Joint Multi-Modal Longitudinal Regression and Classification for Alzheimer's Disease Prediction

Lodewijk Brand, Kai Nichols, Hua Wang, Li Shen, and Heng Huang, *for the ADNI*

***Abstract*—Alzheimer's disease (AD) is a serious neurodegenerative condition that affects millions of individuals across the world. As the average age of individuals in the United States and the world increases, the prevalence of AD will continue to grow. To address this public health problem, the research community has developed computational approaches to sift through various aspects of clinical data and uncover their insights, among which one of the most challenging problem is to determine the biological mechanisms that cause AD to develop. To study this problem, in this paper we present a novel *Joint Multi-Modal Longitudinal Regression and Classification* method and show how it can be used to identify the cognitive status of the participants in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort and the underlying biological mechanisms. By intelligently combining clinical data of various modalities (*i.e.*, genetic information and brain scans) using a variety of regularizations that can identify AD-relevant biomarkers, we perform the regression and classification tasks simultaneously. Because the proposed objective is a non-smooth optimization problem that is difficult to solve in general, we derive an efficient iterative algorithm and rigorously prove its convergence. To validate our new method in predicting the cognitive scores of patients and their clinical diagnosis, we conduct comprehensive experiments on the ADNI cohort. Our promising results demonstrate the benefits and flexibility of the proposed method. We anticipate that our new method is of interest to clinical communities beyond AD research and have open-sourced the code of our method online.[1][2]**

***Index Terms*—Alzheimer's disease, biomarker identification, joint regression-classification, longitudinal, multi-modal, multi-task.**

I. Introduction

Manuscript received May 5, 2019; revised September 26, 2019; accepted December 5, 2019. L. Brand, K. Nichols and H. Wang were partially supported by the National Science Foundation (NSF) under the grants of IIS 1652943, IIS 1849359 and CNS 1932482; L. Shen was partially supported by the National Institutes of Health (NIH) under the grants of R01 EB022574 and RF1 AG063481 and by the NSF under the grant of IIS 1837964; H. Huang was partially supported by the NIH under the grant of R01 AG049371 and by the NSF under the grants of IIS 1836938, DBI 1836866, IIS 1845666, IIS 1852606, IIS 1838627 and IIS 1837956.

L. Brand, K. Nichols, and H. Wang are with the Department of Computer Science, Colorado School of Mines, Golden, CO 80401, U.S.A. All correspondence should be addressed to H. Wang. (email: lbrand@mymail.mines.edu, nichols1@mymail.mines.edu, huawangcs@gmail.com)

L. Shen is with the Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, PA 19104, U.S.A. (email: li.shen@pennmedicine.upenn.edu)

H. Huang is with the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A (email: heng.huang@pitt.edu)

[2]The code package for the proposed *Joint Multi-Modal Longitudinal Regression and Classification* model have been made publicly available online at https://github.com/minds-mines/jmmlrc.

ALZHEIMER'S disease (AD) is a serious neurodegenerative disorder that can lead to grievous mental and financial consequences for those affected and their families. AD is characterized by progressive memory and cognitive decline. According to the *Alzheimer's Association*, 5.5 million people in the United States currently suffer from AD related dementia. It is forecasted that by 2050, the number of people suffering from AD is expected to surpass 13.8 million people. In 2017 alone, the total financial cost associated with health care, long-term care, and hospice services for patients suffering from dementia was estimated to be $259 billion in the United States. With the projected increase of individuals and families affected by AD, it is becoming increasingly important for the scientific community to develop novel methods for the diagnosis and treatment of AD.

A central thread, within the AD research community, has focused on discovering characteristic biomarkers associated with the development of the disease. A key component of this work has been driven by the successful development of a variety of non-invasive clinical observations such as magnetic resonance imaging (MRI) scans, positron emission tomography (PET), and genetic analysis through the identification of single nucleotide polymorphisms (SNPs). Through a collection of public-private partnerships, clinical data from each of theses modalities, paired with clinical diagnoses and evaluations, are publicly available to the scientific community through the Alzheimer's Disease Neuroimaging Initiative (ADNI) [1].

Through the effective analysis of various AD data sources, we are able to build models that have the potential to help clinical researchers narrow down the array of phenotypic and genetic measures that are predictive of a future AD diagnosis. Furthermore, as the library of relevant phenotypic and genetic biomarkers is built and verified, the future research performed by clinical research teams can be more focused on important indicators of AD. These types of data-driven methods are geared towards making it easier for clinicians to focus their time on a handful of the most predictive genetic variations and phenotypic changes relevant to AD.

Identifying important genetic and phenotypic changes from clinical data, like those from the ADNI, provides a few algorithmic challenges from the machine learning perspective. First, it is not always clear how to incorporate relationships in data over time. For example, if a potential AD patient is evaluated once a year, the diagnosis, from a learned model, should depend on all the previous data collected; many state-of-the-art machine learning models do not explicitly incorporate this kind of relationship. Algorithms that effectively leverage longitudinal data are an important research tool that have the

potential to transform the way we handle disease diagnoses and treatment. Combined with institution-wide initiatives, this class of algorithms can help us identify longitudinally-sensitive biomarkers and predict the cognitive trajectories of patients.

Following the body of work done through the ADNI, in this paper we present a new *Joint Multi-Modal Longitudinal Regression and Classification* method that has shown great promise in identifying relevant longitudinal biomarkers in patients with AD. Our proposed method consists of three important regularization terms to capture the temporal and structural relationships of the input data from different perspectives. First, we use an $\ell_{2,1}$-norm regularization [2], [3] to effectively associate input features over-time and generate a sparse solution. Second, we utilize a novel group ℓ_1-norm regularization [4], [5] to globally associate the weights of our input modalities. In biology, a *modality* refers to a single stimulus (*i.e.*, light, sound, touch, *etc.*). In this paper, we use *modality* to indicate a single data grouping (*i.e.*, brain imaging data, genetic data, diagnostic data, *etc.*). The group ℓ_1-norm regularization is able to determine which input modality is most effective at predicting a particular output. Third, we incorporate the trace-norm regularization [6] to account for relationships that occur within and across modalities. Equipped with the three different types of regularizations, our proposed method aims to solve the regression and classification tasks simultaneously, because the joint classification and regression design has shown superior performance when compared to the same tasks performed separately [7], [8].

A. Related Work

Recent research [9] has shown that AD-related brain changes can occur 10-15 years before any symptoms of dementia are observed. These findings underscore the importance of developing models that relate data over time. Understanding the mechanisms behind the development of neurodegenerative diseases such as AD can reveal important opportunities for their early detection and treatment.

The analysis of medical imaging data is the core to understanding AD and its development. The diagnosis of AD through the analysis of MRI scans is difficult, particularly at the early stages of the disease when the brain has sustained less damage. Past researches [10], [11] used various state-of-the-art machine learning models, applied to MRI images, to predict AD diagnoses. This research combined with the continued analysis of MRI images of AD patients, helped provide a foundational understanding on how the disease develops.

Recently, the work on automated AD diagnoses [11] illustrated the effectiveness of voxel-based normalization of MRI images. Their voxel-based technique was able to improve the average accuracy of predicting the disease status of patients with and without AD over previous methods. In addition to the improved prediction performance, the results in [11] also showed that that the grey matter volume, derived from an MRI image, is an effective feature to consider when one differentiates between cognitively normal patients and those who suffer from AD. This work illustrated the importance of being able to investigate the internal structure of the learned model to identify relevant biomarkers, which motivates us to apply linear models instead of more complicated and difficult to interpret models. We will provide a comparison between this elastic logistic-regression model in [11] and our proposed method in our experiments.

More recent works [12], [13] also illustrated the benefits of incorporating longitudinal data into a model that predicts cognitive performance. It is only through the analysis of longitudinal data, combined with an appropriately designed algorithm, that we are able to develop a deeper understanding of the cognitive progression of individuals susceptible to AD. By investigating the computational approaches implemented in these works, we are able to design our own nuanced approach to solve longitudinal prediction problems. Longitudinal approaches form a more complete picture concerning the progress of neurodegenerative diseases. Once inspected, longitudinal models can help provide evidence for the underlying mechanisms that occur during cognitive decline.

