Large-scale classification of major depressive disorder via distributed Lasso

Dajiang Zhu[a], Qingyang Li[b], Brandalyn C. Riedel[a], Neda Jahanshad[a], Derrek P. Hibar[a], Ilya M. Veer[h], Henrik Walter[h], Lianne Schmaal[c,d,e], Dick J. Veltman[c], Dominik Grotegerd[f], Udo Dannlowski[f], Matthew D. Sacchet[g], Ian H. Gotlib[g], Jieping Ye[b], Paul M. Thompson[a]
[a]Imaging Genetics Center, University of Southern California, CA, USA;
[b]Department of Electrical Engineering and Computer Science, University of Michigan, MI, USA
[c]Department of Psychiatry and Neuroscience Campus Amsterdam,
VU University Medical Center, Amsterdam, The Netherlands
[d]Orygen, The National Centre of Excellence in Youth Mental Health, Melbourne, VIC, Australia.
[e]Centre for Youth Mental Health, The University of Melbourne, Melbourne, VIC, Australia
[f]Department of Psychiatry, University of Muenster, Muenster, Germany
[g]Neurosciences Program and Department of Psychology, Stanford University, Stanford, CA, USA
[h]Department of Psychiatry and Psychotherapy, Charité Universitätsmedizin Berlin, Berlin, Germany

ABSTRACT

Compared to many neurological disorders, for which imaging biomarkers are often available, there are no accepted imaging biomarkers to assist in the diagnosis of major depressive disorder (MDD). One major barrier to understanding MDD has been the lack of a practical and efficient platform for collaborative efforts across multiple data centers; integrating the knowledge from different centers should make it easier to identify characteristic measures that are consistently associated with the illness. Here we applied our newly developed "distributed Lasso" method to brain MRI data from multiple centers to perform feature selection and classification. Over 1,000 participants were involved in the study; our results indicate the potential of the proposed framework to enable large-scale collaborative data analysis in the future.

Keywords: ENIGMA, distributed Lasso

1. INTRODUCTION

Major depressive disorder (MDD) is a prevalent psychiatric condition, and affects over 350 million people worldwide [1]. Important advances have been made in understanding abnormal structural and functional brain correlates of emotion processing and regulation in individuals with MDD [2-4]. For example, the ENIGMA-MDD Working Group recently found that compared to healthy adults without a diagnosis of MDD, adults with MDD tend to have thinner cortical gray matter in the anterior and posterior cingulate, insula, orbitofrontal cortices and temporal lobes [2]. A prior study of subcortical structures showed that MDD patients have significantly smaller hippocampal volumes than do controls [3]. Despite this, the identification of specific regions or networks that discriminate between MDD and controls remains an important but unmet goal. Clinically, it is recognized that traditional qualitative radiological methods are not able to distinguish neuroanatomical scans of patients with MDD from healthy controls; therefore, diagnoses are made based on clinical evaluations. However, identifying brain imaging markers of depression could help to identify mechanisms of risk and lead to improvements in evaluating therapeutic interventions and patient outcomes.

There have been several successful attempts to classify MDD patients versus matched healthy controls using imaging-based features [5-7]. Using T1-weighted 3D brain MRI data, one study [5] combined voxel-based morphometry with a

12th International Symposium on Medical Information Processing and Analysis, edited by Eduardo Romero, Natasha Lepore, Jorge Brieva, Ignacio Larrabide, Proc. of SPIE Vol. 10160, 101600Y
 · doi: 10.1117/12.2256935

univariate analysis to select features to classify MDD patients relative to controls, at two institutions. Another study [6] developed an unsupervised machine learning algorithm to identify functional connectivity differences between MDD patients and normal controls. Although both studies achieved a very high (90%) classification accuracy, the MDD sample sizes were small (N<35), limiting generalizability. This reflects a practical obstacle in MDD studies: each center is typically limited in terms of available data and computational resources. An efficient platform is needed for collaborative efforts across multiple data centers to integrate information and increase statistical power for data analysis. In this paper, we applied our newly developed "distributed Lasso" method to data from multiple imaging centers to select features for classifying patients and controls, without compromising individual data privacy. Our data came from 3 cohorts that participate in the ENIGMA MDD consortium, including CODE, Muenster, and Stanford [2-3]. The total sample size is 1072: 370 MDD patients and 702 healthy controls. To the best of our knowledge, this is the largest study to date that has used brain MRI features for MDD classification.

