Improved Prediction of Cognitive Outcomes via Globally Aligned Imaging Biomarker Enrichments Over Progressions

Lyujian Lu, Saad Elbeleidy, Lauren Baker, Hua Wang, Li Shen, Huang Heng, and *for the ADNI*

***Abstract*— *Objective:* Longitudinal neuroimaging data have been widely used to predict clinical scores for automatic diagnosis of Alzheimer's Disease (AD) in recent years. However, incomplete temporal neuroimaging records of the patients pose a major challenge to use these data for accurately diagnosing AD. In this paper, we propose a novel method to learn an enriched representation for imaging biomarkers, which simultaneously captures the information conveyed by both the baseline neuroimaging records of all the participants in a studied cohort and the progressive variations of the available follow-up records of every individual participant. *Methods:* Taking into account that different participants usually take different numbers of medical records at different time points, we develop a robust learning objective that minimizes the summations of a number of not-squared ℓ_2-norm distances, which, though, is difficult to efficiently solve in general. Thus we derive a new efficient iterative algorithm with rigorously proved convergence. *Results:* We have conducted extensive experiments using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Clear performance gains have been achieved when we predict different cognitive scores using the enriched biomarker representations learned by our new method. We further observe that the top selected biomarkers by our proposed method are in perfect accordance with the known knowledge in existing clinical AD studies. *Conclusion:* All these promising experimental results have demonstrated the effectiveness of our new method. *Significance:* We anticipate that our new method is of interest to biomedical engineering communities beyond AD research and have open-sourced the code of our method online.[1]**

***Index Terms*— Alzheimer's Disease, Longitudinal Study, Representation Enrichment, Imaging Biomarker**

Manuscript received February 19, 2021, revised . The work of L. Lu, S. Elbeleidy, L. Baker, and H. Wang was supported in part by the National Science Foundation (NSF) under the grants of IIS 1652943, IIS 1849359, CNS 1932482 and CCF 2029543. The work of L. Shen was supported in part by the National Institutes of Health (NIH) under the grants of R01 EB022574, RF1 AG063481, and R01 LM013463, and by the NSF under the grant of IIS 1837964. The work of H. Huang was supported in part 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. (Corresponding author: Hua Wang.)

L. Lu, S. Elbeleidy, L. Baker and H. Wang are with the Department of Computer Science, Colorado School of Mines, Golden, CO 80401, USA. Email: lyujianlu@mymail.mines.edu, selbeleidy@mymail.mines.edu, laurenzoebaker@mymail.mines.edu, huawangcs@gmail.com.

Li Shen is with the Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: li.shen@pennmedicine.upenn.edu.

Heng Huang is with the Department of Biomedical Informatics and Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15261, USA. Email: heng.huang@pitt.edu.

I. Introduction

ALZHEIMER'S Disease (AD), the most common form of dementia, is characterized by progressive impairment of cognitive and memory functions. A recent research [1] reports that AD is the sixth-leading cause of death in the United States of America, rising significantly every year in terms of the proportion of cause of death. It is also reported that there are 40-50 million AD suffers worldwide, and 1 in 85 people will be affected by AD by 2050 [2]. As a result, an effective presymptomatic diagnosis and treatment of AD would have enormous public health benefits.

Over the past decade, neuroimaging measures have been widely studied to predict disease status of AD and/or cognitive performance. However, there exist several critical limitations in existing predictive models, because many of them routinely perform learning at every time point *separately*, ignoring the longitudinal variations characterized by the temporal brain phenotypes. First, since AD is a progressive neurodegenerative disorder, multiple consecutive neuroimaging records are usually required to monitor the disease progressions. It is apparently beneficial to explore the temporal relations among the longitudinal measurements of the biomarkers for AD studies. Second, the records of neuroimaging biomarkers are often missing at some time points for some participants during the period when AD develops, because it is difficult to conduct medical scans consistently across a large group of participants. This is because higher mortality risk and cognitive impairment hinder older adults from staying in the studies that require multiple visits, which thereby results in incomplete data.

To overcome the first limitation, many studies [3]–[6] explored the temporal data structures of brain phenotypes over time. However, these models often formulate the longitudinal data as a tensor, which inevitably complicates the prediction problem in mathematics. To address the second limitation of data inconsistency, most longitudinal models for AD studies [5]–[7] only make use of the samples with complete temporal

[1]The code package of this paper have been made publicly available online at `https://github.com/lyujian/Improved-Prediction-of-Cognitive-Outcomes`.

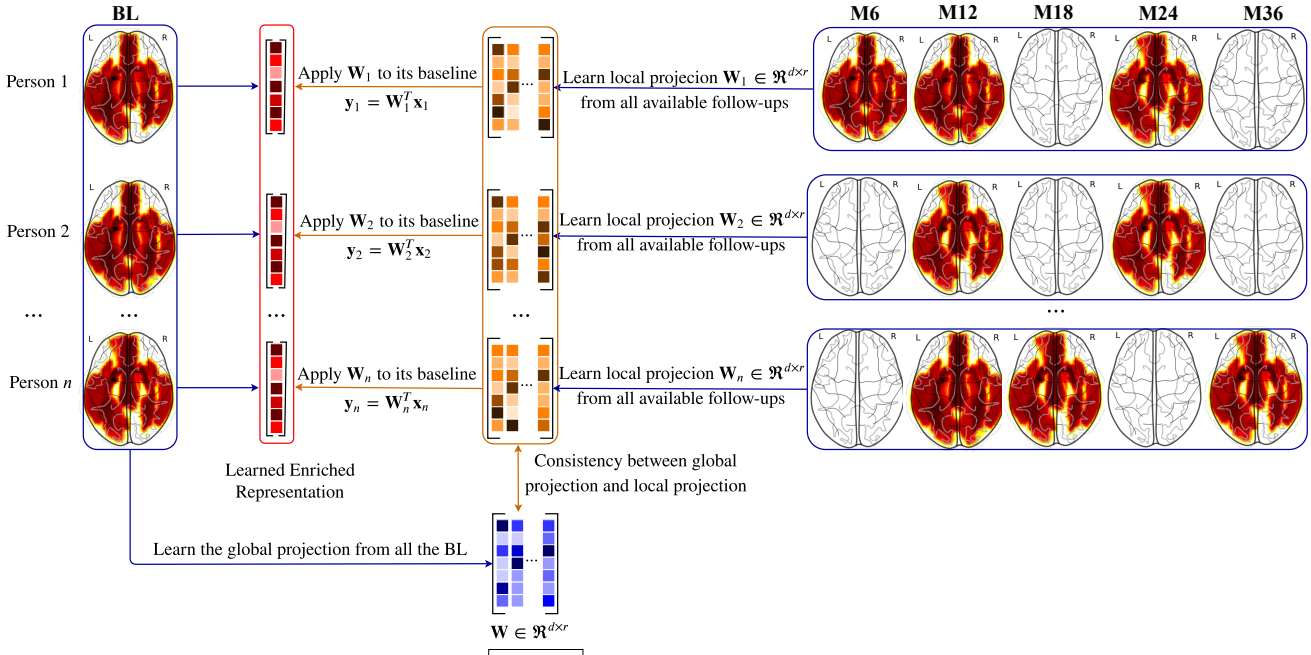

Fig. 1. Illustration of the proposed model to learn the enriched biomarker representations. First, our model learns a global projection $\mathbf{W}$ from the baseline imaging records of all the participants in a studied cohort. Second, it learns a local projection $\mathbf{W}_i$ from the available follow-up records of the i-th participant, which is repeated for every participant. The blank brain plots denote the absence of the brain scans of a participant. Third, the global projection and local projections learned in the above two steps are aligned via a soft constraint. Finally, we get the enriched biomarker representations by projecting the original baseline biomarker representations into subspaces by computing $\{\mathbf{y}_i = \mathbf{W}_i^T\mathbf{x}_i\}_{i=1}^n$, which are a set of fixed-length vectors and can be readily used in traditional machine learning models.

records and ignore the samples with fewer time points, which, however, may potentially neglects substantially valuable information in the data. To solve this problem, data imputation methods [8], [9] were developed to generate missing records over AD progressions. Then the completed data are used for temporal regression analyses. However, missing data imputation methods may introduce undesirable artifacts, which in turn can worsen the predictive power of the longitudinal models.

To fully exploit longitudinal data with incomplete temporal records, in this paper we propose a novel method to learn an enriched biomarker representation to integrate the baseline records of the neuroimaging biomarkers of all the participants in a studied cohort and the dynamic records of each individual taken across the follow-up time points. Instead of solving the missing data problem using imputation, we tackle this challenging problem from a brand new perspective by learning a set of fixed-length vector representations of the imaging biomarkers from varied number of brain scans of the participants over time, which is schematically illustrated in Fig. 1. First, our model learns a global projection from the baseline records of the biomarkers of all the participants to preserve as much information of a studied cohort as possible. Second, we learn a local projection from the available follow-up medical records of every participant in the later couple of years to maintain the local data structures. Finally, a soft constraint is used to ensure that the global and local projections are well aligned. Using the learned projections, we can transform the medical records with inconsistent sizes in a neuroimaging dataset into a set of fixed-length vectors, which can be readily used by conventional machine learning models to predict cognitive outcomes for automatic diagnosis of AD.