Various regularizations, leveraged by the methods proposed in [2], [4], further developed strategies for discovering relationships within longitudinal datasets. These advances in longitudinal regularization, applied to multi-modal datasets, are the key to discovering relationships across features over time. Our proposed method applies a collection of regularizations to assist in the discovery of AD-predictive biomarkers and any associated relationships across different data streams.

B. Scientific Contributions of this Paper

Despite the heavy focus on using machine learning in biological analyses, many methods within the field do not take advantage of newer advances in machine learning. Due to the vast size and complexity of imaging and genetic data, using efficient and robust algorithms is critical to providing patients with accurate information concerning their health. As such, being able to apply new knowledge in the field of machine learning is incredibly valuable. Through the development of new machine learning algorithms applied to the biomedical field, the authors aim to build upon past research from both the computational biology and mathematical perspectives.

Our proposed *Joint Multi-Modal Longitudinal Regression and Classification* model is of clinical significance, because it can effectively identify a small number of important phenotypic features that are consistent over time. As reported in the *Experiments* section, from the empirical perspective, our proposed algorithm is designed to provide clinical researchers with a small set of genetic and phenotypic features on which clinicians should focus. In addition, from the theoretical perspective, we provide a rigorous analysis that guarantees the convergence of the proposed algorithm.

This paper is an extension of our recent work [14] originally reported in the *21st International Conference on Medical Image Computing and Computing Assisted Intervention (MICCAI 2018)*. In this extended journal manuscript, we provide the following expansions over its conference version:

- Rigorously prove in mathematics the convergence of the solution algorithm of the proposed *Joint Multi-Modal Longitudinal Regression and Classification* method.

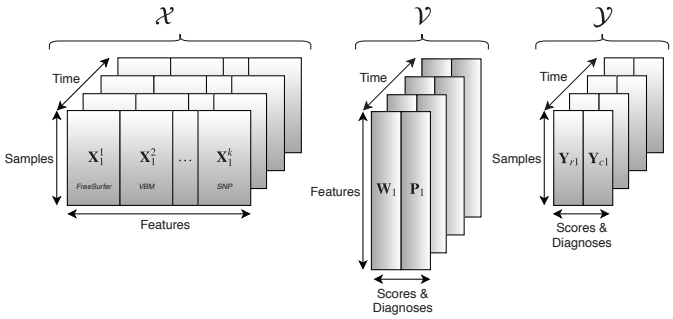

Fig. 1. Visualization of the input ($\mathcal{X}$), parameter ($\mathcal{V}$) and output ($\mathcal{Y}$) tensors. In each time-step of $\mathcal{X}$, the k modalities are explicitly defined to facilitate calculating the group ℓ_1-norm. Note the boundaries contained within $\mathcal{V}$ and $\mathcal{Y}$ to separate the classification and regression tasks. The goal of the *Joint Multi-Modal Longitudinal Regression and Classification* method is to learn the most effective mapping from $\mathcal{X}$ to $\mathcal{Y}$ through the variation of $\mathcal{V}$.

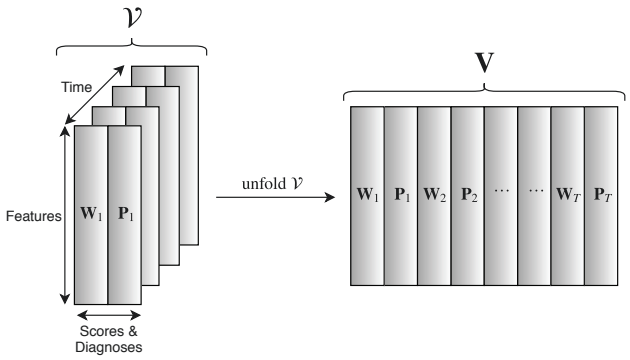

Fig. 2. A visualization of the unfolding operation on $\mathcal{V}$. The objective in (2) contains various regularizations, applied to the matrix $\mathbf{V}$ which is unfolded from the tensor $\mathcal{V}$. This is designed to find phenotypic correlations that are consistent over time and correspond to the modal structure of $\mathcal{X}$.

- Improved mathematical notations in order to unambiguously communicate the algorithm's implementation.
- Significantly expand the experiments to provide additional insights into the benefit of our proposed method.
- Describe the motivations behind the interface and implementation details within our code. This effort is designed to make it easier for future researchers to use the proposed method and its solution algorithm.

II. Methodology

The input imaging and genetic features can be represented by a set of matrices: $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_T\} \in \mathbb{R}^{n \times d \times T}$. Each $\mathbf{X}_t$ represents the input observations for n patients with d features at a given time t $(1 \leq t \leq T)$. The output diagnoses and cognitive scores can be represented by another set of matrices: $\mathcal{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_T\} \in \mathbb{R}^{n \times c \times T}$. Each $\mathbf{Y}_t = [\mathbf{Y}_{rt}\ \mathbf{Y}_{ct}]$ is a concatenation of the cognitive scores (for regression) and diagnoses (for classification) for n patients at time t. We define $c = c_r + c_c$ where c_r is the number of regression targets and c_c is the number of possible diagnoses. Obviously, both $\mathcal{X}$ and $\mathcal{Y}$ are tensors. The goal of our proposed algorithm is to learn a joint regression and classification model represented by the parameter tensor $\mathcal{V} = [\mathcal{W}\ \mathcal{P}]$: $\mathcal{V} = \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_T\} = \{[\mathbf{W}_1\ \mathbf{P}_1], [\mathbf{W}_2\ \mathbf{P}_2], \ldots, [\mathbf{W}_T\ \mathbf{P}_T]\} \in \mathbb{R}^{d \times c \times T}$. Note that here we explicitly separate $\mathbf{W}_t \in \mathbb{R}^{d \times c_r}$ and $\mathbf{P}_t \in \mathbb{R}^{d \times c_c}$ in each $\mathbf{V}_t$, which provides us with convenient notation to explicitly separate the learned coefficient matrices for regression ($\mathbf{W}_t$ $(1 \leq t \leq T)$) and classification ($\mathbf{P}_t$ $(1 \leq t \leq T)$). For easier understanding, the input, output, and parameter tensors are visually illustrated in Figure 1.

For the remainder of this manuscript, we will write matrices as bold uppercase letters, vectors as bold lowercase letters, and scalars as lower case letters. Given a matrix $\mathbf{M}$, its i-th row and j-th column are denoted as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively.

A. Our Objective

The key idea of our proposed approach is to perform the regression and classification tasks at the same time. Joint regression and classification can help discover more robust patterns than those discovered when classification and regression are performed separately [7], [8]. These robust patterns can arise when the learned parameters for one task become outliers for the other. To achieve simultaneous classification and regression, we minimize the following joint objective:

$$\min_{\mathcal{W},\mathcal{P}} \ \mathcal{L}_r(\mathcal{W}) + \mathcal{L}_c(\mathcal{P}) + \mathcal{R}(\mathcal{V}) \ , \tag{1}$$

where $\mathcal{L}_r$ and $\mathcal{L}_c$ are the prescribed loss functions for the respective regression and classification tasks, and $\mathcal{R}(\mathcal{V})$ is the regularization term for better numerical stability and capturing data structures as detailed in the remainder of this subsection.

First, from the clinical research perspective, it is desirable that our model identifies specific features that are *consistent over time*. We want our model to be longitudinally consistent in order to identify specific phenotypic changes which we should pay attention to and investigate further. In order to associate the longitudinal imaging and genetic markers to predict cognitive scores and diagnoses over time, we apply the widely used $\ell_{2,1}$-norm [2], [3] to the parameter matrix $\mathbf{V} = [\mathbf{W}_1\ \mathbf{P}_1\ \mathbf{W}_2\ \mathbf{P}_2 \ldots \mathbf{W}_T\ \mathbf{P}_T]$, which is unfolded from the parameter tensor $\mathcal{V}$ along the time mode as illustrated in Figure 2. Specifically, we use $\mathcal{R}(\mathcal{V}) = \|\mathbf{V}\|_{2,1} = \sum_{i=1}^{d} \left\|\mathbf{v}^i\right\|_2$.

