outside the Amygdala; (ii) Matching each frame from the previous step to the most similar frame from the second session; (iii) Learning to predict the Amygdala activity in the matched frame of the second session, based on the information from the matching frame in the first session and the brain activity outside the Amygdala in the second session; (iv) constructing a signature based on the prevalence of each cluster and the error of prediction for the frames of each cluster; and (v) a linear classifier on top of the obtained signature.

In our setting, $M = 2, T = 18$. We use three datasets in our experiments: (i) **PTSD**- 51 subjects, (ii) **Fibromyalgia**- 24 subjects, and (iii) **Healthy Control**- 87 subjects. The Amygdala parameters and rest-of-brain parameters are identical for all three datasets: $\mathcal{H}_\mathcal{A} = 6$, $\mathcal{W}_\mathcal{A} = 5$, $\mathcal{D}_\mathcal{A} = 6$ and $\mathcal{H}_\mathcal{R} = 91$, $\mathcal{W}_\mathcal{R} = 109$, $\mathcal{D}_\mathcal{R} = 91$, respectively. Further information on the data acquisition scheme is provided in Appendix C.

Clinical data Clinical information about each subject s, denoted as $y_s \in \mathbb{R}^l$ was available in addition to the fMRI sequences, consisting of the following information: (1) **Toronto Alexithymia Scale (TAS-20)**, which is a self-report questionnaire measuring difficulties in expressing and identifying emotions (Bagby et al., 1994), (2) **State-Trait Anxiety Inventory (STAI)**, which is measured using a validated 20-item inventory (Spielberger and Gorsuch, 1983), and (3) **Clinician Administered PTSD Scale (CAPS-5) 1**, which is the outcome of a clinical assessment by a trained psychologist based on this widely-used scale for PTSD diagnosis (Weathers et al., 2013). For the healthy controls, the following demographic information was also available: (1) **Age** and (2) **Past experience in neurofeedback tasks**, presented as a binary label (i.e, experienced / inexperienced subject).

4. Method

Our network receives processed fMRI samples as inputs, and uses them to predict subjects' demographic and psychiatric criteria. We consider two types of fMRI signals: Amygdala, and rest-of-brain, and train our networks in five steps, as depicted in Fig. 1: **(i)** identification of prototypical brain states, C, by applying a k-means clustering scheme to the fMRI signals; **(ii)** consider the rest of the brain regions, and identify for each fMRI frame from the first session the most similar frame from the second session; **(iii)** given a subject's complete

brain state (Amygdala and greater-brain) at each time-step from the first session, we train a neural network, f, to predict the subject's Amygdala state in the matched closest frame from the second session;**(iv)** create a subject signature by aggregating the prediction error of f in each of the k prototypes; **(v)** train a linear regression network, ρ, to predict subjects' criteria, based on the obtained signature.

Data structure For every subject $s \in S$, the dataset contains a series of $M = 2$ active sessions, each with $T = 18$ samples, denoted as $I_s = \left\{I_s^m[t]\right\}_{t=1,m=1}^{T,M}$. We treat each sample as a pair: (i) an Amygdala sample, $A_s^m[t]$, which is cropped out of $I_s^m[t]$ using a fixed ROI matrix, as explained in Sec. 3, and (ii) a sample of other brain parts, $R_s^m[t]$. For computational reasons, we use box-shaped data and constructed $R_s^m[t]$ by extracting the maximal box from $I_s^m[t]$, such that all of its voxels are not part of the Amygdala.

K-means (step i) We learn a set of k cluster centroids, $C = \{\mu_1, \ldots, \mu_k\}$, based on rest-of-brain training samples from the first session, minimizing the within-cluster-sum-of-squares: $\underset{C}{\operatorname{argmin}} \sum_{s,t} \min_{\mu \in C} ||R_s^1[t] - \mu||^2$. Note that clustering as well as cluster assignment are carried out independently of the Amygdala samples, and only once for the entire training set.

Each cluster represents a different prototypical brain state and the number of clusters is selected such that for most training subjects, no cluster is underutilized. See Sec. 5.

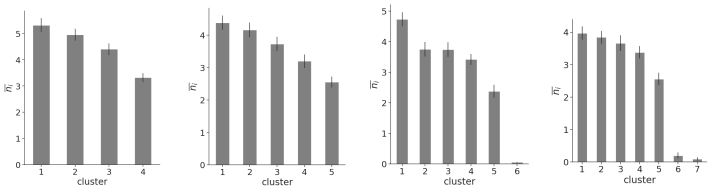
Amygdala state prediction (steps ii and iii) For every subject s and every time step $t \leq T$ in the first session, we identify u_t, the time step during the second session in which the most similar sample was taken: $u_t = \underset{t' \leq T}{\operatorname{argmin}} ||R_s^1[t] - R_s^2[t']||^2$.

The obtained pairs $\{(t, u_t)\}$ indicate tuples of similar samples $\{(I_s^1[t], I_s^2[u_t])\}$, which we use to train a neural-network, f, aimed to predict $A_s^2[u_t]$ (see implementation details in App.A): $\hat{A}_s^2[u_t] = f\big(I_s^1[t], R_s^2[u_t]\big) = f\big(R_s^1[t], A_s^1[t], R_s^2[u_t]\big)$. f is trained independently of the centroids, minimizing the MSE loss $\sum_{s \in S} \sum_{t \leq T} ||f\big(I_s^1[t], R_s^2[u_t]\big) - A_s^2[u_t]||^2$.