2. METHODS

2.1 Overview

The general idea of the method is shown in **Figure 1.** Suppose we have multiple data centers (center-*1* to center-*m*) that may be geographically distributed across the world. Each center manages its own raw imaging data and computes standard measures from brain MRI including regional measures of cortical thickness, cortical surface areas, and subcortical volumes. Through our distributed Lasso algorithm, local gradient information is first computed based on local data. In other words, the models that predict diagnosis at each site, are fitted using an algorithm in which model parameters are iteratively updated by assessing the gradient of the classification error with respect to the model parameters. The gradients of the model parameters at each site are then combined to generate the overall gradient that is then sent back to each data center for updating. After 1,000 runs, a smaller set of brain measures (features) is selected from the Lasso regression model; these are used for classification. During this Lasso regression and classification process, each center does not need to share its raw data with other centers. This requirement can be advantageous, especially when datasets are too large to send to a central site, or when data privacy laws do not allow individual data transfer.

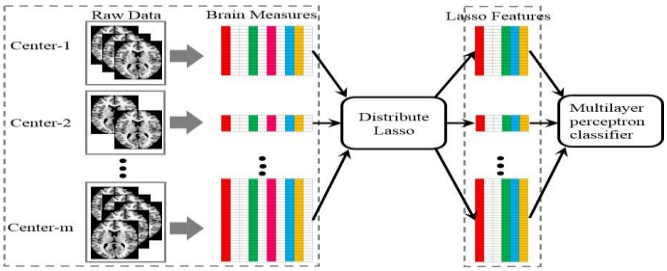

Figure 1. Overview of our proposed framework. Images from each center are consistently processed to compute a large set of standard brain measures. This set is reduced by using a distributed Lasso method that takes into account information from multiple centers, without sharing individual level data. The selected features are then combined using a multi-layer perceptron to perform classification of individuals as MDD or healthy controls.

2.2 Subjects and brain measures

Based on data from 3 ENIGMA-MDD cohorts: CODE, Muenster and Stanford [2-3], 54 of 156 brain measures and 251 of 1323 subjects were removed due to missing data. Demographic data for each center's participants are summarized in **Table 1**. Detailed image acquisition, pre-processing and FreeSurfer-derived brain segmentation and quality control methods have been previously published and may be found in [2-3].

Table 1. Demographics of three cohorts participating in the ENIGMA-MDD consortium.

		Adult samples					
Study	**Sample**	**Age of Controls**	**Age of MDD patient**	**% Female Controls**	**% Female MDD**	**Total N Controls**	**Total N MDD**

		(Mean ± SD)	(Mean ± SD)				
1	**CODE**	40.3 ± 13.0	41.1 ± 12.2	57	69	61	87
2	**Muenster cohort**	34.9 ± 12.0	38.2 ± 12.0	58	58	588	233
3	**Stanford**	37.5 ± 10.8	37.6 ± 10.1	62	58	53	50
	Combined					**702**	**370**

2.3 Classical Lasso formulation

Lasso is one of the most widely-used high-dimensional regression techniques for variable selection; it uses a sparse representation to enhance prediction accuracy. The general idea of Lasso is to minimize the sum of squared errors by forcing a proportion of the predictors' coefficients to be zero and restricting the sum of absolute value of all the regression coefficients to be less than a fixed value, e.g., 1. Classical Lasso regression is defined by the following equation:

$$\min_x \frac{1}{2} \|Ax - y\|_2^2 + \lambda \|x\|_1 : \mathrm{x} \in \mathbb{R}^p \quad (1)$$

In this paper, A represents the brain measures (after regressing out effects of age, sex, site, and ICV for the cortical surface area and subcortical volume measures), which are distributed across different data centers. y is the response vector indicating if the participant is an MDD patient or control. x is the regression coefficient shared between centers. λ is a positive regularization parameter. After optimization using equation (1), the coefficients for some brain measures will shrink to zero, and they will be excluded from the classification task in the next stages.