Second, as we combine different modalities (*i.e.*, VBM, FreeSurfer, and SNP) together, it is important for our model to differentiate the impact that each modality has on each task. This is critical since the features of a specific input modality can have a larger impact in predicting a particular label. For example, features associated with the brain imaging modality may be more useful in determining cognitive scores than the corresponding genetic modality, and vice versa. In order to capture these relationships inherent to the input modalities, we leverage the group ℓ_1-norm (G_1-norm) proposed by [8], [4], [5] (note that k is the number of input modalities): $\|\mathbf{V}\|_{G_1} = \sum_{j=1}^{k} \left\|\mathbf{V}^j\right\|_2$, where $\mathbf{V}^j$ is a horizontal block of coefficients in $\mathbf{V}$ that corresponds to the jth modality in $\mathcal{X}$. Using this group ℓ_1-norm we further develop the regularization term of (1) as $\mathcal{R}(\mathcal{V}) = \gamma_1 \|\mathbf{V}\|_{2,1} + \gamma_2 \|\mathbf{V}\|_{G_1}$.

Third, we know that as AD develops, many cognitive measures are related to one another. In order to account for this inner-modal and inter-modal relationship, we leverage the trace norm regularization of $\mathbf{V}$ [15], [16]: $\|\mathbf{V}\|_* = \sum \sigma_i(\mathbf{V})$, where $\sigma_i(\mathbf{V})$ are the singular values of $\mathbf{V}$. This can develop correlations across each of the learning tasks at different time

points. It is well known [15] that the trace-norm is the best convex approximation of the rank of a matrix. This rank minimization will develop joint correlations across each of the learning tasks at different time points, by which we finally propose our *Joint Multi-Modal Longitudinal Regression and Classification* objective as follows:

$$\begin{aligned}\min_{\mathcal{V}} J = &\sum_{t=1}^{T} \|\mathbf{X}_t\mathbf{W}_t - \mathbf{Y}_{rt}\|_F^2 \\ &+ \sum_{t=1}^{T} (\mathbf{1} - (\mathbf{X}_t\mathbf{P}_t + \mathbf{b}_t) \odot \mathbf{Y}_{ct})_+ \\ &+ \gamma_1 \|\mathbf{V}\|_{2,1} + \gamma_2 \|\mathbf{V}\|_{G_1} + \gamma_3 \|\mathbf{V}\|_* \quad ,\end{aligned} \tag{2}$$

where the function $(a)_+$ is defined as $(a)_+ = \max(0, a)$ and $\odot$ is the Hadamard product. Here the least square loss function in the first term of (2) is used for the regression tasks and the hinge loss function in the second term, where $\mathbf{b}_t$ is the intercept for the t-th multi-class support vector machine (SVM), is used as a penalty for the classification tasks.

B. Derivation of the Solution Algorithm and its Convergence

Despite its clear intuitions, the proposed objective J in (2) is a non-smooth convex problem. To solve this optimization problem efficiently, we derive an iterative algorithm as summarized in Algorithm 1, whose convergence can be rigorously guaranteed by Theorem 2.1 because we employ the iteratively reweighted method [2], [17], [18] to deal with the non-smooth regularization terms.

Algorithm 1: The algorithm to minimize J in (2).

Data: $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_T\} \in \mathbb{R}^{n\times d\times T}$,
$\mathcal{Y} = [\mathcal{Y}_r\ \mathcal{Y}_c] = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_T\} \in \mathbb{R}^{n\times c\times T}$.

1. Initialize $\mathcal{V} = [\mathcal{W}\ \mathcal{P}] \in \mathbb{R}^{d\times c\times T}$ where $\mathcal{W} \in \mathbb{R}^{d\times c_r\times T}$ is generated using the regression results ($\mathcal{Y}_r$) at each individual time point and $\mathcal{P} \in \mathbb{R}^{d\times c_c\times T}$ is derived from T multi-class SVMs fit to $\mathcal{Y}_c$.

while *not converges* **do**

2. Unfold the joint coefficient matrix $\mathcal{V}$: $\mathbf{V} = [\mathbf{V}_1\ \mathbf{V}_2\ ...\ \mathbf{V}_T] \in \mathbb{R}^{d\times cT}$.

3. Calculate the diagonal matrix $\mathbf{D}$ where the i-th diagonal element is computed as: $\mathbf{D}_i^i = \frac{1}{2\|\mathbf{v}^i\|_2}$.

4. Calculate the k block-diagonal matrix $\bar{\mathbf{D}}$. The size of each k-th block is determined via user-defined groups along $\mathbb{R}^d$: $\bar{\mathbf{D}}^k = \frac{1}{2\|\mathbf{V}^k\|_2}\mathbf{I}_k$.

5. Calculate the diagonal matrix $\hat{\mathbf{D}}$: $\hat{\mathbf{D}} = \frac{1}{2}\left(\mathbf{V}\mathbf{V}^T\right)^{-\frac{1}{2}}$.

6. Update $\mathcal{W}$:

$$\mathcal{W} = \left[\left(\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D} + \gamma_2\bar{\mathbf{D}} + \gamma_3\hat{\mathbf{D}}\right)^{-1}\mathbf{X}^T\mathbf{Y}r\right]_{t=1}^{T} .$$

7. Using a SVM solver, update $\mathcal{P}$:

$$\mathcal{P} = \left[\arg\min_{\mathbf{P},\mathbf{b}} \left(\mathbf{1} - \left(\tilde{\mathbf{X}}\mathbf{P} + \mathbf{b}\right) \odot \mathbf{Y}_c\right)_+\right]_{t=1}^{T} ,$$

where $\tilde{\mathbf{X}} = \left(\gamma_1\mathbf{D} + \gamma_2\bar{\mathbf{D}} + \gamma_3\hat{\mathbf{D}}\right)^{-\frac{1}{2}}\mathbf{X}$.

8. Update $\mathcal{V}$: $\mathcal{V} = [\mathcal{W}\ \mathcal{P}]$.

end

Result: $\mathcal{V} = \{\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_T\} \in \mathbb{R}^{d\times c\times T}$

Theorem 2.1: Algorithm 1 monotonically decreases the objective of the problem in (2) in each iteration, and converges to the globally optimal solution.

Proof: We denote the updated-unfolded $\mathbf{W}$ in each iteration as $\tilde{\mathbf{W}}$, the updated-unfolded $\mathbf{P}$ as $\tilde{\mathbf{P}}$, and the updated $\mathbf{V}$ as $\tilde{\mathbf{V}} = [\tilde{\mathbf{W}}_0\ \tilde{\mathbf{P}}_0\ \ldots\ \tilde{\mathbf{W}}_T\ \tilde{\mathbf{P}}_T]$. The least square loss in the g-th iteration is represented by $\mathcal{L}_r^{(g)} = \sum_{t=0}^{T}\|\mathbf{X}\mathbf{W} - \mathbf{Y}_r\|_F^2$ and the hinge loss by $\mathcal{L}_c^{(g)} = \sum_{t=0}^{T}(1 - (\mathbf{X}\mathbf{P} + \mathbf{b}) \odot \mathbf{Y}_c)_+$. Here we drop the time subscripts for brevity and better readability.