Building a signature matrix (step iv) and predicting personal traits (step v) With the group of centroids, C, and the Amygdala-state predictor, f, we construct a signature matrix, $E_s \in \mathbb{R}^{|C| \times 2}$, for every subject s. Each row of this signature corresponds to a specific brain state prototype, and the columns correspond to the number of samples in a cluster and the mean prediction error for that cluster. For every subject $s \in S$ and cluster $C_i \in C$, we define C_i^s as the set of the subject's samples associated with the cluster: $C_i^s = \{t | i = \underset{j}{\operatorname{argmin}} \|R_s^1[t] - \mu_j\|\}$.

We then calculate the average prediction loss of f with respect to each (s, C_i) pair: $\bar{\mathcal{L}}_f(s, C_i) = \frac{1}{|C_i^s|} \cdot \sum_{t \in C_i^s} ||f(R_s^1[t], R_s^2[u_t], A_s^1[t]) - A_s^2[u_t]||^2$, where $|C_i^s|$ indicates the number of visits of subject s in cluster i. The signature matrix E_s has rows of the form $E_s[i] = \left[|C_i^s|, \bar{\mathcal{L}}_f(s, C_i)\right]$. E_s is then fed into ρ, a linear regression network with an objective to predict y_s. Since E_s is a matrix, we use a flattened version of it, denoted as $e_s \in \mathbb{R}^{2 \cdot |C|}$. We then predict $\hat{y}_s = \rho(e_s) = G^\top e_s + b$, where G is a matrix and b is a vector. G, b are learned using the least squared loss over the training set.

Using a low-capacity linear classifier is meant to reduce the effect of overfitting in this step, which is the only one with access to the target prediction labels. The neural network f is trained on a considerably larger dataset, with samples of a much higher dimension (as



(a) $k = 4$ (b) $k = 5$ (c) $k = 6$ (d) $k = 7$

Figure 2: Sorted $\bar{n}_i$ per cluster on healthy controls for $k = 4, 5, 6, 7$. The error bars depict the standard deviation between the training subjects of a typical cross validation split.

every fMRI frame from every subject is a sample). Additionally, its task is a self-supervised one, and is therefore less prone to overfitting to the target labels.

The inference pipeline Given an unseen subject r, we assign each fMRI frame, t, in the first session, to a cluster by finding the prototypes in C closest to $R_r^1[t]$. We also match a frame u_t from the second session to each frame t by minimizing $||R_r^1[t] - R_r^2[u_t]||^2$. A signature e_r is then constructed. Finally, the linear predictor ρ is applied to this signature. Implementation details are provided in App. A.

5. Experiments

Each experiment was repeated five times on random splits. We report the mean and SD. The data is partitioned using a cross-validation scheme between train, validation and test sets, each composed of *different subjects*, with a 60-20-20 split. The same partition holds for all training stages in each dataset. For each dataset (i.e, Healthy, PTSD, Fibromyalgia), we trained a separate network.

Building the representation In order to assure that all k chosen brain-state-prototypes are visited by all subjects, we evaluate the ratio $\bar{n}_i = \frac{\sum_{s \in S} |C_i^s|}{T \cdot |S|}$ of brain states assigned to each of the clusters, when applying k-means on the training data.

We chose the largest k for which the variance between all $\{\bar{n}_i\}_{i=1}^k$ is relatively small. As shown in Fig. 2, for $k \in \{4, 5\}$, all clusters are relatively evenly populated. However, for $k \in \{6, 7\}$, one or two clusters are seldom visited ("cluster starvation"). The results in the figure depict one k-means partitioning performed on healthy control subjects, but are typical of all three datasets, and for many random seeds. We, therefore, choose $k = 5$. In order to show that the clusters we got hold meaningful information, we performed two experiments:(a) train an LSTM based network, which receives as input the temporal signal of transitions between clusters, which proved to have predictive power w.r.t subjects' traits (results are shown in Tab. 1); and (b) train a similar LSTM network to predict the next cluster brain state, given past visited clusters. The network accurately predicted the cluster for 41% of the frames, compared to guessing the mean cluster which yields accuracy of 21%.

Amygdala state prediction We trained the network f until its validation loss converged for each of the three datasets - healthy, PTSD and Fibromyalgia. The MSE error obtained is presented in App.A. The network is quite successful in performing its prediction task, compared to the simple baseline of predicting the network's input, $A_s^1[t]$.

Predicting a subject's psychiatric and demographic criteria We test whether our learned representation, trained only with fMRI images, has the ability to predict a series

of psychiatric and demographic criteria not directly related to the neurofeedback task. We used our method to predict (i) STAI and (ii) TAS-20 for PTSD, Fibromyalgia and control subjects, (iii) CAPS-5 for PTSD subjects. Demographic information, (iv) age, and (v) past neuroFeedback experience were predicted for the control subjects.

Our linear regression scheme, applied to the learned signature vectors, is compared to the following baselines, which all receive the fMRI sequence as input, denoted as x: (1) **Mean prediction** simply predicts a constant value, (2) **Conditional LSTM-** The Amygdala sections of the passive and active temporal signals are fed, as is, to a neural network, which learns a personal representation for each subject (intuitively; the relation between respective active and passive samples represents a subject's ability to regulate their Amygdala). This learned representation is later used to predict the subject's criteria (Osin et al., 2020). In contrast to our method, the learned representation of the conditional-LSTM also employs the passive "watch" data, which our method ignores.

