Multisite Disease Classification with Functional Connectomes via Multitask Structured Sparse SVM

Takanori Watanabe[1], Daniel Kessler[2], Clayton Scott[1], Chandra Sripada[2]

[1]Department of EECS, [2]Department of Psychiatry
University of Michigan, Ann Arbor, MI 48109, USA
{takanori,kesslerd,clayscot,csripada}@umich.edu

Abstract. There is great interest in developing imaging-based methods for diagnosing neuropsychiatric conditions. To this end, multiple data-sharing initiatives have been launched in the neuroimaging field, where datasets are collected across multiple imaging sites. While this enables researchers to study the disorders of interest with substantial sample size, it also creates new challenges since the data aggregation process introduces various sources of site-specific heterogeneities. To address this issue, we introduce a multitask structured sparse support vector machine (SVM) that uses resting state functional connectomes (FCs) as the features for predicting diagnostic labels. Specifically, we employ a penalty that accounts for the following two-way structure that exists in a multisite FC dataset: (1) the 6-D *spatial structure* in the FCs captured via either the GraphNet, fused Lasso, or the isotropic total variation penalty, and (2) the *inter-site* structure captured via the multitask ℓ_1/ℓ_2-penalty. To solve the resulting high dimensional optimization problem, we introduce an extension to a recently proposed algorithm based on the alternating direction method. The potential utility of the proposed method is demonstrated on the multisite ADHD-200 dataset.

Keywords: Multitask learning, structured sparsity, support vector machine, resting-state fMRI, alternating direction method

1 Introduction

In this work, we are interested in a supervised classification problem, where the goal is to predict the diagnostic status of an individual using functional connectomes (FCs) derived from resting-state fMRI (rs-fMRI) [4]. Fortunately, with various data sharing projects emerging in the neuroimaging community [12, 15], we have access to training data of unprecedented sample size. However, such community-wide collaborative efforts typically involve aggregating data from multiple imaging sites, which introduces several sources of systematic confounds, such as variability in the scanner quality, image acquisition protocol, subject demographics, etc. In order to effectively make use of these multisite datasets, it is important to train the classifiers in a way that accounts for these site-specific heterogeneities. To this end, we propose a classification framework that adopts a multitask learning (MTL) approach [5, 8, 10, 13].

The idea behind MTL is to *jointly* train multiple tasks in order to improve classification performance, under the assumption that the tasks are related to each other in some sense. Recently, MTL methods have been successfully applied in brain decoding [8,13], where the *participants* from a multi-subject fMRI study are treated as the tasks. The underlying assumption here is that the brain regions that are activated from a stimulus will share similar patterns across different tasks/subjects. In contrast to these works, the method we propose in this work treats the *sites* from which the rs-fMRI scans are collected as the tasks.

2 Material and Methods

To generate the FCs, we used the grid-based parcellation scheme adopted by Watanabe *et al.* in [16], which involves 347 nodes defined on the standard MNI template; Fig. 1 provides a schematic representation of this parcellation scheme. Each nodes represents a 15mm diameter sphere with 33 voxels, and is placed throughout the entire brain with a spacing of 18×18×18mm (voxel resolution is 3×3×3mm). A regional time-series is assigned on each node by spatially averaging the BOLD signals, and FCs of size $p=\binom{347}{2}=60,031$ are obtained by computing all pairwise Pearson correlations between the time-series of the nodes.

2.1 Supervised Learning and the Multitask Framework

Šuppose we are given K supervised learning tasks, where for each task $k=1,\dots,K$, we are given n_k input/output pairs $\big\{(\boldsymbol{x}_i^k, y_i^k)\big\}_{i=1}^{n_k} \in (\mathbb{R}^p \times \{\pm 1\})^{n_k}$. In the context of our work, $\boldsymbol{x}_i^k$ and y_i^k represent the FC and the diagnostic label of the i-th subject from the k-th site, respectively. The goal is to jointly learn K linear classifiers of the form $f_k(\boldsymbol{x}) = \operatorname{sign}(\langle \boldsymbol{w}^k, \boldsymbol{x}\rangle)$, where $\boldsymbol{w}^1,\dots,\boldsymbol{w}^K \in \mathbb{R}^p$ are task-specific weight vectors obtained by solving the following optimization problem:

$$\operatorname*{arg\,min}_{\boldsymbol{w}^1,\dots,\boldsymbol{w}^K \in \mathbb{R}^p} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell\left(y_i^k \left\langle \boldsymbol{w}^k, \boldsymbol{x}_i^k \right\rangle\right) + \mathcal{R}(\boldsymbol{w}^1,\dots,\boldsymbol{w}^K)\,.$$