According to Step 6 of Algorithm 1, we know that the following inequality holds:

$$\begin{aligned}&\mathcal{L}_r{}^{(g+1)} + \gamma_1\,\mathbf{tr}\left(\tilde{\mathbf{W}}^T\mathbf{D}\tilde{\mathbf{W}}\right) + \gamma_2\,\mathbf{tr}\left(\tilde{\mathbf{W}}^T\bar{\mathbf{D}}\tilde{\mathbf{W}}\right) + \\ &\quad\gamma_3\,\mathbf{tr}\left(\tilde{\mathbf{W}}^T\hat{\mathbf{D}}\tilde{\mathbf{W}}\right) \le \\ &\mathcal{L}_r{}^{(g)} + \gamma_1\,\mathbf{tr}\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right) + \gamma_2\,\mathbf{tr}\left(\mathbf{W}^T\bar{\mathbf{D}}\mathbf{W}\right) + \\ &\quad\gamma_3\,\mathbf{tr}\left(\mathbf{W}^T\hat{\mathbf{D}}\mathbf{W}\right) \quad .\end{aligned} \tag{3}$$

Similarly, according to Step 7 of Algorithm 1, the following inequality holds:

$$\begin{aligned}&\mathcal{L}_c{}^{(g+1)} + \gamma_1\,\mathbf{tr}\left(\tilde{\mathbf{P}}^T\mathbf{D}\tilde{\mathbf{P}}\right) + \gamma_2\,\mathbf{tr}\left(\tilde{\mathbf{P}}^T\bar{\mathbf{D}}\tilde{\mathbf{P}}\right) + \\ &\quad\gamma_3\,\mathbf{tr}\left(\tilde{\mathbf{P}}^T\hat{\mathbf{D}}\tilde{\mathbf{P}}\right) \le \\ &\mathcal{L}_c{}^{(g)} + \gamma_1\,\mathbf{tr}\left(\mathbf{P}^T\mathbf{D}\mathbf{P}\right) + \gamma_2\,\mathbf{tr}\left(\mathbf{P}^T\bar{\mathbf{D}}\mathbf{P}\right) + \\ &\quad\gamma_3\,\mathbf{tr}\left(\mathbf{P}^T\hat{\mathbf{D}}\mathbf{P}\right) \quad .\end{aligned} \tag{4}$$

According to [2, Lemma 1], we know that $\|\tilde{\mathbf{v}}\|_2 - \frac{\|\tilde{\mathbf{v}}\|_2^2}{2\|\mathbf{v}\|_2} \le \|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}\|_2}$, by which we can derive the following set of inequalities applied to $\mathbf{V}$:

$$\sum_{i=1}^{d}\left\|\tilde{\mathbf{v}}^i\right\|_2 - \sum_{i=1}^{d}\frac{\left\|\tilde{\mathbf{v}}^i\right\|_2^2}{2\left\|\mathbf{v}^i\right\|_2} \le \sum_{i=1}^{d}\left\|\mathbf{v}^i\right\|_2 - \sum_{i=1}^{d}\frac{\left\|\mathbf{v}^i\right\|_2^2}{2\left\|\mathbf{v}^i\right\|_2} \quad , \tag{5}$$

$$\sum_{j=1}^{k}\left\|\tilde{\mathbf{V}}^j\right\|_F - \sum_{j=1}^{k}\frac{\left\|\tilde{\mathbf{V}}^j\right\|_F^2}{2\left\|\mathbf{V}^k\right\|_2}\mathbf{I}_k \le \sum_{j=1}^{k}\left\|\mathbf{V}^j\right\|_F - \sum_{j=1}^{k}\frac{\left\|\mathbf{V}^j\right\|_F^2}{2\left\|\mathbf{V}^k\right\|_2}\mathbf{I}_k \quad , \tag{6}$$

According to [12, Lemma 2], we know that $\mathbf{tr}\left(\mathbf{B}^{\frac{1}{2}}\right) - \frac{1}{2}\mathbf{tr}\left(\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\right) \le \mathbf{tr}\left(\mathbf{A}^{\frac{1}{2}}\right) - \frac{1}{2}\mathbf{tr}\left(\mathbf{A}\mathbf{A}^{-\frac{1}{2}}\right)$, by which we can derive the following inequality:

$$\begin{aligned}&\left\|\tilde{\mathbf{V}}\right\|_F - \mathbf{tr}\left(\tilde{\mathbf{V}}\tilde{\mathbf{V}}^T\frac{1}{2}\left(\mathbf{V}\mathbf{V}^T\right)^{-\frac{1}{2}}\right) \le \\ &\|\mathbf{V}\|_F - \mathbf{tr}\left(\mathbf{V}\mathbf{V}^T\frac{1}{2}\left(\mathbf{V}\mathbf{V}^T\right)^{-\frac{1}{2}}\right) \quad .\end{aligned} \tag{7}$$

Then, by using the definitions of $\mathbf{D}$, $\bar{\mathbf{D}}$ and $\hat{\mathbf{D}}$, we can rewrite (5–7) as:

$$\begin{aligned}&\gamma_1\sum_{i=1}^{d}\left\|\tilde{\mathbf{v}}^i\right\|_2 - \gamma_1\,\mathbf{tr}\left(\tilde{\mathbf{V}}^T\mathbf{D}\tilde{\mathbf{V}}\right) \le \\ &\gamma_1\sum_{i=1}^{d}\left\|\mathbf{v}^i\right\|_2 - \gamma_1\,\mathbf{tr}\left(\mathbf{V}^T\mathbf{D}\mathbf{V}\right) \quad ,\end{aligned} \tag{8}$$

$$\gamma_2 \sum_{j=1}^{k} \left\| \tilde{\mathbf{V}}^j \right\|_F - \gamma_2 \, \mathbf{tr}\left(\tilde{\mathbf{V}}^T \bar{\mathbf{D}} \tilde{\mathbf{V}} \right) \leq \gamma_2 \sum_{j=1}^{k} \left\| \mathbf{V}^j \right\|_F - \gamma_2 \, \mathbf{tr}\left(\mathbf{V}^T \bar{\mathbf{D}} \mathbf{V} \right) \quad , \tag{9}$$

$$\gamma_3 \left\| \tilde{\mathbf{V}} \right\|_F - \gamma_3 \, \mathbf{tr}\left(\tilde{\mathbf{V}}^T \hat{\mathbf{D}} \tilde{\mathbf{V}} \right) \leq \gamma_2 \left\| \mathbf{V} \right\|_F - \gamma_3 \, \mathbf{tr}\left(\mathbf{V}^T \mathbf{D} \mathbf{V} \right) \quad . \tag{10}$$

Finally, using the fact that $\mathbf{tr}\left(\mathbf{V}^T\mathbf{D}\mathbf{V}\right) = \mathbf{tr}\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right) + \mathbf{tr}\left(\mathbf{P}^T\mathbf{D}\mathbf{P}\right)$, we add (3-4) to (8-10):

$$\begin{aligned} &\mathcal{L}_r^{(g+1)} + \mathcal{L}_c^{(g+1)} + \\ &\gamma_1 \sum_{i=1}^{d} \left\| \tilde{\mathbf{v}}^i \right\|_2 + \gamma_2 \sum_{j=1}^{k} \left\| \tilde{\mathbf{V}}^j \right\|_F + \gamma_3 \left\| \tilde{\mathbf{V}} \right\|_F \leq \\ &\mathcal{L}_r^{(g)} + \mathcal{L}_c^{(g)} + \\ &\gamma_1 \sum_{i=1}^{d} \left\| \mathbf{v}^i \right\|_2 + \gamma_2 \sum_{j=1}^{k} \left\| \mathbf{V}^j \right\|_F + \gamma_3 \left\| \mathbf{V} \right\|_F \quad . \end{aligned} \tag{11}$$

Therefore, our algorithm decreases the objective value of (2) after each iteration in Algorithm 1. Since the objective in (2) is a convex optimization problem and apparently lower-bounded, Algorithm 1 will converge to a globally optimal solution. □

III. Experiments

Data. We downloaded 1.5T MRI scans, SNP genotypes, and demographic information for 821 ADNI-1 participants. We performed voxel-based morphometry (VBM) and FreeSurfer automated parcellation on the MRI data following [19], and extracted mean modulated gray matter (GM) measures for 90 target regions of interest (ROIs). We followed the SNP quality control steps discussed in [20]. We also downloaded the longitudinal scores of the participants' Rey's Auditory Verbal Learning Test (RAVLT) and their clinical diagnoses: healthy control (HC), mild cognitive impairment (MCI), and Alzheimer's disease (AD). The details of these cognitive assessments can be found in the ADNI procedure manuals. All feature data have been normalized to have zero mean and unit-variance. The time points examined in this study for both imaging markers and cognitive assessments included baseline (BL), Month 6 (M6), Month 12 (M12) and Month 24 (M24). All the participants with no missing BL/M6/M12/M24 MRI measurements, SNP genotypes, and cognitive measures were included in this study, which resulted in a set of 412 subjects. The patient diagnoses at each time point are provided in Table I. The authors note that if more time points were utilized, specifically M18, the resulting training/testing dataset would be too small due to the significant amount of missing data at that particular time point. In the following experiments $\mathcal{X} \in \mathbb{R}^{412\times1404\times4}$, $\mathcal{Y} \in \mathbb{R}^{412\times6\times4}$, and $\mathcal{V} \in \mathbb{R}^{1404\times6\times4}$.

