Short Papers

Semi-Supervised Discriminative Classification Robust to Sample-Outliers and Feature-Noises

Ehsan Adeli, *Member, IEEE*, Kim-Han Thung, Le An, Guorong Wu, *Member, IEEE*, Feng Shi, Tao Wang, and Dinggang Shen, *Fellow, IEEE*

Abstract—Discriminative methods commonly produce models with relatively good generalization abilities. However, this advantage is challenged in real-world applications (e.g., medical image analysis problems), in which there often exist outlier data points (*sample-outliers*) and noises in the predictor values (*feature-noises*). Methods robust to both types of these deviations are somewhat overlooked in the literature. We further argue that denoising can be more effective, if we learn the model using all the available labeled and unlabeled samples, as the intrinsic geometry of the sample manifold can be better constructed using more data points. In this paper, we propose a semi-supervised robust discriminative classification method based on the least-squares formulation of linear discriminant analysis to detect sample-outliers and feature-noises simultaneously, using both labeled training and unlabeled testing data. We conduct several experiments on a synthetic, some benchmark semi-supervised learning, and two brain neurodegenerative disease diagnosis datasets (for Parkinson's and Alzheimer's diseases). Specifically for the application of neurodegenerative diseases diagnosis, incorporating robust machine learning methods can be of great benefit, due to the noisy nature of neuroimaging data. Our results show that our method outperforms the baseline and several state-of-the-art methods, in terms of both accuracy and the area under the ROC curve.

Index Terms—Linear discriminant analysis, semi-supervised learning, robust classification, feature selection, sample outlier detection, Alzheimer's disease, Parkinson's disease, biomarker identification, disease diagnosis, nuclear norm, regularization

1 Introduction

Discriminative methods learn a mapping from the input feature space to the output label space for a task of classification (or regression). Such methods usually achieve good classification (or regression) results, compared to the generative methods, when there is enough number of training samples. But they carry out limited abilities when there are a small number of labeled data [1]. On the other hand, when noise contaminates the data, discriminative models usually fail to find an optimal mapping. In many real-world applications, the data are usually contaminated by different levels of noise. In some cases, a whole bunch of samples are affected (e.g., deviations in neuroimaging data due to radiation or patient movements during the imaging process), and therefore not useful for the learning task. These types of deviations are often denoted as *sample-outliers*. On the other hand, sometimes only some specific predictor values or features are infected, known as intra-sample-outliers (or *feature-noises*).

Various efforts have been made to add robustness to different learning methods. For instance, Suzumaura et al. [2] and Xu et al. [3] introduced robustness to the conventional support vector machine formulation by proposing various regularization terms or suppressing the influence of the outliers. In other works, Kim et al. [4] and Croux et al. [5] proposed robust variations of Fisher/Linear Discriminant Analysis (LDA) method, and Li et al. [6] introduced a worst-case LDA, by minimizing the upper bound of the LDA cost function. These methods are all robust to *sample-outliers*. On the other hand, some methods were proposed to deal with the *feature-noises*, such as [7], [8]. Many previous methods use Robust Principal Component Analysis (RPCA) [9], to deal with feature-noises in an unsupervised manner. Furthermore, many robust approaches that denoise the data while training the model do not offer straightforward strategies to deal with the testing data. Often, the denoising procedure of the training and the testing data are conducted separately (e.g., in [10]), which might induce a bias to the whole learning process. One solution is to denoise the training and the testing data together, provided that the testing data are available. Therefore, we propose to take advantage of them as unlabeled data during the training phase. Under such semi-supervised setting, the constructed discriminative model can be more reliable, particularly for the cases with the small-sample-size problem. This could be attributed to the fact that more samples are being used to model the intrinsic geometry of the sample manifold.

The main application we are anticipating in this paper is the diagnosis of neurodegenerative diseases, based on neuroimaging data. This is a challenging problem, as the data is pretty much prone to noise and often there is a limited number of samples. Hence, there is a calling need for robust machine learning methods for such applications. Neurodegenrative diseases are debilitating and incurable conditions caused by progressive degeneration or death of the cells in the brain nervous system. These diseases affect millions of people around the world. Alzheimer's Disease (AD) and Parkinson's Disease (PD) are among the most common types. Although neurodegenerative diseases manifest with diverse pathological features, the cellular level processes resemble similar structures. For instance, in AD, deposits of tiny protein plaques result into brain damage and progressive loss of memory [11], while PD is mainly initiated by a selective loss of dopaminergic neurons in the Substantia Nigra (SN) brain region, leading to declining in the generation of a chemical messenger, dopamine. Lack of dopamine yields loss of ability to control body movements, along with several non-motor problems (e.g., depression, and anxiety) [12]. These diseases are often incurable; thus, early diagnosis and treatment are crucial to slow down their progression in the initial stages.

The challenges for building reliable diagnosis models include: (1) It is usually burdensome to acquire noise-free imaging data from the patients. Different sources of noise may affect the acquired data, including a wide variety of noises in the neuroimage acquisition procedure, the imposed artifacts due to preprocessing, and the large amount of inter-subject variabilities; (2) To build a good diagnosis

- *E. Adeli is with Stanford University, Stanford, CA 94305, and with the Biomedical Research Imaging Center (BRIC), Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599. E-mail: eadeli@stanford.edu.*
- *K.-H. Thung, L. An, and G. Wu are with the Biomedical Research Imaging Center (BRIC), Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599. E-mail: khthung@email.unc.edu, {le_an, grwu}@med.unc.edu.*
- *F. Shi is with the Biomedical Imaging Research Institute, Cedars Sinai Medical Center, Los Angeles, CA 90048. E-mail: fengshi@med.unc.edu.*
- *T. Wang is with the Department of Geriatric Psychiatry, Shanghai Mental Health Center and the Alzheimer's Disease and Related Disorders Center, Shanghai Jiao Tong University, Shanghai 200000, China. E-mail: wtshhwy@163.com.*
- *D. Shen is with the Biomedical Research Imaging Center (BRIC), Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, and with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea. E-mail: dgshen@med.unc.edu.*

Manuscript received 17 June 2016; revised 8 Dec. 2017; accepted 7 Jan. 2018. Date of publication 16 Jan. 2018; date of current version 16 Jan. 2019.
(Corresponding author: Dinggang Shen.)
Recommended for acceptance by J. Ye.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2018.2794470

model, through learning a classifier, we need a sufficiently large number of labeled subjects. However, acquiring reliably enough labeled data is costly and time-consuming. Therefore, models that can take advantage of unlabeled data (subjects that we are not certain about their disease) could be of great interest; (3) Different neurodegenerative diseases often affect different regions of the brain, i.e., only certain regions of the brain are associated with the disease. Thus, using all features can undermine the diagnosis performance, and we need to identify the imaging biomarkers for each specific disease while learning the diagnosis model.