The first term here is the *pooled empirical risk* of a convex margin-based loss $\ell : \mathbb{R} \to \mathbb{R}_+$ and the second term $\mathcal{R} : \mathbb{R}^{pK} \to \mathbb{R}_+$ is a penalty function that enforces certain kind of structure on the weight vectors. In this work, we employ the *hinge-loss* $\ell(t) = \max(1-t, 0)$ from the well known support vector machine (SVM) classifier, although other convex margin-based losses can be used as well.

For brevity, we define a functional $\mathcal{L}(\boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k) := \sum_{i=1}^{n_k} \ell(y_i^k \langle \boldsymbol{w}^k, \boldsymbol{x}_i^k\rangle)$ which aggregates the empirical loss from the k-th task, where $\boldsymbol{X}^k \in \mathbb{R}^{n_k \times p}$ denotes the design matrix for the k-th task and $\boldsymbol{Y}^k \in \{\pm 1\}^{n_k \times n_k}$ is defined as $\boldsymbol{Y}^k := \operatorname{diag}(y_1^k,\dots,y_{n_k}^k)$. Also for conciseness, let $\underline{\boldsymbol{w}} \in \mathbb{R}^{pK}$ denote the vector obtained by stacking the weight vectors $\{\boldsymbol{w}^k\}_{k=1}^K$ together. In this work, we focus on convex penalty functions of the form: $\mathcal{R}(\underline{\boldsymbol{w}}) = \gamma \sum_{k=1}^K \mathcal{R}_1(\boldsymbol{w}^k) + \lambda \mathcal{R}_2(\underline{\boldsymbol{w}})$, where $\gamma, \lambda \geqslant 0$ are hyperparameters. Thus the objective function can be written as:

$$\operatorname*{arg\,min}_{\underline{\boldsymbol{w}} \in \mathbb{R}^{Kp}} \sum_{k=1}^{K} \frac{1}{n_k} \mathcal{L}(\boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k) + \gamma \sum_{k=1}^{K} \mathcal{R}_1(\boldsymbol{w}^k) + \lambda \mathcal{R}_2(\underline{\boldsymbol{w}})\,. \qquad (1)$$

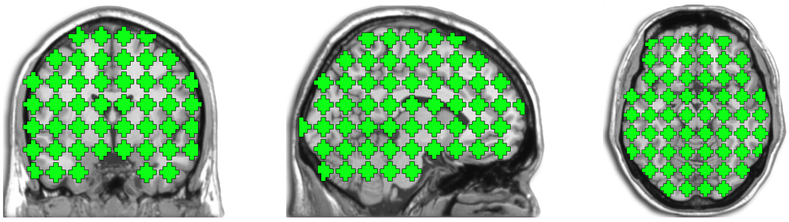


Fig. 1: The brain parcellation scheme adopted in this work. The green regions represent (pseudo)-spherical nodes each encompassing 33 voxels.

The first penalty $\mathcal{R}_1$ allows us to encode prior knowledge about the *intra-task* structure of the data. While various penalties such as GraphNet (GN), fused Lasso (FL), and isotropic total variation (TV) have been applied successfully in the fMRI literature [1, 6, 9, 16], these penalties by themselves do not account for the *inter-task* structure of the dataset (FL is also known as anisotropic total variation). Thus a second penalty $\mathcal{R}_2$ is included in (1), which allows us incorporate a notion of "task-relatedness" by enforcing some form of structure on $\underline{\boldsymbol{w}}$.

For the intra-task penalty $\mathcal{R}_1$, following the recent work of [16], we account for the 6-D spatial structure of FCs (defined by pairs of points in 3-D) by employing either the GN or FL penalty, which can be expressed in the following form:

$$\mathcal{R}_1(\boldsymbol{w}^k) = \frac{1}{q}\left\|\boldsymbol{C}\boldsymbol{w}^k\right\|_q^q = \begin{cases}\text{GraphNet} & \text{if } q = 2\\ \text{Fused Lasso} & \text{if } q = 1\,,\end{cases}$$

where $\boldsymbol{C}$ denotes a 6-D finite differencing matrix. The idea behind GN and FL is to promote spatial contiguity by penalizing the differences among neighboring coordinates of the FC. Similarly, the TV penalty, which is a rotationally invariant counterpart of the FL penalty, can also be used to encourage spatial contiguity; see [9] for its closed form expression.