Settings. For all the results reported in the following experiments, we performed a reasonable grid search for each of the compared methods designed to provide a fair comparison between our proposed method and the other competing methods. The optimal tuning parameters are chosen by the model that provides the best regression or classification performance using a five-fold cross-validation strategy. This approach involves randomly breaking $\mathcal{X}$ and $\mathcal{Y}$ into five approximately-equal groups along n. The first group (or *fold*) is used for validation whereas the remaining four folds are used to train the model; this process is repeated five times for each validation/training pairs. We iterate each five-fold experiment one-hundred times and randomly shuffle $\mathcal{X}$ and $\mathcal{Y}$ in between each iteration. The hyper parameters that result in the best average performance for a given model are used for comparison. The standard deviations for each performance metric during the iterated five-fold experiments are provided with our results. In choosing the parameters for the proposed *Joint Multi-Modal Longitudinal Regression and Classification* method, we fine tuned γ_1, γ_2 and γ_3 in (2) by searching a grid of powers of 10 between 10^{-5} and 10^5 and choose the best model based on the average multitask performance, which are thereby set as $\gamma_1 = 10^{-5}, \gamma_1 = 10^{-2}$ and $\gamma_1 = 10^2$ in our experiments.

Table I
PATIENT DIAGNOSES AT EACH TIME POINT DERIVED FROM THE ADNI.

	AD	MCI	HC
Baseline	79	190	143
Month 6	86	180	146
Month 12	111	155	146
Month 24	155	110	147

Implementation Details. To facilitate the comparisons of the proposed *Joint Multi-Modal Longitudinal Regression and Classification* method to other state-of-the-art algorithms, the authors have made a concerted effort to adhere to a standardized machine learning interface. We follow an implementation interface that is identical to sklearn's `model.fit(X, y)` and `model.predict(X)` API [21]. This is designed to make our code trivial for the reader to understand, run experiments, and perform hyper-parameter searches when applying our new method to their own datasets. The code that implements our new model has been open-sourced online at https://github.com/minds-mines/jmmlrc.

A. Performance

In this section, we explore the performance of the proposed *Joint Multi-Modal Longitudinal Regression and Classification* method from two different perspectives. First, we will investigate the overall performance of the regression and classification tasks averaged over all the time points used in the ADNI study (BL, M6, M12, and M24). Second, using the data gathered from this experiment, we will compare the longitudinal performance of our model at each *individual* time point, where our model will be compared to the second best performing algorithm derived from the first experiment. We will measure performance of the regression task by calculating the root-mean-squared error (RMSE) values and determine

Table II

REGRESSION. RMSE RESULTS, AVERAGED ACROSS ALL TIME POINTS, OF THE PROPOSED METHOD COMPARED TO A COLLECTION OF BROADLY USED REGRESSION METHODS AND THE DEGENERATE VERSIONS OF OUR OWN METHOD. STANDARD DEVIATION RESULTS FROM THE FIVE-FOLD TESTING SCHEME ARE ALSO REPORTED. NOTE THAT RAVLT MEMORY SCORES RANGE FROM 0 TO 74.

Model	RAVLT_TOT	RAVLT30	RAVLT30_RECOG	All
Linear	3.71e11±8.50e11	4.37e11±9.62e11	6.29e11±1.37e12	5.23e12±1.07e11
Ridge	18.8±0.538	20.4±0.625	19.5±0.591	19.6±0.469
Lasso	19.2±0.659	21.1±0.721	19.9±0.627	20.1±0.553
MLP	19.2±0.676	20.9±0.697	19.8±0.675	20.0±0.562
ELM	19.5±0.669	21.3±0.71	20.4±0.697	20.4±0.572
Ours ($\ell_{2,1}$-norm only)	12.0±0.620	18.4±0.803	18.6±0.860	16.6±0.517
Ours (group ℓ_1-norm only)	12.6±0.837	19.7±1.140	19.7±1.190	17.7±0.844
Ours (trace-norm only)	12.1±0.707	18.4±1.160	18.5±0.838	16.6±0.710
Ours (regression only)	12.9±0.938	19.7±1.250	20.1±0.872	17.9±0.871
Ours	**11.7±0.836**	**18.3±0.530**	**18.1±0.792**	**16.2±0.758**

Table III

CLASSIFICATION. F_1 AND BALANCED ACCURACY (BACC) SCORES OF THE PROPOSED METHOD, AVERAGED ACROSS ALL TIME POINTS, COMPARED AGAINST A COLLECTION OF WIDELY USED CLASSIFICATION METHODS AND THE DEGENERATE VERSIONS OF OUR OWN METHOD. EACH F_1 SCORE, AND THEIR STANDARD DEVIATIONS, ARE ASSOCIATED WITH THE PERFORMANCE ON IDENTIFYING THE THREE CLASSIFICATION LABELS REFERENCED IN THE ADNI DATASET. BALANCED ACCURACY IS CALCULATED ACROSS ALL CLASSES. NOTE THAT *ElasticNet* [11], *LinearSVM* [10], AND *ELM* [22] HAVE ALL BEEN USED PREVIOUSLY TO DIAGNOSE AD.

Model	F_1 (AD)	F_1 (MCI)	F_1 (HC)	F_1 (All)	BACC (All)
Logistic	0.282±0.043	0.511±0.026	0.351±0.045	0.416±0.025	0.394±0.024
RandomForest	0.347±0.047	0.403±0.042	0.396±0.040	0.379±0.028	0.373±0.028
SVM	0.272±0.033	0.462±0.030	0.384±0.038	0.394±0.025	0.378±0.024
KNN	0.334±0.051	0.467±0.030	0.400±0.035	0.414±0.025	0.402±0.026
MLP	0.310±0.048	0.422±0.054	0.414±0.040	0.396±0.023	0.385±0.023
ElasticNet	0.301±0.052	0.485±0.048	0.369±0.090	0.415±0.028	0.396±0.027
LinearSVM	0.287±0.081	0.455±0.022	0.354±0.084	0.392±0.021	0.353±0.015
ELM	0.186±0.062	0.472±0.031	0.334±0.062	0.376±0.031	0.351±0.026
Ours ($\ell_{2,1}$-norm only)	0.551±0.048	0.496±0.045	0.660±0.048	0.574±0.038	0.577±0.042
Ours (group ℓ_1-norm only)	0.477±0.044	0.473±0.046	0.524±0.057	0.493±0.038	0.492±0.038
Ours (trace-norm only)	0.550±0.040	0.505±0.048	0.619±0.041	0.546±0.032	0.547±0.033
Ours (classification only)	0.548±0.035	0.506±0.042	0.663±0.054	0.559±0.033	0.574±0.030
Ours	**0.566±0.047**	**0.513±0.043**	**0.683±0.044**	**0.576±0.033**	**0.584±0.033**

the performance of the classification task through calculating class-specific F_1 scores:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad ,$$

and Balanced Accuracy (BACC) [23]. Here we note that, through our recent considerable efforts on fine tuning the parameters for all compared methods, better performances on both regression and classification tasks are reported in this manuscript compared to those reported in its conference version in [14].

Regression. In Table II we report the regression performance results of a collection of broadly used machine learning methods against our new method in standard five-fold cross-validations. We compare our algorithm against multivariate linear regression (*Linear*), ℓ_2-norm regularized linear regression (*Ridge*), ℓ_1-norm regularized linear regression (*Lasso*) [24], multi-layer perceptron regression (*MLP*) [25], extreme learning machine regression (*ELM*) [22]. We also compare our algorithm against its degenerate versions, *i.e.*, regularization using ℓ_{21}-norm, group ℓ_1-norm, trace norm only, and regularizing on the regression coefficients only. We can see from the results presented in Table II that the proposed algorithm shows promising regression performance when compared to its competing counterparts. The optimized hyper parameters for the competing methods are provided with our code.