2.4 Distributed Lasso

Recently, we proposed a novel "distributed Lasso" algorithm to learn a consistent model to rank genomic variants in terms of their predictive value across different institutions without compromising individual data privacy [8]. In the current paper, we applied a similar strategy to perform feature selection. As illustrated in **Fig. 1**, we have m data centers. For the i^{th} center, we use (A_i, y_i) to represent the data set that it provides, where $A_i \in \mathbb{R}^{n_i \times p}$ and $y_i \ \mathbb{R}^{n_i \times 1}$. n_i is the number of participants at this center and p is the number of brain measures (all subjects are assumed to have the same number, p). Our goal is to learn the predictor weightings or coefficients – x by solving problem (1) on distributed data sets - (A_i, y_i).

In order to solve equation (1), we applied Iterative Shrinkage/Thresholding Algorithm (ISTA) [9]. The core step of ISTA is updating x:

$$x^{k+1} = \Gamma_{\lambda t_k} (x^k - t_k \nabla(x^k; \mathrm{A}, \mathrm{y})) \quad (2)$$

Here k is the iteration number, t_k is the step size and Γ is the soft threshold operator [8]. However, we are not able to compute $\nabla(x^k; \mathrm{A}, \mathrm{y})$ as each data center maintains its own data - (A_i, y_i). The key concept of distributed Lasso relies on the following decomposition:

$$\nabla g = A^T (\mathrm{Ax-y}) = \sum_{i=1}^m A_i^T (A_i x - y_i) = \sum_{i=1}^m \nabla g_i \quad (3)$$

The principle behind formula (3) is that it is possible to decompose the gradient computation of all the data into computing local gradients separately, which relate only to local data. For example, the i^{th} center is responsible for calculating $\nabla g_i = A_i^T (A_i x - y_i)$. Only the local gradient information - ∇g_i will be collected to generate ∇g. The output for this measure is then sent back to each center to update x. This process is done iteratively until the total loss is less than the preset threshold.

2.5 Identifying the best Lasso features

By applying (3) to the distributed dataset, some brain measures will "survive", following Lasso regression. The features that are retained are therefore considered more effective for prediction as they contribute more than other features to predicting the disease status - y during the regression. Nevertheless, how to pick an appropriate λ, and thus decide the sparsity of x, is still a challenging and open problem. In this paper, we adopted a practical strategy to guarantee the stability of the sparsity. We defined the following criterion:

$$\sum_{i=1}^{N} L(x, \lambda_i)/N \geq \mathrm{r} \tag{4}$$

$$L(x, \lambda_i) = \begin{cases} 1, if\ x \neq 0\ during\ the\ Lasso\ regression\ with\ \lambda_i \\ 0, otherwise \end{cases} \tag{5}$$

We solved problem (1) N times with a wide range of λ. For each coefficient, if its non-zero possibility is larger than a pre-defined threshold – r, it will be considered as a safe feature to preserve.

3. RESULTS

3.1 Ranking of Lasso features

In this paper, we used three desktop computers to simulate three data centers, although the approach holds regardless of the location of the data, and number of centers participating. Each dataset is stored on one machine without access to the others. To select the most stable features, we applied the strategy described in Section 2.5 and tried λ with 1000 different values (0.0001-0.1). The distribution of non-zero frequency of retention in the predictive model for the 102 brain measures we included is presented in **Figure 2**. The horizontal and vertical axes represent, respectively, the index of the feature and the number of times that feature was non-zero during 1000 Lasso regression procedures. By setting r = 1 in (4), we identified 30 brain measures (left side) that were non-zero through all regression procedures. These 30 selected features were used for classification in Section 3.2.