We have conducted extensive experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [10]. We compare the predictive power of the baseline biomarker measurements against the enriched biomarker representations learned by our new method, using four different broadly used prediction models in statistical learning, including Ridge Regression (RR), Lasso regression (Lasso), Support Vector Regression (SVR) and Convolutional Neural Network (CNN). In our experiments, we have achieved clear performance gains when we predict ten cognitive scores by using both Voxel-Based Morphometry (VBM) biomarkers [11] and FreeSurfer biomarkers [12] as inputs. In addition, top 10 weighted biomarker features are selected through the project matrix learned by our proposed formulation, which are highly suggestive and nicely agree with the existing clinical research findings.

This paper is an extension of our recent work [13] originally reported in the Proceedings of the Twenty-Second International Conference on Medical Image Computing and Computing Assisted Intervention (MICCAI 2019). In this extended journal manuscript, we provide the following expansions over its conference version:

1) We present a complete optimization framework of the smoothed iterative reweighted method, by which our proposed objective can be efficiently solved with theoretically guaranteed convergence. (Section II-B.1)
2) We provide the mathematical details to derive the algorithm to solve our objective. (Section II-B.2)
3) Experimental evaluations have been significantly expanded for demonstrating the benefits of using the enriched biomarker representations learned by our new

method. (Section III)

a) We report new experimental results by using 1 additional type of neuroimaging markers (the FreeSurfer biomarkers) as input and 9 additional cognitive scores as output predictive targets. (Section III)
b) We compare our proposed method against two different longitudinal feature based methods in the tasks of predicting cognitive scores, where we experiment with both VBM and FreeSurfer biomarkers. (Section III-B)
c) We provide a thorough analysis of the identified disease relevant biomarkers to justify the correctness of our new method from the clinical perspective, which is new in this extended journal manuscript. (Section III-C)

II. Method

As the participants in a studied cohort usually take different numbers of brain scans at different time points, different participant has to be represented by different number of vectors. Specifically, we denote the observed imaging measurements of a participant as: $\mathcal{X}_i = \{\mathbf{x}_i, \mathbf{X}_i\}$, where $i = 1, 2, \cdots, n$ is the index of the participant in a longitudinal dataset. In every $\mathcal{X}_i$, $\mathbf{x}_i \in \Re^d$ denotes the imaging measurements of the i-th participant at the baseline time point, where d counts the number of the imaging biomarkers; and $\mathbf{X}_i = \left[\mathbf{x}_1^i, \ldots, \mathbf{x}_{n_i}^i\right] \in \Re^{d\times n_i}$ collects all available follow-up imaging records of the same participant after the baseline time point, where n_i denotes the number of available follow-up imaging records. Here we emphasize that n_i *varies* with respect to i in a given dataset due to the presence of missing medical records. As a result, one cannot directly use traditional machine learning models, such as RR, Lasso, SVR, CNN, *etc.*, to perform data analyses, because these models can only work with the datasets in which every subject sample is represented by one single fixed-length vector. To tackle this difficulty, our goal for biomarker representation learning is to learn *a fixed-length vector* for every participant that can integrate the baseline record and all available follow-up records of the participant.

A. Our Objective

First, because the neuroimaging measures usually reside in a high-dimensional space, they can be redundant and noisy [3]–[5], [14], [15]. To address this, we aim to learn a compact representation of these measures via a global projection to keep the most useful information in all the baseline records of a given input dataset. To achieve this, we can use the Principal Component Analysis (PCA) [16] method that learns from the input data a linear projection to preserve as much information as possible in a low-dimensional projected subspace. Mathematically, PCA minimizes the reconstruction error via a projection $\mathbf{W} \in \Re^{d\times r}$ (usually $r \ll d$) by minimizing the following objective [16]:

$$\begin{aligned} \mathcal{J}_{\text{Global}}\left(\mathbf{W}\right) &= \sum_{i=1}^{n} \left\|\mathbf{x}_i - \mathbf{W}\mathbf{W}^T\mathbf{x}_i\right\|_2^2, \\ s.t. \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I}. \end{aligned} \tag{1}$$

Second, because the neuroimaging measurements of every participant usually do not develop drastic change over a short interval, we need maximize the data smoothness in the projected data space for every participant, for which Locality Preserving Projections (LPP) [17] is the right tool to use. Given the pairwise similarity matrix $\mathbf{S}_i \in \Re^{n_i\times n_i}$ of the i-th subject, LPP preserves the local relationships and maximizes the smoothness of the data in the embedding space by minimizing the following objective [17]:

$$\begin{aligned} \mathcal{J}_{\text{Local}}\left(\mathbf{W}_i\right) &= \sum_{\mathbf{x}_j^i, \mathbf{x}_k^i \in \mathbf{X}_i} s_{jk}^i \left\|\mathbf{W}_i^T\mathbf{x}_j^i - \mathbf{W}_i^T\mathbf{x}_k^i\right\|_2^2, \\ s.t. \quad & \mathbf{W}_i^T\mathbf{W}_i = \mathbf{I}, \end{aligned} \tag{2}$$

where s_{jk}^i assesses the pairwise affinity between the available records of the i-th participant at the j-th and k-th time points.

Now we integrate the global and local projections learned above by developing a combined objective that minimizes:

$$\begin{aligned} \mathcal{J}_{\ell_2^2}(\mathcal{W}) = & \sum_{i=1}^{n} \left\|\mathbf{x}_i - \mathbf{W}\mathbf{W}^T\mathbf{x}_i\right\|_2^2 \\ & + \alpha \sum_{i=1}^{n} \sum_{\mathbf{x}_j^i, \mathbf{x}_k^i \in \mathbf{X}_i} s_{jk}^i \left\|\mathbf{W}_i^T\mathbf{x}_j^i - \mathbf{W}_i^T\mathbf{x}_k^i\right\|_2^2 \\ & + \beta \sum_{i=1}^{n} \left\|\mathbf{W} - \mathbf{W}_i\right\|_2^2, \\ s.t. \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I}, \ \mathbf{W}_i^T\mathbf{W}_i = \mathbf{I}, \end{aligned} \tag{3}$$

where we denote $\mathcal{W} = \{\mathbf{W}, \mathbf{W}_1, \cdots, \mathbf{W}_n\}$. Through the third term of Eq. (3), the projections $\{\mathbf{W}_i\}_{i=1}^n$ learned from the each individual participant separately are aligned with the projection $\mathbf{W}$ learned globally from the baseline measurements of all the participants of the entire dataset. As a result, the information encoded by the global projection $\mathbf{W}$ learned from all the participants as a whole can be transferred to the biomarker representations of each individual participant, which we call as "*enrichment*". Here we note that our model can also benefit from the subject with less than two longitudinal observations. If the subject has less than two longitudinal observations, the second term $\alpha \sum_{\mathbf{x}_j^i, \mathbf{x}_k^i \in \mathbf{X}_i} s_{jk}^i \left\|\mathbf{W}_i^T\mathbf{x}_j^i - \mathbf{W}_i^T\mathbf{x}_k^i\right\|_2^2$ of our objective in Eq. (3) will become 0. Namely, the enriched the representation of this participant will be learned from the global projection of the entire dataset. R2.2

Finally, because the follow-up neuroimaging records of different participants may be taken at different time points, as illustrated in Fig. 1, it can happen that one participant have the imaging scans at the 12th and 24th months, while another participant visits the doctor at the 6th, 12th and 36th months. That is, the follow-up medical records of the participants in a studied cohort are not well aligned in terms of time by nature. Therefore, it is critical to improve our model for better robustness. To this end, we substitute the first two squared ℓ_2-norm terms in Eq. (3) by their *not-squared* counterparts for

improved robustness against outliers [18]–[20] as follows:

$$\begin{aligned}\mathcal{J}_{\ell_2}(\mathcal{W}) = &\sum_{i=1}^{n}\left\|\mathbf{x}_i - \mathbf{W}\mathbf{W}^T\mathbf{x}_i\right\|_2 \\ &+\alpha\sum_{i=1}^{n}\sum_{\mathbf{x}_j^i,\mathbf{x}_k^i\in\mathbf{X}_i} s_{jk}^i\left\|\mathbf{W}_i^T\mathbf{x}_j^i - \mathbf{W}_i^T\mathbf{x}_k^i\right\|_2 \\ &+\beta\sum_{i=1}^{n}\left\|\mathbf{W}-\mathbf{W}_i\right\|_2^2, \\ s.t.\quad &\mathbf{W}^T\mathbf{W}=\mathbf{I},\ \mathbf{W}_i^T\mathbf{W}_i=\mathbf{I}.\end{aligned} \tag{4}$$

By solving Eq. (4), we can obtain the fixed-length biomarker representation for every participant computing $\{\mathbf{y}_i = \mathbf{W}_i^T\mathbf{x}_i\}_{i=1}^n$, which can be readily fed into traditional machine learning models for subsequent data analyses.