In addition to the regression performance improvements afforded by our proposed method, we can see that each of the degenerate versions provide a significant boost in performance when compared to the basic regularizations utilized in ridge and lasso multivariate regressions. Therefore, the results in Table II show that the regularizations described in (2) have the capacity to improve the RAVLT score predictions associated with participants in the ADNI study. While our method improves the average prediction performance of all three RAVLT scores as shown in Table II, the authors recognize that the performance of the fully regularized version of our new method only slightly improves upon its degenerate versions. We will provide deeper insight into the differences between the degenerate methods and the fully regularized objective later when we investigate the *Biomarker Identification* properties of our method.

Classification. In Table III we report the classification performance results of our method compared to a variety of broadly used classification algorithms in five-fold cross-validations. We compare our method against logistic regression (*Logistic*), random forest classifier (*RandomForest*), support vector machine using a sigmoid-kernel (*SVM*), k-nearest neighbors classifier (*KNN*), logistic regression with elastic net regularization (*ElasticNet*) [11], linear support vector machine (*LinearSVM*) [10] and an extreme learning machine (*ELM*) [22]. From Table III we can see that our algorithm, and its assorted degenerate versions, show the most improvement

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2958943, IEEE Transactions on Medical Imaging

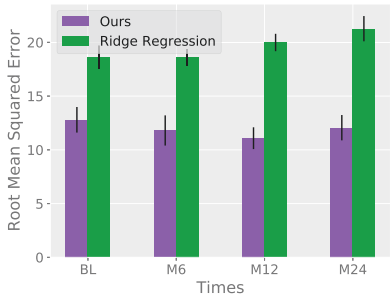

Fig. 3. RAVLT score prediction: RMSE of our model at each individual time point compared against *Ridge* regression. The black lines superimposed on each bar represent the standard deviations derived from the five-fold cross-validation experiments.

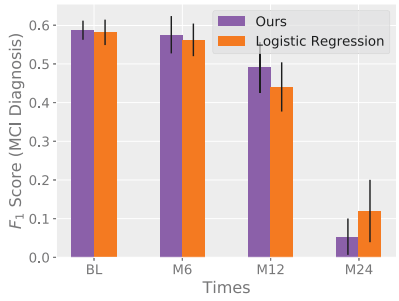

Fig. 4. F_1 scores and the standard deviations derived from the five-fold cross-validation experiments, separated by time point, predicting MCI by our method compared against *Logistic* regression: Our method slightly outperforms logistic regression at all time points except for M24.

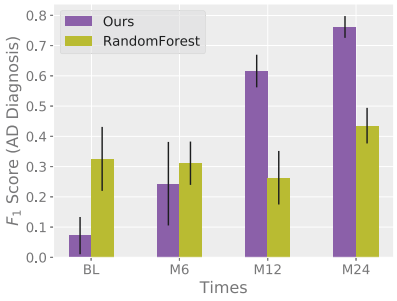

Fig. 5. F_1 scores and the standard deviations derived from the five-fold cross-validation experiments, separated by time point, predicting AD diagnoses of our method compared against *RandomForest*: Our method is more effective at detecting cognitive decline at later time points.

when predicting HC and AD diagnoses. The relative performance improvement of predicting MCI patients is limited when compared to more traditional methods such as logistic regression. We also notice that the results of the algorithms in [10], [11], and [22] perform significantly worse than our new method. This is likely because these two algorithms were only designed to differentiate between two classes (HC and AD) instead of three (HC, MCI, and AD.)

The results in Table III again demonstrate that the regularizations used in our method are able to exploit the longitudinal and grouped relationships that are not leveraged in any of the competing algorithms. On average, our new method significantly outperforms the detection of HC and AD in ADNI participants when compared to all the competing machine learning models. The authors again note that, similar to the regression results, the fully regularized version of our new method shows modest improvement over the degenerate versions of our method.

Longitudinal Regression (RAVLT). An important performance consideration for evaluating the effectiveness of our proposed method is how well it performs at each individual time point. This longitudinal analysis is a critical component for understanding what effects the regularization on $\mathcal{V}$ has on the accuracy of our model at each time point.

In Figure 3 we provide the *regression* performance values at each individual time point contained in the ADNI study. We compare our method against ridge regression at each time point. Here we chose ridge regression for this study, because it has the best performance among the tested competing methods as reported in Table II. We can see from Figure 3 that the proposed method outperforms ridge regression at every time point. This success can be attributed to the fact that the regularizations on $\mathcal{V}$ help the algorithm take into account the longitudinal patterns evident in the ADNI dataset.

Longitudinal Classification (Diagnosis of AD and MCI). Recent research [26] suggests that the early identification and diagnosis of patients with AD is a key component for reducing the financial burden associated with the disease. The earlier a patient is diagnosed with AD, and by extension MCI, the earlier caregiver interventions and pharmacological treatments can occur. Early intervention has been shown to slow down the progression of AD. In Figure 4 and Figure 5, we report the time-separated MCI and AD classification performance of the proposed method compared against the *Logistic* regression and *RandomForest* classifiers. We can see from Figure 4 that during BL and M6 our new method slightly outperforms *Logistic* regression in predicting MCI diagnoses in ADNI participants. Then between M12 and M24, according to Figure 5, our method is able to significantly outperform the *RandomForest* algorithm in identifying AD. Although, this improvement is not consistent over all time points for predicting AD and MCI.

This observed decrease in predictive performance at M24 is likely due to a significant change in disease status between M6 and M12. In Table I we can see that between M6 and M12 the number of patients that are diagnosed with MCI drops while the number of AD patients significantly increases. Nonetheless, our method is able to effectively identify MCI early and AD late. This performance boost, paired with the longitudinal regularizations proposed in our method, provides insight into biomarkers that are predictive of cognitive decline.

B. Biomarker Identification

From the clinical research perspective, one *key* contribution of our *Joint Multi-Modal Longitudinal Regression and Classification* method is its capability to discover phenotypic features that have the largest impact on our regression and classification predictions *over time*. This "biomarker identification" capacity is only possible because we can inspect the internal structure of our learned model, which contrasts significantly from "black-box" algorithms like deep neural networks where it can be difficult to determine which specific input features impact the resulting prediction. In our case, once the proposed algorithm optimizes (2), we can analyze the magnitude of each row in $\mathbf{V}$ to determine which feature is the most important feature.

Brain Imaging Modalities. In Figure 6 and Figure 7 we plot the magnitudes, derived from $\mathbf{V}$, of coefficients associated with the FreeSurfer and VBM features contained in $\mathcal{X}$. We can clearly see that the biomarkers discovered across all four time points are all longitudinally consistent. Visually, the brain heat-map images from BL to M24 look almost identical, which

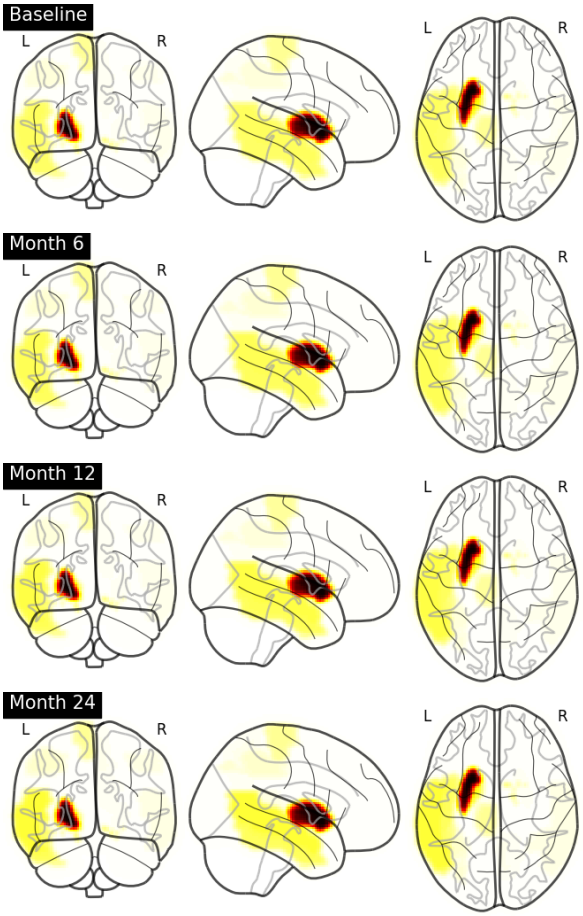

Fig. 6. Visualization of the FreeSurfer modality coefficients derived from $\mathcal{V}$ at various times (BL/M6/M12/M24). The top ten brain AAL[27] regions are as follows (largest to smallest): *Left Hippocampus, Left Entorhinal Cortex, Right Entorhinal Cortex, Left Amygdala, Left Middle Temporal Gyrus, Left Inferior Temporal Gyrus, Left Parahippocampal Gyrus, Left Inferior Parietal Gyrus, Left Banks of Superior Temporal Sulcus, Right Inferior Parietal Gyrus, Right Middle Temporal Gyrus.* Created using the nilearn software package [28].