Another baseline predicts the target label without building a signature: (3) **CNN-** A convolutional network with architecture identical to our f network, except for two modifications: (a) the input signal to this network is the entire second sample, $I_s^2[u_t]$ (instead of $R_s^2[u_t]$). This way, the network has access to the same signals our proposed method has; and (b) the decoder is replaced with a fully connected layer, which predicts the label.

We also perform comparison with an alternative framework that ignores the fMRI signals and considers only the clinical data: (4) **clinical prediction-** an SVM regression with the RBF kernel performed on every trait, according to the other traits (leave-one-trait-out, where the data contains all psychiatric traits and the two demographic traits).

To perform an ablation study, we also compared the performance of regression networks trained only with a mixture of partial data from E_s and fMRI signals: (5) **Raw difference-** A similar signature matrix, with the $\bar{\mathcal{L}}_f$ value replaced by the average norm of differences between amygdala signals for every pair $\{(t, u_t)\}$. Rows of the resulting matrix are: $\tilde{E}_s[i] = \left[|C_i^s|, \frac{1}{|C_i^s|} \cdot \sum_{t \in C_i^s} ||A_s^1[t] - A_s^2[u_t]||^2\right]$; and (6, 7) **partial** E_s**-** A network trained using only one of the signature matrix columns (i.e, either $|C_i^s|$ or $\bar{\mathcal{L}}_f$).

(8) **ClusterLSTM-** an LSTM based network, which receives as input the sequence of cluster memberships per frame, and predicts the subject's traits according to it.

To show that the neurofeedback learning that occurs across sessions is what is important, rather than the expected state of amygdala based on "Rest-of-Brain" state, we implemented (9) **No Feedback-** which predicts the Amygdala state given "Rest-of-Brain" state for the sample, without pairing samples from different sessions. The clustering step is performed without considering the Amygdala, in order to reserve this region for the prediction task. A dedicated ablation compares this to the alternative of using the entire fMRI frame: (10) **Alternative clustering-** In step (i), the clustering objective is changed, such that it depends on the subjects' complete brain state: $\underset{C}{\operatorname{argmin}} \sum_{s,t} \min_{\mu \in C} ||I_s^1[t] - \mu||^2$. Lastly, to demonstrate the importance of using the Amygdala itself, we run baseline (11) **Alternative ROI-** a framework identical to ours, but with neural area of focus shifted from the Amygdala to the primary motor cortex, an area of dimensions $\mathcal{H} = \mathcal{W} = \mathcal{D} = 8$, which is presumed not to take part in the performance of the NF task.

The full results are shown in Tab. 1 for performing regression on age, TAS, STAI and CAPS-5. Despite the dataset size, our results are statistically significant (see App. B for

	Healthy			Fibromyalgia		PTSD		
	Age↓	TAS↓	STAI↓	TAS↓	STAI↓	TAS↓	STAI↓	CAPS-5↓
Mean	13.7	121.7	79.3	98.6	78.5	153.0	159.5	119.3
C.LSTM	10.1	81.6	67.4	44.0	73.1	99.2	132.7	85.4
CNN	17.2	110.3	81.6	65.2	90.1	105.0	166.3	98.4
SVM	13.0	100.0	78.2	74.3	77.9	148.3	150.0	117.4
Ours, old-ROI	**9.2**	**74.8**	61.2	**31.2**	72.0	**88.0**	124.5	**79.3**
Ours, new-ROI	9.3	79.0	**55.2**	35.3	**62.3**	89.1	**97.0**	84.3
Ablation $\tilde{E}_s$	13.2	98.4	73.2	77.8	77.4	102.2	194.8	136.0
Ablation $\mathcal{L}_f$	10.3	91.6	70.0	84.4	81.9	119.6	141.0	119.2
Ablation $\|C_i\|$	11.0	84.2	68.1	113.6	109.4	137.5	137.8	151.5
ClusterLSTM	12.1	96.8	68.8	75.7	90.0	118.1	155.0	108.0
No Feedback	13.7	125.0	75.0	90.0	77.7	165.3	167.6	139.0
Alt. clusters	12.3	114.0	72.5	70.2	80.3	98.3	174.0	103.6
Alt. ROI	13.5	125.5	79.0	96.3	84.5	135.6	135.8	128.7

Table 1: Traits prediction comparison (MSE). Old-ROI is presented only for comparison with the previous version.

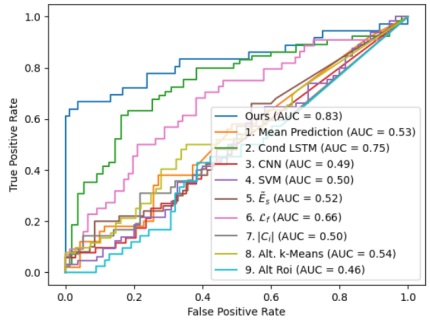


Figure 3: Past Experience ROC

more data), and the p value of the corrected re-sampled t-test between our method and the baseline methods is always lower than 0.01. As can be seen, the baseline of (Osin et al., 2020) greatly outperforms the mean prediction and both the CNN classifier and the one based on the clinical data. However, it is evident that across all three datasets our method outperforms this method, as well as all ablations, by a very significant margin in predicting the correct values for both demographic and psychiatric traits. The analysis of the relative error is given in App.A.