To deal with the aforementioned challenges, we propose a semi-supervised discriminative classifier, to take advantage of the available unlabeled testing data. This leads to a more substantial number of samples, which can yield better modeling of the intrinsic geometry of the sample manifold. As a result, our model jointly estimates the noise model (both sample-outliers and feature-noises) on the whole labeled training and unlabeled testing data and simultaneously builds a discriminative model upon the denoised training data. Unlike many previous works on denoising medical images, we do not define the problem of denoising separately from the analysis part. In the sense that if a sample (or a feature value) does not act in accordance with others in building the model, it should be counted as a sample-outlier (or a feature-noise). This observation suggests that intertwining the denoising procedure with the learning framework will help to identify the sample-outliers and feature-noises more efficiently while learning a robust classification model. It is important to note that denoising and outlier detection has a long history in the area of medical image analysis and computing. The inter- and intra-subject variabilities, the noise sourced from the images devices, and the pre-processing errors emerge the study of robust methods for analyzing medical imaging data. For instance, in the recent years, several attempts have been made for denoising the medical images [13], [14], [15], [16] or detecting outliers [17], [18], as a preprocessing step to any analysis on medical images.

1.1 Background and Overview of the Proposed Method

In this paper, we introduce a novel classification model based on LDA, which is robust against both sample-outliers and feature-noises, referred to as robust feature-sample linear discriminant analysis (RFS-LDA). The original LDA formulation finds the mapping between the sample space and the label space through a linear transformation matrix, maximizing a so-called Fisher discriminant ratio [4]. In practice, the major drawback of LDA is the small-sample-size problem, which arises when the number of available training samples is much less than the dimensionality of the feature space [19]. Original LDA finds the mapping by incorporating covariance matrices of the input feature matrices. In cases where the number of samples is much less than the number of features, these matrices are probably rank-deficient [20]. A reformulation of LDA based on the reduced-rank least-squares problem (known as LS-LDA) [20] tackles this problem. LS-LDA finds the mapping $\boldsymbol{\beta} \in \mathbb{R}^{l \times m}$ by solving the following problem:

$$\min_{\boldsymbol{\beta}} \|(\mathbf{Y}_{tr}\mathbf{Y}_{tr}^{\top})^{-1/2}(\mathbf{Y}_{tr} - \boldsymbol{\beta}\mathbf{X}_{tr})\|_{\mathrm{F}}^{2}, \tag{1}$$

where $\mathbf{Y}_{tr} \in \mathbb{R}^{l \times N_{tr}}$ is a binary class label indicator matrix, for l different classes (or labels), and $\mathbf{X}_{tr} \in \mathbb{R}^{m \times N_{tr}}$ is the matrix containing N_{tr} m-dimensional training samples. $(\mathbf{Y}_{tr}\mathbf{Y}_{tr}^{\top})^{-1/2}$ is a normalization factor that compensates for the different number of samples in each class [20]. As a result, the mapping $\boldsymbol{\beta}$ is a reduced rank transformation matrix [8], [20], which could be used to project a test data $\mathbf{x}_{tst} \in \mathbb{R}^{m \times 1}$ onto an l-dimensional space. Note that directly minimizing (1) avoids the small-sample-size problem by not using the covariance matrices. After it projects the samples to the output space, we need a simple step to infer the class labels. LDA maximizes inter-class variance, while minimizing the intra-class variance, in the mapped space. Thus, we expect that in the mapped space, same-class samples to be closer to each other. The class labels could, therefore, be simply determined using a k-NN strategy.

To make LDA robust against noisy data, Fidler et al. [7] estimate a robust basis, which consists all the discriminative information for classification or regression. In the testing phase, the estimated basis identifies the outliers in samples (images in their case) and then calculates the coefficients using a subsampling approach. On the other hand, Huang et al. [8] proposed a general formulation for Robust Regression (RR) and classification (i.e., Robust LDA or RLDA), where, they first denoise the training feature values using a strategy similar to RPCA [9], and then build the above LS-LDA model using the denoised data. In the testing stage, they denoise the testing samples using the denoised training data. This separate denoising procedure could not effectively form the underlying geometry of sample space to denoise the data. Furthermore, RR [8] only accounts for feature-noises by imposing a sparse noise model constraint on the features matrix, despite the fact that the least-squares data fitting term in (1) is vulnerable to large sample-outliers.

Recently, in robust statistics, it is found that ℓ_1 functions are able to make more reliable estimations [21] than ℓ_2 least-squares fitting functions. This has been previously adopted in many applications, including robust face recognition [22] and robust dictionary learning [23]. Reformulating the objective in (1) with ℓ_1 loss entails the following problem:

$$\min_{\boldsymbol{\beta}} \|(\mathbf{Y}_{tr}\mathbf{Y}_{tr}^{\top})^{-1/2}(\mathbf{Y}_{tr} - \boldsymbol{\beta}\mathbf{X}_{tr})\|_{1}. \tag{2}$$

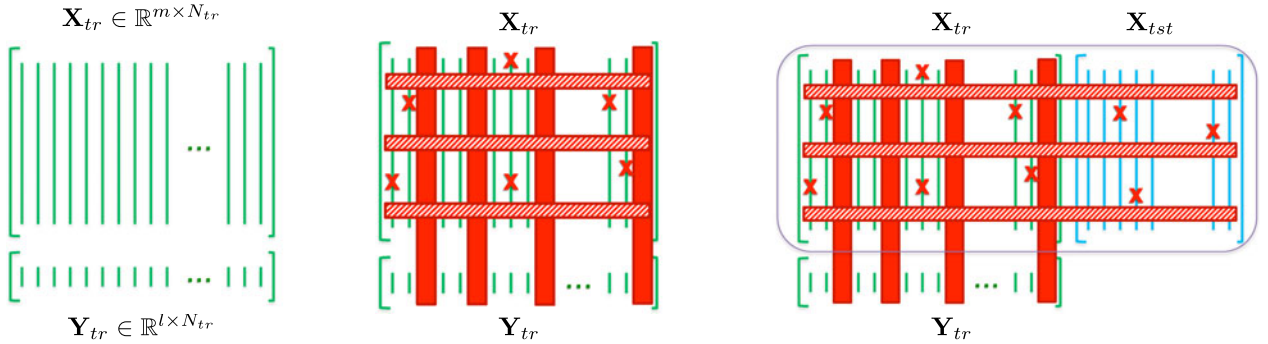
We incorporate this fitting function to deal with the sample-outliers, in this paper. We also adopt a strategy to simultaneously denoise the data from feature-noises. This is done through a semi-supervised setting to take advantage of all labeled and unlabeled data, and build the structure of the sample space more robustly. Fig. 1 illustrates this idea, in which Fig. 1a shows a traditional learning problem. However, if the data contains sample-outliers or some samples suffer from noise in their feature values (Fig. 1b), traditional methods usually fail to build reliable models.

