frontiers in Computational Neuroscience

Classifying cognitive profiles using machine learning with privileged information in Mild Cognitive Impairment

Hanin H. Alahmadi[1], Yuan Shen[1*], Shereen Fouad[1], Caroline Luft[2], Peter Bentham[1], Zoe Kourtzi[3], Peter Tino[1*]

[1]The University of Birmingham, United Kingdom, [2]University of London Queen Mary, United Kingdom, [3]The University of Cambridge, United Kingdom

Submitted to Journal:
Frontiers in Computational Neuroscience

ISSN:
1662-5188

Article type:
Original Research Article

Received on:
30 Jun 2016

Accepted on:
31 Oct 2016

Provisional PDF published on:
31 Oct 2016

Frontiers website link:
www.frontiersin.org

Citation:
Alahmadi HH, Shen Y, Fouad S, Luft C, Bentham P, Kourtzi Z and Tino P(2016) Classifying cognitive profiles using machine learning with privileged information in Mild Cognitive Impairment. *Front. Comput. Neurosci.* 10:117. doi:10.3389/fncom.2016.00117

frontiers

Classifying cognitive profiles using machine learning with privileged information in Mild Cognitive Impairment

Hanin Al-Alahmadi [1], Yuan Shen [1,*], Shereen Fouad [2], Caroline Di Bernardi Luft [3], Peter Bentham [4], Zoe Kourtzi [5] and Peter Tino [1,*]

[1] *School of Computer Science, The University of Birmingham, Birmingham, United Kingdom*
[2] *School of Dentistry, The University of Birmingham, United Kingdom*
[3] *School of Biological and Chemical Sciences, University of London Queen Mary, United Kingdom*
[4] *School of Clinical and Experimental Medicine, The University of Birmingham, United Kingdom*
[5] *Department of Psychology, The University of Cambridge, United Kingdom*

Correspondence*:
Yuan Shen
y.shen.2@cs.bham.ac.uk

Peter Tino
pxt@cs.bham.ac.uk

ABSTRACT

Early diagnosis of dementia is critical for assessing disease progression and potential treatment. State-or-the-art machine learning techniques have been increasingly employed to take on this diagnostic task. In this study, we employed Generalised Matrix Learning Vector Quantization (GMLVQ) classifiers to discriminate patients with Mild Cognitive Impairment (MCI) from healthy controls based on their cognitive skills. Further, we adopted a "Learning with privileged information" approach to combine cognitive and fMRI data for the classification task. The resulting classifier operates solely on the cognitive data while it incorporates the fMRI data as privileged information (PI) during training. This novel classifier is of practical use as the collection of brain imaging data is not always possible with patients and older participants.

MCI patients and healthy age-matched controls were trained to extract structure from temporal sequences. We ask whether machine learning classifiers can be used to discriminate patients from controls based on the learning performance and whether differences between these groups relate to individual cognitive profiles. To this end, we tested participants in four cognitive tasks: working memory, cognitive inhibition, divided attention, and selective attention. We also collected fMRI data before and after training on the learning task and extracted fMRI responses and connectivity as features for machine learning classifiers.

Our results show that the PI guided GMLVQ classifiers outperform the baseline classifier that only used the cognitive data. In addition, we found that for the baseline classifier, divided attention is the only relevant cognitive feature. When PI was incorporated, divided attention remained the

most relevant feature while cognitive inhibition became also relevant for the task. Interestingly, this analysis for the fMRI GMLVQ classifier suggests that (1) when overall fMRI signal for structured stimuli is used as inputs to the classifier, the post-training session is most relevant; and (2) when the graph feature reflecting underlying spatiotemporal fMRI pattern is used, the pre-training session is most relevant. Further analysis reveals that for MCI patients, training may alter brain activation level as well as local brain connectivity pattern. Taken together these results suggest that brain connectivity before training and overall fMRI signal after training are both diagnostic of cognitive skills in MCI.

Keywords: discriminative feature extraction, supervised metric learning, learning with privileged information, learning vector quantization, linear discriminant analysis, fMRI graph feature

1 INTRODUCTION

Alzheimer's Disease (AD) is the most common neurodegenerative disease in ageing. It is characterised by the progressive impairment of neurons and their connections. Mild Cognitive Impairment (MCI) is the prodromal stage of AD. Thus, accurate diagnosis of MCI (i.e. the early stage of AD) is very important for timely treatment and delay of disease progression. As MCI results in detectable loss of cognitive function, cognitive test scores have been used diagnostically (Albert et al., 2010). Further, MCI is known to cause changes in brain activation patterns as well as in brain connectivity. Therefore, fMRI has been increasingly used as a diagnostic tool of MCI patients (Challis et al., 2015; Chen et al., 2015). In machine learning terms, diagnosis of MCI patients can be formulated as a classification task to discriminate MCI patients from healthy controls. In this paper, we present a novel classifier using cognitive test scores as inputs to the classifier and using fMRI data as privileged information.

In the recent literature on the classification tasks related to AD, we observe a clear trend: state-of-the-art machine learning techniques have been increasingly employed to take on new tasks. For example, a classification task should also provide insights into the relevance of the input features used for the task. In Challis et al. (2015), Gaussian process classifiers have been employed for the discrimination between healthy controls and MCI patients as well as the the discrimination between MCI and AD patients. More importantly, Gaussian process classifiers have been used to automatically determine the relevant input features when training the classifier. In Chen et al. (2015), a challenging classification task was tested, that is, discrimination of two subgroups of MCI patients. Patients in one subgroup will likely progress to AD but those in another group will not convert to AD. In the literature, this classification task is referred to as MCI-AD conversion prediction. This work incorporates data from both healthy subjects and AD patients for classification of MCI patients using the transfer learning framework. Transfer learning is a (relatively) new development in machine learning that aims to boost the performance of a classifier operating in one domain (e.g. MCI patients) by incorporating data from other domains (e.g. healthy subjects and AD patients).

Here we ask whether MCI patients differ in their cognitive skills from controls. Our task is to classify cognitive profiles in patients vs. controls based on cognitive scores and fMRI data. Furthermore, we address the case when fMRI data are not available for classifying a new subject. To utilise the fMRI data for the task, we train our classifier on participants for whom both cognitive and fMRI data are available. After that, the trained classifier will classify a new subject solely based on his/er cognitive test scores. This case is of relevance in practice because (1) When compared to cognitive data, the collection of neuroimaging data is much more time-consuming and expensive; (2) Many older individuals (e.g. those with a cardiac pacemaker) may not be safe for imaging such as fMRI scanning. On the other hand, neuroimaging data

have more diagnostic power than cognitive data and thus should be used when available. In our work, the classifier is trained by adopting a "metric learning" based approach to *Learning with Privileged Information* (LPI) (Fouad, 2013). As transfer learning, LPI is also a new development in machine learning. In our context, cognitive data are the inputs to the classifier. In contrast, fMRI data act as privileged information that is used only for training the classifier (along with the cognitive data). As most classifiers operate based on a distance/similarity measure between pairs of input vectors, the metric tensor used to compute such distance is therefore crucial for the classification task. In the model of Fouad (2013), the privileged information (in our case fMRI data) is used to modify the metric tensor (and hence the metric) in the original space (in our case cognitive test scores) to improve the classification accuracy in the original space. Intuitively, if cognitive test scores of two participants appear "similar", but their fMRI data shows different characteristics, the distance between the two cognitive test score vectors should be increased (and vice-versa). As the scale parameter in Challis et al. (2015), the diagonal elements of the discriminative metric tensor can be used to automatically determine the relevant cognitive features.

2 MATERIALS

The cognitive and fMRI data used in this study were collected in the context of two behavioral & fMRI studies (Baker et al., 2015; Luft et al., 2015, 2016) in which the participants were asked to predict the orientation of a test stimulus following exposure to structured sequence of leftwards and rightwards oriented gratings, and no feedback were given. Both studies aimed to (1) test whether training on structured temporal sequences improves the ability to predict upcoming sensory events and (2) identify brain regions that support the ability of using implicit knowledge about the past for predicting future. In particular, Baker et al. (2015) and Luft et al. (2015) investigated how MCI patients differ from healthy controls in terms of (1) their ability to learn predictive structures as well as (2) their learning-dependent brain activation patterns. The diagnosis of MCI patients was made by an experienced consultant psychiatrist (PB) using the National Institute of Ageing and Alzheimer's association working group criteria (Albert et al., 2010).

In both studies, participants took part in two fMRI scans before and after behavioural training (i.e. pre- and post-training session) during which they completed 5–8 independent runs of the prediction task in each scanning session. Each run comprised 5 blocks of structured and 5 blocks of random sequences (3 trials per block) presented in a random counterbalanced order. In each trial, the participant was presented with a sequence of eight left and rightward oriented gratings (in rapid succession, 250ms + fixation 200ms) followed by a repeat of the same sequence. The participant was instructed to pay attention to the sequence and respond whether the test grating (randomly chosen grating during the second repeat) was correct or incorrect given that presented sequence. Even though the participants could not tell what exactly was the sequence structure, they learn how to correctly predict whether the grating has the correct orientation given the presented sequence. In random sequence trials, the grating's orientations were randomly generated so the participant could not correctly predict them.

The fMRI data used in this study were acquired in a 3T Achieva Philips scanner at the Birmingham University Imaging Centre using a thirty two-channel head coil. Anatomical images were obtained using a sagittal three dimensional T1-weighted sequence with 175 slices (voxel size = $1 \times 1 \times 1$ mm^3) for localisation and visualisation of functional data. Functional data were acquired using a T2-weighted EPI sequence with 32 slices (whole-brain coverage; TR = 2 s; TE = 35 ms; flip angle = 73; voxel size = $2.5 \times 2.5 \times 4$ mm^3).

In Luft et al. (2016), regions-of-interest (ROIs) were identified by applying whole-brain general linear model analysis with a voxel-wise mixed-design three-way ANOVA, that is,

session (pre- vs. post-training) × sequence (structured vs. random) × group (MCI vs. controls).

Statistical maps were cluster threshold corrected ($p < 0.05$). Table 1 in Luft et al. (2016) listed all brain regions showing significant interaction between session, sequence, and group. For the study presented in this paper, we combined two ROIs in the frontal region (Superior Frontal Gyrus, SFG, on the right hemisphere and Medial Frontal Gyrus, MFG, on the left hemisphere) and two ROIs in the cerebellar region (Cerebellar Lingual and Cullmen ROIs in both hemispheres). This resulted in a frontal ROI of size 126 and a cerebellar ROI of size 82. Also, a subcortical ROI (that is, the parahippocampal gyrus ROI of size 32) was selected for the study.

All 60 participants involved in this study had undergone cognitive skill tests (including working memory, cognitive inhibition and attentional skills). These tests provide four quantitative measures of different cognitive skills for each participant:

1. In the working memory task, a number of coloured dots are on display for half second. Then, they disappear for 1 second and reappear with some dots having changed their colour. A participant is asked to judge whether a given dot has changed its colour or not. The participant's working memory skill can be measured by the maximal number of coloured dots on display for achieving a 70.7% test performance (denoted by n_{dots});
2. To quantify a participant's attention skill, the following cognitive task was performed: two objects are on display, one located at the display centre, another located on the periphery of the display. The peripheral object can only take one of eight equally distributed radial directions (with respect to the display center). The central object could be either car or truck silhouette, whereas the peripheral object must always be the truck silhouette. The participant was asked to identify the type of the central object (car vs. truck) and the location of the peripheral stimulus before the display was masked by white visual noise. This skill is measured by the minimal display time required for the participant to achieve 70% task performance. Depending on whether or not there are distractors on the display, the skill of divided or selective attention is measured (denoted by t^d_{disp} and t^s_{disp}, respectively);
3. The skill of inhibition is measured in a stop-signal test. A participant is first cued to perform a motor task. This is followed by a tone with some time delay, which signals task abortion. The quantity measuring the inhibition skill, t_{delay}, is given by the minimum delay time for achieving a 70.7% test performance.

Sixty participants are involved in this study. Thirty-four of them have both cognitive and fMRI data. Among these participants, nine MCI patients and nine healthy controls come from the cohort reported in Luft et al. (2015). The remaining sixteen healthy controls come from the cohort reported in Luft et al. (2016). The size of that cohort is twenty. Four of them are not included in this study because their cognitive data were missing. Note that for these thirty-four subjects having both cognitive and neuroimaging data for training of classifiers, MCI patients and healthy controls were age matched: mean age of MCI patients was 68.9 , and mean age of controls was 68.3. The remaining twenty-six participants have cognitive data only. Among them, four MCI patients and five healthy controls come from Baker et al. (2015) and Luft et al. (2015). The remaining seventeen participants are from unpublished studies but they participated exactly the same experiments as other participants. Note that all neuroimaging data used in this study are reported either in Luft et al. (2015) or in Luft et al. (2016).

3 METHODS

3.1 Generation of fMRI features

3.1.1 fMRI signal features

For each ROI and each (pre- and post training) session, we calculated percent signal change (PSC) by subtracting fMRI responses to random sequences from fMRI response to structured sequences and dividing by averaged fMRI response to both stimulus sequences. Let n_r and n_s denote the number of volumes scanned during the trials with random and structured sequences, respectively. For a ROI of size V, its PSC value is computed as follows:

$$PSC = \frac{1}{V}\sum_{v=1}^{V}\frac{\frac{1}{n_s}\sum\limits_{i\in I_s} y_{vi} - \frac{1}{n_r}\sum\limits_{j\in I_r} y_{vj}}{\frac{1}{n_s}\sum\limits_{i\in I_s} y_{vi} + \frac{1}{n_r}\sum\limits_{j\in I_r} y_{vj}} \tag{1}$$

where i and j denote volume index, v voxel index, $I_s = \{i_1, ..., i_{n_s}\}$ the collection of "structured" volumes and $I_r = \{j_1, ..., j_{n_s}\}$ the collection of "random" volumes. The above definition implies that PSC measures scaled fMRI-response to temporally structured stimuli and it is an overall measure averaged over both volumes and voxels.

3.1.2 fMRI graph features

3.1.2.1 Graph matrix

Graph structure characterises the connectivity between nodes of a graph. In this study, the graph structure of a single ROI is represented by so-called graph matrix G of size $V \times V$ where V denotes the ROI size. The value of G_{ij} measures the functional connectivity between voxel i and voxel j, and is computed as (linear) cross-correlation between two fMRI time series of length n on the voxel pair (denoted by $\mathbf{y}_i = (y_{i1}, ..., y_{in})^\intercal$ and $\mathbf{y}_j = (y_{j1}, ..., y_{jn})^\intercal$, respectively), that is,

$$G_{ij} = \frac{1}{n}\cdot\frac{\sum\limits_{k=1}^{n}(y_{ik}-\mu_i)\cdot(y_{jk}-\mu_j)}{\sigma_i\cdot\sigma_j} \tag{2}$$

where μ and σ stand for the mean and standard deviation of individual fMRI time series. In the case of $i = j$, we obtain $G_{ij} = 1$. Note that G_{ij} is a connectivity measure independent of the activation intensity on each of two voxels.