2.2 Structured Sparsity with Group Variable Selection

We propose to integrate the *structured sparsity* framework introduced in [16] with the popular multitask ℓ_1/ℓ_2-penalty [5, 10]. Specifically, for the inter-task penalty $\mathcal{R}_2$, we use $\mathcal{R}_2(\underline{\boldsymbol{w}}) = \sum_{j=1}^{p} \|\boldsymbol{w}_j\|_2$, which is the so-called ℓ_1/ℓ_2-penalty. Here $\boldsymbol{w}_j \in \mathbb{R}^K$ is a vector formed by stacking the j-th weight vector coefficients across the K tasks. This penalty has the appealing *group variable selection* property [5, 10], which promotes learning features that are relevant across all sites, thereby simplifying interpretation of the selected features. At the same time, the actual weights associated with a given correlation can vary across site, in contrast to training a single classifier over a pooled dataset.

2.3 Optimization Algorithm

To solve the proposed large scale optimization problem, we apply the *alternating direction method of multipliers* (ADMM) algorithm [2] introduced in [16], but with a minor modification. The complete algorithm is outlined in Alg. 1. We note that this section focuses on GN and FL, but the ADMM algorithm for TV differs only in line 5 of Alg. 1, but the details are omitted for lack of space.

Alg. 1 ADMM for Multitask Structured Sparse SVM

```
 1: Initialize variables, assign hyperparameters $\lambda, \gamma \geqslant 0$
 2: repeat
 3:     for $k = 1, \dots, K$ do
 4:         $\boldsymbol{w}^k \leftarrow ((\boldsymbol{X}^k)^T \boldsymbol{X}^k + 2\boldsymbol{I}_p)^{-1} \{ (\boldsymbol{Y}^k \boldsymbol{X}^k)^T \left(\boldsymbol{v_1^k} - \boldsymbol{u_1^k}\right) \left(\boldsymbol{v_2^k} - \boldsymbol{u_2^k}\right) + \boldsymbol{A}^T \left(\boldsymbol{v_4^k} - \boldsymbol{u_4^k}\right) \}$
                                                        ▻ solve using matrix inversion Lemma
 5:         $\boldsymbol{v_3^k} \leftarrow \begin{cases} \text{apply Equation (3)} & \text{if } q = 1 \text{ (FL)} \\ \rho(\gamma \boldsymbol{B} + \rho \boldsymbol{I})^{-1} \widetilde{\boldsymbol{C}}(\boldsymbol{v_4^k} - \boldsymbol{u_3^k}) & \text{if } q = 2 \text{ (GN)} \end{cases}$
 6:         $\boldsymbol{v_1^k} \leftarrow \operatorname{Prox}_{\ell/(\rho n_k)}(\boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k + \boldsymbol{u_1^k})$   ▻ $\operatorname{Prox}_{\tau\ell}(t) := \begin{cases} t & \text{if } t > 1 \\ 1 & \text{if } 1 - \tau \leqslant t \leqslant 1 \\ t + \tau & \text{if } t < 1 - \tau \end{cases}$
 7:         $\boldsymbol{v_4^k} \leftarrow \left(\widetilde{\boldsymbol{C}}' \widetilde{\boldsymbol{C}} + \boldsymbol{I}_{\tilde{p}}\right)^{-1} \left(\widetilde{\boldsymbol{C}}' [\boldsymbol{v_3^k} + \boldsymbol{u_3^k}] + \boldsymbol{A}\boldsymbol{w}^k + \boldsymbol{u_4^k}\right)$   ▻ solve using FFT
 8:     end for
 9:     for $j = 1, \dots, p$ do
10:         $\boldsymbol{v}_{2,j} \leftarrow \operatorname{vsoft}_{\lambda/\rho}(\boldsymbol{w}_j + \boldsymbol{u}_{2,j})$   ▻ $\operatorname{vsoft}_\tau(\boldsymbol{t}) := \max(1 - \frac{\tau}{\|\boldsymbol{t}\|_2}, 0)\,\boldsymbol{t}, \ \boldsymbol{t} \in \mathbb{R}^K$
11:     end for
12:     for $k = 1, \dots, K$ do                              ▻ dual variable update
13:         $\boldsymbol{u_1^k} \leftarrow \boldsymbol{u_1^k} + \boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k - \boldsymbol{v_1^k}$
14:         $\boldsymbol{u_2^k} \leftarrow \boldsymbol{u_2^k} + \boldsymbol{w}^k - \boldsymbol{v_2^k}$
15:         $\boldsymbol{u_3^k} \leftarrow \boldsymbol{u_3^k} + \boldsymbol{v_3^k} - \widetilde{\boldsymbol{C}} \boldsymbol{v_4^k}$
16:         $\boldsymbol{u_4^k} \leftarrow \boldsymbol{u_4^k} + \boldsymbol{A}\boldsymbol{w}^k - \boldsymbol{v_4^k}$
17:     end for
18: until stopping criterion is met
```