Semi-supervised learning has long been of great interest in different fields, because it can make use of unlabeled or poorly labeled data to achieve better prediction models [24], [25]. For instance, Joulin and Bach [26] introduced a convex relaxation and used their model in different semi-supervised learning scenarios. In another work, Cai et al. [27] proposed a semi-supervised discriminant analysis, where the separation between different classes is maximized using the labeled data points, while the unlabeled data points estimate the structure of the data. Belkin et al. [28] similarly used the unlabeled data for regularization. In contrast, we incorporate the unlabeled testing data in our formulation to better estimate the intrinsic geometry of the sample manifold and denoise the data, while building the discriminative model upon the labeled training data. By incorporating the unlabeled testing data (Fig. 1c), we learn the classification model, while denoising both training and testing data and detecting sample-outliers.

We apply our method for the diagnosis of neurodegenerative brain disorders. Specifically, in this study, we use two popular databases: PPMI [29] and ADNI [30]. The former aims at investigating PD and its related disorders, while the latter is designed for diagnosing AD and its prodromal stage, known as Mild-Cognitive Impairment (MCI). In addition, to validate the proposed method, we further conduct experiments on synthetic data, as well as some benchmark datasets for semi-supervised learning.

1.2 Contributions

The contributions of this paper are multi-fold: (1) We propose an approach to dealing with the sample-outliers and feature-noises



(a) A traditional supervised learning task, defined as finding a mapping from training samples, $\mathbf{X}_{tr}$, to their respective labels, $\mathbf{Y}_{tr}$.

(b) A robust classifier needs to deal with both sample-outliers (red bars) and feature-noises (red crosses). In some applications (like ours), discarding non-discriminant features (shaded bars) is also crucial.

(c) In order to form the underlying geometry of the sample manifold with largely enough number of samples, we incorporate the unlabeled testing data. This could contribute to denoising both the training and testing samples, while learning the mapping from the training samples to their respective labels.

Fig. 1. Overview of the proposed semi-supervised learning framework, robust to both sample-outliers and feature-noises.

simultaneously and build a robust discriminative classifier. The sample-outliers are penalized through an ℓ_1 fitting function. (2) Our proposed model operates under a semi-supervised setting, where the whole data (i.e., labeled training, and unlabeled testing samples) are incorporated to build the intrinsic geometry of the sample space, which leads to better data denoising. (3) We further select the most discriminative features for the learning process through regularizing the weights matrix with an ℓ_1 norm. This is especially of great interest for the neurodegenerative disease diagnosis, where the features from different regions of the brain are extracted, but not all the regions are associated with a certain disease. Thus, the most discriminative regions associated with the disease would be identified, leading to a more reliable diagnosis model.

2 The Proposed Method: RFS-LDA

Suppose we have N_{tr} training and N_{tst} testing samples, each with a m dimensional feature vector, which leads to a set of $N = N_{tr} + N_{tst}$ total samples. Let $\mathbf{X} \in \mathbb{R}^{m \times N}$ denote the set of all samples (both training and testing), in which each column indicates a single sample, and also let $\mathbf{y}_i \in \mathbb{R}^{1 \times N}$ their corresponding ith labels. In general, with l different labels, we can define $\mathbf{Y} \in \mathbb{R}^{l \times N}$. Thus, $\mathbf{X}$ and $\mathbf{Y}$ are composed by stacking up the training and testing data as: $\mathbf{X} = [\mathbf{X}_{tr}\ \mathbf{X}_{tst}]$ and $\mathbf{Y} = [\mathbf{Y}_{tr}\ \mathbf{Y}_{tst}]$. Our goal is to determine the labels of the test samples, $\mathbf{Y}_{tst} \in \mathbb{R}^{l \times N_{tst}}$.

Note that, throughout the paper, bold capital letters denote matrices (e.g., $\mathbf{A}$), while bold lowercase letters denote vectors (e.g., $\mathbf{a}$). All non-bold letters denote scalar variables. a_{ij} is the scalar in the row i and column j of $\mathbf{A}$. $\langle \mathbf{a}_1, \mathbf{a}_2 \rangle$ denotes the inner product between $\mathbf{a}_1$ and $\mathbf{a}_2$. $\|\mathbf{a}\|_2^2 = \langle \mathbf{a}, \mathbf{a} \rangle = \sum_i a_i^2$ and $\|\mathbf{a}\|_1 = \sum_i |a_i|$ represent the squared euclidean norm and the ℓ_1 norm of $\mathbf{a}$, respectively. $\|\mathbf{A}\|_{\mathrm{F}}^2 = (\mathbf{A}^\top \mathbf{A}) = \sum_{ij} a_{ij}$, $\|\mathbf{A}\|_{1,1} = \sum_j \sum_i |a_{ij}|$ and $\|\mathbf{A}\|_*$ designate the squared Frobenius norm, $\ell_{1,1}$ norm and the nuclear norm (sum of singular values) of $\mathbf{A}$, respectively. $\mathbf{I}_K \in \mathbb{R}^{K \times K}$ denotes the identity matrix.

2.1 Formulation

All the available samples, both labeled and unlabeled, are arranged into a matrix, $\mathbf{X} \in \mathbb{R}^{m \times N}$, each of whose columns represents the feature vector of a sample. To achieve a robust classifier, we seek to denoise this matrix. Following [31], [32], this could be done by assuming that $\mathbf{X}$ can be spanned on a low-rank subspace and therefore should be rank-deficient. This assumption supports the fact that samples from same classes are more correlated [8], [32] and linearly-dependent. Accordingly, the original matrix $\mathbf{X}$ is decomposed into the summation of two counterparts, $\mathbf{D} \in \mathbb{R}^{m \times N}$ and $\mathbf{E} \in \mathbb{R}^{m \times N}$. The former represents the denoised data matrix, while the latter is the error matrix. This is similar to RPCA [9], used in many computer vision applications. With this decomposition, we can assume that the denoised data matrix shall be rank-deficient and the error matrix sparse.