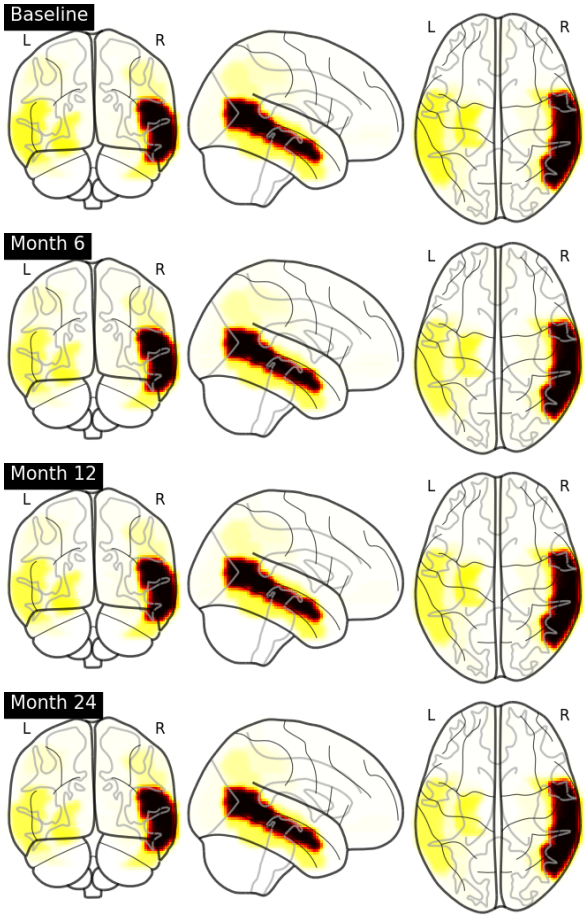

Fig. 7. Visualization of the voxel-based morphometry modality coefficients derived from $\mathcal{W}$ at various times (BL/M6/M12/M24). The top ten AAL[27] regions are as follows (largest to smallest): *Right Middle Temporal Gyrus, Left Hippocampus, Left Middle Temporal Gyrus, Right Inferior Temporal Gyrus, Right Angular Gyrus, Left Inferior Temporal Gyrus, Left Inferior Parietal Gyrus, Left Amygdala, Right Hippocampus, Left Supramarginal Gyrus.* Created using the nilearn software package [28].

illustrates the power of the $\ell_{2,1}$-norm regularization that provides our algorithm with the ability to identify longitudinally consistent biomarkers. This consistency is especially important from the clinical perspective, because we can leverage this longitudinal consistency to identify which parts of the brain the medical community should focus on with regards to AD. We can also verify the performance of our method by determining whether the ranked features listed in the description of Figure 6 and Figure 7 are consistent with previous AD research.

For example, Mu *et al*. [29] provided a review that documented how the hippocampus is affected by the early stages of AD, which is in perfect accordance with our experimental results in that this region of the brain is one of the top features discovered by our model in both FreeSurfer and VBM modalities. Besides, Van Hoesen *et al*. [30] reported a strong evidence that a severely damaged entorhinal cortex (Broadmann's area 28) is observed in patients suffering from AD, which confirms our experimental finding that the thickness of the entorhinal cortex is incorporated as a feature in our FreeSurfer dataset and its coefficient is also ranked highly. In addition, Poulin *et al*. [31] analyzed the impact of amygdala atrophy and determined that it was highly predictive of AD severity during the early clinical stages of AD, which nicely supports the biomarkers identified by our model. Finally, Convit *et al*. [32], using a logistic regression model, determined that the medial, inferior and middle temporal gyri are some of the first areas in the brain affected by AD, whose importance are reinforced by the ranked features discovered by our method.

The combination of longitudinal consistency of the MRI biomarkers discovered by our new method, supported by previous AD research focused on identifying important brain biomarkers, indicate the promise that our new method holds in discovering biomarkers relevant to AD.

The Case for All Three Regularizations. We observe in Table II and Table III that the fully regularized version of our method has stronger performance than its degenerate versions, although this performance improvement over those regularized by either the $\ell_{2,1}$-norm or the trace norm is not big. The authors recognize that the small performance improvements afforded by the extra regularizations introduced by our method may not provide sufficient evidence to increase the complexity of the proposed algorithm. Here, we aim to convince the

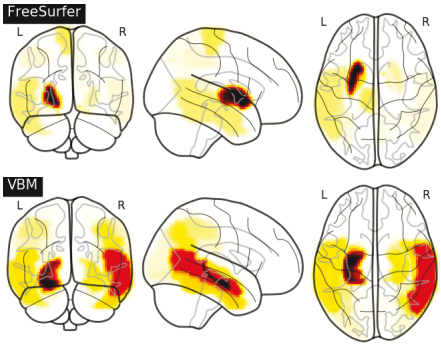

Fig. 8. **Top:** FreeSurfer coefficients derived from $\mathcal{V}$ trained via an objective with the $\ell_{2,1}$ norm regularization *only*. **Bottom:** VBM coefficients derived from $\mathcal{V}$ trained via an objective with the $\ell_{2,1}$ norm regularization *only*.

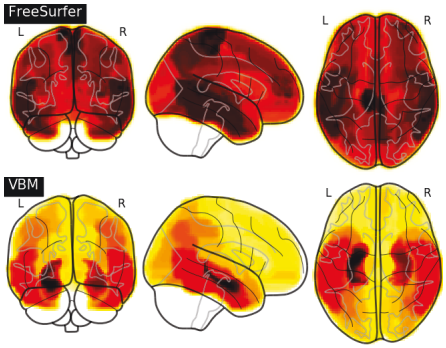

Fig. 9. **Top:** FreeSurfer coefficients derived from $\mathcal{V}$ trained via an objective with the ℓ_1-group norm *only*. **Bottom:** VBM coefficients derived from $\mathcal{V}$ trained via an objective with the ℓ_1-group norm regularization *only*.

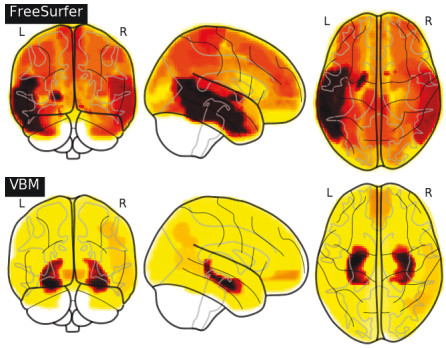

Fig. 10. **Top:** FreeSurfer coefficients derived from $\mathcal{V}$ trained via an objective with the trace norm regularization *only*. **Bottom:** VBM coefficients derived from $\mathcal{V}$ trained via on objective with the trace norm regularization *only*.

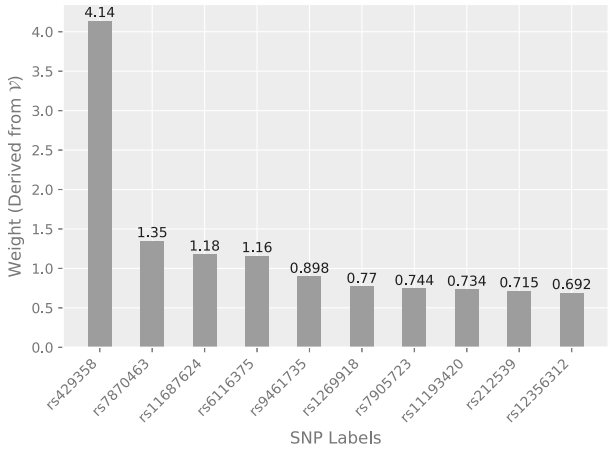

Fig. 11. The top-10 SNP weights derived from the learned matrix $\mathcal{V}$. Each of the values reported on the top of each bar represent the linear weights associated with each SNP. A higher value indicates that SNP is more predictive of a particular AD diagnosis or RAVLT score prediction.

readers that the *Biomarker Identification* properties of our method clearly illustrate the benefits of the fully regularized method when compared to its degenerate versions.