3.1.2.2 Discriminative feature extraction

Often, a classifier's inputs are not those raw data to be classified but the features extracted from the raw data. This can significantly reduce the input dimension, which tackles both "curse of dimensionality" and the small sample-size problem. Therefore, a good choice of feature vector plays an important role in classification. This is the motivation for extraction of discriminative features. The discriminative features are suitable because they are extracted in a task-driven & supervised manner. Linear Discriminant Analysis (LDA) is a machine learning technique for discriminative feature extraction. The assumption of LDA is that the feature vectors of each class are Gaussian-distributed. In LDA, high-dimensional feature vectors are projected into a lower-dimensional space and the projection matrix is optimized so that the classes are maximally separated in the projection space. To this end, the empirical covariance matrices need to

be estimated using the feature vectors from individual classes. If the number of feature vectors is small and their dimension is high, the empirical estimates of covariance matrices are not accurate. Thus, LDA suffers from the same problem as classifiers do. So-called 2D-LDA has been proposed by Sato et al. (2008) for the cases where data items are matrices (e.g. graph matrices in this study) and a direct application of standard LDA with vectorised matrices could fail due to the above-mentioned problem. In the following, we summarise both standard LDA and 2D-LDA with the dimension of the projection space fixed to one.

For standard LDA, assume that we have N d-dimensional feature vectors, $\{\mathbf{x}_n : n = 1, ..., N\}$, for training in which N_1 feature vectors are from *Class 1* and $N_2 = N - N_1$ from *Class 2*. Denote these two subsets by $\mathscr{C}_1$ and $\mathscr{C}_2$, respectively. The mean vectors of *Class 1* and *Class 2* are given by $\mathbf{m}_1 = \frac{1}{N_1} \sum_{\mathbf{x}_n \in \mathscr{C}_1} \mathbf{x}_n$ and $\mathbf{m}_2 = \frac{1}{N_2} \sum_{\mathbf{x}_n \in \mathscr{C}_2} \mathbf{x}_n$, respectively. Define the between-class covariance matrix $\mathbf{S}_B$ and the total within-class covariance matrix $\mathbf{S}_W$ as

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\mathsf{T} \tag{3}$$

and

$$\mathbf{S}_W = \sum_{\mathbf{x}_n \in \mathscr{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\mathsf{T} + \sum_{\mathbf{x}_n \in \mathscr{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\mathsf{T}. \tag{4}$$

The projection matrix $\mathbf{w}$ of size $d \times 1$ is optimized by maximizing the Fisher criterion defined by

$$J(\mathbf{w}) = \frac{\mathbf{w}^\mathsf{T}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^\mathsf{T}\mathbf{S}_W\mathbf{w}} = \frac{\mathbf{D}_B}{\mathbf{D}_W}. \tag{5}$$

$\mathbf{D}_B$ and $\mathbf{D}_W$ are referred to as the between-class distance and the total within-class distance. Denote the optimized $\mathbf{w}$ by $\mathbf{w}_{\text{opt}}$ and the extracted features are given as $\{f_n = \mathbf{w}_{\text{opt}}^\mathsf{T}\mathbf{x}_n : n = 1, ..., N\}$.

For 2D-LDA, assume that we have N graph matrices of size $d \times d$, $\{\mathbf{X}_n : n = 1, ..., N\}$, for training in which N_1 feature vectors are from *Class 1* and $N_2 = N - N_1$ from *Class 2*. Denote these two subsets by $\mathscr{C}_1$ and $\mathscr{C}_2$, respectively. For *Class 1* and *Class 2*, their mean matrices are given by $\mathbf{M}_1 = \frac{1}{N_1} \sum_{\mathbf{X}_n \in \mathscr{C}_1} \mathbf{X}_n$ and $\mathbf{M}_2 = \frac{1}{N_2} \sum_{\mathbf{X}_n \in \mathscr{C}_2} \mathbf{X}_n$. In contrast to standard LDA, we need two (left and right) projection matrices (or vectors), denoted by $\mathbf{a}$ and $\mathbf{b}$ of size $d \times 1$ projecting the matrices into real numbers. Similarly, the between-class distance and the total within-class distance are defined as

$$\begin{aligned} \mathbf{D}_B &= \mathbf{a}^\mathsf{T}(\mathbf{M}_2 - \mathbf{M}_1)\mathbf{b}\mathbf{b}^\mathsf{T}(\mathbf{M}_2 - \mathbf{M}_1)\mathbf{a} && (6) \\ &= \mathbf{b}^\mathsf{T}(\mathbf{M}_2 - \mathbf{M}_1)\mathbf{a}\mathbf{a}^\mathsf{T}(\mathbf{M}_2 - \mathbf{M}_1)\mathbf{b} && (7) \end{aligned}$$

and

$$\begin{aligned} \mathbf{D}_W &= \sum_{\mathbf{X}_n \in \mathscr{C}_1} \mathbf{a}^\mathsf{T}(\mathbf{X}_n - \mathbf{M}_1)\mathbf{b}\mathbf{b}^\mathsf{T}(\mathbf{X}_n - \mathbf{M}_1)\mathbf{a} + \sum_{\mathbf{X}_n \in \mathscr{C}_2} \mathbf{a}^\mathsf{T}(\mathbf{X}_n - \mathbf{M}_2)\mathbf{b}\mathbf{b}^\mathsf{T}(\mathbf{X}_n - \mathbf{M}_2)\mathbf{a} && (8) \\ &= \sum_{\mathbf{X}_n \in \mathscr{C}_1} \mathbf{b}^\mathsf{T}(\mathbf{X}_n - \mathbf{M}_1)\mathbf{a}\mathbf{a}^\mathsf{T}(\mathbf{X}_n - \mathbf{M}_1)\mathbf{b} + \sum_{\mathbf{X}_n \in \mathscr{C}_2} \mathbf{b}(\mathbf{X}_n - \mathbf{M}_2)\mathbf{a}\mathbf{a}^\mathsf{T}(\mathbf{X}_n - \mathbf{M}_2)\mathbf{b}. && (9) \end{aligned}$$

Note that $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{X}_n$, n = 1, 2, ..., N, are all symmetric matrix. The projection vectors $\mathbf{a}$ and $\mathbf{b}$ are optimized by maximizing $J(\mathbf{a}, \mathbf{b}) = \mathbf{D}_B/\mathbf{D}_W$ iteratively. At each iteration, we optimize $\mathbf{a}$ or $\mathbf{b}$ while

keeping $\mathbf{b}$ or $\mathbf{a}$ fixed. This procedure is repeated until J has converged. Denote the optimized $\mathbf{a}$ and $\mathbf{b}$ by $\mathbf{a}_{\text{opt}}$ and $\mathbf{b}_{\text{opt}}$. The extracted features are given as $\{f_n = \mathbf{a}_{\text{opt}}^{\intercal}\mathbf{X}_n\mathbf{b}_{\text{opt}} : n = 1, ..., N\}$.

Note that the number of free parameters to be optimised is d^2 for standard LDA operating on vectorised graph matrices and $2d$ for 2D-LDA operating on graph matrices directly.

3.1.2.3 Small sample-size problem

The main idea of this study is using costly but informative fMRI measurements as valuable privileged information in a classification task operating on cognitive features only. To do so the complex spatial-temporal structure in fMRI signals will need to be transformed into a set of indexes (scalars) that best discriminate between the classes.

In our approach we first capture the spatial-temporal structure of fMRI signals within an ROI as a cross-correlation graph. An ROI of V voxels will be represented as a full undirected graph with n nodes (one for each voxel) and the edge between nodes i and j is weighted by the value of the correlation coefficient between fMRI signals in the two voxels. Each such graph will in turn be represented by an $V \times V$ symmetric matrix $\mathbf{X}$ collecting the edge weights.

In this study we have two classes of N subjects - N_p patients and N_c healthy controls (that is $N = N_p + N_c$). The graph matrices of patients and controls are collected in matrix sets $\mathscr{C}_p$ and $\mathscr{C}_c$. Given the two sets of matrices, we propose to extract the discriminating feature f through a quadratic form applied to graph matrix $\mathbf{X}$: $f = \mathbf{a}^{\intercal}\mathbf{X}\mathbf{b}$. Both $\mathbf{a}$ and $\mathbf{b}$ are a V-dimensional vectors determined via an optimization problem expressing the need to maximally separate the two classes, while keeping the within-class variability minimal. To find the projection vectors $\mathbf{a}$ and $\mathbf{b}$ we used 2D-LDA (Ye et al., 2004).

For an ROI with V voxels, the discriminative features $\mathbf{a}$ and $\mathbf{b}$ are V-dimensional vectors, meaning that when determining $\mathbf{a}$ and $\mathbf{b}$ we have $2V$ free parameters. As the number of subjects N is smaller than $2V$, in order to avoid overfitting, the size of the graph representing spatial-temporal structure of cortical activations in that ROI needs to be reduced. Note that in our original formulation, each element a_i of $\mathbf{a}$ corresponds to a particular voxel i whose spatial position is $\mathbf{r}_i$. It is natural to expect that spatially close voxels will have similar activation patterns. We therefore introduce a set of K spatially smoothing Gaussian kernels $\mathcal{N}(\mathbf{r}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, 2, ..., K$, in the voxel space, positioned at $\boldsymbol{\mu}_k$, shape determined by the covariance matrix $\boldsymbol{\Sigma}_k$. This leads to a decomposition:

$$a_i = \sum_{k=1}^{K} \tilde{a}_k \mathcal{N}(\mathbf{r}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{10}$$

The values of the smoothing kernels k at each voxel i can be collected in the smoothing matrix.

$$\mathbf{P}_{i,k} = \mathcal{N}(\mathbf{r}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{11}$$

The feature vectors $\mathbf{a}$ and $\mathbf{b}$ can then be written as $\mathbf{a} = \mathbf{P}\tilde{\mathbf{a}}$ and $\mathbf{b} = \mathbf{P}\tilde{\mathbf{b}}$, respectively. We have:

$$f = \mathbf{a}^{\intercal}\mathbf{X}\mathbf{b} = \tilde{\mathbf{a}}^{\intercal}\mathbf{P}^{\intercal}\mathbf{X}\mathbf{P}\tilde{\mathbf{b}} \tag{12}$$

The $V \times V$ graph matrix $\mathbf{X}$ is thus reduced to the $K \times K$ matrix

$$\tilde{\mathbf{X}} = \mathbf{P}\mathbf{X}\mathbf{P}^{\mathsf{T}} \tag{13}$$

and

$$f = \tilde{\mathbf{a}}^{\mathsf{T}}\tilde{\mathbf{X}}\tilde{\mathbf{b}} \tag{14}$$

For a given number K of Gaussian kernels, their position is determined by k-means clustering in the voxel space and the covariance matrices of each cluster were estimated from the voxel positions within the corresponding clusters.

The number of smoothing kernels K in the three ROIs with 32, 82 and 126 voxels was set to 3, 4 and 8, respectively. The largest ROI is contained in both hemispheres. Hence, the sub-ROIs within each hemisphere were clustered independently into 4 clusters. Spatial smoothing with Gaussian kernels described above expresses the assumption that nearby voxels should have similar functionality. We refer to this approach as Spatial Grouping (SG) and to the resulting feature as SGF. An alternative approach would be to identify groups of voxels that are not only spatially close but also exhibit similarity in the activation time series (as quantified through cross-correlation) (Carpineto and Romano, 2012). We thus obtain N functional clusterings of the voxel space, one for each subject. These groupings at the subject level are then merged into a single population based functional clustering of voxels through Consensus Clustering (Carpineto and Romano, 2012). Given the resulting K voxel clusters, we calculated their means $\boldsymbol{\mu}_k$ and covariance matrices $\boldsymbol{\Sigma}_k$, thus obtaining a set of K "functionally informed" smoothing Gaussian kernels $\mathcal{N}(\mathbf{r}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. The reduced graph matrix $\tilde{\mathbf{X}}$ is then calculated as in eqs: (11) and (13). We refer to such functional voxel clustering as Functional grouping (FG)and to the resulting feature as FGF.

3.1.3 Feature generation pipeline

Figure 1 illustrates the flow of fMRI feature generation. We obtain three fMRI features (PSC, FGF, SGF) independently from fMRI data $\mathbf{Y} \in \mathbb{R}^{V \times T}$. Recall that V is number of voxels and T is the number of volumes. Feature PSC is computed directly from $\mathbf{Y}$. To compute other two features, we first transform $\mathbf{Y}$ to a graph matrix $\mathbf{X}$ of size $V \times V$ and reduce $\mathbf{X}$ to $\tilde{\mathbf{X}}$ of size $K \times K$ with $(K < V)$ either through spatial projection or through functional clustering. Finally, we extract SGF from $\tilde{\mathbf{X}}$ obtained by spatial projection and FGP from $\tilde{\mathbf{X}}$ obtained by functional clustering.

3.2 Classification Tools

3.2.1 Generalized Matrix Learning Vector Quantization (GMLVQ)

The classification algorithms of Learning Vector Quantization (LVQ) (Arbib, 2003) are supervised learning paradigms which work iteratively to modify the quantization prototypes to find the boundaries of the class. LVQ classifiers are represented by a set of vectors, so-called prototypes, embodying classes in the input space, and a distance metric on the input data. During training, prototypes are adapted in an iterative manner to define class borders. For each training point, the algorithm determines two closest prototypes, one with the same class as the training point, and another with a different class. The position of the two closet prototypes are then updated, where same class prototype is moved closer to the data point, while different class prototype is pushed away from the data point. During testing, an unknown point is assigned to the class represented by the closest prototype with respect to the given distance.

The LVQ scheme, which is originally introduced by Kohonen in 1986, applies Hebbian online learning in order to adapt prototype with training data. Subsequent, researchers proposed a number of modifications

to the basic learning scheme. Such variations utilize an explicit cost functionality, whereas others allow for incorporating adaptive distance measures (Schneider, 2010; Schneider et al., 2009).

Given training data $(\mathbf{x}_i, y_i) \in \mathbb{R}^m \times \{1, \cdots, K\}, i = 1, 2, \cdots, n$, where m denotes the dimensionality of data and K signifies the number of different classes. Typically, a LVQ network will include L prototypes $\mathbf{w}_q \in \mathbb{R}^m, q = 1, 2, 3, ..., L$, which is characterized according to their location available in the input space and their class $c(\mathbf{w}_q) \in \{1, ..., K\}$. At least one prototype in each class needs to be present. The overall number of prototypes is a model hyper-parameter that is to be optimized. The (squared) Euclidean distance $d(\mathbf{x}, \mathbf{w}) = (\mathbf{x} - \mathbf{w})^\intercal(\mathbf{x} - \mathbf{w})$ within $\mathbb{R}^m$ quantifies the distance between the input vectors and prototypes. The classification performed using the winner-takes all scheme: the data point $\mathbf{x}_i \in \mathbb{R}^m$ belongs to the label $c(\mathbf{w}_j)$ of the prototype $\mathbf{w}_j$ if and only if with $d(\mathbf{x}, \mathbf{w}_j) < d(\mathbf{x}, \mathbf{w}_q), \forall j \neq q$. For every prototype $\mathbf{w}_j$ with class $c(\mathbf{w}_j)$ a receptive field is defined within the input space. According to the LVQ model, points located in the respective field [1] will be assigned to the class $c(\mathbf{w}_j)$.

The aim of learning is to adapt prototypes automatically in such a way that the gap between data points of class $c \in \{1, ..., K\}$ and the corresponding prototypes with label c (the one that the data are belonging to) will be reduced to a minimum distance. During the stage of training for each data point $\mathbf{x}_i$ with class label $c(\mathbf{x}_i)$, the most proximal prototype with the same label is rewarded by pushing closer towards the training input; the most closest prototype with a different label will be disallowed by moving pattern $\mathbf{x}_i$ away.