B. The Solution Algorithm and Its Convergence Analysis

Although the motivations of the proposed objective in Eq. (4) are clearly justified, it is a non-smooth optimization problem. Thus, it is difficult to efficiently solve in general. To tackle this difficulty, in this section we derive an efficient iterative algorithm to solve our objective.

1) The Smoothed Iterative Reweighted Method: First, we introduce a general optimization framework, which solve the following general optimization problem:

$$\min_x f(x) + \sum_i \left\|g_i(x)\right\|_2, \tag{5}$$

where $g_i(x)$ is a vector output function. Apparently, our objective in Eq. (4) is a special case of the problem in Eq. (5).

Because the problem in Eq. (5) is non-smooth, we turn to solve the following smooth problem:

$$\min_x f(x) + \sum_i \sqrt{g_i^T(x)g_i(x)+\delta}, \tag{6}$$

where $\delta > 0$ is a small positive constant. It can be verified that, when $\delta \to 0$, Eq. (6) is reduced to Eq. (5). By setting the derivative of Eq. (6) with respect to x to zero, we have:

$$f'(x) + \sum_i \frac{g_i(x)}{\sqrt{g_i^T(x)g_i(x)+\delta}} = 0. \tag{7}$$

Denote

$$s_i = \frac{1}{2\sqrt{g_i^T(x)g_i(x)+\delta}}, \tag{8}$$

we can rewrite Eq. (7) as follows:

$$f'(x) + \sum_i 2s_i g_i(x) = 0. \tag{9}$$

Because s_i is dependent on x, Eq. (9) is generally difficult to solve. However, if s_i is given for a specific i, solving Eq. (9) is equivalent to solving the following optimization problem:

$$\min\ f(x) + \sum_i s_i g_i^T(x)g_i(x). \tag{10}$$

With the above observation, we propose an iterative algorithm, as summarized in Algorithm 1, to find the solution of Eq. (7), which is also the optimal solution of the problem in Eq. (6).

Algorithm 1: The algorithm to solve the problem in Eq. (6).

Initialize x;
while *not converge* **do**
1. For each i, calculate s_i according to Eq. (8);
2. Update x by solving the problem Eq. (10);

Because Algorithm 1 is an iterative algorithm, it is critical to rigorously prove its convergence in mathematics, for which we first prove the following lemma.

Lemma 1: For any vectors x, $\tilde{x}$ with the same size, the following inequality holds:

$$\sqrt{\tilde{x}^T\tilde{x}+\delta} - \frac{\tilde{x}^T\tilde{x}}{2\sqrt{x^Tx+\delta}} \le \sqrt{x^Tx+\delta} - \frac{x^Tx}{2\sqrt{x^Tx+\delta}}. \tag{11}$$

Proof. We begin with an obvious inequality as follows:

$$-(\sqrt{\tilde{x}^T\tilde{x}+\delta} - \sqrt{x^Tx+\delta})^2 \le 0, \tag{12}$$

by which we can derive:

$$\begin{aligned}&-(\sqrt{\tilde{x}^T\tilde{x}+\delta} - \sqrt{x^Tx+\delta})^2 \le 0 \\ \Rightarrow &2\sqrt{\tilde{x}^T\tilde{x}+\delta}\sqrt{x^Tx+\delta} - \left(\tilde{x}^T\tilde{x}+\delta\right) \le x^Tx+\delta \\ \Rightarrow &\sqrt{\tilde{x}^T\tilde{x}+\delta} - \frac{\tilde{x}^T\tilde{x}+\delta}{2\sqrt{x^Tx+\delta}} \le \frac{\sqrt{x^Tx+\delta}}{2} \\ \Rightarrow &\sqrt{\tilde{x}^T\tilde{x}+\delta} - \frac{\tilde{x}^T\tilde{x}+\delta}{2\sqrt{x^Tx+\delta}} \le \sqrt{x^Tx+\delta} - \frac{x^Tx+\delta}{2\sqrt{x^Tx+\delta}} \\ \Rightarrow &\sqrt{\tilde{x}^T\tilde{x}+\delta} - \frac{\tilde{x}^T\tilde{x}}{2\sqrt{x^Tx+\delta}} \le \sqrt{x^Tx+\delta} - \frac{x^Tx}{2\sqrt{x^Tx+\delta}},\end{aligned} \tag{13}$$

which completes the proof. □

Equipped with Lemma 1, we can now prove the following theorem that guarantees the convergence of Algorithm 1.

Theorem 1: Algorithm 1 monotonically decreases the objective in Eq. (6) in each iteration.

Proof. In step 2 of Algorithm 1, we denote the updated x as $\tilde{x}$. According to step 2, we know that:

$$f(\tilde{x}) + \sum_i s_i g_i^T(\tilde{x})g_i(\tilde{x}) \le f(x) + \sum_i s_i g_i^T(x)g_i(x). \tag{14}$$

According to Eq. (8), we have:

$$\begin{aligned}&f(\tilde{x}) + \sum_i \frac{g_i^T(\tilde{x})g_i(\tilde{x})}{2\sqrt{g_i^T(x)g_i(x)+\delta}} \\ \le &f(x) + \sum_i \frac{g_i^T(x)g_i(x)}{2\sqrt{g_i^T(x)g_i(x)+\delta}}.\end{aligned} \tag{15}$$

According to Lemma 1, we have:

$$\begin{aligned}&\textstyle\sum_i \sqrt{g_i^T(\tilde{x})g_i(\tilde{x})+\delta} - \sum_i \frac{g_i^T(\tilde{x})g_i(\tilde{x})}{2\sqrt{g_i^T(x)g_i(x)+\delta}} \\ \le &\textstyle\sum_i \sqrt{g_i^T(x)g_i(x)+\delta} - \sum_i \frac{g_i^T(x)g_i(x)}{2\sqrt{g_i^T(x)g_i(x)+\delta}}.\end{aligned} \tag{16}$$

By summing the above two equations on the both sides, we have:

$$\begin{aligned} & f(\tilde{x}) + \sum_i \sqrt{g_i^T(\tilde{x}) g_i(\tilde{x}) + \delta} \\ \leq & f(x) + \sum_i \sqrt{g_i^T(x) g_i(x) + \delta}, \end{aligned} \tag{17}$$

which completes the proof that Algorithm 1 monotonically decreases the objective of the problem in Eq. (6) in each iteration, until the algorithm converges. □

Here we note that the iterative reweighted method introduced [21], [22] solves the nonsmooth ℓ_1-norm or $\ell_{2,1}$-norm minimization problems. However, the method described in [21], [22] do not explicitly use the smoothness constant (*i.e*, σ in Eq. (6)). Without this smoothness term, the algorithm is heavily impacted by the singularity problem due to inverted matrices that divide 0s, which result in inferior performances of the learning models. To improve the numerical stability, we formally add a smoothness term (*i.e*, σ in Eq. (6)) and theoretically prove the convergence of our algorithm in which the smoothness term leads to much more numerically stable solutions. We call Algorithm 1 as the proposed *Smoothed Iterative Reweighted Method*, which can be broadly used to solve a variety of difficult machine learning problems that minimize the objectives using ℓ_1-norm or $\ell_{2,1}$-norm minimization problems.