To apply Alg. 1, we employ the *data augmentation+masking* strategy that was proposed in [16]. In brief, the idea behind this method is that as it stands, the ADMM algorithm for solving the objective function (1) with the GN, FL, or TV penalty will require the inversion of the Laplacian matrix $\boldsymbol{C}^T\boldsymbol{C}$, which is prohibitively large. Thus we rewrite the GN/FL penalty as $\mathcal{R}_1(\boldsymbol{w}^k) = \|\boldsymbol{B}\widetilde{\boldsymbol{C}}\boldsymbol{A}\boldsymbol{w}^k\|_q^q$, where $\boldsymbol{A}$ is an *augmentation matrix*, $\widetilde{\boldsymbol{C}}$ is the finite differencing matrix for the augmented $\boldsymbol{w}^k$, and $\boldsymbol{B}$ is a diagonal masking matrix that ensures the penalty remains unaffected, *i.e.*, $\|\boldsymbol{B}\widetilde{\boldsymbol{C}}\boldsymbol{A}\boldsymbol{w}^k\|_q^q = \|\boldsymbol{C}\boldsymbol{w}^k\|_q^q$. This results in a new Laplacian matrix $\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}}$, which possesses a special structure known as *block-circulant with circulant-blocks*, whose matrix inverse can be evaluated efficiently via the fast Fourier Transform (FFT) (line 7, Alg. 1; see [16] for more details).

Using this augmentation+masking strategy, we can rewrite the objective as:

$$\min_{\underline{\boldsymbol{w}}} \sum_{k=1}^{K} \frac{1}{n_k} \mathcal{L}(\boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k) + \frac{\gamma}{q} \sum_{k=1}^{K} \left\|\boldsymbol{B}\widetilde{\boldsymbol{C}}\boldsymbol{A}\boldsymbol{w}^k\right\|_q^q + \lambda \sum_{j=1}^{p} \|\boldsymbol{w}_j\|_2 \ ,$$

which can be converted into the following canonical ADMM form [2]:

$$\min_{\{\boldsymbol{w}^k, \boldsymbol{v_1^k}, \boldsymbol{v_2^k}, \boldsymbol{v_3^k}, \boldsymbol{v_4^k}\}} \sum_{k=1}^{K} \frac{1}{n_k} \mathcal{L}(\boldsymbol{v_1^k}) + \frac{\gamma}{q} \sum_{k=1}^{K} \left\|\boldsymbol{B}\boldsymbol{v_3^k}\right\|_q^q + \lambda \sum_{j=1}^{p} \|\boldsymbol{v_{2,j}}\|_2$$

$$\text{s.t. } \boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k = \boldsymbol{v_1^k}, \boldsymbol{w}^k = \boldsymbol{v_2^k}, \widetilde{\boldsymbol{C}}\boldsymbol{v_4^k} = \boldsymbol{v_3^k}, \boldsymbol{A}\boldsymbol{w}^k = \boldsymbol{v_4^k} \quad \forall k = 1, \dots, K. \tag{2}$$