Fig. 3 presents classification results for past-experience information, which is only available for the healthy control subjects. Here, too, our method outperforms the baselines and ablation methods. Specifically, it obtains an AUC of 0.83 ± 0.03, while the method presented by (Osin et al., 2020) obtains an AUC of 0.75 ± 0.03.

The ablation experiments provide insights regarding the importance of the various components. First, modeling based on an irrelevant brain region, instead of the Amygdala, leads to results that are sometimes worse than a mean prediction. Similarly, predicting using raw differences in the Amygdala activity (without performing prediction), is not effective. It is also important to remove the Amygdala from the clustering procedure, keeping this region and outside regions separate. The variant based on the prediction error alone seems to be more informative than that based only on cluster frequency. However, only together do they outperform the strong baseline of (Osin et al., 2020).

6. Conclusion

NF data offers unique access to individual learning patterns. By aggregating the deviation between actual and predicted learning success across clusters of brain activities, we obtain a signature that is highly predictive of the history of a person, as well as of their clinical test scores and psychiatric diagnosis. The presented method provides a sizable improvement in performance over previous work.

Perhaps even more importantly, the obtained signature is based on explicit measurements that link brain states to the difference between actual and expected learning success, while previous work was based on an implicit embedding that is a by-product of training a network to predict a loosely related task of predicting transient signal dynamics.

References

Phillip L Ackerman. Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of experimental psychology: General*, 117(3):288, 1988.

Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *arXiv preprint arXiv:1806.06877*, 2018.

YM Asano, C Rupprecht, and A Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image. In *International Conference on Learning Representations*, 2019.

R Michael Bagby, James DA Parker, and Graeme J Taylor. The twenty-item toronto alexithymia scale—i. item selection and cross-validation of the factor structure. *Journal of psychosomatic research*, 38(1):23–32, 1994.

Maya Bleich-Cohen, Shahar Jamshy, Haggai Sharon, Ronit Weizman, Nathan Intrator, Michael Poyurovsky, and Talma Hendler. Machine learning fmri classifier delineates subgroups of schizophrenia patients. *Schizophrenia research*, 160(1-3):196–200, 2014.

Vince D Calhoun, Robyn Miller, Godfrey Pearlson, and Tulay Adalı. The chronnectome: time-varying connectivity networks as the next frontier in fmri data discovery. *Neuron*, 84(2):262–274, 2014.

Matthew JC Crump, Joaquín MM Vaquero, and Bruce Milliken. Context-specific learning and control: The roles of awareness, task relevance, and relative salience. *Consciousness and cognition*, 17(1):22–36, 2008.

Nicha C Dvornek, Pamela Ventola, Kevin A Pelphrey, and James S Duncan. Identifying autism from resting-state fmri using long short-term memory networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 362–370. Springer, 2017.

Tom Fruchtman-Steinbok, Jackob N Keynan, Avihay Cohen, Iman Jaljuli, Shiri Mermelstein, Gadi Drori, Efrat Routledge, Michael Krasnoshtein, Rebecca Playle, David EJ Linden, et al. Amygdala electrical-finger-print (amygefp) neurofeedback guided by individually-tailored trauma script for post-traumatic stress disorder: Proof-of-concept. *NeuroImage: Clinical*, 32:102859, 2021.

Shachar Gal, Niv Tik, Michal Bernstein-Eliav, and Ido Tavor. Predicting individual traits from unperformed tasks. *NeuroImage*, page 118920, 2022.

Talma Hendler, Gal Raz, Solnik Shimrit, Yael Jacob, Tamar Lin, Leor Roseman, Wahid Madah Wahid, Ilana Kremer, Marina Kupchik, Moshe Kotler, et al. Social affective context reveals altered network dynamics in schizophrenia patients. *Translational psychiatry*, 8(1):1–12, 2018.

Y Jacob, O Shany, PR Goldin, JJ Gross, and T Hendler. Reappraisal of interpersonal criticism in social anxiety disorder: A brain network hierarchy perspective. *Cerebral Cortex*, 29(7):3154–3167, 2019.

Jackob N Keynan, Avihay Cohen, Gilan Jackont, Nili Green, Noam Goldway, Alexander Davidov, Yehudit Meir-Hasson, Gal Raz, Nathan Intrator, Eyal Fruchter, et al. Electrical fingerprint of the amygdala guides neurofeedback training for stress resilience. *Nature human behaviour*, 3(1):63–73, 2019.

Yulia Lerner, Maya Bleich-Cohen, Shimrit Solnik-Knirsh, Galit Yogev-Seligmann, Tamir Eisenstein, Waheed Madah, Alon Shamir, Talma Hendler, and Ilana Kremer. Abnormal neural hierarchy in processing of verbal information in patients with schizophrenia. *NeuroImage: Clinical*, 17:1047–1060, 2018.

Nitzan Lubianiker, Noam Goldway, Tom Fruchtman-Steinbok, Christian Paret, Jacob N Keynan, Neomi Singer, Avihay Cohen, Kathrin Cohen Kadosh, David EJ Linden, and Talma Hendler. Process-based framework for precise neuromodulation. *Nature human behaviour*, 3(5):436–445, 2019.

Michael Marxen, Mark J Jacob, Dirk K Müller, Stefan Posse, Elena Ackley, Lydia Hellrung, Philipp Riedel, Stephan Bender, Robert Epple, and Michael N Smolka. Amygdala regulation following fmri-neurofeedback without instructed strategies. *Frontiers in human neuroscience*, 10:183, 2016.