Using the proposed smoothed iterative reweighted method to solve our objective in Eq. (4), we need solve the optimization problem in Step 2 of Algorithm 1 (*i.e.*, the problem in Eq. (10)) in every iteration, which, in our case, is to minimize the following objective:

$$\begin{aligned} \mathcal{J}^{\mathrm{R}}_{\ell_2}(\mathcal{W}) = & \operatorname{tr}\left(\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\right)\mathbf{\Gamma}\left(\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\right)^T \\ & + \alpha \sum_{i=1}^{n} \operatorname{tr}\left(\mathbf{W}_i^T\mathbf{X}_i\mathbf{L}_i\mathbf{X}_i^T\mathbf{W}_i\right) + \beta \sum_{i=1}^{n} \|\mathbf{W} - \mathbf{W}_i\|_2^2, \\ s.t. \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I},\ \mathbf{W}_i^T\mathbf{W}_i = \mathbf{I}, \end{aligned} \tag{18}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m] \in \Re^{d\times n}$ summarizes the baseline imaging measurements of all the participants, and $\mathbf{\Gamma} \in \Re^{n\times n}$ is a diagonal matrix and its i-th element is $\gamma^i = \frac{1}{2\sqrt{\|\mathbf{x}_i - \mathbf{W}\mathbf{W}^T\mathbf{x}_i\|_2^2 + \delta}}$. Defining $\theta^i_{jk} = \frac{1}{2\sqrt{\left\|\mathbf{W}_i^T\mathbf{x}_j^i - \mathbf{W}_i^T\mathbf{x}_k^i\right\|_2^2 + \delta}}$ and $\tilde{\mathbf{S}}_i \in \Re^{n_i\times n_i}$ whose element value is $\tilde{s}^i_{jk} = \theta^i_{jk}s^i_{jk}$, in Eq. (18) we compute $\mathbf{L}^i = \mathbf{D}^i - \tilde{\mathbf{S}}^i$, where $\mathbf{D}^i$ is a diagonal matrix whose diagonal entries are the column (or row) sums of $\tilde{\mathbf{S}}_i$, *i.e.*, $d_{jj} = \sum_j \tilde{s}_{jk}$.

2) The Algorithm to Minimize the Objective in Eq. (18): Now we derive the algorithm to solve the problem in Eq. (18) using the Alternating Direction Method of Multipliers (ADMM), which was proposed in [23] to solve convex optimization problems by breaking them into smaller pieces that are easier to handle. Specifically, given the following objective with the equality constraint:

$$\min_{x,z} f(x) + g(z), \qquad s.t. \quad h(x,z) = 0, \tag{19}$$

Algorithm 2 solves the problem in Eq. (19) by decoupling it into subproblems and optimizing each variable while fixing the others, where y is the Lagrangian multiplier to the constraint function h. It is worth noting that Algorithm 2 has been proved to converge Q-linearly to the optimal solution [23].

Algorithm 2: The ADMM algorithm.

Set $1 < \rho < 2$ and initialize $\mu > 0$ and y;
while *not converge* **do**
- **1.** Update x by solving $x^{k+1} = \arg\min_x (f(x) + \frac{\mu}{2}\|h(x, z^k) + \frac{y^k}{\mu}\|^2)$;
- **2.** Update z by solving $z^{k+1} = \arg\min_z (g(z) + \frac{\mu}{2}\|h(x^{k+1}, z) + \frac{y^k}{\mu}\|^2)$;
- **3.** Update y by $y^{k+1} = y^k + \mu h(x^{k+1}, z^{k+1})$;
- **4.** Update μ by $\mu = \rho\mu$.

Using the ADMM framework in Algorithm 2, we can rewrite our objective in Eq. (18) as follows:

$$\begin{aligned} \mathcal{J}^{\mathrm{ADMM}}_{\ell_2}(\mathcal{W}, \mathcal{P}) = & \operatorname{tr}\left(\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\right)\mathbf{\Gamma}\left(\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\right)^T \\ & + \alpha \sum_{i=1}^{n} \operatorname{tr}\left(\mathbf{W}_i^T\mathbf{X}_i\mathbf{L}_i\mathbf{X}_i^T\mathbf{W}_i\right) + \beta \sum_{i=1}^{n} \|\mathbf{P} - \mathbf{P}_i\|_F^2 \\ & + \frac{\mu}{2}\left\|\mathbf{W} - \mathbf{P} + \frac{1}{\mu}\mathbf{\Lambda}\right\|_2^2 + \sum_{i=1}^{n}\frac{\mu}{2}\left\|\mathbf{W}_i - \mathbf{P}_i + \frac{1}{\mu}\mathbf{\Lambda}_i\right\|_2^2, \\ s.t. \quad & \mathbf{P}^T\mathbf{P} = \mathbf{I},\ \mathbf{P}_i^T\mathbf{P}_i = \mathbf{I}, \end{aligned} \tag{20}$$

where $\mathcal{P} = \{\mathbf{P}, \mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_n\}$, $\mathbf{\Lambda} \in \Re^{d\times r}$ is the Lagrangian multiplier for the constraint of $\mathbf{W} = \mathbf{P}$, and $\mathbf{\Lambda}_i \in \Re^{d\times r}$ is the Lagrangian multiplier for the constraint of $\mathbf{W}_i = \mathbf{P}_i$. Now we solve the problem in Eq. (20) as following.

Step 1. When $\mathbf{W}$, $\mathbf{P}$, $\mathbf{P}_i$, $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_i$ are fixed, the objective in Eq. (20) with respect to $\mathbf{W}_i$ can be rewritten as follows:

$$\begin{aligned} \mathcal{J}^{\mathrm{ADMM}}_{\ell_2}(\mathbf{W}_i) = & \alpha \sum_{i=1}^{n} \operatorname{tr}\left(\mathbf{W}_i^T\mathbf{X}_i\mathbf{L}_i\mathbf{X}_i^T\mathbf{W}_i\right) \\ & + \frac{\mu}{2}\left\|\mathbf{W} - \mathbf{P} + \frac{1}{\mu}\mathbf{\Lambda}\right\|_2^2 + \sum_{i=1}^{n}\frac{\mu}{2}\left\|\mathbf{W}_i - \mathbf{P}_i + \frac{1}{\mu}\mathbf{\Lambda}_i\right\|_2^2. \end{aligned} \tag{21}$$

Taking the derivative of Eq. (21) with respect to $\mathbf{W}_i$ and setting it to $\mathbf{0}$, we can get the solution by computing the following:

$$\mathbf{W}_i = \left(2\alpha\mathbf{X}_i\mathbf{L}_i\mathbf{X}_i^T + \mu\mathbf{I}\right)^{-1}\left(\mu\mathbf{P}_i - \mathbf{\Lambda}_i\right). \tag{22}$$

Step 2. When $\mathbf{W}$, $\mathbf{W}_i$, $\mathbf{P}$, $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_i$ are fixed, the objective in Eq. (20) with respect to $\mathbf{P}_i$ can be rewritten as:

$$\max_{\mathbf{P}_i} \ \operatorname{Tr}(\mathbf{P}_i\mathbf{N}_i), \quad s.t. \quad \mathbf{P}_i^\top\mathbf{P}_i = \mathbf{I}, \tag{23}$$

where $\mathbf{N}_i = 2\beta\mathbf{P} + \mu\mathbf{W}_i + \mathbf{\Lambda}_i$. The problem in Eq. (20) can be solved by computing SVD of $\mathbf{N}_i$: if $\operatorname{svd}(\mathbf{N}_i) = \mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^T$, the solution of Eq. (23) is given by $\mathbf{U}_i\mathbf{V}_i^T$ [24, Theorem 1].

Step 3. When $\mathbf{W}_i$, $\mathbf{P}$, $\mathbf{P}_i$, $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_i$ are fixed, the objective in Eq. (20) with respect to $\mathbf{W}$ can be rewritten as follows:

$$\begin{aligned} \mathcal{J}^{\mathrm{ADMM}}_{\ell_2}(\mathbf{W}) = & \operatorname{tr}\left(\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\right)\mathbf{\Gamma}\left(\mathbf{X} - \mathbf{W}\mathbf{W}^T\mathbf{X}\right)^T \\ & + \frac{\mu}{2}\left\|\mathbf{W} - \mathbf{P} + \frac{1}{\mu}\mathbf{\Lambda}\right\|_2^2. \end{aligned} \tag{24}$$

Algorithm 3: Solve the optimization problem in Eq. (20).

Initialization: $\mathbf{W}$, $\mathbf{W}_i$, $\mathbf{P}$, $\mathbf{P}_i$, $\mathbf{\Lambda}$, $\mathbf{\Lambda}_i$, $1 < \rho < 2$, $\mu, \alpha, \beta > 0$;
while *not converge* **do**
1. Update $\mathbf{W}_i$ by $\mathbf{W}_i = \left(2\alpha \mathbf{X}_i \mathbf{L}_i \mathbf{X}_i^T + \mu \mathbf{I}\right)^{-1} (\mu \mathbf{P}_i - \mathbf{\Lambda}_i)$;
2. Update $\mathbf{P}_i$ by $\mathbf{P}_i = \mathbf{U}_i \mathbf{V}_i^T$, where $\mathbf{N}_i = 2\beta \mathbf{P} + \mu \mathbf{W}_i + \mathbf{\Lambda}_i$ and $\text{svd}(\mathbf{N}_i) = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^T$;
3. Update $\mathbf{W}$ by $\mathbf{W} = \left(\mu \mathbf{I} - 2\mathbf{X}\mathbf{\Gamma}\mathbf{X}^T\right)^{-1} (\mu \mathbf{P} - \mathbf{\Lambda})$;
4. Update $\mathbf{P}$ by $\mathbf{P} = \mathbf{U}\mathbf{V}^T$, where $\mathbf{N} = 2\beta \sum_{i=1}^{n} \mathbf{P}_i + \mu \mathbf{W} + \mathbf{\Lambda}$ and $\text{svd}(\mathbf{N}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$;
5. Update $\mathbf{\Lambda}_i$ by $\mathbf{\Lambda}_i = \mathbf{\Lambda}_i + \mu (\mathbf{W}_i - \mathbf{P}_i)$;
6. Update $\mathbf{\Lambda}$ by $\mathbf{\Lambda} = \mathbf{\Lambda} + \mu (\mathbf{W} - \mathbf{P})$;
7. Update μ by $\mu = \rho\mu$;
Output: $\mathbf{W}$, $\mathbf{W}_i$.