The Generalized Matrix LVQ (GMLVQ) is a recent extension of the LVQ that employs a full matrix tensor for a better measure of distance between two feature vectors. The new distance measure not only is capable of scaling individual features but also accounts for pairwise correlations between the features. Assuming $\Lambda \in \mathbb{R}^{m \times m}$ is a positive definite matrix, $\Lambda \succ 0$, the generalized form of the squared Euclidean distance is defined as

$$d_\Lambda(\mathbf{x}_i, w) = (\mathbf{x}_i - \mathbf{w})^\intercal \Lambda (\mathbf{x}_i - \mathbf{w}) \tag{15}$$

The positive definiteness of Λ is guaranteed by imposing $\Lambda = \Omega^\intercal \Omega$, where $\Omega \in \mathbb{R}^{m \times m}$ is a full-rank matrix. Furthermore, to prevent the degeneration of the algorithm, Λ is trace normalized after each learning step (i.e. $\sum_i \Lambda_{ii} = 1$) so that the summation of eigenvalues is kept fixed in the learning process. The model is trained in an online-learning fashion and the steepest descent method is employed to minimize the cost function given as:

$$f_{GMLVQ} = \sum_{i=1}^{n} \phi(\mu_\Lambda(\mathbf{x}_i)) \tag{16}$$

with

$$\mu_\Lambda(\mathbf{x}_i) = \frac{d_\Lambda(\mathbf{x}_i, \mathbf{w}^+) - d_\Lambda(\mathbf{x}_i, \mathbf{w}^-)}{d_\Lambda(\mathbf{x}_i, \mathbf{w}^+) + d_\Lambda(\mathbf{x}_i, \mathbf{w}^-)}, \tag{17}$$

where ϕ is a monotonic function (the identity function $\phi(l) = l$ is a common choice). The main advantage of the GMLVQ framework is that (unlike LVQ (Schneider, 2010; Schneider et al., 2009)), it allows us to naturally incorporate privileged information through metric learning.

3.2.2 Privileged information (PI) guided GMLVQ

This paper employs the Information Theoretic Metric Learning (ITML) approach (Davis et al., 2007) in order to incorporate privileged information into the learning phase of the GMLVQ.

[1] The set of points in the input space is defined by the receptive field of prototype $\mathbf{w}$, where this prototype is picked as their winner.

Given a training dataset, we have one space where the original training data live and another space where the privileged training data live. They are denoted by $\mathscr{X}$ and $\mathscr{X}^*$, respectively, and their corresponding global metric tensors are denoted by $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^*$. The distances between the privileged training points in $\mathscr{X}^*$ are first computed using $\mathbf{\Lambda}^*$ and then are sorted in ascending order. Based on the closeness information in $\mathscr{X}^*$, the original training points are tagged in a categorical manner (similar and dis-similar). After that, the ITML approach is adopted to impose similarity constraints in the original space. The main goal is to learn a new metric in the original space (denoted by $\mathbf{\Lambda}_{\text{new}}$) so that under the new metric, the distance between two original training points is small if their counterparts in the privileged space are similar (close), and vice versa. Implementation of the above concept is described in the following.

The training dataset is given as $\{(\mathbf{x}_i, \mathbf{x}_i^*, y_i) : \mathbf{x}_i \in \mathscr{X}, \mathbf{x}_i^* \in \mathscr{X}^*, i = 1, 2, ..., N\}$. Recall that y represents class label. For each pair of two training examples, $1 \leq i < j \leq N$, we compute three different squared Mahalanobis distances as follows

$$d_{\mathbf{\Lambda}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\intercal \mathbf{\Lambda}(\mathbf{x}_i - \mathbf{x}_j), \mathbf{x}_i, \mathbf{x}_j \in \mathscr{X} \quad (18)$$

$$d_{\mathbf{\Lambda}^*}(\mathbf{x}_i^*, \mathbf{x}_j^*) = (\mathbf{x}_i^* - \mathbf{x}_j^*)^\intercal \mathbf{\Lambda}^*(\mathbf{x}_i^* - \mathbf{x}_j^*), \mathbf{x}_i^*, \mathbf{x}_j^* \in \mathscr{X}^* \quad (19)$$

$$d_{\mathbf{\Lambda}_{\text{new}}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\intercal \mathbf{\Lambda}_{\text{new}}(\mathbf{x}_i - \mathbf{x}_j), \mathbf{x}_i, \mathbf{x}_j \in \mathscr{X} \quad (20)$$

Note that $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^*$ are both given whereas $\mathbf{\Lambda}_{\text{new}}$ needs to be learned. The metric tensor $\mathbf{\Lambda}_{\text{new}}$ should be optimized in a supervised manner so that $d_{\mathbf{\Lambda}_{\text{new}}}(\mathbf{x}_i, \mathbf{x}_j)$ will be shrunk if $\mathbf{x}_i^*$ and $\mathbf{x}_j^*$ are similar. Otherwise, $d_{\mathbf{\Lambda}_{\text{new}}}(\mathbf{x}_i, \mathbf{x}_j)$ will be enlarged. To this end, we form two sets of pairs of the training data points in the original space $\mathscr{X}$: S_+ is a set of similar pairs and S_- a set of dissimilar pairs. These two sets are formed using the proximity information in the privileged space $\mathscr{X}^*$ as follows:

1. If $d_{\mathbf{\Lambda}^*}(\mathbf{x}_i^*, \mathbf{x}_j^*) \leq l^*$and $y_i = y_j$(same class label), then $(\mathbf{x}_i, \mathbf{x}_j) \in S_+$;
2. If $d_{\mathbf{\Lambda}^*}(\mathbf{x}_i^*, \mathbf{x}_j^*) \geq u^*$and $y_i \neq y_j$(different class label), then $(\mathbf{x}_i, \mathbf{x}_j) \in S_-$.

Here, l^* and u^* represent the upper and lower bound for the distances of similar and dissimilar pairs, respectively, in the privileged space. The value of l^* is chosen as the upper bound for the $< a^*$ percentile of all $d_{\mathbf{\Lambda}^*}(\mathbf{x}_i^*, \mathbf{x}_j^*)$ values, $1 \leq i < j \leq N$. Similarly, the value of u^* is chosen as the lower bound for the $> 1 - b^*$ percentile of all $d_{\mathbf{\Lambda}^*}(\mathbf{x}_i^*, \mathbf{x}_j^*)$ values, $1 \leq i < j \leq N$. At the same time, the choice of l^* and u^* is subject to the constraint $u^* > l^*$. Also, a^* and b^* are pre-determined with $0 < a^* < b^* < 1$.

In the GMLVQ framework, the privileged information is incorporated by fusing the metric $\mathbf{\Lambda}^*$ in the privileged space $\mathscr{X}^*$ with the metric $\mathbf{\Lambda}$ in the original space $\mathscr{X}$ (for more details, see Fouad et al. (2013)).

3.2.3 Imbalanced class problem

Class imbalance occurs when there is a mismatch between sample sizes representing different classes. Class imbalance is one of the most common issues in classification. Unless explicitly treated, the classifier can be biased towards the majority class. In general, model fitting algorithms of various forms of classifiers assume balanced class distribution. A variety of methods have been proposed to tackle the class imbalance problem [e.g. Garcia et al. (2007)]. For example, the imbalance problem can be addressed by either upsampling the minority class(es) (Perez-Ortiz et al., 2015), or downsampling the majority class(es) (Elrahman and Abraham, 2013), so that the training set becomes balanced.

Since the data sets available for our study are relatively small, instead of upsampling small minority class, we decided to downsample the majority class, and repeat the downsampling $N_d = 100$ times. Training portion of the minority class remains fixed and each time the majority class is downsampled we construct

a classifier based on balanced classes. We thus obtain a collection of N_d classifiers trained on different versions of downsampled majority class. These classifiers are then combined in an ensemble to form a single classifier using majority voting over the ensemble members.

3.2.4 Employing Different Types of PI

We have two different kinds of features extracted from fMRI signals and used as privileged information, namely percent change (PSC) in overall ROI activation and graph based features described above.

The PSC feature quantifies the relative activation difference in the whole ROI when subjects were shown structured vs. random stimuli. This is calculated both from both pre- and post-training fMRI data. We consider 3 ROIs, hence there are 6 PSC privileged information features. Analogously, for the graph-based spatial-temporal features, there is a single feature for each ROI, measured both pre- and post-training, yielding a totality of 6 graph-based privileged information features.

An obvious combination of PSC and graph-based features would be to concatenate them into 12-dimensional vector. However, given the small sample size of participants, such an approach might lead to overfitting. Therefore we constructed an alternative way of combining privileged information features, as outlined below.

We independently construct two classifiers operating in the original space, but trained with the two different kinds of privileged information. Given a test input, if both classifiers predict the same class label, that label is used as the model output. If, on the other hand, they disagree, we output the class label that is predicted with "more confidence" - i.e. smaller distance between the test input and the closest class prototype.

However, note that for the classification purposes, the metric tensor in a single classifier can be arbitrarily scaled, since only the relative relations between distances of test point to the class prototypes are relevant. Hence, in order to compare distances of the test point to the closest prototype in the two classifiers, we need to normalize the learnt metrics. We do this by eigen-decomposing the two metric tensors Λ_1 and Λ_2 and normalizing their eigenvalues to sum to 1. In particular, the eigen-decomposition of Λ_i, $i = 1, 2$, reads $\Lambda_i = \mathbf{U}_i \, \mathbf{diag}(\lambda_1^i, \lambda_2^i, ...\lambda_d^i)\mathbf{U}_i^\intercal$. The normalized metric tensor is obtained as

$$\hat{\Lambda}_i = \mathbf{U}_i \, \mathbf{diag}(\hat{\lambda}_1^i, \hat{\lambda}_2^i, ..., \hat{\lambda}_d^i)\mathbf{U}_i^\intercal, \tag{21}$$

where the normalized eigenvalues are

$$\hat{\lambda}_j^i = \frac{\lambda_j^i}{\sum_{k=1}^d \lambda_k^i}. \tag{22}$$

Given a test input, when combining two ensemble classifiers C_1 and C_2, if they agree on the predicted label, we output that label as the overall label estimate. If, however, C_1 and C_2 disagree on the label, we prefer the label produced with "more certainty" - in our context - small average distance to the closest prototype. In particular, if C_1 is claiming class +1, we calculate the mean distance of the test input to the closest prototype of class +1 across those ensemble members that output class +1 (e.g. their closest prototype to the test input has label +1). Analogously, for C_2 claiming class -1, we record the mean distance of the test input to the closest prototype of class -1 across ensemble members outputting class -1. The overall class label of the combined classifier for the test input is the label with the minimal average distance to the closest prototype.

3.3 Experimental Design

The value of using brain imaging data as privileged information in our setting can be evaluated through two extreme cases:

- No privileged information is available - the models (classifiers) are constructed purely based on the cognitive data. We will refer to this case as M-CD;
- Privileged brain imaging data is always available and is used directly as input data in the classifier construction and testing, without the need to resort to learning with privileged information. We will refer to this case as M-PD. The classifiers obtained in this regime with the PSC, FGF and SGF representations of brain imaging data are referred to as M-PSC, M-FGF and M-SGF, respectively.

When the classifiers are constructed in the framework of learning with privileged information, with cognitive data serving as classifier inputs and brain imaging data used as privileged information, depending on what representation of brain imaging data is used, we denote the resulting classifiers by M^+-CD-PSC, M^+-CD-FGF and M^+-CD-SGF.

As explained above, PSC representation of spatial-temporal structure of cortical activations within an ROI is the simplest one, integrating out both the spatial and temporal structures. In contrast, a more subtle representation is obtained in the graph based features FGF and SGF, integrating over time, but preserving aspects spatial structure. The PSC and graph based features may contain complementary information for the classification task and hence we further combine the classifiers obtained using brain imaging data into composite ones, in particular M^+-CD-PSC and M^+-CD-FGF are combined into a single classifier M^+-CD-PSC+FGF and the combination of M^+-CD-PSC and M^+-CD-SGF is referred to as M^+-CD-PSC+SGF. Analogously, M-PD-PSC and M-FGF are combined to form M-PSC+FGF and combination of M-PSC with M-SGF results in M-PSC+SGF. The overall model structure setup is illustrated in Figure 2.

4 EXPERIMENTS

This section assesses the classification performance of the proposed methodology that incorporates fMRI as privileged information (PD) in the training phase, against baseline algorithms trained without PD, or trained solely with PD. Since we expect that the brain imaging fMRI data carry lot of information regarding possible MCI, the classier trained directly on fMRI (M-PD) will provide a lower bound on the classification error that a classifier trained solely on cognitive data (M-CD) (carrying less information on possible MCI) cannot achieve. We expect that the power of learning with privileged information will boost the classification performance, so that the classifier trained with CD as inputs, but able to incorporate fMRI indirectly in the training process (M^+-CD-PD), will have classification performance between the two extremes M-PD and M-CD, even though in the test phase, both M-CD and M^+-CD-PD classify solely based on CD. The methodology is formulated in the framework of prototype-based classification (GMLVQ) with metric learning (Fouad et al., 2013; Schneider, 2010; Schneider et al., 2009). In this experiment, the original and privileged features correspond to cognitive profiles and brain imaging data, respectively. The overall experimental design is explained in section 3.3.

4.1 Experimental Setup

In the M-PD case, we have in total a set of 34 subjects having both cognitive and brain imaging data, consisting of 9 patients and 25 controls. We create 50 training-test set splits by randomly sampling 6 and 17 patients and controls, respectively, to form the training set (the rest is in the test set). In the M-CD

case we have 60 subjects having cognitive data, consisting 13 patients and 47 controls. Again, we created 50 training-test set splits by randomly sampling 9 and 33 patients and controls, respectively, to form the training set. We made sure that in each resampled training and test set there is an equal balance between subjects with and without PD.

As explained in section Section 3.2.3, to deal with class imbalance in the M-PD case, we construct ensemble classifiers by using the same set of 6 patients and repeatedly sampled 6 controls from the 17 training ones. Analogous setting was used in the M-CD case, this time with 9 patients and 33 controls.

In all experiments, the (hyper-)parameters of the ensemble classifiers were tuned via cross-validation on the training set of the first sub-split only. The found values were then fixed across the remaining 99 classifiers. In the GMLVQ classifier, data classes are represented by one prototype per class. The class prototypes are initialized as means of random subsets of training samples selected from the corresponding class. In the IT metric learning settings given in Fouad et al. (2013), lower (a, a^*) and upper (b, b^*) percentile bounds for the privileged and original spaces were tuned over the values of $5, 10, 15$ and of $85, 90, 95$, respectively.

Throughout the experiments we had one data set in the original space of CD. However, experiments were repeated for three different fMRI PD: PSC, SGF and FGF. PD of each subject is represented by 6 features, 3 pre-training and 3 post-training, corresponding to 3 ROIs. Due to the imbalanced nature of our classes we utilized the following below evaluation measures:

1. Confusion Matrix: it is a popular performance indicator for machine learning algorithms. It is organized along the the actual classes (rows) and the predicted ones columns) (Elrahman and Abraham, 2013). In this study positive and negative examples represent patients and controls, respectively. In the confusion matrix, True Positive (TP) denotes the number of positive examples correctly classified, True Negatives (TN) is the number of negative examples correctly classified , False Positives (FP) is the number of negative examples incorrectly classified, False Negatives (FN) is the number of positive examples incorrectly classified as negative. The true positive rate ($TPR = \frac{TP}{TP+FN}$) measures the percentage of patients who are correctly classified, whereas the true negative rate ($TNR = \frac{TN}{TN+FP}$) measures the proportion of the correctly identified controls. False positive rate ($FPR = \frac{FP}{FP+TN}$), refers to the probability of falsely classifying the patients, whereas the false negative rate ($FNR = \frac{FN}{FN+TP}$) refers to the probability of falsely classifying the controls.