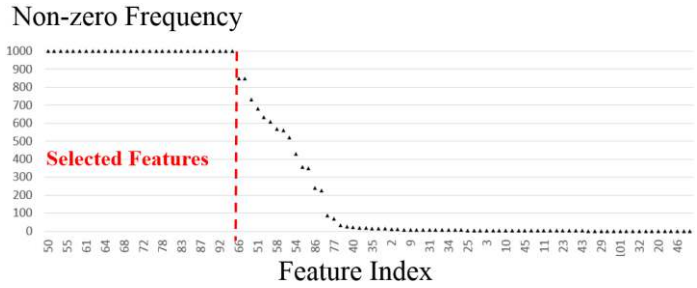

Figure 2. Non-zero frequency distribution of all 102 features over 1000 runs, ranking how often each was retained in predictive models for diagnostic classification of MDD versus controls.

One interesting finding is that all of the selected features are thickness measures. This result is consistent with our prior study [2] showing a consistent and significant difference in cortical thickness, but not cortical surface area, between adult MDD patients and controls. In addition, the 30 chosen brain measures show considerable overlap with the "most significant" cortical regions reported in [2], including the isthmus of the cingulate cortex in the left hemisphere, the bilateral rostral anterior cingulate cortex and the bilateral insula. The brain regions corresponding to the 30 cortical thickness measures are displayed in **Figure 3**.

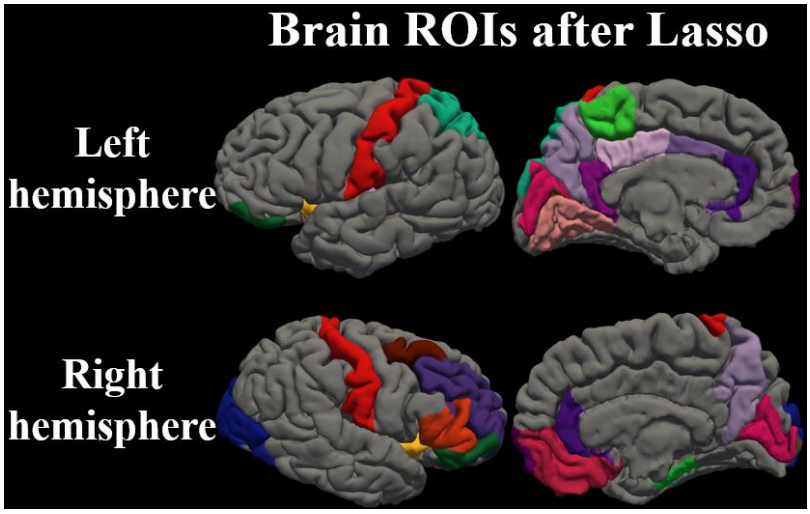

Figure 3. Cortical thickness measures in these 30 regions were retained by a distributed machine learning model for classifying individuals as having major depression or not. Corresponding brain areas in each hemisphere are shown with the same color [10]. The retention of limbic regions is in line with expectation, and consistent with the symptoms of depressed mood in MDD.

3.2 MDD classification

After feature selection with the distributed Lasso method, each data center received a list of selected features for classification. Here we adopted the Multilayer perceptron (MLP) classifier to categorize in individuals into the two groups– MDD versus control. One advantage of MLP is that it does not make any assumption regarding the underlying probability density functions of the input data. We performed 10-fold cross-validation using all the data at all 3 sites. **Table 2** summarizes classification results using these 30 brain measures. To evaluate the robustness of our method, we also compared the classification results using *all* the features without the Lasso regression process. All the classification results are reported with the best performance using the same classifier.