Taking the derivative of Eq. (24) with respect to $\mathbf{W}$ and seting it to $\mathbf{0}$, we can get the solution by computing the following:

$$\mathbf{W} = \left(\mu \mathbf{I} - 2\mathbf{X}\mathbf{\Gamma}\mathbf{X}^T\right)^{-1} (\mu \mathbf{P} - \mathbf{\Lambda}). \tag{25}$$

Step 4. When $\mathbf{W}$, $\mathbf{W}_i$, $\mathbf{P}_i$, $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_i$ are fixed, the objective in Eq. (20) with respect to $\mathbf{P}$ can be rewritten as follows:

$$\max_{\mathbf{P}} \ \text{Tr}(\mathbf{P}\mathbf{N}), \quad s.t. \quad \mathbf{P}^\top \mathbf{P} = \mathbf{I} \tag{26}$$

where $\mathbf{N} = 2\beta \sum_{i=1}^{n} \mathbf{P}_i + \mu \mathbf{W} + \mathbf{\Lambda}$. Similar to Step 2, we compute the SVD of $\mathbf{N}$: if $\text{svd}(\mathbf{N}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, the solution of Eq. (26) is given by $\mathbf{U}\mathbf{V}^T$ [24, Theorem 1].

Step 5. Update $\mathbf{\Lambda}_i$ by $\mathbf{\Lambda}_i = \mathbf{\Lambda}_i + \mu (\mathbf{W}_i - \mathbf{P}_i)$.

Step 6. Update $\mathbf{\Lambda}$ by $\mathbf{\Lambda} = \mathbf{\Lambda} + \mu (\mathbf{W} - \mathbf{P})$.

Step 7. Update μ by $\mu = \rho\mu$.

Finally, we summarize the solution algorithm to minimize the objective in Eq. (20) in Algorithm 3. R2.1 The most computationally intensive steps in our algorithm are computing SVDs in step 2 and step 4 in our algorithm, with a complexity of $\mathcal{O}(dr)$ [25]. Thus the overall complexity of our proposed algorithm is $\mathcal{O}(ndrm)$, where m denotes the iteration times. Empirically our optimization algorithm converge very fast, usually within less than 50 iterations.

III. Experiment Results

The data used in our experiments were obtained from the ADNI database [10]. We downloaded 1.5T MRI scans and demographic information for 821 ADNI-1 participants. We performed VBM and FreeSurfer on the MRI data following [26] and extracted mean modulated gray matter measures for 90 target regions of interest. These measures are adjusted for the baseline intracranial volume using regression weights derived from the Health Control (HC) participants at the baseline. We also downloaded the longitudinal scores of the participants in five independent cognitive assessments including Alzheimer's Disease Assessment Scale (ADAS-cog), Mini-Mental State Examination (MMSE), Fluency test (FLU), Rey's Auditory Verbal Learning Test (RAVLT) and Trail making test (TRAILS). Details about these cognitive assessments are available in the ADNI procedure manuals.

We use 10 cognitive scores as predictive targets in our studies: (1) ADAS TOTAL scores from ADAS-cog; (2) FLU ANIM and (3) FLU VEG scores from FLU; (4) MMSE score from MMSE; (5) RAVLT TOTAL, (6) RAVLT 30 and (7) RAVLT 30 RECOG scores from RAVLT; (8) TRAIL A, (9) TRAIL B and (10) TRAIL B-A scores from TRAILS. The time points examined in this study for both imaging biomarkers and cognitive assessments include the baseline (BL), the 6th month (M6), the 12th month (M12), the 18th month (M18), the 24th month (M24) and the 36th month (M36). All the participants' data used in studying our enriched biomarker representation are required to have a BL MRI measurement, BL cognitive score and at least two available measures from M6/M12/M18/M24/M36. As a result, a total of 544 sample subjects are selected in our study, among which 92 samples are diagnosed with AD, 205 samples are diagnosed to be with Mild Cognitive Impairment (MCI), and 177 samples are HC.

A. Predictive Power of the Enriched Biomarker Representations

We first experimentally evaluate the proposed method by applying it to the ADNI database, where we compare the predictive power of the enriched biomarker representations learned by our new method against the BL MRI measurements using both VBM and FreeSurfer biomarkers respectively.

To validate the effectiveness of our proposed method, we compare the predictive capabilities of the two types of biomarker representations, the enriched representations learned by our new method and the BL biomarker measurements, in the tasks of predicting cognitive outcomes. In our experiments, we implement four most broadly used regression models, including RR, Lasso, SVR, and CNN, to evaluate the predictive power of the two compared biomarker representations. RR is a regularized version of linear regression that uses regularization for the better generalization capability. Lasso regression performs both variable selection and regularization in order to enhance the prediction accuracy. SVR is the regression version of support vector machine, which is broadly used in may different real-world applications. When CNN is used to perform regression, it has demonstrated the superior performance compared to many classical machine learning models.

R2.6 R2.7 For all these regression models, we randomly select 70% samples as the training set, 20% samples as the validation set, and the remaining 10% samples as the testing set. The validation set in our experimental setting is designed to to tune the hyperparameters of the model. The test dataset is used to provide an unbiased evaluation of a final model fit on the training dataset. For RR and Lasso models, we fine tune the regularization parameters by searching the grid of $\{10^{-10}, \ldots, 10^{-1}, 1, 10, \cdots, 10^{10}\}$. For SVR model, the Gaussian kernel is used and we fine tune the parameters via a grid search in $\{10^{-5}, \ldots, 10^{-1}, 1, 10, \cdots, 10^{5}\}$. In the CNN