But as one can easily infer, this process of denoising does not incorporate the label information and is, therefore, unsupervised. Nevertheless, recall that we are also seeking a mapping between the denoised training samples and their respective labels. So, matrix $\mathbf{D}$ should be spanned on a low-rank subspace that would lead to a good classification model of its sub-matrix, $\mathbf{D}_{tr}$. We incorporate the regression model in (2) as the fitting function to compute a mapping $\boldsymbol{\beta}$. A schematic illustration of the proposed method is depicted in Fig. S1 of the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2018.2794470.

To ensure the rank-deficiency of the matrix $\mathbf{D}$, like many previous works [9], [31], [32], we approximate the rank function using the nuclear norm (i.e., the sum of the singular values of the matrix). The noise is modeled using the ℓ_1 norm of the matrix, which ensures a sparse noise model on the feature values. Accordingly, the objective function for RFS-LDA under a semi-supervised setting would be formed as

$$\begin{aligned} \min_{\boldsymbol{\beta}, \mathbf{D}, \hat{\mathbf{D}}, \mathbf{E}} \quad & \frac{\eta}{2} \|\mathbf{Y}_{tr} - \boldsymbol{\beta} \hat{\mathbf{D}}\|_1 + \|\mathbf{D}\|_* + \lambda_1 \|\mathbf{E}\|_1 + \lambda_2 \mathcal{R}(\boldsymbol{\beta}), \\ s.t. \quad & \mathbf{X} = \mathbf{D} + \mathbf{E}, \hat{\mathbf{D}} = [\mathbf{D}_{tr}; \mathbf{1}^\top], \end{aligned} \tag{3}$$

where the first term is the ℓ_1 regression model introduced in (2). This term only operates on the denoised training samples from matrix $\mathbf{D}$ with a row of all 1's added to it (denoted as $\hat{\mathbf{D}}$), to counter for the bias in the linear model. The second and third terms, together with the first constraint, are similar to the RPCA formulation [9]. They denoise the labeled training and unlabeled testing data together, and in combination with the first term, we ensure that the denoised data also specifies a favorable regression. The last term is a regularization on the learned mapping coefficients, to avoid trivial or unexpectedly large values. The hyperparameters η, λ_1 and λ_2 are the scalar regularization hyperparameters, which will be discussed in detail later.

The regularization on the coefficients could be posed as a simple norm of the matrix, $\boldsymbol{\beta}$. But, in many applications, like ours (disease diagnosis), many of the features in the feature vectors are redundant. This is because we extract features from different brain regions, but not all the regions contribute to a certain disease.

Therefore, it is desirable to determine which features are the most relevant and the most discriminative for the task. Following [11], [22], [33], we are seeking a sparse set of weights that ensures incorporating the most discriminative features. Therefore, we propose a regularization on the weights matrix as a combination of the ℓ_1 and Frobenius norms

$$\mathcal{R}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_{1,1} + \gamma\|\boldsymbol{\beta}\|_{\mathrm{F}}. \tag{4}$$

Evidently, the solution to the objective function in (3) is not easy to achieve. This is because it contains a quadratic term, and the minimization of the ℓ_1 fitting function is not straightforward, due to its indifferentiability. To this end, we formalize the solution with a similar strategy as in Iteratively Re-weighted Least Squares (IRLS) [21]. The ℓ_1 fitting term is approximated by a conventional ℓ_2 least-squares, in which each of the samples in the $\hat{\mathbf{D}}$ matrix is weighted with the reverse of their regression residual. Additionally, since we regularize the weights $\boldsymbol{\beta}$ using a combination of ℓ_1 and ℓ_2 norms, the non-zero elements would represent the selected features by the algorithm. In order to reflect this to feature denoising scheme, we define a projection operator $\mathcal{P}_{\boldsymbol{\beta}}(.)$. This operator projects the values of the non-selected features (respective to zero values in $\boldsymbol{\beta}$) to zero, to decrease their effect in minimizing the rank of the matrix $\mathbf{D}$ (in the second term). Therefore, the new problem would be

$$\begin{aligned} \min_{\boldsymbol{\beta},\mathbf{D},\hat{\mathbf{D}},\mathbf{E}} \quad & \frac{\eta}{2}\|(\mathbf{Y}_{tr} - \boldsymbol{\beta}\hat{\mathbf{D}})\hat{\boldsymbol{\alpha}}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\boldsymbol{\beta}}(\mathbf{D})\|_* + \lambda_1\|\mathbf{E}\|_1 + \lambda_2\mathcal{R}(\boldsymbol{\beta}), \\ s.t. \quad & \mathbf{X} = \mathbf{D} + \mathbf{E}, \hat{\mathbf{D}} = [\mathbf{D}_{tr}; \mathbf{1}^\top], \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\alpha}}$ is a diagonal matrix, the ith diagonal element of which is the ith sample's weight

$$\hat{\boldsymbol{\alpha}}_{ii} = \frac{1}{\sqrt{(\mathbf{y}_i - \boldsymbol{\beta}\hat{\mathbf{d}}_i)^2 + \delta}}, \forall\, i,j \in \{0,\ldots,N_{tr}\}, i \neq j, \hat{\boldsymbol{\alpha}}_{ij} = 0. \tag{6}$$

Hyperparameter δ is a small positive number (10^{-4} in our experiments), to prevent from any division by zeros in (6). In the next section, we introduce an algorithm to solve this optimization problem.

Our work is closely related to the RR formulations in [8], where the authors impose a low-rank assumption on the training data feature values and an ℓ_1 assumption on the noise model. The discriminant model is learned similarly to LS-LDA, as described in (1). Whereas, we observed that to have a more robust regression model, we need to establish a strategy where we can weight the samples. This is because the ℓ_1 noise model in [8] can only discard a controlled amount of sparse noise in the feature values, not the whole samples. On the other hand, our model operates under a semi-supervised setting, where both labeled training and unlabeled testing samples are denoised simultaneously, leading to a more robust denoising model. Also, our model further selects the most discriminative features to learn the regression model, by regularizing the learned weights and enforcing a sparsity condition on them.