In Figure 8-10 we can see that all the three degenerate versions of the proposed method do not sparsify the parameter tensor $\mathcal{V}$ to the same degree as the fully regularized objective visualized in Figures 6 and 7. When all three regularizations are incorporated, we find that our model unambiguously selects certain areas of the brain. The results of this strong feature selection appear to benefit the performance of the final model. In addition, this feature selection provides our method with a significant clinical advantage: when our method is fully regularized, it clearly identifies specific areas of the brain. This feature selection property, afforded by the fully regularized objective in (2), warrants the potential of our method for identifying AD-relevant biomarkers for future research.

Genetic Modality. In Figure 11 we rank the top-10 features contained within the SNP modality derived from $\mathbf{V}$. As expected, the highest impact SNP discovered by our algorithm is *rs429358*. This SNP, frequently known as the APOE-$\varepsilon 4$ allele, has been found [33] to be predictive of early-onset AD. The coefficient associated with *rs420358* is approximately three times larger than the second largest coefficient displayed in Figure 11. Besides *rs420358*, other AD research [34] also mentions the following high-impact SNPs discovered by our algorithm: *rs11687624*, *rs1269918* and *rs11193420*.

IV. Conclusion

Developing effective methods for modeling the relationship between a variety of different input and output modalities is an important topic in AD research. In the presented *Joint Multi-Modal Longitudinal Regression and Classification* method, we show how an appropriately regularized regression and classification model can provide state-of-the-art performance in predicting the cognitive progression of participants within the ADNI. In addition to the performance improvements enabled, the convergence of our method is rigorously proven. The potential for our method to be used to identify relevant biomarkers across time, perhaps its biggest asset, is supported by a widely recognized machine learning interface that makes it more straightforward for other groups to test and incorporate our method with their own datasets. Our promising experimental results, read in conjunction with the greater collection of AD research, illustrate the utility of the proposed method.

V. Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research

& Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

[1] M. W. Weiner, P. S. Aisen, C. R. Jack Jr, W. J. Jagust, J. Q. Trojanowski, L. Shaw *et al.*, "The alzheimer's disease neuroimaging initiative: progress report and future plans," *Alzheimer's & Dementia*, vol. 6, no. 3, pp. 202–211, 2010.

[2] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.

[3] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A. J. Saykin *et al.*, "Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 557–562.

[4] H. Wang, F. Nie, and H. Huang, "Multi-view clustering and feature learning via structured sparsity," in *International conference on machine learning*, 2013, pp. 352–360.

[5] H. Wang, F. Nie, H. Huang, and C. Ding, "Heterogeneous visual features fusion via sparse multimodal machine," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3097–3102.

[6] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[7] H. Wang, F. Nie, H. Huang, S. Risacher, A. J. Saykin, L. Shen *et al.*, "Identifying ad-sensitive and cognition-relevant imaging biomarkers via joint classification and regression," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2011, pp. 115–123.

[8] H. Wang, F. Nie, H. Huang, S. L. Risacher, A. J. Saykin, L. Shen *et al.*, "Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning," *Bioinformatics*, vol. 28, no. 12, pp. i127–i136, 2012.

[9] E. M. Reiman and W. J. Jagust, "Brain imaging in the study of alzheimer's disease," *Neuroimage*, vol. 61, no. 2, pp. 505–516, 2012.

[10] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer *et al.*, "Automatic classification of mr scans in alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[11] R. Casanova, B. Wagner, C. T. Whitlow, J. D. Williamson, S. A. Shumaker, J. A. Maldjian *et al.*, "High dimensional classification of structural mri alzheimer's disease data based on large scale regularization," *Frontiers in neuroinformatics*, vol. 5, p. 22, 2011.

[12] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher *et al.*, "High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction," in *Advances in neural information processing systems*, 2012, pp. 1277–1285.

[13] E. Adeli, Y. Meng, G. Li, W. Lin, and D. Shen, "Multi-task prediction of infant cognitive scores from longitudinal incomplete neuroimaging data," *NeuroImage*, vol. 185, pp. 783–792, 2019.

[14] L. Brand, H. Wang, H. Huang, S. Risacher, A. Saykin, L. Shen *et al.*, "Joint high-order multi-task feature learning to predict the progression of alzheimer's disease," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 555–562.

[15] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[16] H. Wang, F. Nie, and H. Huang, "Low-rank tensor completion with spatio-temporal consistency," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[17] Y. Liu, Y. Guo, H. Wang, F. Nie, and H. Huang, "Semi-supervised classifications via elastic and robust embedding," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[18] H. Yang, K. Liu, H. Wang, and F. Nie, "Learning strictly orthogonal p-order nonnegative laplacian embedding via smoothed iterative reweighted method," Ph.D. dissertation, Colorado School of Mines. Arthur Lakes Library.

[19] S. L. Risacher, L. Shen, J. D. West, S. Kim, B. C. McDonald, L. A. Beckett *et al.*, "Longitudinal mri atrophy biomarkers: relationship to conversion in the adni cohort," *Neurobiology of aging*, vol. 31, no. 8, pp. 1401–1418, 2010.

[20] L. Shen, S. Kim, S. L. Risacher, K. Nho, S. Swaminathan, J. D. West *et al.*, "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in mci and ad: A study of the adni cohort," *Neuroimage*, vol. 53, no. 3, pp. 1051–1063, 2010.

[21] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel *et al.*, "Api design for machine learning software: experiences from the scikit-learn project," *arXiv preprint arXiv:1309.0238*, 2013.

[22] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang *et al.*, "Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment," *Frontiers in neuroscience*, vol. 12, 2018.

[23] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3121–3124.

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[25] G. E. Hinton, "Connectionist learning procedures," in *Machine learning*. Elsevier, 1990, pp. 555–610.

[26] D. L. Weimer and M. A. Sager, "Early identification and treatment of alzheimer's disease: social and fiscal outcomes," *Alzheimer's & Dementia*, vol. 5, no. 3, pp. 215–226, 2009.

[27] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix *et al.*, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[28] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi *et al.*, "Machine learning for neuroimaging with scikit-learn," *Frontiers in neuroinformatics*, vol. 8, p. 14, 2014.

[29] Y. Mu and F. H. Gage, "Adult hippocampal neurogenesis and its role in alzheimer's disease," *Molecular neurodegeneration*, vol. 6, no. 1, p. 85, 2011.

[30] G. W. Van Hoesen, B. T. Hyman, and A. R. Damasio, "Entorhinal cortex pathology in alzheimer's disease," *Hippocampus*, vol. 1, no. 1, pp. 1–8, 1991.

[31] S. P. Poulin, R. Dautoff, J. C. Morris, L. F. Barrett, B. C. Dickerson, A. D. N. Initiative *et al.*, "Amygdala atrophy is prominent in early alzheimer's disease and relates to symptom severity," *Psychiatry Research: Neuroimaging*, vol. 194, no. 1, pp. 7–13, 2011.

[32] A. Convit, J. De Asis, M. De Leon, C. Tarshish, S. De Santi, and H. Rusinek, "Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to alzheimer's disease," *Neurobiology of aging*, vol. 21, no. 1, pp. 19–26, 2000.

[33] M. Kamboh, F. Demirci, X. Wang, R. Minster, M. M. Carrasquillo, V. Pankratz *et al.*, "Genome-wide association study of alzheimer's disease," *Translational psychiatry*, vol. 2, no. 5, p. e117, 2012.

[34] G. Laumet, V. Chouraki, B. Grenier-Boley, V. Legry, S. Heath, D. Zelenika *et al.*, "Systematic analysis of candidate genes for alzheimer's disease in a french, genome-wide association study," *Journal of Alzheimer's disease*, vol. 20, no. 4, pp. 1181–1188, 2010.