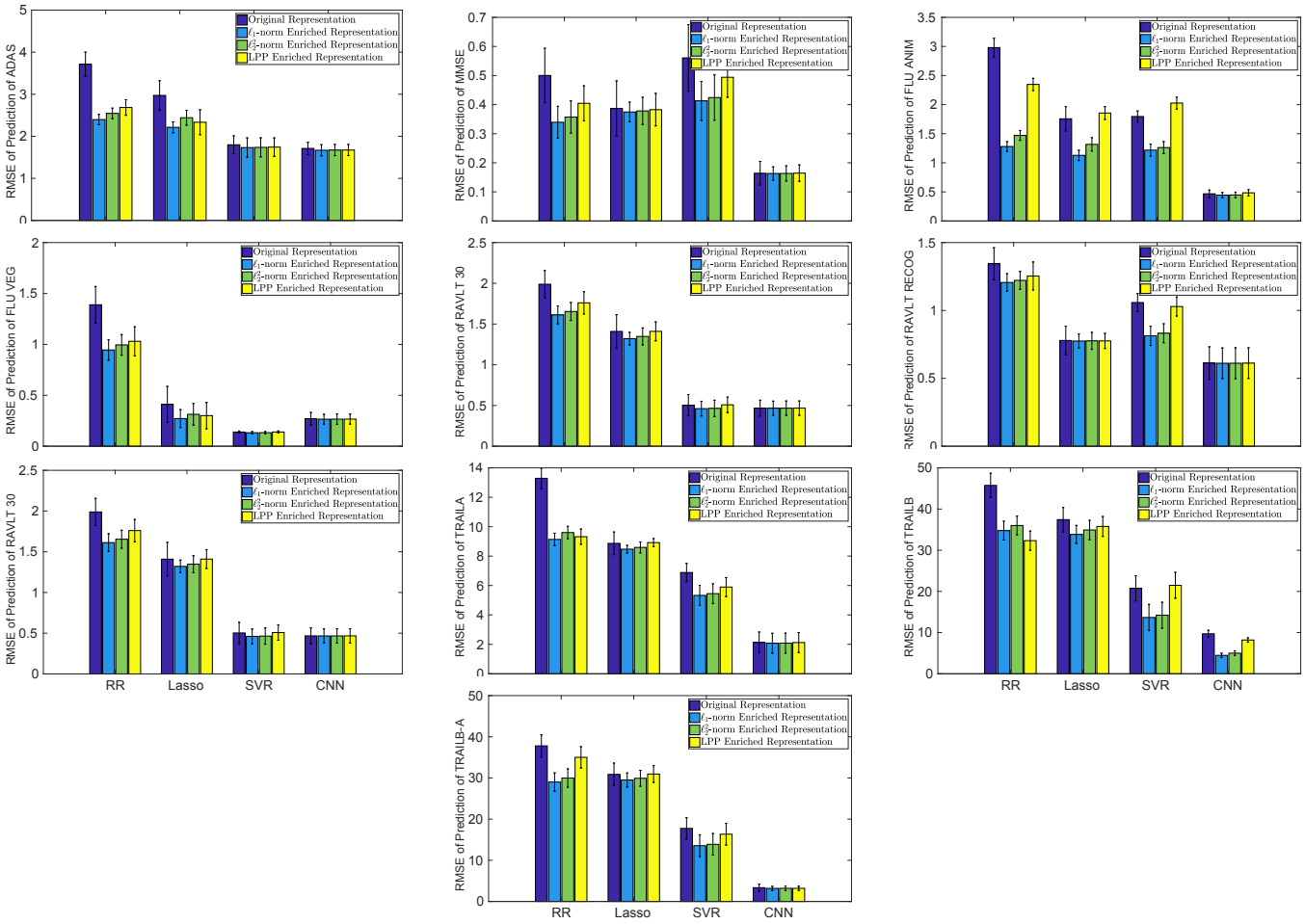

Fig. 2. Comparisons of the predictive performances of the original representations at the baseline time point, ℓ_1-norm enriched representations, squared ℓ_2-norm enriched representations and LPP enriched representations of the VBM biomarkers, when they are used to predict the 10 different baseline cognitive outcomes using the 4 different regression models (RR, Lasso, SVR, and CNN). The RMSEs (smaller is better ↓) for predicting each cognitive outcome by each type of representations are shown for comparison, where the vertical bars show the standard deviations.

model, we construct a two layer convolution architecture for the cognitive outcomes prediction: (1) 16 1×5 convolutions (unpadded convolutions), followed by a Rectified Linear Unit (ReLU) and a 1×2 max pooling operation; (2) 32 1×10 convolutions (unpadded convolutions) with ReLU and a 1×2 max pooling operation. The dropout technique is leveraged to reduce overfitting in the CNN model and prevent complex co-adaptations on training data. The dropout probability is set to be 0.3 and the batch size is set to be 16. The reported performance are based on the results on the testing set.

To evaluate the predictive power of the enriched biomarker representations learned by our new method, we use them as input to predict the 10 cognitive scores by the 4 regression models as listed above. We compare the prediction performances of the enriched biomarker representations against the original biomarker representations at the baseline time point and ℓ_1-norm enriched biomarker representations. Beside the comparison between enriched biomarker representations against its degenerative counterparts, we also compare our proposed enriched biomarker representation with enriched biomarker representations learned from LPP [17]. As a result, for each of the two types of input neuroimaging biomarkers, VBM and FreeSurfer, we end up with prediction 160 tasks. The detailed prediction performance comparisons are reported in Fig. 2 for VBM biomarkers and in Fig. 3 for FreeSurfer biomarkers, which show that our the learned biomarker representations with enrichments outperform the their counterparts in al prediction tasks. The comparisons in these figures show that the predictive capability of the biomarkers have been apparently improved by the enrichments using the information of temporal developments of AD, sometimes very significantly. For example, when we use VBM biomarkers to predict FLU ANIM by RR, the predictive performance of the enriched representations are better than the original ones by about 55%.

While it is exciting to see the clearly improved predictive capability of the enriched biomarker representations, it is more important to study why the temporal enrichments learned by our new method can improve the performance for predicting cognitive outcomes, for which we attribute the enhanced predictive capability of the learned biomarker representations with enrichments to the following two reasons. Firstly, the original baseline neuroimaging biomarker representations are static measurements at one single time point, therefore they cannot benefit from the longitudinal correlations of the neuroimaging

R1.1 R1.2 R1.3

R2.4

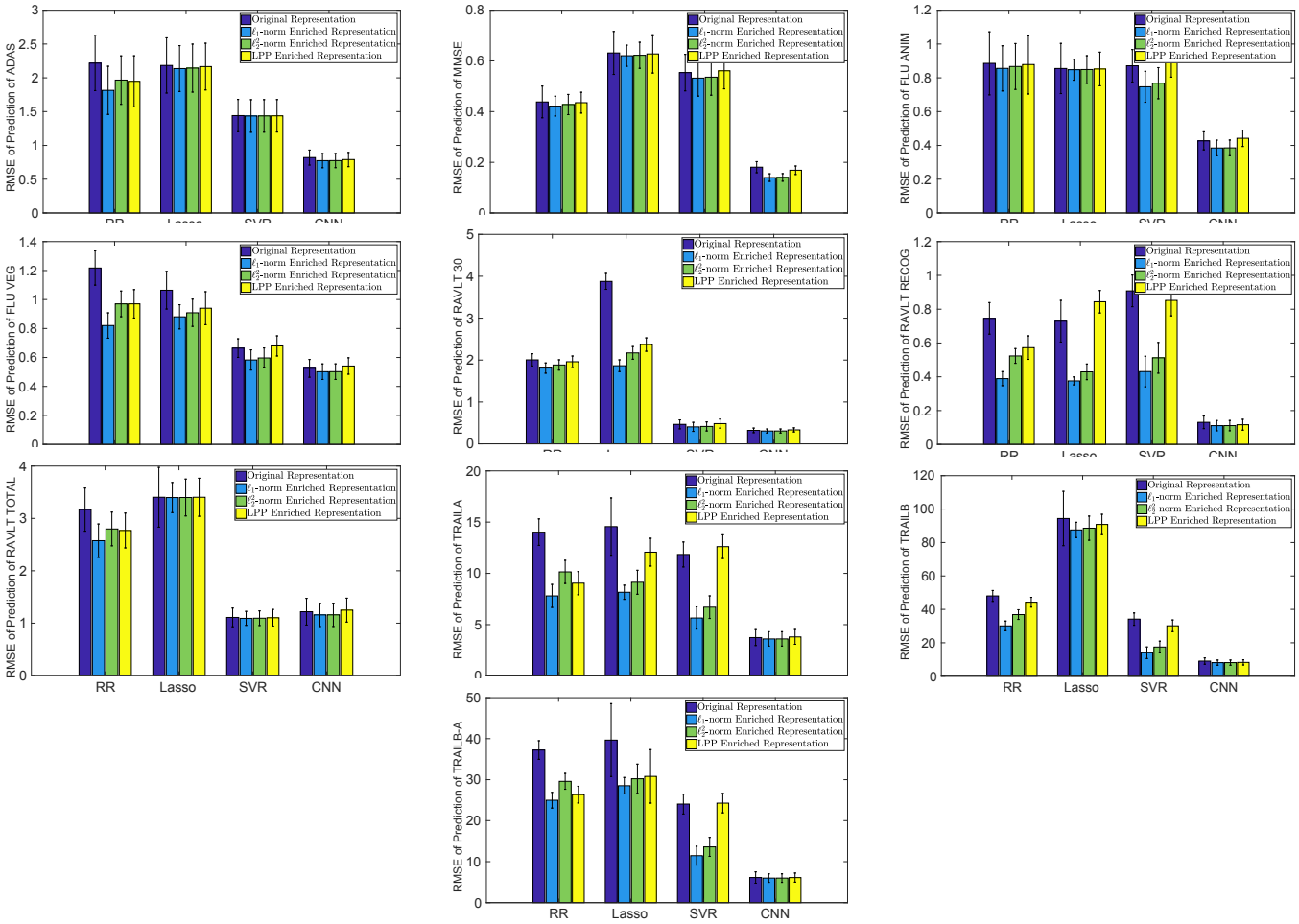

Fig. 3. Comparisons of the predictive performances of the original representations at the baseline time point, ℓ_1-norm enriched representations, squared ℓ_2-norm enriched representations and LPP enriched representations of the FreeSurfer biomarkers, when they are used to predict the 10 different baseline cognitive outcomes using the 4 different regression models (RR, Lasso, SVR, and CNN). The RMSEs (smaller is better ↓) for predicting each cognitive outcome by each type of representations are shown for comparison, where the vertical bars show the standard deviations.

R2.4 measurements when they change over time. In contrast, the enriched representations learned by our new method aim to characterize not only the brain statuses of the participants atthe baseline time point, but also the temporal variations of the same set of measures in AD progressions. Because the cognitive capabilities of AD patients progressively degenerate, integrating longitudinal information of the subjects by our new method is critical for developing prediction models and can improve the predictive power of enriched representations of the biomarkers. Secondly, the original baseline neuroimaging measurements may contain redundant and potentially noisy information [3], [5], [7], [15]. Therefore, by transforming the raw biomarker representations over time via using the projections learned by our new method, we map the baseline cognitive measurements into a lower-dimensional subspace that can mitigate the issues of raw neuroimaging data. This hypothesis is confirmed by all our experimental results in that, compared to the original higher-dimensional neuroimaging measurements, our enriched representations in the projected subspace can achieve significantly better results for predicting cognitive outcomes.