It is straightforward to show that the above two problems are equivalent, and Alg. 1 follows from applying the standard ADMM iteration on (2). We emphasize that all the updates in Alg. 1 can be carried out efficiently in analytical form. For example, line 5 in Alg. 1 is a simple diagonal matrix inversion in the case of GN, and for the FL case we have the following closed form update:

$$\left[\boldsymbol{v}_{\mathbf{3}}^{k}\right]_s \leftarrow \begin{cases} \text{soft}_{\gamma/\rho}\left(\left[\widetilde{\boldsymbol{C}}(\boldsymbol{v}_{\mathbf{4}}^{k} - \boldsymbol{u}_{\mathbf{3}}^{k})\right]_s\right) & \text{if } \boldsymbol{B}_{s,s} = 1 \\ \left[\widetilde{\boldsymbol{C}}(\boldsymbol{v}_{\mathbf{4}}^{k} - \boldsymbol{u}_{\mathbf{3}}^{k})\right]_s & \text{if } \boldsymbol{B}_{s,s} = 0, \end{cases} \tag{3}$$

where $\text{soft}_\tau(t){:=}\max(1 - \frac{\tau}{|t|}, 0) \cdot t$ denotes the *soft-threshold operator* and $[\cdot]_s$ indexes the s-th element of a vector. Finally, we note $\text{Prox}_{\tau\ell}(t)$ in line 6 is an elementwise update corresponding to the proximal operator of the hinge-loss.

3 Experiments

The ADHD-200 Dataset. We used the publicly available ADHD-200 competition dataset [15], which contains rs-fMRI scans of subjects diagnosed as either typically developing (TD) or with ADHD. The dataset is collected across seven sites and consists of two parts: a training set and a validation test set (Brown site excluded from our study as the subject labels are not released). Analyses were limited to participants with: (1) MPRAGE anatomical images with consistent near-full brain coverage with successful registration; (2) complete phenotypic information for main phenotypic variables (diagnosis, age, handedness); (3) mean framewise displacement (FD) within two standard deviation (SD) of the sample mean; (4) full IQ within two SDs of the ADHD-200 sample mean. After applying these sample selection criteria, we analyzed resting state scans from 628 individuals (TD=416, ADHD=212) in the training set and 106 subjects (TD=65, ADHD=41) in the test set. Functional images were reconstructed, slice-time corrected, motion corrected, and co-registered to the MNI space using SPM8.

Experimental Results. To assess the validity of the proposed method, we compared the performance of various SVM-based classifiers using the ADHD-200 dataset, where resting-state FCs were produced using the parcellation scheme described in Sec. 2. For the intra-task penalty $\mathcal{R}_1$, we compared four different regularization schemes: Elastic-net (EN) [5] with $\mathcal{R}_1(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$, GN, FL, and TV. For the inter-task penalty $\mathcal{R}_2$, we compared three different approaches:

1. **Pooled** ℓ_1: a single classifier is trained on the entire ADHD-200 dataset ($\mathcal{R}_2(\underline{\boldsymbol{w}}) = \|\underline{\boldsymbol{w}}\|_1$ with $\underline{\boldsymbol{w}} \in \mathbb{R}^p$ as $K = 1$).
2. **Single-task** ℓ_1/ℓ_1: equivalent to training separately across sites due to the separability of the penalty across sites ($\mathcal{R}_2(\underline{\boldsymbol{w}}) = \sum_{j=1}^{p} \|\boldsymbol{w}_j\|_1$).
3. **Multitask** ℓ_1/ℓ_2: *jointly* train the classifiers by solving (1).

The regularization parameters $\{\lambda, \gamma\}$ are tuned by conducting a 5-fold cross-validation (CV) on the training set over the following two-dimensional grid: $\lambda, \gamma {\in} \{2^{-13}, 2^{-12}, \ldots, 2^{-3}\}$. The final weight vector estimate is obtained by retraining the classifiers on the entire training set using the $\{\lambda, \gamma\}$ values that maximized the CV classification accuracy; for validation, we predicted the labels of the test set subjects using this weight vector. All methods were solved using ADMM with the algorithm terminated when the condition $\left\|\underline{\boldsymbol{w}}^{\text{new}} - \underline{\boldsymbol{w}}^{\text{old}}\right\|_2 \leqslant 5{\cdot}10^{-3} \times \left\|\underline{\boldsymbol{w}}^{\text{old}}\right\|_2$ was met or the iteration count reached 400.