Ilkay Oksuz, Gastao Cruz, James Clough, Aurelien Bustin, Nicolo Fuin, Rene M Botnar, Claudia Prieto, Andrew P King, and Julia A Schnabel. Magnetic resonance fingerprinting using recurrent neural networks. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1537–1540. IEEE, 2019.

Jhonathan Osin, Lior Wolf, Guy Gurevitch, Jackob Nimrod Keynan, Tom Fruchtman-Steinbok, Ayelet Or-Borichev, and Talma Hendler. Learning personal representations from fmri by predicting neurofeedback performance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–478. Springer, 2020.

Christian Paret, Noam Goldway, Catharina Zich, Jackob Nimrod Keynan, Talma Hendler, David Linden, and Kathrin Cohen Kadosh. Current progress in real-time functional magnetic resonance-based neurofeedback: Methodological challenges and achievements. *NeuroImage*, 202:116107, 2019.

Gal Raz, Lavi Shpigelman, Yael Jacob, Tal Gonen, Yoav Benjamini, and Talma Hendler. Psychophysiological whole-brain network clustering based on connectivity dynamics analysis in naturalistic conditions. *Human brain mapping*, 37(12):4654–4672, 2016.

Ranganatha Sitaram, Tomas Ros, Luke Stoeckel, Sven Haller, Frank Scharnowski, Jarrod Lewis-Peacock, Nikolaus Weiskopf, Maria Laura Blefari, Mohit Rana, Ethan Oblak, et al. Closed-loop brain training: the science of neurofeedback. *Nature Reviews Neuroscience*, 18(2):86–100, 2017.

Charles Donald Spielberger and Richard L Gorsuch. *State-trait anxiety inventory for adults: Manual and sample: Manual, instrument and scoring guide.* Consulting Psychologists Press, 1983.

James Sulzer, Sven Haller, Frank Scharnowski, Nikolaus Weiskopf, Niels Birbaumer, Maria Laura Blefari, Annette B Bruehl, Leonardo G Cohen, R Christopher DeCharms, Roger Gassert, et al. Real-time fmri neurofeedback: progress and challenges. *Neuroimage*, 76:386–399, 2013.

Vincent Taschereau-Dumouchel, Cody Cushing, and Hakwan Lau. Real-time functional mri in the treatment of mental health disorders. *Annual Review of Clinical Psychology*, 18, 2022.

Christopher H Utman. Performance effects of motivational state: A meta-analysis. *Personality and Social Psychology Review*, 1(2):170–182, 1997.

FW Weathers, DD Blake, PP Schnurr, DG Kaloupek, BP Marx, and TM Keane. The clinician-administered ptsd scale for dsm-5 (caps-5). interview available from the national center for ptsd, 2013.

Susan Whitfield-Gabrieli and Alfonso Nieto-Castanon. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain connectivity*, 2(3):125–141, 2012.

Weizheng Yan, Vince Calhoun, Ming Song, Yue Cui, Hao Yan, Shengfeng Liu, Lingzhong Fan, Nianming Zuo, Zhengyi Yang, Kaibin Xu, et al. Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site fmri data. *EBioMedicine*, 47:543–552, 2019.

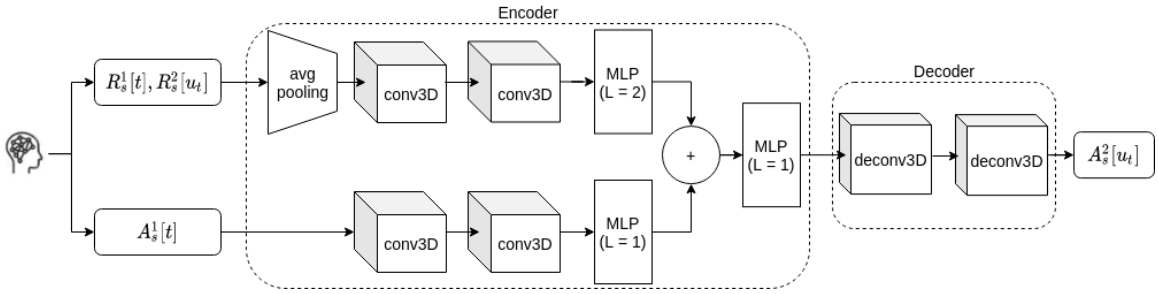


Figure 4: Architecture of f, our Amygdala prediction Neural Network

Table 2: The MSE error of network f in comparison to two simple baselines

Method	Healthy	Fibromyalgia	PTSD
Mean Prediction	$0.1151 \pm 3 \cdot 10^{-3}$	$0.0780 \pm 1 \cdot 10^{-3}$	$0.0770 \pm 1 \cdot 10^{-3}$
Predicting $A_s^1[t]$	$0.1087 \pm 2 \cdot 10^{-3}$	$0.0833 \pm 4 \cdot 10^{-3}$	$0.0781 \pm 3 \cdot 10^{-3}$
$f(I_s^1[t], R_s^2[u_t])$	$\mathbf{0.0735} \pm 2 \cdot 10^{-3}$	$\mathbf{0.0561} \pm 2 \cdot 10^{-3}$	$\mathbf{0.0550} \pm 1 \cdot 10^{-3}$

Appendix A. Implementation details and additional results

The architecture details of network f The network architecture of f is illustrated in Fig. 4, and comprised of an encoder and a decoder. The encoder receives the preprocessed fMRI signals; Rest-of-brain signals, $\left(R_s^1[t], R_s^2[u_t]\right)$, and Amygdala signals, $A_s^1[t]$, are processed separately. First, the rest-of-brain signals are going through an average pooling layer, with a kernel of size $2X2X2$. Next, both signals are processed using two independent 3D-convolutional layers (separated by a ReLU layer), followed by an MLP layer. The outputs of the MLPs are then concatenated and fed into another MLP layer, with an output of size 18.