2. Macroaveraged Mean Absolute Error (MMAE): it is a macroaveraged version of Mean Absolute Error and it is a weighted sum of the classification errors across classes (Fouad, 2013). It measures the per-class accuracy of class predictions $\hat{y}$ with respect to true class y on a test set:

$$MMAE = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{y_i=N}|y_i - \hat{y}_i|}{T_n} \tag{23}$$

Where N is the number of classes and T_n is the number of test points whose true class is n.

4.2 Classification Results

We are primarily interested in classification performance of M^+-CD-PD classifiers, that is, classifiers using cognitive data as their inputs and incorporating brain imaging data as privileged information. this classification performance will be put in the context of performances when no brain imaging information is available (M-CD) and when the full brain imaging is available as input (M-PD). This will allow us

to quantitatively investigate how much performance improvement over M-CD could be obtained by incorporating privileged information through metric learning. Following our experimental setup, we obtained 50 MMAE estimates for each classifier summarised by the mean, standard deviation, median and the (25%, 75%) percentiles. The results are summarised in Tables 1 and 2.

Table 1 shows that for all five types of PD, M-PD outperforms M-CD. Recall that we have extracted three different features from the brain imaging data, namely PSC, SGF, and FGF, and all of them can be used as PD. For PSC, which is related to brain activation level, the corresponding median MMAE is reduced by relatively 39.6% when compared to that of M-CD. The other two types of PD, SGF and FGF, are related to brain connectivity pattern. When compared to the baseline classifier, the relative reduction of their median MMAE is about 24% and 40%, respectively. The above results indicate that PSC is at least as useful as the graph feature (FGF), or even more useful (SGF). Im principle, the activation level and connectivity pattern are two independent fMRI features. Therefore, PSC could be used as PD along with SGF or FGF. Row 6–7 in Table 1 show that the resulting classifier can either attain the classification performance of M-PSC in the case of SGF, or improve on it in the case of FGF. In summary, brain imaging data contain more information that are relevant to the task than cognitive data.

Models	Mean	Std-Dev	Median	(25%, 75%) Percentile	p-value
M-CD	0.3992	0.0949	0.3942	(0.3173, 0.4423)	–
M-PSC	0.2357	0.1655	0.2381	(0.1429, 0.3333)	0.00
M-SGF	0.2666	0.1151	0.2995	(0.2143, 0.4048)	0.00
M-FGF	0.2376	0.1231	0.2381	(0.1429, 0.3333)	0.00
M-PSC+SGF	0.2438	0.1067	0.2381	(0.2143, 0.3095)	0.00
M-PSC+FGF	0.2200	0.1245	0.2143	(0.1429, 0.3095)	0.00

Table 1. Classification performance measured by Macroaveraged Mean Absolute Error (MMAE) for the baseline classifier, M-CD, and five different M-PD classifiers (see Column 1). For each classifier, we report both mean MMAE, its standard deviation, median MMAE and its (25%, 75%) percentile in Column 2 – 5, respectively. They were computed using the MMAE estimates obtained from 50 randomly created training-test splits. Each p-value displayed in Column 6 was obtained from one-sided sign-rank test against the null hypothesis that the corresponding M-PD classifier is inferior to the baseline classifier.

Table 2 shows that for all five types of PD, M^+-CD-PD outperforms M-CD. In particular, PSC and SGF are the best two among the five PD types that are used as the privileged information along with CD as GMLVQ's inputs. Compared to M-CD, both M^+-CD-PSC + M^+-CD-SGF show a reduction of their median MMAE by relatively 20%. This relative improvement is shrunk to 13.4%, 9.7%, and 3.4% for M^+-PSC+FGF, for M^+-PSC+SGF, and for M^+-FGF (respectively).

Models	Mean	Std-Dev	Median	(25%, 75%) Percentile	p-value
M^+-CD-PSC	0.3448	0.0988	0.3173	(0.2788, 0.4038)	0.00
M^+-CD-SGF	0.3128	0.0804	0.3153	(0.2308, 0.3942)	0.56
M^+-CD-FGF	0.3925	0.1211	0.3810	(0.2885, 0.4135)	0.12
M^+-CD-PSC+SGF	0.3426	0.1116	0.3558	(0.2788, 0.4038)	0.00
M^+-CD-PSC+FGF	0.3553	0.1157	0.3413	(0.2788, 0.4808)	0.04

Table 2. The same as in Table 1 but for evaluation of the classification performance of five different M^+-CD-PD classifiers, that is, the classifiers using CD as their inputs and PD as privileged information.

Table 3 presents the results of average TPR and TNR of the models. The best two TPR results (0.83 and 0.80) were achieved by M-SGF and M-PSC-SGF (respectively), whereas the best two TNR result (0.88 and 0.87) were attained by M-PSC and M^+-CD-FGF (respectively). Overall, M-FGF emerges as the classifier with most balance performance.

Model	TPR	TNR
M-CD	0.60	0.60
M-PSC	0.64	0.88
M^+-CD-PSC	0.69	0.63
M-SGF	0.83	0.51
M^+-CD-SGF	0.53	0.67
M-PSC+SGF	0.80	0.68
M^+-CD-PSC+SGF	0.56	0.70
M-FGF	0.74	0.76
M^+-CD-FGF	0.38	0.87
M-PSC+FGF	0.56	0.70
M^+-CD-PSC+FGF	0.61	0.70

Table 3. Overall true positive rates (TPR) and true negative rates (TNR) on hold-out sets

4.3 Further Analysis

GMLVQ is a fully adaptive algorithm to learn global metric tensor which accounts for different importance weighting of individual features and pairwise interplay between the features, with respect to the given classification task. Hence, it allows us to study the task-dependent relevance of the input features by using the diagonal elements of the GMLVQ metric tensor matrix. Moreover, the global metric can be further optimized adaptively by incorporating privileged information into the GMLVQ model via the distance relations revealed in the privileged space (Fouad, 2013). In the following we analyse the learned classification models in terms of the learned metric tensor and discuss possible implications regarding the cognitive and brain imaging fMRI features used in this study.

4.3.1 Cognitive features only

We first present a procedure to study the relevance of four cognitive features (working memory, cognitive inhibition, divided attention, and selective attention) using the GMLVQ metric (tensor) matrices obtained from the experiments whose classification results are discussed in Section 4.2. Each of these experiments resulted in 50×100 GMLVQ classifiers with the associated metric (tensor) matrices Λ obtained by training GMLVQ classifiers on 50×100 (small) data sets independently. Recall that these data sets were generated by first randomly splitting the whole training set into 50 smaller sets of equal size and then randomly downsampling the majority class to the size of the minority class in each split 100 times. However, many of the 50×100 classifiers performed poorly and they should not be included in the analysis of the relevant cognitive features. We therefore discard the data split producing the ensemble classfier whose N_b-th best ensemble member (classifier) produced error larger than a threshold value denoted by E_{max}, and pool all ensemble members from each of the remaining splits for further analysis. This procedure is applied to three experiments as follows: M-CD, M^+-CD-PSC and M^+-CD-FGF. We found out that N_b = 15 and $E_{max} = 25\%$ worked universally across these data sets.

Each of the four cognitive features is associated with one of the four diagonal element in the metric (tensor) matrix. For each cognitive feature, its importance is measured by the frequency of its associated

diagonal elements in $> 90\%$ percentile of the set of all diagonal elements from the metric (tensor) matrices selected by the above procedure. The left panel in Figure 3 shows that the divided attention (i.e. t^d_{disp}) is the most discriminative feature for the classification task (MCI patients vs. healthy controls).

Next, we studied the off-diagonal elements of those metric (tensor) matrices. Each off-diagonal element controls the interplay between two associated cognitive features. To illustrate how this interplay works, we provide a toy example as follows: Denote a two-dimensional feature vector by (x, y) and a 2×2 metric tensor by $\begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}$. The distance between two feature vectors indexed by i and j is given by

$$d_{ij} = \underbrace{\alpha^2 \cdot (x_i - x_j)^2 + \beta^2 \cdot (y_i - y_j)^2}_{d^M_{ij}} + \underbrace{2\gamma \cdot (x_i - x_j)(y_i - y_j)}_{d^2_{ij}}. \tag{24}$$

The first two terms of d_{ij} is actually so-called Mahalanobis distance between the i-th and j-th feature vectors (denoted by d^M_{ij}). In the case of $\gamma = 0$, the diagonal term α and β are optimized by maximizing between-class Mahalanobis distances while minimizing within-class ones. When the metric matrix has non-zero off-diagonal elements, the distance measure has additional contribution d^2_{ij} which can either enhance or collapse the total distance measure depending on *(i)* the sign of γ and *(ii)* the sign of *between-class* correlation (i.e. correlation between class-conditional means of x and y). For example, in the case of negative *between-class* correlation, negative γ can further enhance the class separation and vice versa.

To test whether the interplay between two cognitive features, indexed by i and j, is positive or negative, we performed two one-sided sign-rank tests for the hypotheses $\Lambda_{ij} > 0$ and $\Lambda_{ij} < 0$ (respectively) using the corresponding off-diagonal element from the selected GMLVQ metric (tensor) matrices. The upper-left panel of Figure 4 shows that there exists statistically significant, negative interplay between divided attention and two following cognitive features: (1) working memory (n_{dots}) and (2) cognitive inhibition (t_{delay}). From the lower-left panel, we found statistically significant, positive interplay between three cognitive features as follows: (1) working memory, (2) cognitive inhibition, and (3) selective attention (t^s_{disp}). Finally, note that there is no significant interplay between divided attention and selective attention.

To examine the relation between the interplay and *between-class* correlation revealed by Eq. 24, we need to determine whether or not there exists statistically significant *between-class* correlation between two of the four cognitive features. To this end, we first used one-sided sign-rank test to determine, for each of the four features, whether its values for MCI patients are significantly larger or significantly smaller than those for healthy controls. For each pair of the cognitive features, if the outcomes of their tests are both statistically significant and are consistent with (or in opposite to) each other, then their *between-class correlation* is considered as positive (or negative). Otherwise, the *between-class* correlation is insignificant. From this analysis we observe (1) the class-conditional mean of working memory is positively correlated with that of cognitive inhibition; and (2) the class-conditional mean of divided attention is negatively correlated with that of working memory as well as that of cognitive inhibition. These observations agree with the observation of the interplay between the corresponding cognitive features, which can enhance the class separation. For the remaining pairs of the cognitive features, their *between-class correlation* is not significant. In Figure 5, we graphically illustrate the presence or absence of these correlations.

In summary, though the divided attention is the most relevant feature among the four cognitive features, all four features are indispensable for maximising the classification performance. This is because these exists *between-class correlation* between the features.

4.3.2 fMRI features

We carried out the same relevance analysis for M-PSC, M-SGF, and M-FGF as for M-CD in Section 4.3.1. Recall that in these three experiments, the inputs to GMLVQ classifiers are comprised of six fMRI features as follows: **(i)** PSC-Cerebellar-Pre, PSC-Cerebellar-Post, PSC-Frontal-Pre, PSC-Frontal-Post, PSC-Subcortical-Pre, PSC-Subcortical-Post; **(ii)** SGF-Cerebellar-Pre, SGF-Cerebellar-Post, SGF-Frontal-Pre, SGF-Frontal-Post, SGF-Subcortical-Pre, SGF-Subcortical-Post; and **(iii)** FGF-Cerebellar-Pre, FGF-Cerebellar-Post, FGF-Frontal-Pre, FGF-Frontal-Post, FGF-Subcortical-Pre, FGF-Subcortical-Post (respectively). The fMRI feature "PSC-Cerebellar-Pre" denotes PSC feature that is derived from fMRI data measured in the cerebellar ROI and during the pre-training session. and the remaining fMRI features are abbreviated in the same way. Recall that PSC is referred to as Percent Signal Change, SGF as Spatially grouped Graph Feature and FGF as Functionally grouped Graph Feature.

Figure 6 shows that PSC-Frontal-Post and FGF-Frontal-Pre are the most discriminative fMRI feature in Experiment M-PSC and M-FGF (respectively). We first note that the most relevant feature in both cases is derived from the frontal ROI (that is, the largest ROI among the three ROIs used in this study). It is more interesting to address two following questions: (1) why is the post-training session is more relevant than the pre-training one, when PSC is used for the task; and (2) why is the opposite true when the graph feature is used for the task.

The left panel in Figure 7 shows that before training, the PSC level for MCI patients and healthy controls are on average comparable. However, training caused a remarkable increase of the PSC level for MCI patients but not for healthy controls. As a result, these two participant groups differ in their PSC level after the training. This is why PSC-Frontal-Post is identified as the most relevant feature for Experiment M-PSC. The right panel in Figure 7 shows that the graph feature FGF differs between MCI patients and healthy controls before training. This could be related to the suggestions that MCI may have caused changes in brain connectivity. We further observe that for both participant groups, training increased their FGF values but to different extents. After training, the difference between MCI patients and healthy controls became much less significant. This is why FGF-Frontal-Pre is identified as the most relevant feature for Experiment M-FGF. This observation allows us to speculate that training could "mitigate" the changes in brain connectivity caused by MCI.

The above analysis suggests that brain connectivity may have changed after training and this is significant particularly for MCI patients. In the following, we address the question whether a sub-network rather than the entire (local) network within the frontal ROI has changed. Recall that all 128 voxels in the frontal ROI are grouped into 7 spatially contiguous clusters. This results in a local brain network consisting of 7 nodes and 21 edges. Each off-diagonal element of the graph matrix G quantifies the connectivity between two nodes and measures the strength of the corresponding edge. Recall that the graph features FGF were extracted by applying 2D-LDA. To this end, 2D-LDA provides two feature-generating vectors $\mathbf{a}$ and $\mathbf{b}$ from which we can derive a task-dependent importance matrix denoted by I as follows:

$$I = \frac{1}{2}(\mathbf{a}\mathbf{b}^{\mathsf{T}} + \mathbf{b}\mathbf{a}^{\mathsf{T}}). \tag{25}$$

Each off-diagonal element of I measures the importance of the corresponding edge in terms of discriminating MCI patients from healthy controls. To identify possible sub-networks that have significantly changed after training, we are first to identify the edges whose importance measure has significantly changed after training. To this end, we generated an ensemble of the selected importance matrices using the procedure that was used to generate an ensemble of the selected GMLVQ metric (tensor) matrices

for the relevance feature analysis. Subsequently, we conducted two one-sided sign rank tests for each of the 21 edges to find those edges whose importance values have significantly increased or reduced after training. Denote the edge connecting node i and j by E_{ij}. This analysis revealed that the importance measure of three following edges has significantly increased: E_{17}, E_{16} and E_{64}. A significant reduction of its importance measure was observed for E_{65}. Figure 9 highlighted a subtle difference between the sub network (i.e. E_{17}, E_{16} and E_{64}) and the single edge E_{65}. For the three-node sub-network, the connectivity strength is highest for MCI patients before training. For the single edge E_{65}, the connectivity strength is lowest for healthy controls before training. This suggests that FGF-Frontal-Pre, the most relevant feature in M-FGF, could be related to these three-node and single-node sub-networks.