To optimize the objective function in (5), we use the Alternating Direction Method of Multipliers (ADMM) [34]. The detailed optimization steps, along with the comprehensive analysis of the algorithm, its convergence properties and an upper bound for the time complexity of the proposed algorithm are provided in the supplementary material, available online.

3 Experiments

To evaluate the proposed approach, we compare our method against several baselines and state-of-the-art methods in different scenarios. The first experiment evaluates our method on a synthetic set of data, which highlights how the proposed method is robust against sample-outliers or feature-noises separately, or when they occur at the same time. Then we employ some benchmark semi-supervised learning datasets and report results in comparisons with some baseline and state-of-the-art methods. The results of these two experiments (i.e., on *synthetic* and *benchmark data*) are reported in the *supplementary material*, available online. We then apply the proposed RFS-LDA method to the problem of neurodegenerative brain disorder and disease diagnosis.

For the choice of hyperparameters, a set of possible values are first predefined, and the best hyperparameters are selected through 10-fold cross-validation, for all the competing methods. The RFS-LDA hyperparameters (as in Eq. (5)) are set with the same strategy as in [8]

$$\lambda_1 = \frac{\Lambda_1}{(\sqrt{\min(m,N)})},\ \lambda_2 = \frac{\Lambda_2}{\sqrt{m}},\ \eta^k = \frac{\Lambda_3\|\mathbf{X}\|_*}{\|\mathbf{Y}_{tr} - \boldsymbol{\beta}^k\hat{\mathbf{D}}^k\|_{\mathrm{F}}^2}, \tag{7}$$

and ρ (controlling the $\{\mu\}$s in the iterative optimization algorithm) is set to 1.01. We have set $\Lambda_1, \Lambda_2, \Lambda_3$ and γ through inner-cross-validation grid-search in the range $[10^{-4}, 10]$.

3.1 Datasets

In this study, we use two real-world databases for two different brain neurodegenerative diseases, namely PD and AD. The first set of data is obtained from the Parkinson's Progression Markers Initiative (PPMI) database [29], with the MRI data from 374 PD and 169 normal control (NC) subjects. The second dataset comes from the Alzheimer's disease neuroimaging initiative (ADNI) database, which includes MRI and FDG-PET data. We used 93 AD patients, 202 MCI patients, and 101 NC subjects, each with complete MRI and FDG-PET data. The subjects' brain images are preprocessed and regions of interest (ROI) features are extracted for each subject. For more detailed information about these two datasets and the preprocessing steps for feature extraction refer to the *supplementary material*, available online.

3.2 Baseline Methods

We compare our proposed method with different baseline methods, including the conventional LS-LDA [20], RLDA [8], and linear Support Vector Machine (SVM). Another baseline method can be defined as running the same procedures as in the proposed method but disjointly. Therefore, we apply RPCA on the matrix $\mathbf{X}$ separately to first denoise, and then classify the denoised data using LS-LDA (denoted as RPCA+LS-LDA) [8]. To analyze the effectiveness of the feature selection strategy of the proposed method, we also include baseline methods which use sparse feature selection (SFS) together with SVM (SFS+SVM), and RLDA (SFS+RLDA). Except for RPCA+LDA, the other methods in comparison do not incorporate the testing data. In order to have a fair set of comparisons, we also compare against the transductive Matrix Completion (MC) approach [32] and the semi-supervised formulation of SVM (S^3VM) [35]. These two methods incorporate the unlabeled testing data in the process of training their models. Additionally, in order to further evaluate the effect of the ℓ_1 norm regularization on the weights matrix $\boldsymbol{\beta}$, we also report results for RFS-LDA when regularized by only $\gamma\|\boldsymbol{\beta}\|_{\mathrm{F}}$ (denoted as RFS-LDA*), rather than the regularization term introduced in (4). Finally, we report results using the supervised version of our proposed method, which is denoted as supervised RFS-LDA (S-RFS-LDA). In S-RFS-LDA, we train our model using only the training data, where $\mathbf{X}$ in (5) is replaced with $\mathbf{X}_{tr}$. In this way, we can examine the effect of using unlabeled testing data in the prediction model.

3.3 Disease Diagnosis

We evaluate our method with two popular datasets for neurodegenerative disease diagnosis, PPMI and ADNI, for diagnosis of PD

TABLE 1
Diagnosis Accuracy of the Proposed Method (RFS-LDA) and the Baseline Methods on Both PPMI and ADNI Datasets

	RFS-LDA	RFS-LDA*	S-RFS-LDA	RLDA	SFS+RLDA	RPCA+LS-LDA	LS-LDA	SVM	SFS+SVM	S^3VM	MC
PD versus NC	**79.8**	76.1	74.1	68.3	70.5	61.0†	58.5†	66.1†	69.2	71.5	70.6
AD versus NC	**92.1**	89.8	88.3	87.8	90.0	87.6	82.7	85.4	87.1	**90.5**	88.7
MCI versus NC	**81.9**	80.6	79.1	79.5	**80.9**	76.9	72.3†	74.1	78.3	**80.8**	76.1

The † sign indicates a p-value > 0.05 in a Fisher exact test.

and AD, respectively. These datasets, subject information, preprocessing steps, and feature extraction are explained in Section C of the supplementary material, available online.

Results. The first row in Table 1 shows the diagnostic accuracy of the proposed technique (RFS-LDA) in comparisons with different baseline and state-of-the-art methods using 10-fold cross-validation. The results show that the proposed method outperforms all others. This can be attributed to the fact that our method better deals with feature-noises and sample-outliers. Recall that samples and their corresponding feature vectors extracted from the neuroimaging data are quite prone to noise, as discussed earlier. Therefore, some of the samples might not be useful, and some might be contaminated by a certain amount of noise. Our method can deal with both types of noises, as supported by the results. The second disease diagnosis experiment is conducted on ADNI, in which the goal is to discriminate normal controls from mild cognitive impairment and AD subjects. Therefore, NC subjects form our negative class, while the positive class is defined as AD in one experiment, and MCI in the other. The diagnosis results of the AD versus NC and MCI versus NC are reported in the second and third rows in Table 1, respectively. As it can be seen, in comparison with the state-of-the-art, our method achieves better results in terms of both accuracy and the area under ROC curve.