B. Comparison of Our New Method to State-of-the-Art Longitudinal Learning Models

In the previous empirical studies, we compared the enriched biomarker representations learned by our new method against their BL counterparts. The latter, however, are static measurements and only characterize the brain status at the baseline time point, but do not have the information encoded at the follow-up time points. To further demonstrate the advantage of the our new method, we compare its predictive performance against two recent longitudinal learning models, including (1) the temporal group feature (TGF) method [27]; (2) the longitudinal spatial features (LSF) method [28]; R1.2 R2.9 one multi-task based longitudinal methods: Joint Multi-Modal Longitudinal Regression and Classification for Alzheimer's Disease Prediction (JMMLRC) [29], and one RNN model filling based imputation method (RNNMF) [30]. Different from RR, Lasso, SVR and CNN regression models, these methods are able to use the longitudinal data over all the examined time points. In our experiments, after we learn the enriched biomarker representations by our new method, we use CNN for the regression analyses. For these competing methods, we fine tune their parameters following the procedures described in

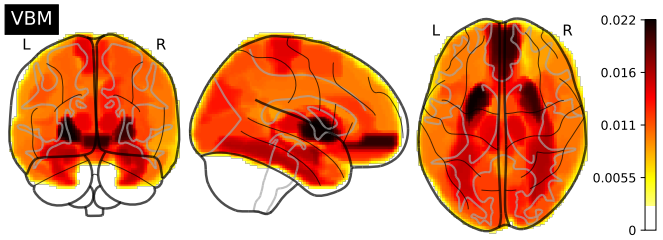

Fig. 4. Visualization of the top 10 VBM biomarkers in the brain map ranked by the relevances to cognitive outcomes learned by our method: LHippocampus and RHippocampus [31], LAmygdala and RAmygdala [32], LPutamen and RPutamen [33], LHeschl and RHeschl [34], LFusiform [33], RParahipp [35].

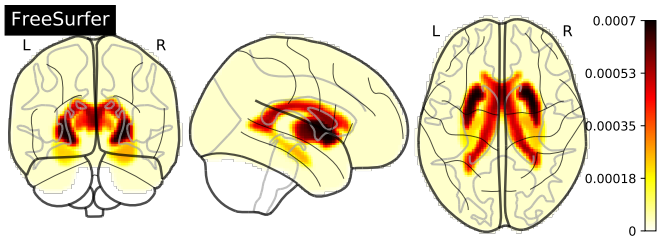

Fig. 5. Visualization of the top 10 FreeSurfer biomarkers in the brain map ranked by the relevances to cognitive outcomes learned by our method: LCerebWM and RCerebWM [36], LCerebCtx and RCerebCtx [37], LCerebellCtx and RCerebellCtx [37], LLatVent and RLatVent [38], LInfLatVent and RInfLatVent [38].

R1.2 R2.9 the respective papers. We report the comparison results in Table I, which show that our new method achieves the best performance when predicting different clinical scores via VBM and FreeSurfer biomarkers respectively, which again firmly confirms the advantage of our new method.

C. Identifying Disease-Relevant Imaging Biomarkers

Apart from better predicting the cognitive outcomes, another important goal of our regression analyses is to identify a subset of neuroimaging markers which are highly correlated to AD progressions. Therefore, we examine the neuroimaging markers identified by our new method. As can be seen in our proposed objective in Eq. (4), while we do not use $\mathbf{W}$ to compute the enriched biomarker representations, it is learned from the baseline biomarker representations of all the participants and is balanced by the projections learned from all the participants individually. Thus, $\mathbf{W}$ encodes the relevances of the input biomarkers to cognitive outcomes and disease status. We use the ℓ_2-norm of every row of $\mathbf{W}$ to quantify the relevances of the biomarkers and visualize them in Fig. 4 for the VBM biomarkers and in Fig. 5 for the FreeSurfer biomarkers.

From Fig. 4 we can see that the VBM biomarkers with highest weights are in perfect accordance with the knowledge documented in existing clinical findings. Specifically, we observe that the bilateral hippocampus are among the top selected biomarkers. The hippocampus is a small organ located within the brain's medial temporal lobe and forms an important part of the limbic system and it is the region that regulates emotions. The hippocampus is mainly associated with memory, in particular long-term memory. This brain region also plays an important role in spatial navigation. Emerging evidence has indicated that altered neurogenesis in the adult hippocampus represents an early critical event in the course of AD [31]. Therefore, this observation firmly confirms the effectiveness of our new method from the clinical perspective. In addition, the bilateral amygdala is also among the top selected biomarkers. We know that the amygdala performs a primary role in the processing of memory, decision-making and emotional response. Thus amygdala is an important subcortical region that is severely and consistently affected by pathology in AD [32]. Finally, We notice that the bilateral putamen are also among the top selected biomarkers. The putamen is a large structure, involved in a very complex feedback loop that prepares and aids in movement of the limbs. It is known in [33] that the volumes of putamen will decrease as AD progresses, which is one more indication of the correctness our new method.

Visualized in Fig. 5 is the significance of the FreeSurfer biomarkers, which again nicely agree with many evidences documented in existing literature. For example, Cerebellar White Matter is among the most significant biomarkers as rated by our new method. It is generally recognized that cerebellar white matter is composed of bundles, which connect various gray matter areas (the locations of nerve cell bodies) of the brain to each other, and carry nerve impulses between neurons. It has been confirmed in existing studies [39] that white matter abnormalities has complex interaction with AD. Another biomarker with top significance is cerebellar cortex, which is supported by [37] in that the processes of aging and AD patients have both differential and partially overlapping effects on specific regions of the cerebellar cortex. Finally, it is broadly accepted in neuroimaging studies that the lateral ventricles are structures within the brain that contain cerebrospinal fluid, a clear, watery fluid that provides cushioning for the brain while also helping to circulate nutrients and remove waste. Fig. 5 shows that the weights of the biomarkers of Lateral Ventricle Volumes are among the top that is consistent with the discovery in [38], in that the lateral ventricle volume will experience a longitudinal change during different periods of AD.

In summary, the identified imaging biomarkers are highly suggestive and strongly agree with existing medical research findings with regard to AD, which warrants the correctness of the discovered imaging cognition associations to reveal the complex relationships between MRI measures and cognitive scores. This is important for both theoretical research and clinical practices for a better understanding of AD mechanism.

IV. Conclusion

Missing data pose a critical challenge in longitudinal AD studies. In this paper, we propose a new method to learn a fixed-length biomarker representation for an input neuroimaging dataset. The enriched biomarker representations simultaneously capture both the global consistency from baseline measurements and local pairwise pattern from available follow-up measurements of each participant. Our experimental results show that the learned biomarker representations with

TABLE I

PERFORMANCE COMPARISONS OF OUR METHOD (CNN IS USED FOR REGRESSION) AGAINST THREE METHODS, TGF [27], LSF [28], JMMLRC [29] AND RNNMF [30] MEASURED BY RMSE (SMALLER IS BETTER ↓), WHERE VBM AND FREESURFER (FS) BIOMARKERS ARE USED AS INPUTS.