To evaluate the quality of the classifiers, we analyzed the following set of performance measures for both the 5-fold CV and the validation test set results:

- Classification accuracy (ACC)
- Area under the ROC curve (AUC)
- Balanced score rate (BSR) = (sensitivity+specificity)/2
- Stability score (Stab.) = a measure of feature selection stability (see [1, 14])
- P-value (PVAL) computed from binomial test.
- Sparsity level (SP%) = $100 \cdot \frac{|\text{\# non-zero features}|}{pK}$

The AUC and BSR are analyzed since ACC by itself can be misleading when the dataset labels are imbalanced (ACC, AUC, and BSR are averaged across the tasks); the ROC curves are constructed by varying the threshold of the classifiers. *Stability score* is a measure introduced in [14] which quantifies the stability of the features selected across the CV folds (see [1, 14] for its precise definition). Classifier performance on the test set was compared to random guessing via a binomial test based on a binomial distribution B(p,n) with p=0.5 and n=109 samples, with $PVAL$ evaluated via an one-sided binomial test [7]; the alternative approach of permutation test was not pursued due to its severe computational cost. Finally, *sparsity level* is the fraction of features selected in the final model.

Table 1 presents the classification results from the 5-fold CV and validation on the test-set, and Fig. 1 displays the corresponding ROC curves. These results demonstrate that training a single classifier via the "pooling" approach yields the worst performance in terms of accuracy, AUC, and BSR, suggesting that blindly aggregating the datasets across different sites can be problematic for accurate disease classification. Comparison between the single-task and the multitask approaches shows that the ℓ_1/ℓ_2-penalized approach yields superior performance in terms of AUC, although no striking difference can be observed in terms of accuracy and BSR.

In addition to the performance gain with the ℓ_1/ℓ_2-penalty, the set of weight vector estimates $\{\hat{\boldsymbol{w}}^k\}_{k=1}^K$ all share a common support of length p with this multitask approach. This is invaluable for interpretation, as the selected features can be viewed as edges that are informative across all sites. For visualization, we grouped the indices of this support according to the network parcellation scheme proposed by Yeo *et al.* in [17], and reshaped them into a 347×347 symmetric matrix with zeroes on the diagonal. The resulting support matrices for the EN+ℓ_1/ℓ_2 and the FL+ℓ_1/ℓ_2-penalized SVM are presented in Fig. 3 (results for GN+ℓ_1/ℓ_2 and TV+ℓ_1/ℓ_2 were very similar to FL+ℓ_1/ℓ_2). An interesting observation here is that the support structure from the FL+ℓ_1/ℓ_2-penalized SVM shows concentrated connectivity patterns in the intra-frontoparietal (6-6) and the intra-default network (7-7) regions; Fig. 3 provides a brain space representation of these connections (figures generated using BrainNet Viewer, `www.nitrc.org/projects/bnv/`). These network regions are frequently reported to exhibit disrupted connectivity patterns in resting state studies of ADHD [3],

Table 1: The classification results from the 5-fold CV and the validation test-set.

	CV (628 subjects)				Test-set (106 subjects)				
	ACC	AUC	BSR	Stab.	ACC	AUC	BSR	PVAL	SP%
EN (ℓ_1)	.689	.687	.630	.277	.557	.617	.476	.143	2.54%
GN (ℓ_1)	.704	.708	.631	.253	.594	.608	.494	.032	28.88%
FL (ℓ_1)	.688	.720	.586	.059	.632	.592	.530	.004	64.85%
TV (ℓ_1)	.701	.715	.620	.005	.623	.608	.521	.007	90.32%
EN (ℓ_1/ℓ_1)	.709	.752	.649	.276	.623	.609	.530	.007	0.28%
GN (ℓ_1/ℓ_1)	.713	.750	.652	.165	.642	.613	.573	.002	67.14%
FL (ℓ_1/ℓ_1)	.715	.750	.659	.329	.632	.634	.547	.004	1.30%
TV (ℓ_1/ℓ_1)	.718	.753	.661	.345	.642	.654	.550	.002	1.61%
EN (ℓ_1/ℓ_2)	.720	.754	.657	.217	.651	.645	.556	.001	0.25%
GN (ℓ_1/ℓ_2)	.720	.766	.657	.320	.642	.668	.546	.002	1.03%
FL (ℓ_1/ℓ_2)	.718	.766	.653	.315	.642	.673	.546	.002	0.79%
TV (ℓ_1/ℓ_2)	.720	.766	.658	.316	.642	.672	.546	.002	0.80%