It is worth noting that running the model using a 10-fold cross-validation for the PD versus NC (543 subjects), AD versus NC (194 subjects), and MCI versus NC (303 subjects) experiments on a PC (Intel Core i7 @ 2.30 GHz and 8.00 GB of memory), with a parallel implementation in MATLAB (i.e., using `parfor` for 4 workers) took approximately 6, 2 and 3.5 hours, respectively. Additionally, to test the statistical significance of the obtained results, we further conducted a Fisher exact test [36] on the accuracy score achieved by each of the methods. This test verifies that the method is significantly more accurate (with a p-value of $p < 0.05$) than randomly assigning the samples to the two classes. The results of this statistical test indicated that the proposed method achieves a p-value of even less than 0.001. This shows that there are no random associations with the obtained results. However, for some of the compared baseline methods, a p-value of $p > 0.05$ was observed, which is not appealing. These methods are marked with a † sign in Table 1. It is important to note that the comparisons between the supervised (S-RFS-LDA) and the semi-supervised (RFS-LDA) versions of the proposed algorithm in both Table 1 and Fig. S5 of the supplementary material, available online, show that including the unlabeled testing data improves the results by a relatively notable margin. This can be because including more samples gives us a better representation of the sample manifold, leading to better denoising of simultaneous training and testing data, in a way that a better classifier is built.

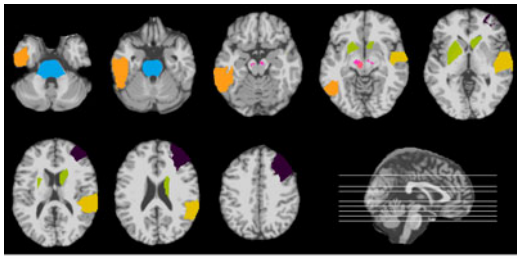
Although the studies on Parkinson's disease using modern machine learning techniques are scarce, there are quite a few studies in the literature for Alzheimer's disease. State-of-the-art machine learning approaches for this purpose either aim at developing feature selection techniques or focus on designing delicate classifiers. The first type usually use sophisticated techniques for feature selection [37], [38], feature learning [39], or feature extraction[40], [41], [42] and then an straightforward classification technique (like SVM) is utilized. The second type develops task-specific classifiers to enhance the classification accuracies, e.g., [43], [44], [45]. In contrast, our method constructs the sample manifold using all labeled and unlabeled data to denoise the features and also selects the best features for classification, with a classification loss robust to sample-outliers. In Table 2, we compare our method with several state-of-the-art methods for Alzheimer's disease diagnosis. The table includes all the information about the dataset and the methods they used for obtaining those results. This is only to show where our method stands among the previous works in the same field.

As discussed earlier, in medical imaging applications many sources of noise contribute to the acquired data, and therefore methods that can deal with noise and outliers are of great interest. Our method enjoys from a single optimization objective that can simultaneously suppress sample-outliers and feature-noises, which, compared to other methods, exhibits a good performance. One of the interesting functions of the proposed method is the regularization on the mapping coefficients with the ℓ_1 norm, which would select a compact set of features to contribute to the learned mapping. The magnitude of the coefficients would show the relevance of the specific features for building the prediction model. In our application, the features from the whole brain regions are extracted, but not all the ROIs are associated with the disease (e.g., AD, MCI or PD). By exploring the learned coefficients by our method, we can determine which brain regions are highly associated with a certain disease.

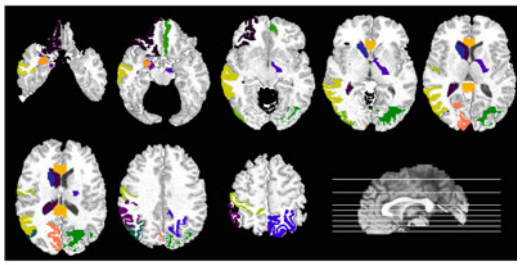
TABLE 2
Comparisons of the Proposed Method with State-of-the-Art Methods for Diagnosis of AD and MCI

Method	Subjects			Methodology	Modalities	AD versus NC (%)	MCI versus NC (%)
	AD	MCI	NC				
Liu et al. [45]	198	N/A	229	Voxel GM+SVM Ensemble	MRI	**92.0**	N/A
Cuingnet et al. [42]	137	N/A	162	Voxel Direct D+SVM	MRI	88.58	N/A
Eskildsen et al. [41]	194	N/A	226	Cortical Thickness+SVM	MRI	84.50	N/A
Duche. et al. [40]	75	N/A	75	Tensor-based Morphometry+SVM	MRI	**92.0**	N/A
Min et al. [37]	97	N/A	128	Multi-Atlas ROI Features+SVM	MRI	91.6	N/A
Gary et al. [44]	37	75	35	Random Forest	MRI+PET+CSF+Gen	89.0	74.6
Tong et al. [43]	35	75	77	Graph Fusion	MRI+PET+CSF+Gen	91.8	79.5
Liu et al. [39]	85	169	77	Deep Feature Learning	MRI+PET	91.4	**82.1**
Ours	93	202	101	RFS-LDA	MRI+PET	**92.1**	**81.9**

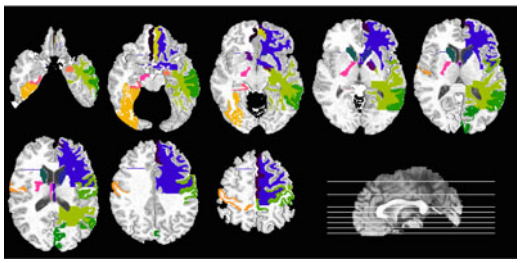
N/A: indicates that the methods did not report results for that experiment; CSF: cerebrospinal fluid; Gen: Categorical genetic information.



(a) PD *vs.* NC



(b) AD *vs.* NC



(c) MCI *vs.* NC

Fig. 2. Top selected regions for each experiment. Selected regions are shown with different colors for clarity.

Identification of Disease Biomarkers. To extract these most relevant ROIs, we select the ROIs that were given larger weights in 50 percent of the ten repetitions of the 10-fold cross-validation tests. Fig. 2a visualizes the most relevant regions for PD on a raw brain template, including the middle frontal gyrus right, pons, substrata nigra left and right, red nucleus left, pallidum left, pautmen left, caudate right, inferior temporal left, and superior temporal gyrus right. As in the previous studies in the literature [46], [47], deep brain and striatum areas are known to play crucial roles for PD. Our study also confirms these clinical findings. Same experimental settings for AD and MCI identifies the top regions selected by our algorithm in AD versus NC and MCI versus NC classification scenarios (Figs. 2b and 2c, respectively). These regions, including middle temporal gyrus, medial front-orbital gyrus, postcentral gyrus, caudate nucleus, cuneus, and amygdala, have also been reported to be associated with AD and MCI in the literature [11], [48]. The analysis of such selection of brain regions can be further incorporated for future clinical studies.