The decoder part of f contains two deconvolutional layers, separated by a ReLU layer. The decoder outputs the network's prediction map of the subjects' Amygdala during the paired time-step, $\hat{A}_s^2[u_t]$.

The prediction model f is implemented in PyTorch, contains 184K parameters, and runs in real-time on an NVIDIA Tesla P100 GPU: a forward pass takes 18 ms on average.

Training details The clusters association and sample matching (step i + ii in Fig.1), are performed before training the networks f, ρ. Results are stored in the dataset, such that each record in it contains: (1) first session signals, $(R_s^1[t], A_s^1[t])$; (2) matched second session rest-of-brain signal, $R_s^2[u_t]$; and (3) the cluster to which the sample was associated. The hyper-parameters (learning rate, batch size, etc.) of both networks f and ρ were selected according to a grid search using the cross validation scores on the validation set. For training, we used an Adam optimizer, with initial learning rates of 0.001 and 0.01, and a batch sizes of 16 and 1, respectively. We trained the network f, until convergence of its validation loss for each of the three datasets - healthy, PTSD and Fibromyalgia. The MSE error reached is presented in Tab. 2. As can be seen, the network is quite successful in performing its prediction task in comparison to the simple baselines of predicting the activations in the network's input $A_s^1[t]$ and predicting the mean activation of each Amygdala's voxels.

Table 3: Relative error for age, TAS, STAI, and CAPS-5. Shown are mean ± Standard Deviation over five random train/test splits.

		Healthy		Fibromyalgia		PTSD		
	Age↓	TAS↓	STAI↓	TAS↓	STAI↓	TAS↓	STAI↓	CAPS-5↓
Mean pred	0.168 ± 0.04	0.250 ± 0.05	0.296 ± 0.08	0.230 ± 0.06	0.180 ± 0.04	0.247 ± 0.06	0.268 ± 0.07	0.474 ± 0.13
(Osin et al., 2020)	0.144 ± 0.04	0.205 ± 0.04	0.273 ± 0.09	0.154 ± 0.05	0.174 ± 0.03	0.198 ± 0.04	0.244 ± 0.06	0.400 ± 0.12
CNN	0.188 ± 0.05	0.238 ± 0.08	0.301 ± 0.08	0.187 ± 0.07	0.193 ± 0.06	0.204 ± 0.06	0.274 ± 0.12	0.430 ± 0.15
Clinical SVM	0.163 ± 0.06	0.227 ± 0.07	0.294 ± 0.07	0.200 ± 0.06	0.181 ± 0.05	0.243 ± 0.07	0.260 ± 0.07	0.460 ± 0.16
Ours- old ROI	$\mathbf{0.137} \pm 0.04$	0.196 ± 0.06	0.260 ± 0.08	0.129 ± 0.06	$\mathbf{0.170} \pm 0.04$	0.187 ± 0.05	0.237 ± 0.06	$\mathbf{0.385} \pm 0.14$
Ours- new ROI	0.140 ± 0.03	$\mathbf{0.192} \pm 0.06$	$\mathbf{0.251} \pm 0.07$	$\mathbf{0.120} \pm 0.04$	0.172 ± 0.05	$\mathbf{0.184} \pm 0.05$	$\mathbf{0.206} \pm 0.04$	0.390 ± 0.11
Ablation $\tilde{E}_s$	0.165 ± 0.04	0.225 ± 0.08	0.285 ± 0.09	0.204 ± 0.06	0.179 ± 0.06	0.201 ± 0.09	0.297 ± 0.05	0.507 ± 0.12
Ablation $\mathcal{L}_f$	0.145 ± 0.06	0.217 ± 0.04	0.278 ± 0.10	0.213 ± 0.08	0.184 ± 0.07	0.218 ± 0.06	0.251 ± 0.07	0.475 ± 0.16
Ablation $\lvert C_i \rvert$	0.150 ± 0.07	0.208 ± 0.06	0.274 ± 0.11	0.247 ± 0.09	0.213 ± 0.09	0.234 ± 0.08	0.249 ± 0.09	0.532 ± 0.11
ClusterLSTM	0.151 ± 0.03	0.225 ± 0.07	0.254 ± 0.10	0.194 ± 0.07	0.190 ± 0.05	0.203 ± 0.07	0.272 ± 0.09	0.411 ± 0.10
No feedback	0.168 ± 0.05	0.253 ± 0.07	0.288 ± 0.08	0.220 ± 0.07	0.179 ± 0.07	0.256 ± 0.08	0.275 ± 0.08	0.512 ± 0.13
Alt. clustering	0.159 ± 0.04	0.242 ± 0.05	0.282 ± 0.09	0.194 ± 0.06	0.182 ± 0.05	0.197 ± 0.06	0.280 ± 0.08	0.436 ± 0.15
Alt. ROI	0.167 ± 0.03	0.254 ± 0.06	0.296 ± 0.10	0.227 ± 0.07	0.187 ± 0.05	0.232 ± 0.08	0.247 ± 0.09	0.491 ± 0.09

Table 4: Regression results (MSE) for age, TAS, STAI, and CAPS-5. Shown are mean ± Standard Deviation over five random train/test splits.