Clinical Score	Biomarker	TGF	LSF	JMMLRC	RNNMF	Our Method
ADAS	VBM	4.524 ± 0.213	2.403 ± 0.126	3.871 ± 0.236	1.703 ± 0.141	**1.669 ± 0.132**
	FreeSurfer	4.300 ± 0.175	2.246 ± 0.223	4.139 ± 0.244	2.027 ± 0.106	**0.775 ± 0.105**
MMSE	VBM	2.178 ± 0.083	1.723 ± 0.147	1.719 ± 0.098	0.185 ± 0.022	**0.163 ± 0.023**
	FreeSurfer	2.074 ± 0.057	0.627 ± 0.086	1.425 ± 0.086	0.854 ± 0.016	**0.140 ± 0.015**
FLU_ANIM	VBM	2.553 ± 0.124	1.509 ± 0.113	1.681 ± 0.083	0.456 ± 0.045	**0.445 ± 0.044**
	FreeSurfer	2.638 ± 0.102	1.358 ± 0.132	1.714 ± 0.089	0.617 ± 0.042	**0.385 ± 0.046**
FLU_VEG	VBM	2.649 ± 0.103	1.439 ± 0.151	1.390 ± 0.082	0.480 ± 0.051	**0.266 ± 0.048**
	FreeSurfer	2.918 ± 0.093	1.513 ± 0.183	1.376 ± 0.069	0.571 ± 0.050	**0.502 ± 0.054**
RAVLT_TOTAL	VBM	3.879 ± 0.241	1.514 ± 0.133	1.481 ± 0.095	0.911 ± 0.200	**0.857 ± 0.202**
	FreeSurfer	3.581 ± 0.182	1.431 ± 0.107	1.486 ± 0.127	1.862 ± 0.206	**1.158 ± 0.222**
RAVLT_30	VBM	2.513 ± 0.063	1.713 ± 0.143	1.749 ± 0.128	0.469 ± 0.092	**0.467 ± 0.087**
	FreeSurfer	2.644 ± 0.051	1.665 ± 0.113	1.467 ± 0.124	0.410 ± 0.048	**0.306 ± 0.048**
RAVLT_RECOG	VBM	3.395 ± 0.097	2.213 ± 0.137	2.706 ± 0.114	0.648 ± 0.103	**0.611 ± 0.113**
	FreeSurfer	2.907 ± 0.217	2.318 ± 0.108	2.825 ± 0.130	0.631 ± 0.041	**0.111 ± 0.030**
TRAILA	VBM	15.183 ± 1.017	18.417 ± 1.617	14.819 ± 1.371	2.259 ± 0.710	**2.077 ± 0.677**
	FreeSurfer	12.983 ± 1.317	18.943 ± 1.471	11.521 ± 1.314	3.749 ± 0.645	**3.605 ± 0.703**
TRAILB	VBM	36.714 ± 4.317	38.137 ± 3.717	28.902 ± 3.891	4.469 ± 0.550	**4.467 ± 0.508**
	FreeSurfer	37.714 ± 4.461	43.614 ± 4.471	31.493 ± 4.176	9.975 ± 1.651	**8.134 ± 1.550**
TRAILB-A	VBM	27.371 ± 2.571	28.253 ± 3.751	19.493 ± 3.295	3.370 ± 0.535	**3.186 ± 0.487**
	FreeSurfer	31.708 ± 3.651	36.572 ± 2.981	24.487 ± 2.263	8.782 ± 1.028	**5.988 ± 1.040**

enrichments outperform the baseline biomarker measurements when we predict the cognitive scores. Furthermore, the identified biomarkers are highly suggestive and strongly agree with the existing research findings, which warrants the correctness of our approach and adds to its values for the usage in clinical practices for a better understanding of AD mechanisms.

ACKNOWLEDGEMENT

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.; 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 Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

[1] A. Association *et al.*, "2018 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[2] E. Nichols, C. E. Szoeke, S. E. Vollset, N. Abbasi, F. Abd-Allah, J. Abdela, M. T. E. Aichour, R. O. Akinyemi, F. Alahdab, S. W. Asgedom *et al.*, "Global, regional, and national burden of alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 18, no. 1, pp. 88–106, 2019.

[3] 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 *MICCAI*, 2018, pp. 555–562.

[4] L. Lu, H. Wang, X. Yao, S. Risacher, A. Saykin, and L. Shen, "Predicting progressions of cognitive outcomes via high-order multi-modal multi-task feature learning," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 545–548.

[5] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen, "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.

[6] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, and ADNI, "From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant SNPs," *Bioinformatics*, vol. 28, no. 18, pp. i619–i625, 2012.

[7] X. Wang, J. Yan, X. Yao, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, H. Huang *et al.*, "Longitudinal genotype-phenotype association study via temporal structure auto-learning predictive model," in *International Conference on Research in Computational Molecular Biology*. Springer, 2017, pp. 287–302.

[8] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, J. Ye, A. D. N. Initiative *et al.*, "Bi-level multi-source learning for heterogeneous block-wise missing data," *NeuroImage*, vol. 102, pp. 192–206, 2014.

[9] Y. Li, L. Wang, J. Zhou, and J. Ye, "Multi-task learning based survival analysis for multi-source block-wise missing data," *Neurocomputing*, vol. 364, pp. 95–107, 2019.

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

[11] J. Ashburner and K. J. Friston, "Voxel-based morphometry—the methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

[12] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[13] L. Lu, S. Elbeleidy, L. Baker, H. Wang, H. Huang, L. Shen *et al.*, "Improved prediction of cognitive outcomes via globally aligned imaging biomarker enrichments over progressions," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 140–148.

[14] J. Yan, T. Li, H. Wang, H. Huang, J. Wan, K. Nho, S. Kim, S. L. Risacher, A. J. Saykin, L. Shen *et al.*, "Cortical surface biomarkers for predicting cognitive outcomes using group l2, 1 norm," *Neurobiology of aging*, vol. 36, pp. S185–S193, 2015.

[15] X. Hao, C. Li, J. Yan, X. Yao, S. L. Risacher, A. J. Saykin, L. Shen, D. Zhang, and A. D. N. Initiative, "Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis," *Bioinformatics*, vol. 33, no. 14, pp. i341–i349, 2017.

[16] I. Jolliffe, "Principal component analysis," in *International encyclopedia of statistical science*. Springer, 2011, pp. 1094–1096.

[17] X. He and P. Niyogi, "Locality preserving projections," in *Advances in neural information processing systems*, 2004, pp. 153–160.

[18] H. Wang, F. Nie, and H. Huang, "Learning robust locality preserving projection via p-order minimization," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[19] 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.

[20] K. Liu, L. Brand, H. Wang, and F. Nie, "Learning robust distance metric with side information via ratio minimization of orthogonally constrained l21-norm distances." in *IJCAI*, 2019, pp. 3008–3014.

[21] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[22] 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.

[23] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Athena Scientific, 1996.

[24] H. Wang, F. Nie, H. Huang, and F. Makedon, "Fast nonnegative matrix tri-factorization for large-scale data co-clustering," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[25] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013, vol. 3.

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

[27] D. Zhang, D. Shen, A. D. N. Initiative *et al.*, "Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers," *PloS one*, vol. 7, no. 3, p. e33182, 2012.

[28] J. Zhang, M. Liu, L. An, Y. Gao, and D. Shen, "Alzheimer's disease diagnosis using landmark-based features from longitudinal structural mr images," *IEEE journal of biomedical and health informatics*, vol. 21, no. 6, pp. 1607–1616, 2017.

[29] L. Brand, K. Nichols, H. Wang, L. Shen, and H. Huang, "Joint multi-modal longitudinal regression and classification for alzheimer's disease prediction," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1845–1855, 2019.

[30] M. Nguyen, T. He, L. An, D. C. Alexander, J. Feng, B. T. Yeo, A. D. N. Initiative *et al.*, "Predicting alzheimer's disease progression using deep recurrent neural networks," *NeuroImage*, vol. 222, p. 117203, 2020.

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

[32] 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.

[33] L. De Jong, K. Van der Hiele, I. Veer, J. Houwing, R. Westendorp, E. Bollen, P. De Bruin, H. Middelkoop, M. Van Buchem, and J. Van Der Grond, "Strongly reduced volumes of putamen and thalamus in alzheimer's disease: an mri study," *Brain*, vol. 131, no. 12, pp. 3277–3285, 2008.

[34] M. M. Esiri, R. Pearson, and T. Powell, "The cortex of the primary auditory area in alzheimer's disease," *Brain Research*, vol. 366, no. 1-2, pp. 385–387, 1986.

[35] C. Echávarri, P. Aalten, H. B. Uylings, H. Jacobs, P. J. Visser, E. Gronenschild, F. Verhey, and S. Burgmans, "Atrophy in the parahippocampal gyrus as an early biomarker of alzheimer's disease," *Brain Structure and Function*, vol. 215, no. 3-4, pp. 265–271, 2011.

[36] J. Acosta-Cabronero, G. B. Williams, G. Pengas, and P. J. Nestor, "Absolute diffusivities define the landscape of white matter degeneration in alzheimer's disease," *Brain*, vol. 133, no. 2, pp. 529–539, 2010.

[37] A. Bakkour, J. C. Morris, D. A. Wolk, and B. C. Dickerson, "The effects of aging and alzheimer's disease on cerebral cortical anatomy: specificity and differential relationships with cognition," *Neuroimage*, vol. 76, pp. 332–344, 2013.

[38] C. DeCarli, J. V. Haxby, J. Gillette, D. Teichberg, S. Rapoport, and M. Schapiro, "Longitudinal changes in lateral ventricular volume in datients with dementia of the alzheimer type," *Neurology*, vol. 42, no. 10, pp. 2029–2029, 1992.

[39] A. Moghekar, M. Kraut, W. Elkins, J. Troncoso, A. B. Zonderman, S. M. Resnick, and R. J. O'Brien, "Cerebral white matter disease is associated with alzheimer pathology in a prospective cohort," *Alzheimer's & Dementia*, vol. 8, no. 5, pp. S71–S77, 2012.