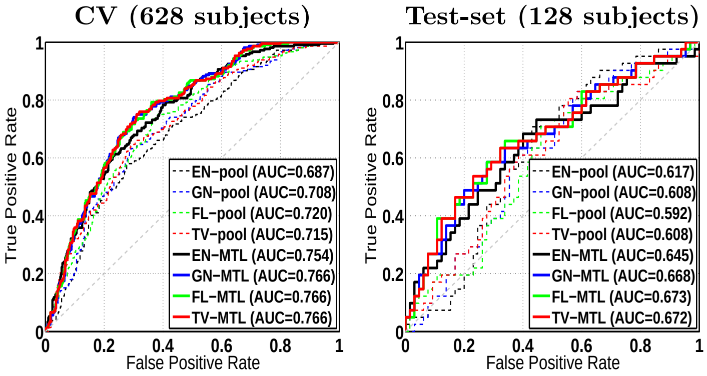


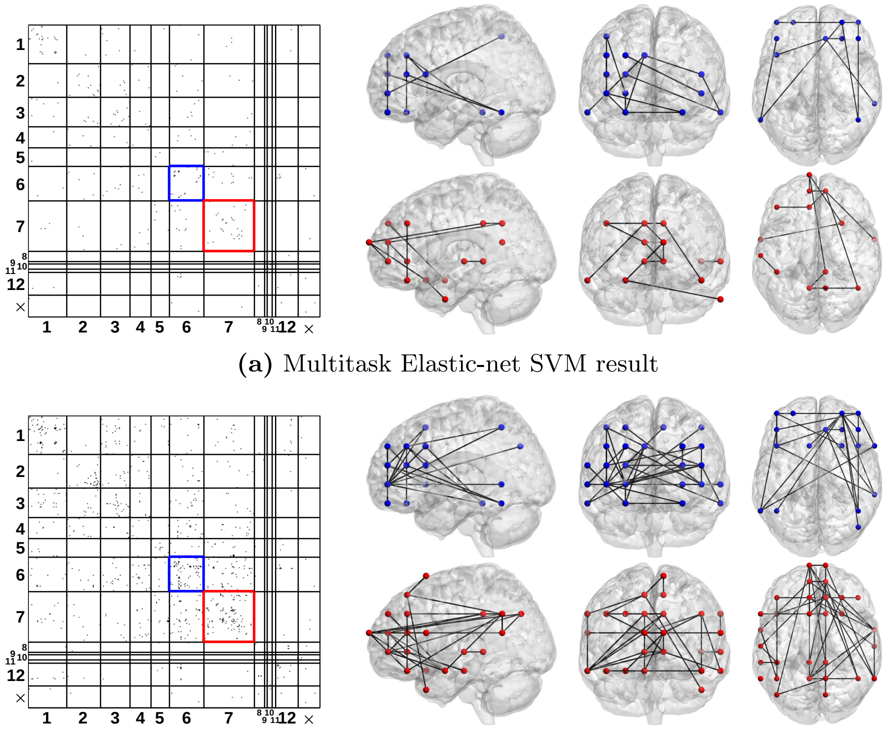
Fig. 2: Table 1 classifiers' ROC (ℓ_1/ℓ_1-curves omitted to improve curve visibility).

although the accuracies obtained from our classifiers are not at the level where the selected features can be interpreted as reliable ADHD biosignatures.

Finally, we note that most of the accuracies reported on the validation test-set in Table 1 exceeded the highest result from the actual ADHD-200 competition (which was 61.54% [15]). However, there are two major caveats: (1) the results in this work cannot be directly compared with the official competition results due to the subject screening procedure we applied on the test set (the criteria such as the FD-based one is important for avoiding confounds from excessive head motion), and (2) the participants in the actual competition were required to predict the labels of 26 subjects from the Brown site, despite the fact that no training data were provided from this site, making it harder to predict the labels for these subjects. The second caveat also implies that most MTL methods, including the ℓ_1/ℓ_2-penalty employed in this work, cannot be applied since there are no means to train a weight vector for a task whose data are not provided. An alternative approach such as *transfer learning* [11] may be considered for this.

Table 2: Network parcellation scheme of the brain proposed by Yeo *et al.* in [17].