Method Discussions. To analyze the effect of the sample-outlier detection in the proposed framework, we employ a dimensionality reduction technique to facilitate the visualization of the data points. We project the samples of the AD versus NC experiment into the 2-D space using t-SNE [49]. The t-SNE projection technique visualizes high-dimensional data by giving each sample a location in a two-dimensional map. The map created by the t-SNE reveals the neighborhood structure of the sample manifold at many different scales [49]. This is particularly important for our application, in which the high-dimensional neuroimaging data lie on several different low-dimensional manifolds since the samples come from different subjects with or without the neurodegenerative disease. Fig. 3 shows

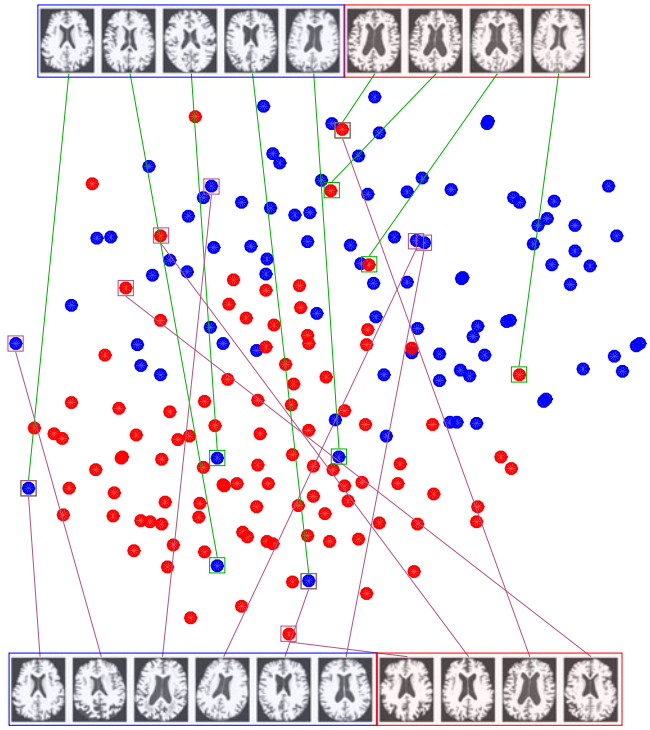


Fig. 3. t-SNE projection of AD versus NC samples (better viewed in color). Top: Samples detected as outliers by our method. Bottom: Samples detected as outliers using RANSAC [50].

the t-SNE projection in the 2-D space. In this figure, the samples, which received the smallest weights in their respective elements in the $\hat{\alpha}$ weight matrix (as in Eq. (5)), are shown in the top part of the figure. We also depict the samples detected as outliers using the RANSAC [50] algorithm in the bottom part of the figure. Notably, as it is obvious in the figure, the samples detected as sample-outliers by our algorithm are those which are more controversial for the task of classification and lie outside the main neighborhood of each class. This is attributed to the fact that we detect them jointly with the classifier learning framework. On the other hand, the outliers detected by RANSAC are not always the best in terms of discriminability. This suggests that unsupervised outlier detection methods might not perform well when the aim is to learn a classifier or a regression model. In other words, in many learning tasks, the definition for sample-outliers might be different based on what the goal is.

One of the important hyperparameters in the proposed RFS-LDA is λ_1, as in Eq. (5), which controls the noise term. Modifying this hyperparameter leads to altered noise levels, detected by our algorithm. To analyze its effect on the learning performance, we fix all other hyperparameters and run the algorithm with different values of Λ_1, and therefore λ_1 (as discussed at the beginning of Section 3). The changes in the AUC for each of our experiments are illustrated in Fig. 4. As can be seen, the proposed method achieves

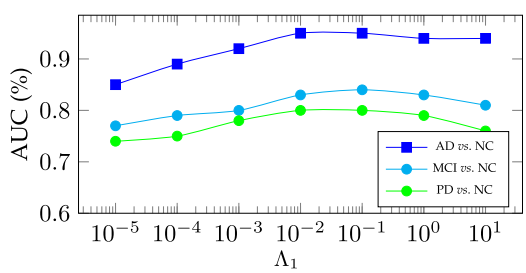


Fig. 4. Area under the ROC curve (AUC) as a function of the RFS-LDA hyperparameter Λ_1, related to λ_1 in (3).

reasonably good results with a wide range of the values of the hyperparameter.

It is worth noting that the proposed method works under a semi-supervised setting, which is interesting for the application of disease diagnosis. When performing the diagnosis for new patients, all subjects whose clinical diagnosis has not been finalized (i.e., they are still in the process of evaluations and clinical monitoring) can yet be included in model building as unlabeled samples, to build a potentially more reliable classifier.

4 Conclusion

In this paper, we proposed a novel approach for discriminative classification, which is robust against both sample-outliers and feature-noises. Our method enjoys a semi-supervised setting, where all the labeled training and the unlabeled testing data are used to detect outliers and are denoised simultaneously. We have applied our method to several datasets, including synthetic, semi-supervised learning benchmark, and neurodegenerative brain disease diagnosis datasets, specifically for Parkinson's disease and Alzheimer's disease. The results showed that our method outperformed all competing techniques. As a direction for the future works, one can develop a multi-task learning reformulation of the proposed method to incorporate diagnosis from multiple modalities of neuroimaging data or extend the approach for the case of incomplete data.

Acknowledgments

Parts of the data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.edu), and Parkinson's Progression Markers Initiative (PPMI) database (http://www.ppmi-info.org). The investigators within the ADNI or PPMI contributed to the design and implementation of the datasets and/or provided data but did not participate in analysis or writing of this paper. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how to apply/ADNI Acknowledgement List.pdf.

References

[1] A. Jordan, "On discriminative versus generative classifiers: A comparison of logistic regression and naive Bayes," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2002, Art. no. 841.

[2] S. Suzumura, K. Ogawa, M. Sugiyama, and I. Takeuchi, "Outlier path: A homotopy algorithm for robust SVM," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1098–1106.

[3] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *J. Mach. Learn. Res.*, vol. 10, pp. 1485–1510, 2009.