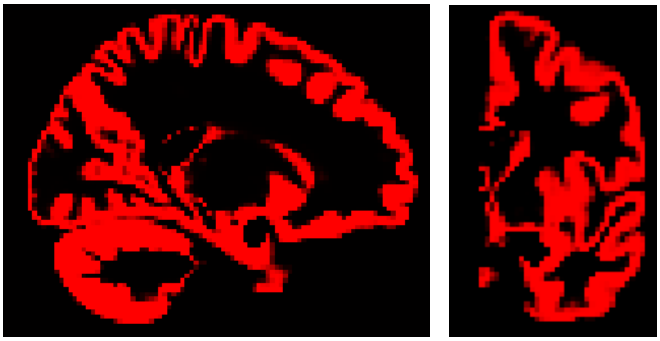
		Healthy		Fibromyalgia		PTSD		
	Age↓	TAS↓	STAI↓	TAS↓	STAI↓	TAS↓	STAI↓	CAPS-5↓
Mean pred	13.7 ± 1	121.7 ± 6	79.3 ± 6	98.6 ± 7	78.5 ± 5	153.0 ± 11	159.5 ± 13	119.3 ± 10
(Osin et al., 2020)	10.1 ± 1	81.6 ± 7	67.4 ± 6	44.0 ± 5	73.1 ± 3	99.2 ± 5	132.7 ± 8	85.4 ± 8
CNN	17.2 ± 1	110.3 ± 12	81.6 ± 7	65.2 ± 10	90.1 ± 9	105.0 ± 11	166.3 ± 33	98.4 ± 16
Clinical SVM	13.0 ± 2	100.0 ± 12	78.2 ± 8	74.3 ± 7	77.9 ± 11	148.3 ± 12	150.0 ± 11	117.4 ± 14
Ours, old-ROI	$\mathbf{9.2} \pm 1$	$\mathbf{74.8} \pm 7$	61.2 ± 6	$\mathbf{31.2} \pm 6$	72.0 ± 4	$\mathbf{88.0} \pm 7$	124.5 ± 9	$\mathbf{79.3} \pm 12$
Ours, new-ROI	9.3 ± 1	79.0 ± 3	$\mathbf{55.2} \pm 7$	35.3 ± 8	$\mathbf{62.3} \pm 7$	89.1 ± 11	$\mathbf{97.0} \pm 10$	84.3 ± 10
Ablation $\tilde{E}_s$	13.2 ± 1	98.4 ± 15	73.2 ± 9	77.8 ± 7	77.4 ± 11	102.2 ± 20	194.8 ± 6	136.0 ± 14
Ablation $\mathcal{L}_f$	10.3 ± 1	91.6 ± 4	70.0 ± 9	84.4 ± 13	81.9 ± 14	119.6 ± 11	141.0 ± 9	119.2 ± 19
Ablation $\lvert C_i \rvert$	11.0 ± 2	84.2 ± 9	68.1 ± 9	113.6 ± 32	109.4 ± 20	137.5 ± 18	137.8 ± 20	151.5 ± 19
ClusterLSTM	12.1 ± 1	96.8 ± 6	68.8 ± 8	75.7 ± 9	90.0 ± 6	118.1 ± 12	155.0 ± 14	108.0 ± 11
No Feedback	13.7 ± 2	125.0 ± 9	75.0 ± 8	90.0 ± 8	77.7 ± 7	165.3 ± 16	167.6 ± 18	139.0 ± 11
Alt. clustering	12.3 ± 1	114 ± 5	72.5 ± 9	70.2 ± 8	80.3 ± 8	98.3 ± 12	174.0 ± 20	103.6 ± 12
Alt. ROI	13.5 ± 1	125.5 ± 7	79.0 ± 10	96.3 ± 9	84.5 ± 8	135.6 ± 16	135.8 ± 19	128.7 ± 5

Appendix B. Relative Error and Standard Deviation

Table 3 contains the relative error (mean and Standard Deviation) of our main experiments. Table 4 is similar only it contains the MSE. it reports the same stats as does Tab. 1 in the main paper, with addition of the Standard Deviation.

Appendix C. fMRI data acquisition and pre-processing

The structural and functional scans were obtained with a 3.0T Siemens MRI system (MAGNETOM Prisma) using a 20-channel head coil. A T1-weighted three-dimensional (3D) sagittal MPRAGE pulse sequence (repetition time/echo time=1,860/2.74 ms, flip angle=8°, pixel size=1×1 mm, field of view=$256 \times 256mm$) was used to increase the resolution of the structural images. The functional scans were performed in an interleaved top-to-bottom order,



(a) (b)

Figure 5: (a) Rest of brain mask; and (b) a 90° flipped version of it

using a T2*-weighted gradient echo planar imaging pulse sequence (repetition time/echo time=3,000/35ms, flip angle=90°, pixel size=1.56mm, field of view=200 × 200mm, slice thickness=3 mm, 44 slices per volume).