Network membership Table (× is "unlabeled")			
1. Visual	2. Somatomotor	3. Dorsal Attention	4. Ventral Attention
5. Limbic	6. Frontoparietal	7. Default	8. Striatum
9. Amygdala	10. Hippocampus	11. Thalamus	12. Cerebellum



(a) Multitask Elastic-net SVM result

(b) Multitask Fused Lasso SVM result

Fig. 3: Weight vectors estimated from the EN+ℓ_1/ℓ_2 and FL+ℓ_1/ℓ_2-penalized SVM. **Left:** support matrices of the selected features (rows/cols grouped by network membership). **Right:** brain space representation of the selected edges in the intra-frontoparietal (6-6: blue) and the intra-default network (7-7: red).

4 Conclusion

We presented a multitask structured sparse SVM, a multitask extension to the connectome-based disease classification method introduced in [16], where the imaging sites are treated as *tasks*. Experimental results on the multisite ADHD-200 dataset suggest that the multitask approach using the ℓ_1/ℓ_2-penalty can provide improvement in classification performance over the naive *pooling approach*, where a single classifier is trained on the entire multisite dataset. In addition, the ℓ_1/ℓ_2-penalty achieved higher AUC scores than the single-task ℓ_1/ℓ_1-penalty, and the *group variable selection* property of the multitask approach gives a more interpretable model by selecting the same set of features across sites, which can be visualized compactly in brain space.

References

1. Baldassarre, L., Mourao-Miranda, J., Pontil, M.: Structured sparsity models for brain decoding from fMRI data. Proc. PRNI pp. 5–8 (2012)
2. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 3, 1–122 (2011)
3. Castellanos, F., Proal, E.: Large-scale brain systems in ADHD: beyond the prefrontalstriatal model. Trends Cogn. Sci. 16, 17 (2012)
4. Castellanos, F.X., Martino, A.D., Craddock, R.C., Mehta, A.D., Milham, M.P.: Clinical applications of the functional connectome. NeuroImage 80(0), 527 – 540 (2013)
5. Chen, X., He, J., Lawrence, Carbonell, J.G.: Adaptive multi-task sparse learning with an application to fMRI study. In: SDM (2012)
6. Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E.: Interpretable whole-brain prediction analysis with GraphNet. NeuroImage 72(0), 304 – 321 (2013)
7. Heinzle, J., Wenzel, M.A., Haynes, J.D.: Visuomotor functional network topology predicts upcoming tasks. J. Neurosci. 32(29), 9960–9968 (2012)
8. Marquand, A., Brammer, M., Williams, S., Doyle, O.: Bayesian multi-task learning for decoding multi-subject neuroimaging data. NeuroImage 92(0), 298 – 311 (2014)
9. Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B.: Total variation regularization for fMRI-based prediction of behavior. IEEE Trans. Med. Imag. 30(7), 1328–1340 (2011)
10. Obozinski, G., Taskar, B., Jordan, M.I.: Joint covariate selection and joint subspace selection for multiple classification problems. Stat. Comput. 20(2), 231–252 (2010)
11. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22(10), 1345–1359 (2010)
12. Poline, J.B., Breeze, J.L., Ghosh, S.S., Gorgolewski, K., Halchenko, Y.O., Hanke, M., Helmer, K.G., Marcus, D.S., Poldrack, R.A., Schwartz, Y., Ashburner, J., Kennedy, D.N.: Data sharing in neuroimaging research. Front. Neuroinformatics 6(9) (2012)
13. Rao, N.S., Cox, C.R., Nowak, R.D., Rogers, T.T.: Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. NIPS pp. 2202–2210 (2013)
14. Rasmussen, P.M., Hansen, L.K., Madsen, K.H., Churchill, N.W., Strother, S.C.: Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition 45(6), 2085 – 2100 (2012)
15. The ADHD-200 Consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Front. Syst. Neurosci. 6(62) (2012)
16. Watanabe, T., Kessler, D., Scott, C., Angstadt, M., Sripada, C.: Disease prediction based on functional connectomes using a scalable and spatially-informed support vector machine. NeuroImage 96(0), 183 – 202 (2014)
17. Yeo, B., Krienen, F., Sepulcre, J., Sabuncu, M., Lashkari, D., Hollinshead, M., Roffman, J., Smoller, J., Zöllei, L., Polimeni, J., Fischl, B., Liu, H., Buckner, R.: The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106(3), 1125–65 (2011)