[4] S.-J. Kim, A. Magnani, and S. Boyd, "Robust fisher discriminant analysis," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 659–666.

[5] C. Croux and C. Dehon, "Robust linear discriminant analysis using S-estimators," *Can. J. Statist.*, vol. 29, no. 3, pp. 473–493, 2001.

[6] H. Li, C. Shen, A. van den Hengel, and Q. Shi, "Worst-case linear discriminant analysis as scalable semidefinite feasibility problems," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2382–2392, Aug. 2015.

[7] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 337–350, Mar. 2006.

[8] D. Huang, R. Cabral, and F. De la Torre, "Robust regression," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 616–630.

[9] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[10] D. Huang, R. Cabral, and F. De la Torre, "Robust regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 363–375, Feb. 2016.

[11] K.-H. Thung, C.-Y. Wee, P.-T. Yap, and D. Shen, "Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion," *Neuroimag.*, vol. 91, pp. 386–400, 2014.

[12] D. Ziegler and J. Augustinack, "Harnessing advances in structural MRI to enhance research on Parkinson's disease," *Imag. Med.*, vol. 5, no. 2, pp. 91–94, 2013.

[13] P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot, "An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 425–441, Apr. 2008.

[14] I. Rodrigues, J. Sanches, and J. Bioucas-Dias, "Denoising of medical images corrupted by poisson noise," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 1756–1759.

[15] H. Bhadauria and M. Dewal, "Medical image denoising using adaptive fusion of curvelet transform and total variation," *Comput. Elect. Eng.*, vol. 39, no. 5, pp. 1451–1460, 2013.

[16] J. Manjón, P. Coupé, A. Buades, D. L. Collins, and M. Robles, "New methods for MRI denoising based on sparseness and self-similarity," *Med. Image Anal.*, vol. 16, no. 1, pp. 18–27, 2012.

[17] V. Fritsch, G. Varoquaux, B. Thyreau, J.-B. Poline, and B. Thirion, "Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators," *Med. Image Anal.*, vol. 16, no. 7, pp. 1359–1370, 2012.

[18] S. Mriaux, A. Roche, B. Thirion, and G. Dehaene-Lambertz, "Robust statistics for nonparametric group analysis in fMRI," in *Proc. 3rd IEEE Int. Symp. Biomed. Imag.: Nano Macro*, 2006, pp. 936–939.

[19] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2003, pp. 97–104.

[20] F. De la Torre, "A least-squares framework for component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1041–1055, Jun. 2012.

[21] N. Bissantz, L. Dümbgen, A. Munk, and B. Stratmann, "Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces," *SIAM J. Optimization*, vol. 19, no. 4, pp. 1828–1845, 2009.

[22] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Towards a practical face recognition system: Robust registration and illumination by sparse representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 597–604.

[23] C. Lu, J. Shi, and J. Jia, "Online robust dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 415–422.

[24] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006.

[25] X. Zhu, "Semi-supervised learning literature survey," Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. 1530, 2005.

[26] A. Joulin and F. Bach, "A convex relaxation for weakly supervised classifiers," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1279–1286.

[27] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–7.

[28] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *Proc. Int. Conf. Comput. Learn. Theory*, 2004, pp. 624–638.

[29] K. Marek, et al., "The parkinson progression marker initiative (PPMI)," *Progress Neurobiol.*, vol. 95, no. 4, pp. 629–635, 2011.

[30] S. G. Mueller, et al., "Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI)," *Alzheimer's Dementia: J. Alzheimer's Assoc.*, vol. 1, pp. 55–66, Jul. 2005.

[31] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[32] A. Goldberg, X. Zhu, B. Recht, J.-M. Xu, and R. Nowak, "Transduction with matrix completion: Three birds with one stone," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 757–765.

[33] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1873–1879.

[34] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[35] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 1998, pp. 368–374.

[36] R. A. Fisher, "The logic of inductive inference," *J. Roy. Statistical Soc.*, vol. 98, no. 1, pp. 39–82, 1935.

[37] R. Min, G. Wu, J. Cheng, Q. Wang, and D. Shen, "Multi-atlas based representations for Alzheimer's disease diagnosis," *Human Brain Mapping*, vol. 35, no. 10, pp. 5052–5070, 2014.

[38] M. Liu, D. Zhang, E. Adeli-Mosabbeb, and D. Shen, "Inherent structure based multi-view learning with multi-template feature representation for Alzheimer's disease diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1473–1482, Jul. 2016.

[39] S. Liu, et al., "Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1132–1140, Apr. 2015.

[40] S. Duchesne, A. Caroli, C. Geroldi, C. Barillot, G. B. Frisoni, and D. L. Collins, "MRI-based automated computer classification of probable AD versus normal controls," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 509–520, Apr. 2008.

[41] S. F. Eskildsen, et al., "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning," *Neuroimag.*, vol. 65, pp. 511–521, 2013.

[42] R. Cuingnet, et al., "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database," *Neuroimag.*, vol. 56, no. 2, pp. 766–781, 2011.

[43] T. Tong, K. Gray, Q. Gao, L. Chen, and D. Rueckert, "Nonlinear graph fusion for multi-modal classification of Alzheimer's disease," in *Machine Learning in Medical Imaging*. Berlin, Germany: Springer, 2015, pp. 77–84.

[44] K. R. Gray, et al., "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease," *Neuroimag.*, vol. 65, pp. 167–175, 2013.

[45] M. Liu, D. Zhang, and D. Shen, "Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis," *Human Brain Mapping*, vol. 35, no. 4, pp. 1305–1319, 2014.
[46] H. Braak, K. Tredici, U. Rub, R. de Vos, E. J. Steur, and E. Braak, "Staging of brain pathology related to sporadic Parkinson's disease," *Neurobiol. Aging*, vol. 24, no. 2, pp. 197–211, 2003.
[47] A. Worker, et al., "Cortical thickness, surface area and volume measures in Parkinson's disease, multiple system atrophy and progressive supranuclear palsy," *PLoS One*, vol. 9, no. 12, 2014, Art. no. e114167.
[48] B. Pearce, A. Palmer, D. Bowen, G. Wilcock, M. Esiri, and A. Davison, "Neurotransmitter dysfunction and atrophy of the caudate nucleus in Alzheimer's disease," *Neurochemical Pathology*, vol. 2, no. 4, pp. 221–32, 1985.
[49] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 2579–2605, 2008, Art. no. 85.
[50] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.