The CONN MATLAB toolbox (Whitfield-Gabrieli and Nieto-Castanon, 2012) was used to realign the functional volumes, motion correction using rigid-body transformations in six axes, normalization to MNI space and spatial smoothing with an isotropic 6-mm full width at half-maximum Gaussian kernel. Subsequently de-noising and de-trending regression algorithms were applied, followed by bandpass filtering in the range of 0.008-0.09 Hz. The frequencies in the bandpass filter reflect the goal of modeling the individual throughout the session, while removing the effects of the fast paced events that occur during the neurofeedback session. This filtering follows previous work (Osin et al., 2020). While it may remove important information from the signal, we adopt it as is for a fair comparison. The Amygdala voxels were defined as the functional cluster centered at coordinates (x=21, y=-2, z=-24).

Amygdala and Rest-of-Brain ROI calculation The Amygdala region of interest was defined in SPM as a 6mm sphere located at MNI coordinates [21, -1, -22]. This region was used for providing the feedback during the real-time fMRI experiments.

See Fig. 5 for a visualisation of the regions included in the Rest-of-Brain signal.

Table 5: The prototypical clusters, sorted by the mean prediction error for the frames that belong to each cluster.

	MSE	Regions
(a)	0.0653	Primary Sensorimotor, Primary Visual Cortex
(b)	0.0691	Posterior Cingulate Cortex, Medial Prefrontal Cortex (Default mode network)
(c)	0.0697	Insular Cortex, Supplementary motor area, Cingulate Cortex, Visual Cortex (Salience Network)
(d)	0.0745	Superior Parietal Cortex, Frontal Eye Fields (Dorsal attention network)
(e)	0.0776	High visual areas (Ventral and Dorsal Stream)

Appendix D. Analysis of the obtained clusters

In order to understand the prototypical brain states that are obtained through the clustering process, we have visualized the cluster centroids which represent the prototypical brain states found during the NF task for the Healthy subgroup. The resulting maps, depicted in Fig. 6, show the main activated nodes in each state.

The fact that the clusters are distinct in their spatial arrangement is supportive of the relevance of using clustering for this purpose. Sorting the clusters according to the mean predicting error can hint on the brain states that are more supportive of learning the NF task, see Tab. 5.

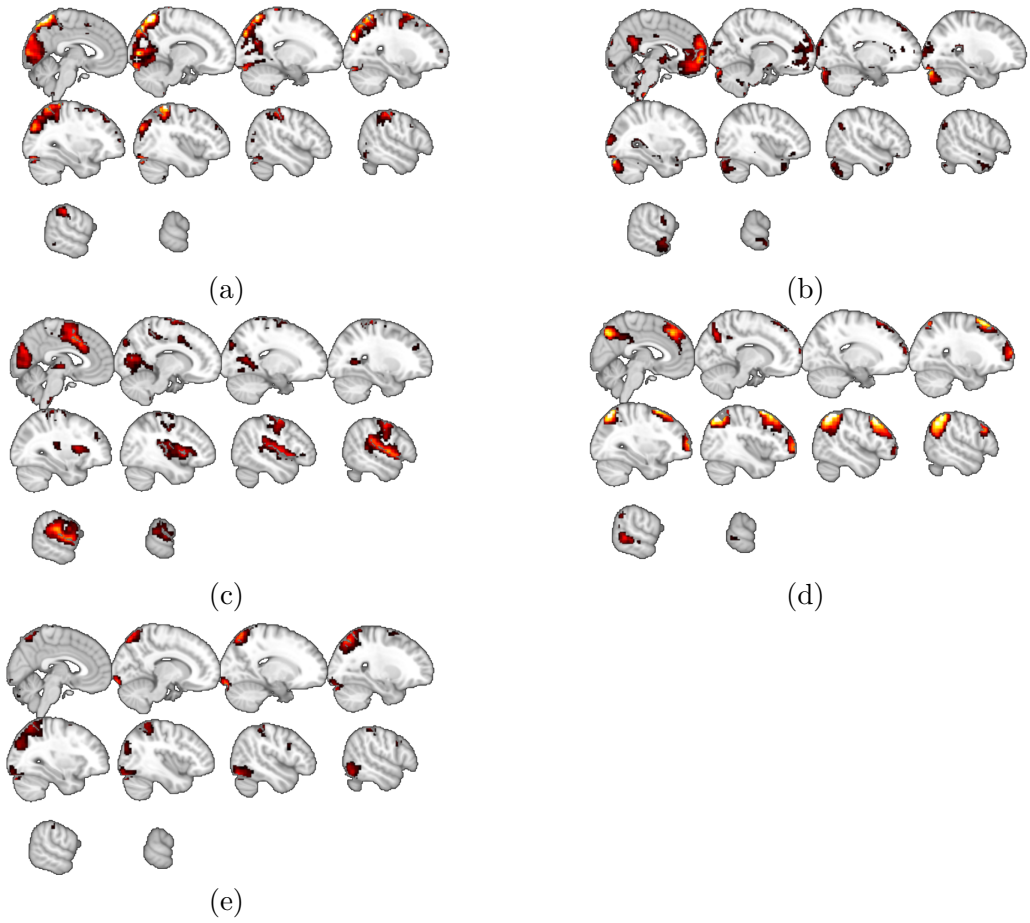


Figure 6: Five prototypical clusters obtained on the healthy individuals dataset. (a) Sensorimotor Network and Visual Cortex. (b) Main nodes of the Default Mode network (c) Main nodes of the Salience Network. (d) Dorsal attention network. (e) High visual areas (Ventral and Dorsal Stream).