4.3.3 Privileged information

In addition to M-CD, M-PSC and M-FGF, M^+-CD-PSC and M^+-CD-FGF were conducted to investigate GMLVQ classification of MCI patients and controls when fMRI features were incorporated as privileged information. The relevance of the four cognitive features in M^+-CD-PSC and M^+-CD-FGF was estimated from the diagonal elements of the metric tensors and displayed in the middle and right panel of Figure 3 (respectively). Though PSC and FGF are two different kinds of fMRI features, we still consistently observed that cognitive inhibition and divided attention are the two most relevant cognitive features. Moreover, the relevance of divided attention is more profound than that of cognitive inhibition. When compared to M-CD, cognitive inhibition did emerge as a relevant feature only when the privileged information was incorporated. Also, Figure 4 shows that when compared to M-CD, the interplay between divided attention and selective attention became significantly positive in M^+-CD-PSC and M^+-CD-FGF, that is, the experiments in which the privileged information was incorporated.

5 CONCLUSION

In this study, we employed GMLVQ classifiers to discriminate cognitive skills in MCI patients vs. healthy controls using cognitive and/or fMRI data. Specially, we have adopted a "Learning with privileged information (PI)" approach to combine cognitive and fMRI data. In this setting, fMRI data as an addition to cognitive data are only used to train GMLVQ classifier and classification of a new participant is solely based on cognitive data. As the inputs to GMLVQ classifier, the cognitive features include working memory, cognitive inhibition, divided attention and selective attention scores. Also, we extracted three different types of fMRI features from fMRI data as follows: PSC (percent signal change), and SGF (spatially grouped graph feature) and (functionally grouped graph feature).

We first tested our baseline GMLVQ classifier with four cognitive features as inputs. Its classification performance is measured by (25%, 75%) percentile of Macro-averaged Mean Absolute Error (MMAE), that is, (0.32, 0.44). The best of the five fMRI GMLVQ classifiers (i.e. the ones using the fMRI features as their inputs) yields a lower bound of classification error, which is (0.14, 0.31). Interestingly, the best of the PI-guided GMLVQ classifiers (i.e. the ones using the four cognitive features as their inputs and using the fMRI features as privileged information) have achieved (0.23. 0.39). This implies that incorporating fMRI features as privileged information can significantly improve the classification performance of a baseline GMLVQ classifier for classification of cognitive skills in MCI patients vs. controls.

Crucially, we have also performed "relevant feature analysis" for all three GMLVQ classifiers: the baseline GMLVQ classifier, the best fMRI-guided GMLVQ classifier, and the fMRI GMLVQ classifier. For the baseline classifier, "divided attention" is the only relevant cognitive feature for the classification task. When the privileged information is incorporated, divided attention remains the most relevant feature

while cognitive inhibition becomes also relevant. The above results suggest that attention-rather than only memory-plays an important role for the classification task. More interestingly, this analysis for the fMRI GMLVQ classifier suggests that (1) among three ROIs used, the frontal ROI is most relevant for the classification task; (2) when the PSC feature as an overall measure of fMRI response to structured stimuli is used as the inputs to the classifier, the post-training session is most relevant; and (3) when the graph feature reflecting underlying spatiotemporal fMRI pattern is used, the pre-training session is most relevant. Further analysis has indicated that training may cause an overall increase of the brain activity only for MCI patients while it may have "mitigated" the difference in brain connectivity pattern between MCI patients and healthy controls. Moreover, these training-dependent changes are most significant for a three-node sub-network in the frontal ROI. Taken together these results suggest that brain connectivity before training and overall fMRI signal after training are both diagnostic of cognitive skills in MCI

Our study employs machine learning algorithms to investigate the neurocognitive factors and their interactions that mediate learning ability in Mild Cognitive Impairment. Our work is not limited to developing and validating machine learning approaches; in contrast it advances our understanding of the neurocognitive mechanisms that mediate learning in health and disease. For example, the role of cognitive inhibition in cognitive profile classification seems to be significantly enhanced when brain imaging information (obtained in a sequence learning prediction task) is provided as privileged information. This opens questions about the possible interplay between circuits involved in cognitive inhibition and those involved in learning sequence prediction tasks. We also observed significant positive interplay between divided and selective attention when brain imaging data is used as privileged information. No such interplay was detected without the privileged information. Again, this raises interesting questions regarding circuitry involved in sequence prediction and the two attention types.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

- Collection of cognitive and fMRI data: ZK, CL;
- Diagnosis of MCI: PB;
- Determination of ROIs: CL;
- Design of the work: PT, ZK, YS;
- Analysis and interpretation: HA, YS, PT, SF;
- Drafting the article: HA, YS, PT, SF, ZK, CL;
- Critical revision of the article: YS, PT, ZK;
- Final approval of the version to be published: HA, YS, SF, CL, PB, ZK, PT.

ACKNOWLEDGEMENTS

PT and YS were supported by EPSRC grant no EP/L000296/1 "Personalised Medicine through Learning in the Model Space".

We thank Joseph Giorgio for his careful reading of the manuscript, his insightful comments and suggestions.

REFERENCES

Albert, S., Dekosky, S., Dickson, D., Dubois, B., Feldman, H., Fox, N., et al. (2010). The diagnosis of mild cognitive impairment due to Alzheimer's disease. *Alzheimers Dement* 7, 270–279

Arbib, A. (2003). *The Handbook of Brain Theory and Neural Networks* (MIT Press), 2nd edn.

Baker, R., Bentham, P., and Kourtzi, Z. (2015). Learning to predict is spared in mild cognitive impairment due to alzheimer's disease. *Exp Brain Res* 233, 2859–2867

Carpineto, C. and Romano, G. (2012). Consensus clustering based on a new probabilistic rand index with application to subtopic retrieval. *IEEE Trans Pattern Anal* 34, 15–26

Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., and Cercignani, M. (2015). Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage* 112, 232–243

Chen, B., Liu, M., Zhang, D., and Shen, D. (2015). Domain transfer learning for mci conversion prediction. *IEEE Transaction on Biomedical Engineering* 62, 232–243

Davis, J., Kulis, B., Jain, P., Sra, S., and Dhillon, I. (2007). Information-theoretic metric learning. *in Proceedings of the 24th International Conference on Machine Learning, ser. ICML 07. New York, NY, USA: ACM* , 209–216

Elrahman, S. A. and Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing* 1, 332–340

Fouad, S. (2013). *Metric Learning for Incorporating Privileged Information in Prototype-based Models*. Thesis of the degree of doctor of philosophy, School of Computer Science

Fouad, S., Tino, P., Raychaudhury, S., and Schneider, P. (2013). Incorporating privileged information through metric learning. *IEEE Transactions on Neural Networks and Learning System* 24, 1086–1098

Garcia, V., Sanchez, J., Mollineda, R., Alejo, R., and Sotoca, J. (2007). The class imbalance problem in patteren classification and learning

Luft, C., Baker, R., Bentham, P., and Kourtzi, Z. (2016). Learning temporal statistics for sensory predictions in mild cognitive impairment. *Neuropsychologia* 75, 368–380

Luft, C., Baker, R., Goldstone, A., Zhang, Y., and Kourtzi, Z. (2015). Learning temporal statistics for sensory predictions in aging. *Journal of Cognitive Neuroscience* 28, 1–15

Perez-Ortiz, M., Gutierrez, P., Tino, P., and Hervas-Martinez, C. (2015). Over-sampling the minority class in the feature space. *IEEE Transaction on Neural Networksand Learning System*

Sato, J., Thomaz, C., Cardoso, E., Fujita, A., Martin, M., and Amaro, E. (2008). Hyperplane navigation: A method to set individual scores in fmri group datasets. *Neuroimage* 42, 1473–1480

Schneider, P. (2010). *Advanced methods for prototype-based classification*. PhD Dissertation, University of Groningen

Schneider, P., Biehl, M., and Hammer, B. (2009). Adaptive relevance matrices in learning vector quantization. *Neural Computation* 21, 3532–3561

Ye, J., Janardan, R., and Li, Q. (2004). Two-dimensional linear discriminant analysis. *Advances in Neural Information Processing Systems* 17, 1569–1576

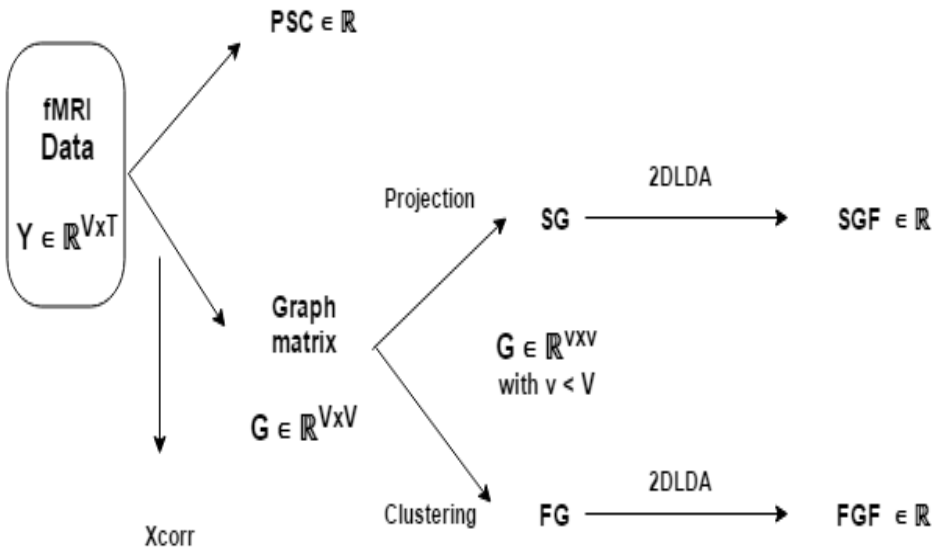


Figure 1. Illustration of fMRI feature generation pipeline: from BOLD signal data $\mathbf{Y}$ to three fMRI features (PSC, FGF, and SGF). $F\mathbf{G}$ and $S\mathbf{G}$ are the reduced version of graph matrix $\mathbf{G}$ via functional grouping and spatial grouping (respectively). Note that FGF and SGF are both discriminative features extracted from $F\mathbf{G}$ and $S\mathbf{G}$ in a supervised manner using 2D-LDA (that is, Linear Discriminant Analysis operating on matrices).

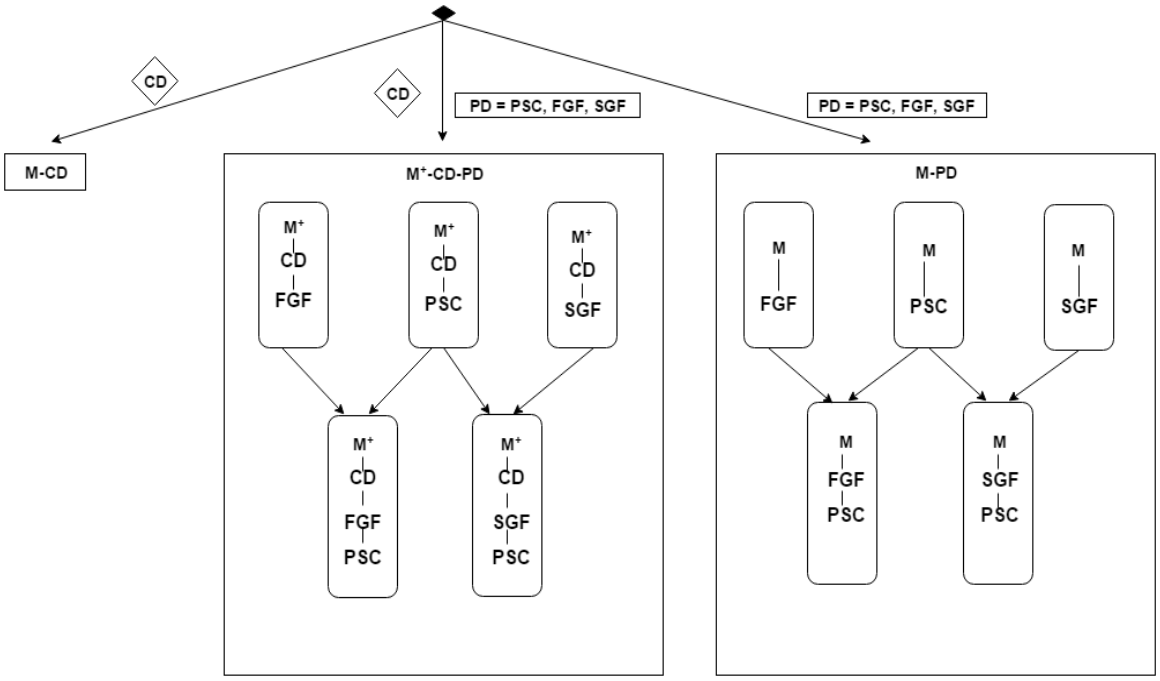


Figure 2. Schematic illustration of the experimental design described in Section3.3. The items in diamond shape denote data: CD for cognitive data, PD for privileged information data, PSC for Percent Signal Change, FGF for functionally grouped graph feature, and SGF for spatially grouped graph feature. M-XXX denotes a GMLVQ classifier that does not use privileged information while XXX denotes the inputs to this classifier. For example, M-PSC means a GMLVQ classifier with PSC features as its inputs. M^+-XXX-YYY denotes a GMLVQ classifier using feature XXX as its inputs and feature YYY as privileged information. For example, M^+-CD-PSC means a GMLVQ classifier using cognitive features as its inputs and PSC features as privileged information. M^+-XXX-YYY-ZZZ denotes a hybrid classifier that combines the classification output of classifier M^+-XXX-YYY and classifier M^+-XXX-ZZZ using a certain rule (e.g. majority voting rule).

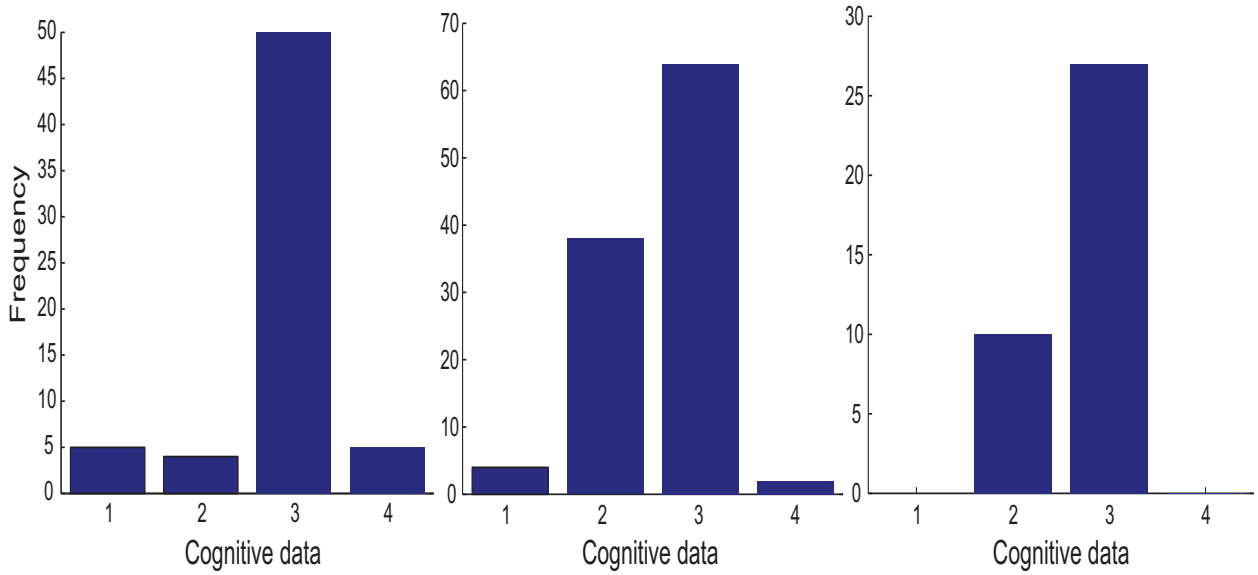


Figure 3. The importance histogram of the four cognitive features as follows: working memory (n_{dots}), cognitive inhibition (t_{delay}), divided attention (t^d_{disp}), and selective attention (t^s_{disp}) (numbered as 1, 2, 3, and 4 in the order). These features are used as the input to the following GMLVQ classifiers: M-CD, M^+-CD-PSC, and M^+-CD-FGF (from left to right). Note that each cognitive feature is associated with a diagonal element of the GMLVQ metric tensor matrix Λ and the importance histogram counts the number of each diagonal element in the $>90\%$ percentile of all diagonal elements from an ensemble of Λs.