Table 2. Summary of classification results.

	Feature Number	General Accuracy	Specificity	Sensitivity
Without Distributed Lasso Regression	102	55.8%	69.1%	30.5%
With Distributed Lasso Regression	30	60.9%	76.4%	31.6%
Improvement		**5.1%**	**7.3%**	**1.1%**

With our distributed Lasso regression, the general classification accuracy increased from 55.8% to 60.9%. Both specificity and sensitivity improved when using only 29% of the original features, with the help of feature selection. This is the largest study for MDD classification to date, including MRI-derived brain measures from over 1,000 participants from three data centers. Note that we only considered T1-weighted MRI-based brain measures including cortical thickness, regional surface areas and subcortical volumes in this study. In the future, we hope to introduce data from more centers, as well as additional features from other modalities, such as DTI and fMRI, including connectivity measures and surface-based shape metrics. With these, we envision that the overall classification performance can be further improved.

4. CONCLUSIONS

In this paper, we applied our newly developed distributed Lasso method to MRI-derived brain data from multiple data centers to perform feature selection in a diagnostic classification task. We showed proof of concept of the proposed method. However, the results should be interpreted with caution, as we did not control for the difference in group size

resulting in relatively good classification of control subjects (high specificity) but not of MDD patients (low sensitivity); only a small proportion of datasets from ENIGMA MDD consortium were involved in this pilot study. In addition, almost one-third of potential features were excluded due to missingness. In the future, we will include more data from these ENIGMA centers, accommodate missing data and handle unequal group sizes, and the selected brain measures and classification accuracy may change slightly.

ACKNOWLEDGMENTS

This work is funded in part by NIH ENIGMA Center grant U54 EB020403, supported by the Big Data to Knowledge (BD2K) Centers of Excellence program.

REFERENCES

[1] World Health Organization. World Health Organization Depression Fact sheet No 369. (2012). Available from: http://www.who.int/mediacentre/factsheets/fs369/en/

[2] Schmaal, L., Hibar, D. P., Sämann, P. G., Hall, G. B., Baune, B. T., Jahanshad, N., ... & Vernooij, M. W., "Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group," **Molecular Psychiatry** (2016).

[3] Schmaal, L., Veltman, D. J., van Erp, T. G., Sämann, P. G., Frodl, T., Jahanshad, N., ... & Vernooij, M. W.," Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group," **Molecular Psychiatry**, 21(6), 806-812 (2016).

[4] Sacchet, M. D., Livermore, E. E., Iglesias, J. E., Glover, G. H., & Gotlib, I. H., "Subcortical volumes differentiate major depressive disorder, bipolar disorder, and remitted major depressive disorder," **Journal of Psychiatric Research**, 68, 91-98 (2015).

[5] Mwangi, B., Ebmeier, K. P., Matthews, K., & Steele, J. D., "Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder," **Brain**, 135(5), 1508-1521 (2012).

[6] Zeng, L. L., Shen, H., Liu, L., & Hu, D., "Unsupervised classification of major depression using functional connectivity MRI," Human Brain Mapping, 35(4), 1630-1641 (2014).

[7] Sacchet, M. D., Prasad, G., Foland-Ross, L. C., Thompson, P. M., & Gotlib, I. H., "Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory," **Frontiers in Psychiatry**, 6 (2015).

[8] Li, Q., Yang, T., Zhan, L., Hibar, D., Jahanshad, N., Wang, Y., Ye, J., Thompson, P., Wang, J., "Large-scale Collaborative Imaging Genetics Studies of Risk Genetic Factors for Alzheimer's Disease Across Multiple Institutions," **MICCAI**, in press (2016).

[9] Daubechies, I., Defrise, M., & De Mol, C., "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," **Communications on Pure and Applied Mathematics**, 57(11), 1413-1457 (2004).

[10] https://surfer.nmr.mgh.harvard.edu/