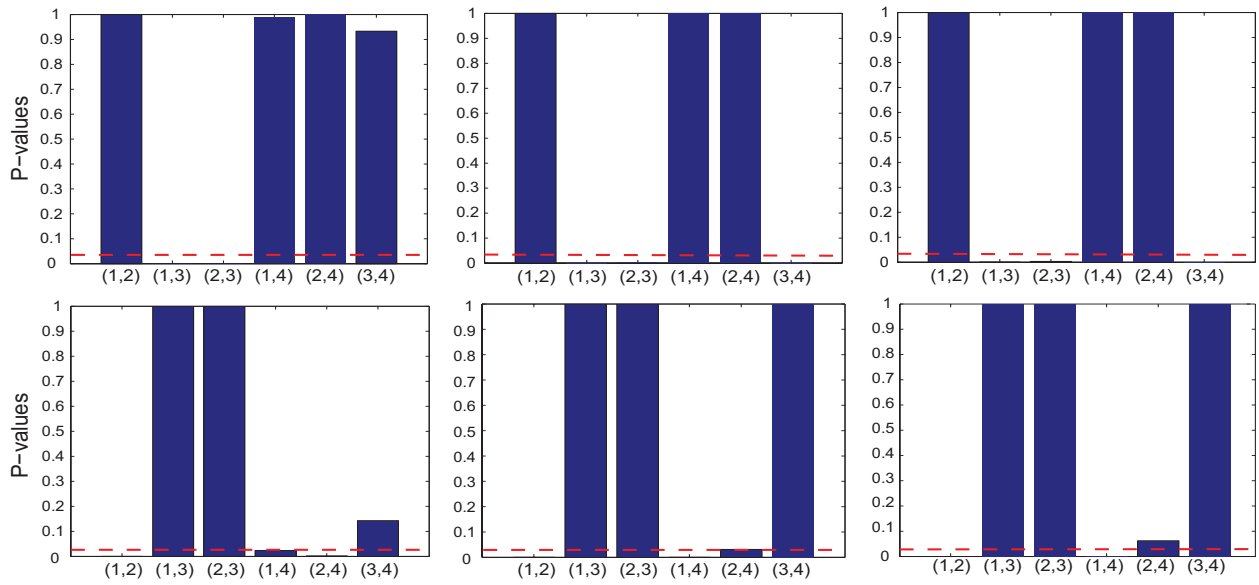


Figure 4. The p values of the one-sided sign-rank tests for studying the interplay between two of the following cognitive features: working memory (n_{dots}), cognitive inhibition (t_{delay}), divided attention (t^d_{disp}), and selective attention (t^s_{disp}) (numbered as 1, 2, 3, and 4 in the order). From each panel in the upper and lower row, one can read that if the p value is smaller than the threshold $p = 0.05$ (indicated by red dashed line), the interplay of two corresponding cognitive features is statistically significant and it takes a negative and positive value (respectively); These features are the inputs to three GMLVQ classifiers as follows: M-CD, M^+-CD-PSC, and M^+-CD-FGF (from left to right). Note that the tests used the off-diagonal elements of the GMLVQ metric tensor matrices.

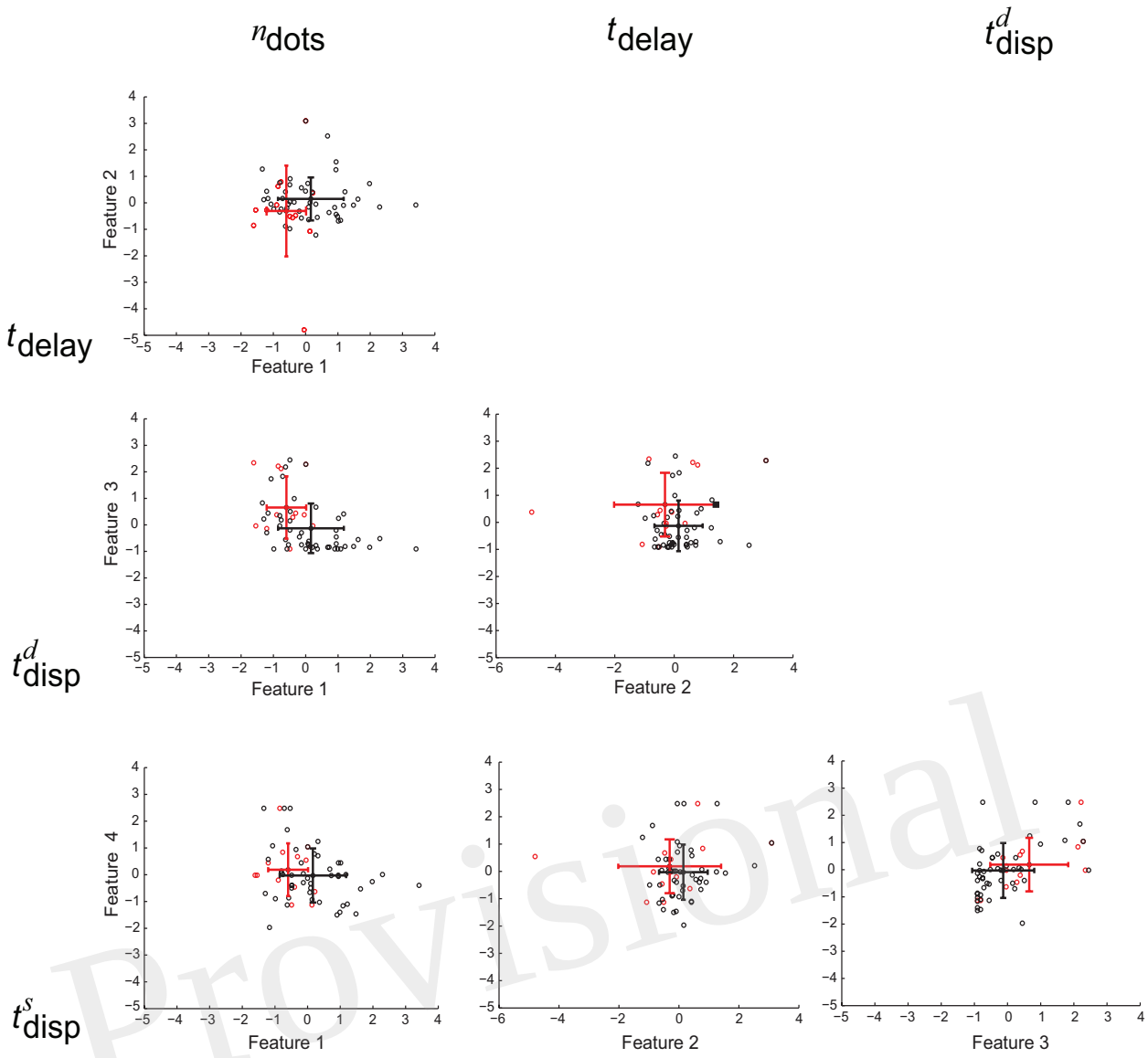


Figure 5. Scatter plot for six possible feature pairs from the four cognitive features as follows: working memory (n_{dots}), cognitive inhibition (t_{delay}), divided attention (t^d_{disp}), and selective attention (t^s_{disp}). For individual MCI patients and healthy controls, their feature pairs (i.e. Feature 1 vs Feature 2) are displayed as red and blue dots (respectively). The corresponding class-conditional means and standard deviations are also displayed by coloured error bars. For each panel, the corresponding Feature 1 and Feature 2 are indicated at the top of each column and on the utmost left of each row (respectively).

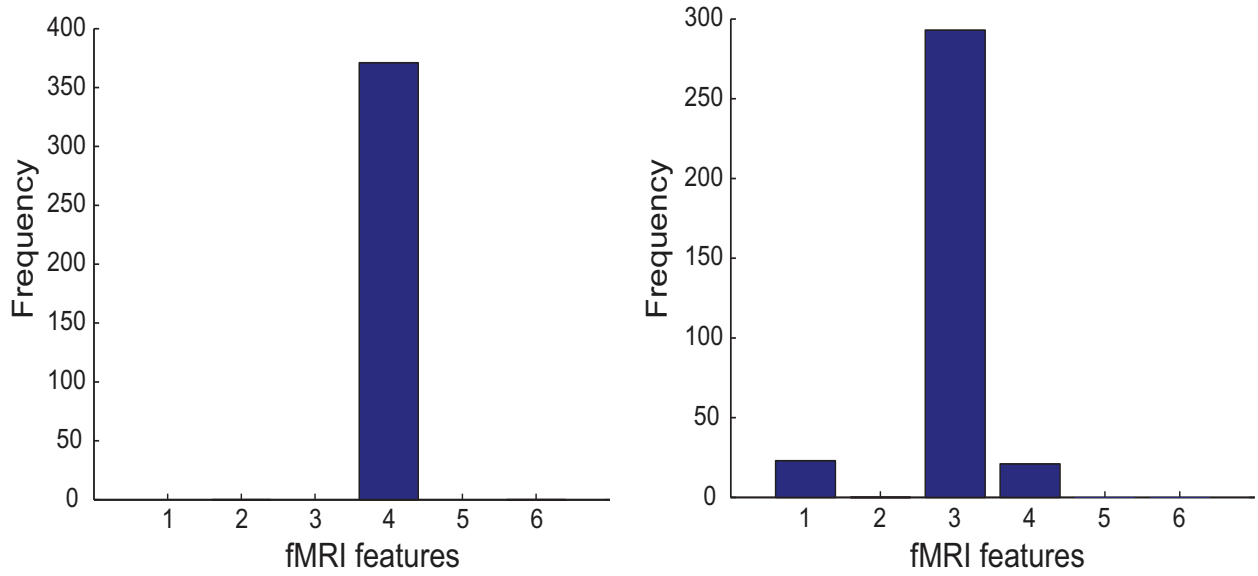


Figure 6. Left panel: The importance histogram of the six fMRI features as follows: PSC-Cerebellar-Pre, PSC-Cerebellar-Post, PSC-Frontal-Pre, PSC-Frontal-Post, PSC-Subcortical-Pre, and PSC-Subcortical-Post. (numbered as 1, ..., and 6 in the order). PSC is referred to as Percent Signal Change, Pre as Pre-training session, Post as Post-training session, Cerebellar (Frontal and Subcortical) as the cerebellar(frontal and subcortical, respectively) ROI. For example, PSC-Cerebellar-Pre means that the fMRI data were acquired before training and PSC feature was extracted from the cerebellar ROI). Right panel: The same as in the left panel but for the following fMRI features: FGF-Cerebellar-Pre, FGF-Cerebellar-Post, FGF-Frontal-Pre, FGF-Frontal-Post, FGF-Subcortical-Pre, and FGF-Subcortical-Post.

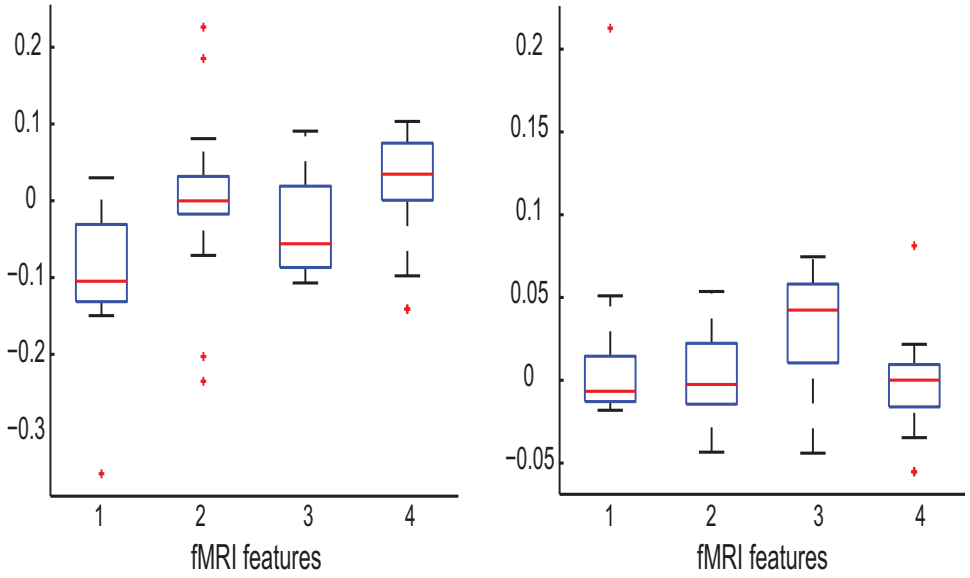


Figure 7. Left: Boxplot of the following fMRI features: FGF-Frontal-Pre for MCI patients, FGF-Frontal-Pre for healthy controls, FGF-Frontal-Post for MCI patients, and FGF-Frontal-Post for healthy controls (numbered as 1, 2, 3 and 4 in the order). Note that the y-axis represents the values of the corresponding fMRI features; Right: Boxplot of the following fMRI features: PSC-Frontal-Pre for MCI patients, PSC-Frontal-Pre for healthy controls, PSC-Frontal-Post for MCI patients, and PSC-Frontal-Post for healthy controls (numbered as 1, 2, 3 and 4 in the order).

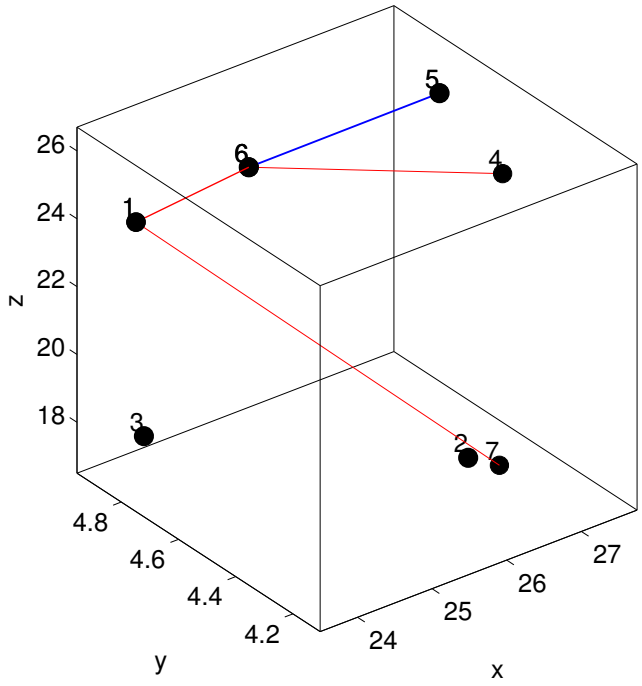


Figure 8. The node configuration for the frontal ROI which includes Superior Frontal Gyrus on the right hemisphere and Medial Frontal Gyrus on the left hemisphere. The straight lines indicate the edges whose importance for discriminating MCI patients from healthy controls has significantly changed. For the three-node subnetwork (indicated by red lines), its importance has increased after training. In contrast, the single-node subnetwork (indicated by blue line), training has reduced its importance.

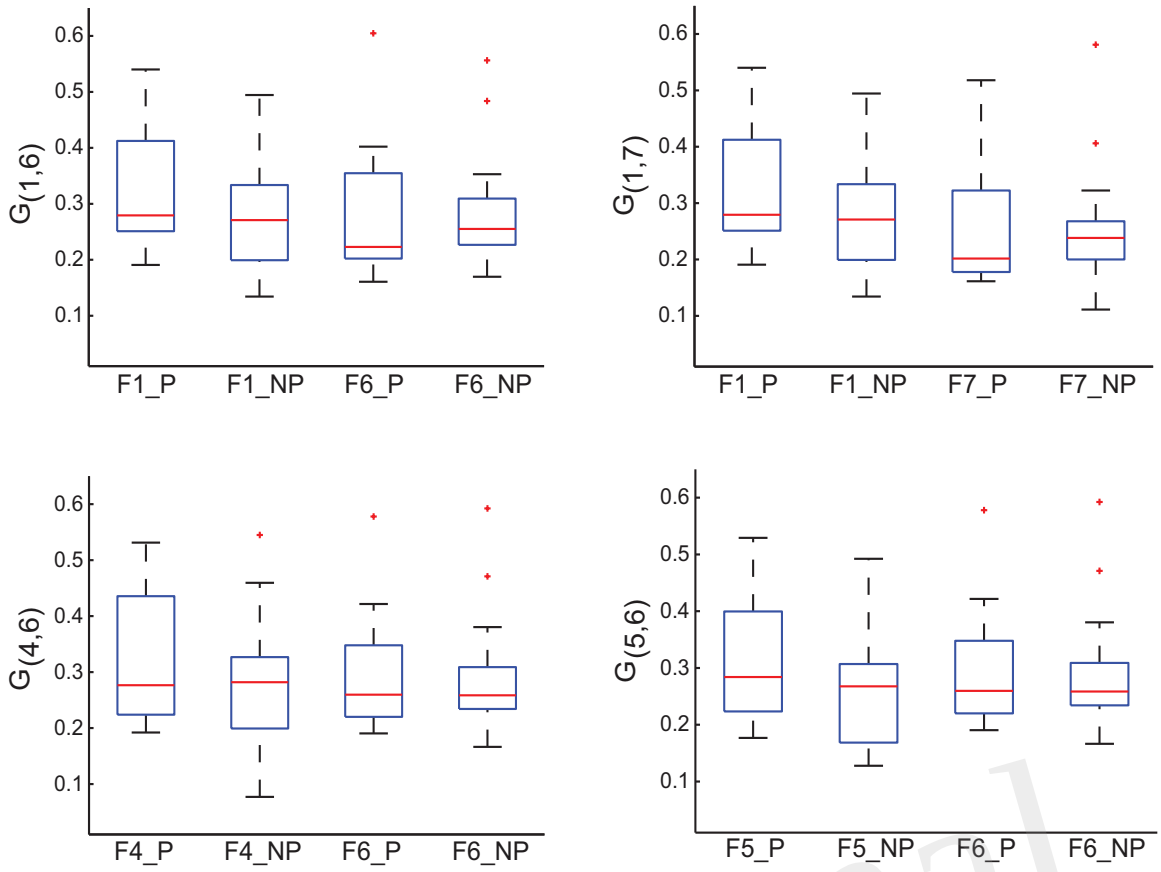


Figure 9. For the graph matrices generated in this study, we display four of their matrix elements which are associated with the four edges highlighted in Figure 8. $G_{1,6}$ in the upper-left panel, $G_{1,7}$ in the upper-right panel, and $G_{4,5}$ in the lower-left panel measure the connectivity of edge $E_{1.6}$, $E_{1,7}$ and $E_{4,5}$ (respectively) that form the three-node sub-network. Recall that the task-related importance of this sub-network has significantly increased after training. In contrast, $G_{5,6}$ in the lower-right panel measures the connectivity of edge $E_{5.6}$ and its task-related importance has significantly reduced after training. The four boxplots in each panel are associated with pre-training session & patient group, pre-training session & control group, post-training session & patient group, and and post-training session & control group (from left to right, numbered as 1, 2, 3, and 4 in the order).

Figure 01.JPEG

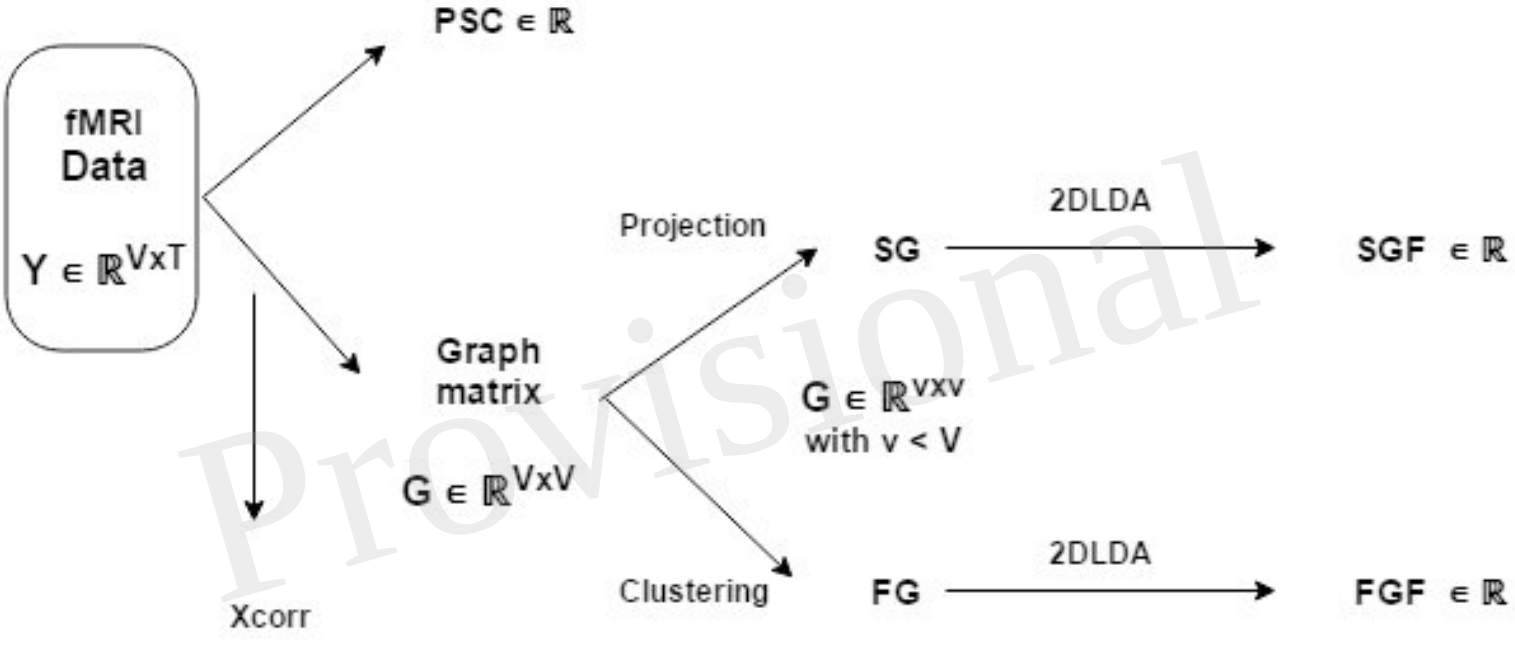


Figure 02.JPEG

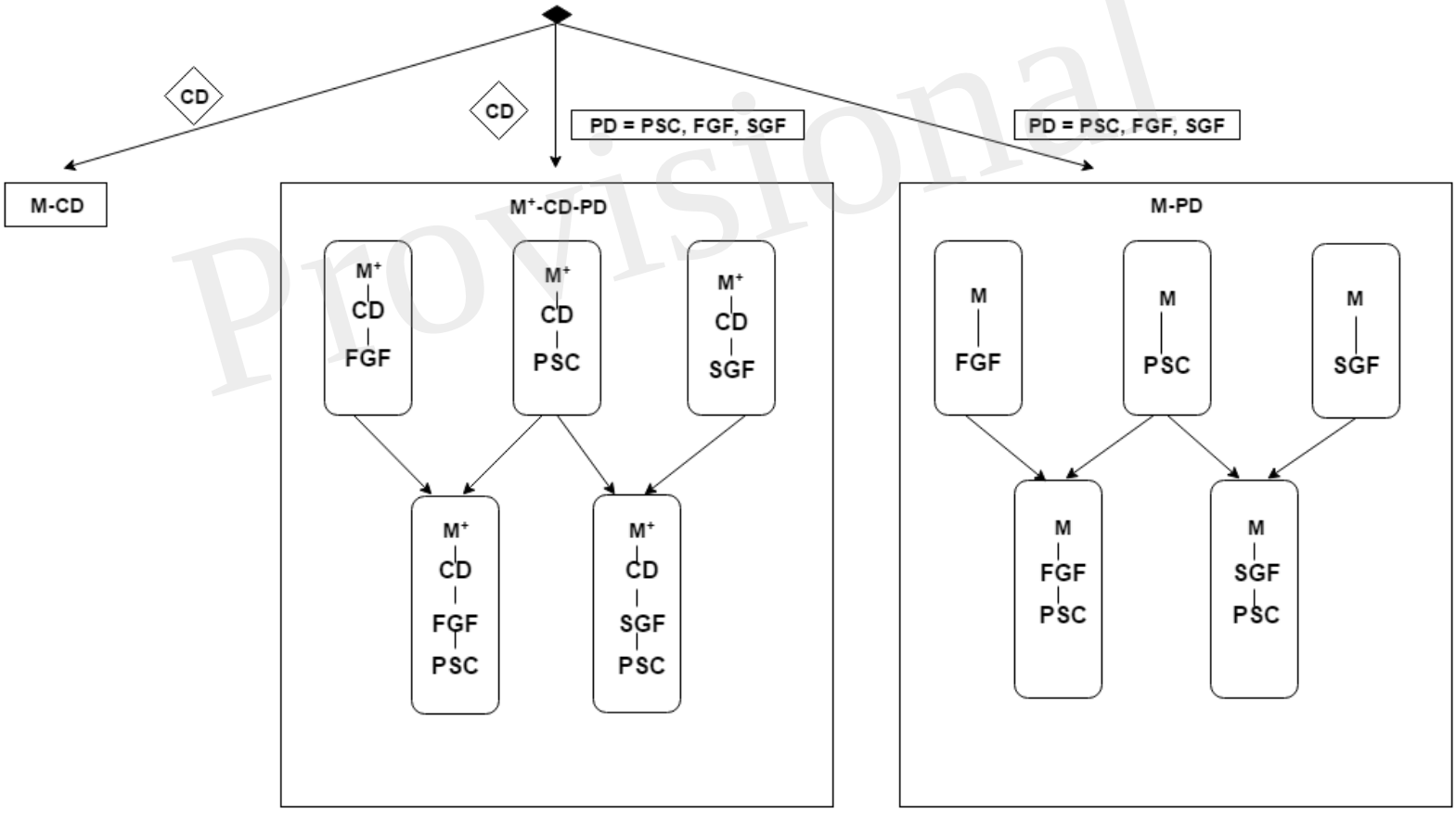


Figure 03.JPEG

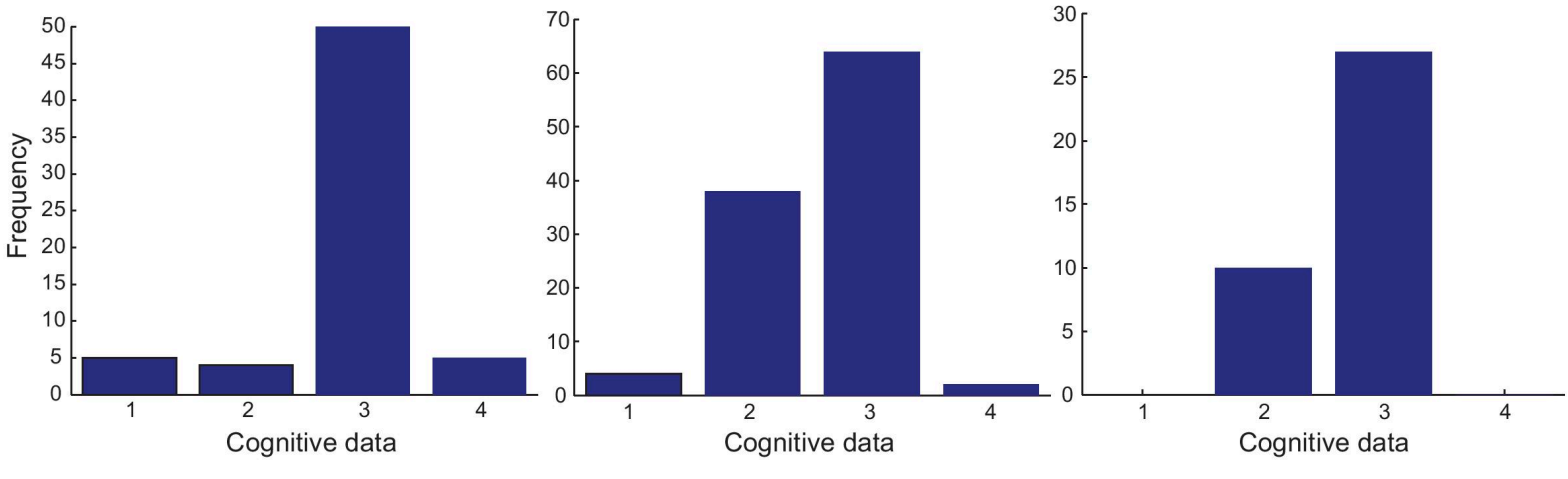


Figure 04.JPEG

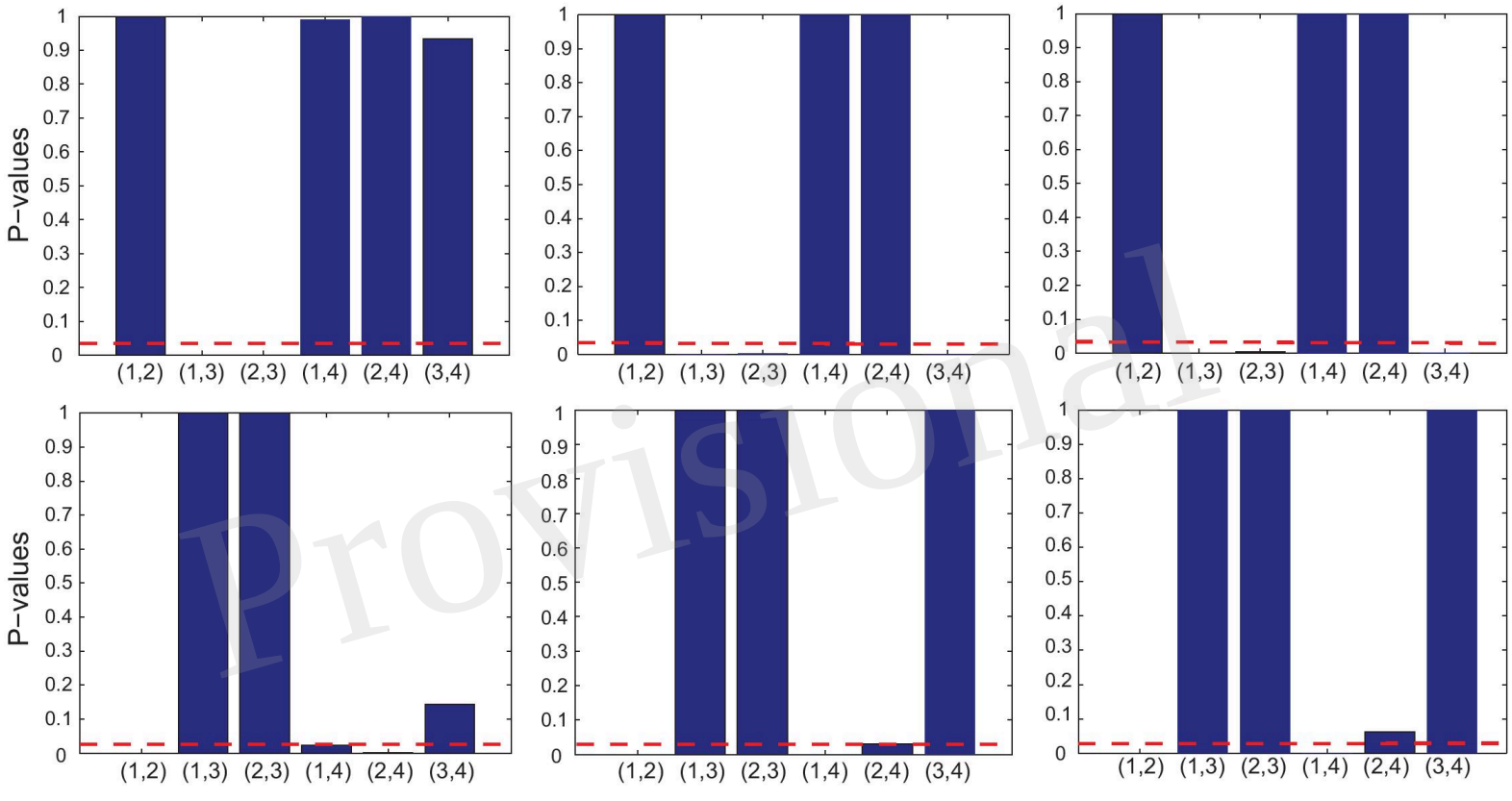


Figure 05.JPEG

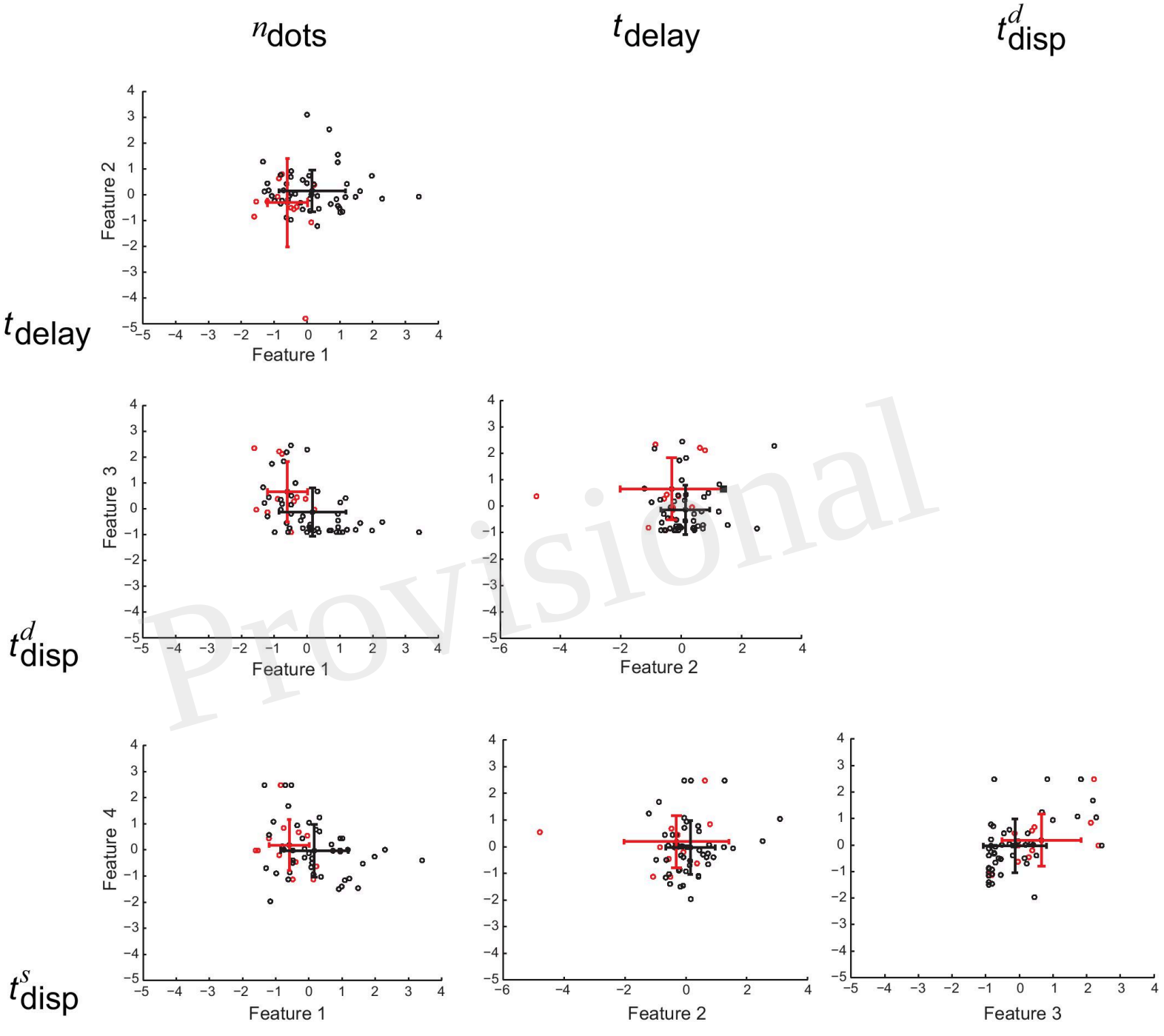
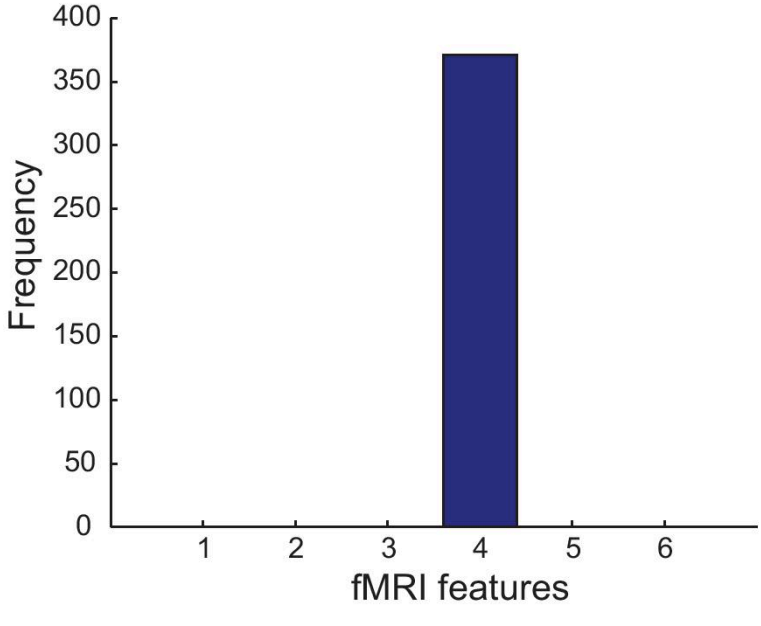


Figure 06.JPEG



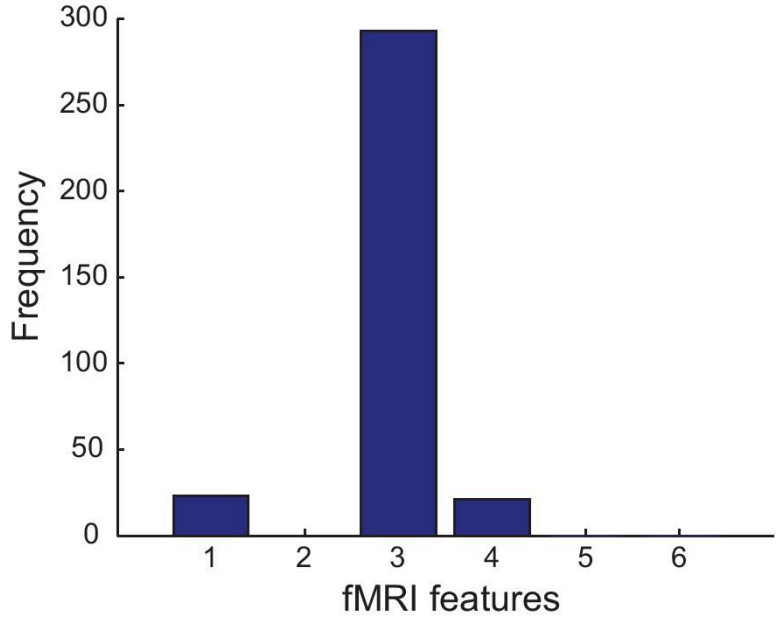
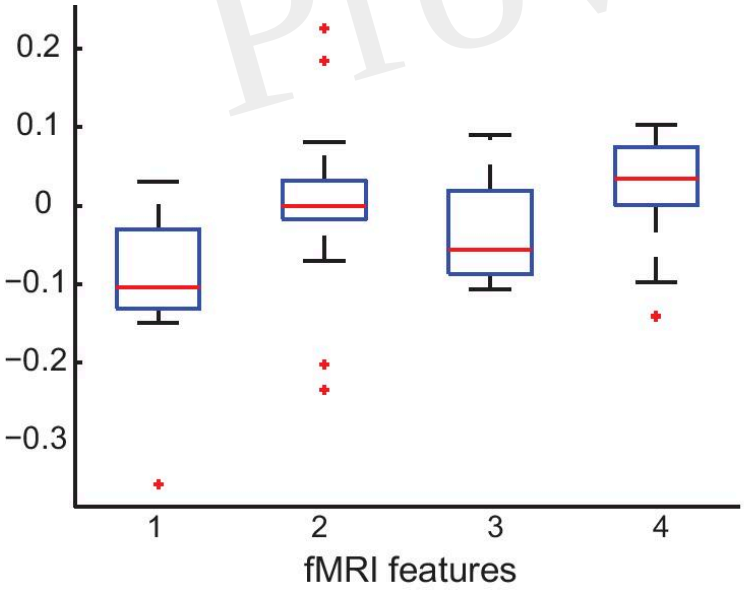


Figure 07.JPEG



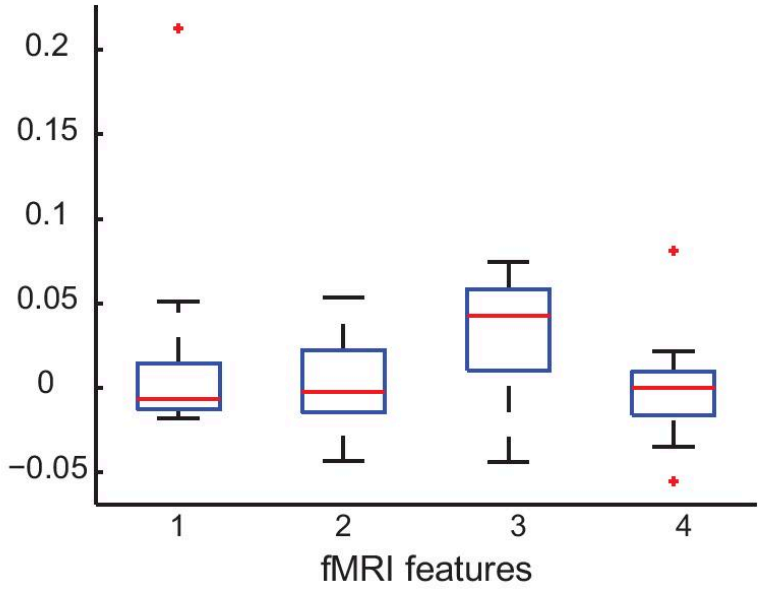


Figure 08.JPEG

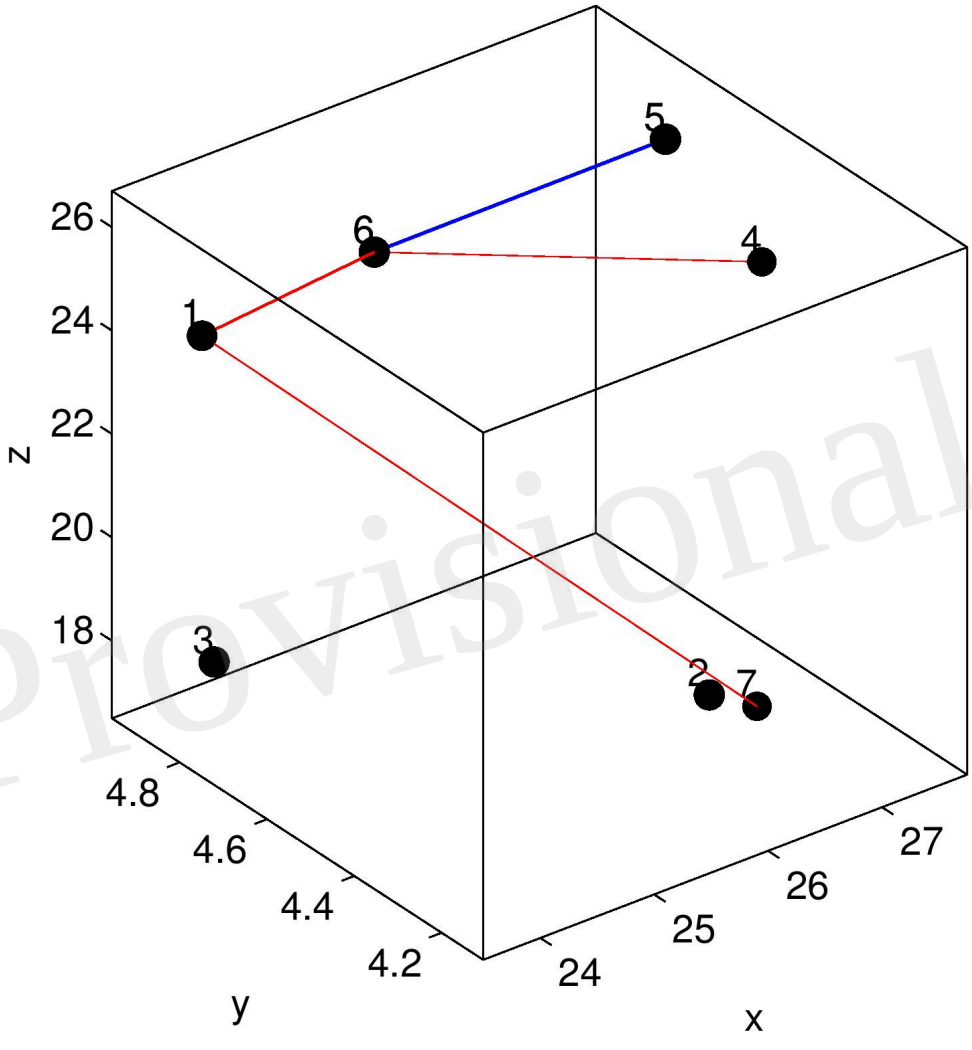


Figure 09.JPEG