Comparison of Single and Multitask Learning for Predicting Cognitive Decline Based on MRI Data

VANDAD IMANI [1], **MITHILESH PRAKASH** [1], **MARZIEH ZARE** [2], **AND JUSSI TOHKA** [1], **for the Alzheimer's Disease Neuroimaging Initiative***

[1] A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, 70211 Kuopio, Finland
[2] Department of Computing Sciences, Tampere University, 33100 Tampere, Finland

Corresponding author: Vandad Imani (vandad.imani@uef.fi)

ABSTRACT The Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) is a neuropsychological tool that has been designed to assess the severity of cognitive symptoms of dementia. Personalized prediction of the changes in ADAS-Cog scores could help in timing therapeutic interventions in dementia and at-risk populations. In the present work, we compared single and multitask learning approaches to predict the changes in ADAS-Cog scores based on T1-weighted anatomical magnetic resonance imaging (MRI). In contrast to most machine learning-based prediction methods ADAS-Cog changes, we stratified the subjects based on their baseline diagnoses and evaluated the prediction performances in each group. Our experiments indicated a positive relationship between the predicted and observed ADAS-Cog score changes in each diagnostic group, suggesting that T1-weighted MRI has a predictive value for evaluating cognitive decline in the entire AD continuum. We further studied whether correction of the differences in the magnetic field strength of MRI would improve the ADAS-Cog score prediction. The partial least square-based domain adaptation slightly improved the prediction performance, but the improvement was marginal. In summary, this study demonstrated that ADAS-Cog change could be, to some extent, predicted based on anatomical MRI. Based on this study, the recommended method for learning the predictive models is a single-task regularized linear regression due to its simplicity and good performance. It appears important to combine the training data across all subject groups for the most effective predictive models.

INDEX TERMS Alzheimer's disease, prediction, MRI, ADAS, heterogeneity reduction, transfer learning, domain adaptation, machine learning algorithms, biomedical imaging, engineering in medicine and biology.

I. INTRODUCTION

Alzheimer's disease (AD) is a chronic neurodegenerative disorder and a major health burden, with 152 million people expected to suffer from AD by 2050 [1]. Pathophysiological changes in AD begin many years prior to clinical manifestations of disease, and the spectrum of AD spans from clinically asymptomatic to severely impaired [2]. Because of this, there is an appreciation that AD should not only be viewed with discrete and defined clinical stages, but also as a multifaceted process moving along a continuum. Mild cognitive impairment (MCI) is an essential concept along this continuum, representing a transitional stage between healthy elderly individuals and AD [3]. Approximately 10% to 20%

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

of MCI patients tend to progress to AD annually, whereas others will continue with cognitive decline or even revert to normal cognition (NC) [4]. Many treatment strategies have been proposed to decelerate AD, with limited success [5]; one problem is that treatments are not administered within the correct time window along the AD continuum. Therefore, early prediction of disease progression is a crucial step towards better therapies, unburdening the health care system, and preventing adverse events caused by AD [6], [7].

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of the ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Cognitive test batteries have been developed to assess the cognitive decline of individuals. Two of the most commonly used standards are the Mini-Mental State Examination (MMSE) [8] and the Alzheimer's Disease Assessment Scale-cognition subscale (ADAS-Cog) [9], which are important criteria for the clinical diagnosis of AD. In the evaluation of cognitive decline due to dementia, ADAS-Cog is considered to be more sensitive and reliable than MMSE [10]. The ADAS-Cog is widely used to evaluate the cognitive state of patients with mild to advanced AD. The modified ADAS-Cog 13 contains 13 items for assessing cognitive dysfunction, with a total score of 0–85, with higher scores indicating greater dysfunction [11].

Due to the above-stated reasons *machine learning* (ML) for predicting ADAS-Cog scores, as opposed to the diagnosis (a recent review in [12]), has been gaining research interest. For example, Utsumi *et al.* [13] found that the combination of two variants of personalized Gaussian process models can improve the accuracy of predicting future ADAS-Cog13 scores using a limited set of subjects with multimodal data, which included imaging biomarkers (MRI, diffusion tensor imaging), cerebrospinal fluid (CSF) biomarkers, demographics, genetics, and cognitive scores. Zhu *et al.* [14] concluded that the canonical feature selection method had a significant effect on improving the performance of sparse multitask learning (MTL) to predict the clinical scores of ADAS-Cog. Prakash *et al.* [15] utilized multivariate regression techniques and determined that longitudinal prediction of AD progression is possible with multimodal data from the baseline, which included MRI, positron emission tomography (PET), CSF biomarkers, cognitive scores, and APOE. Tsao *et al.* [16] found that leveraging hippocampal surface features together with multimodal data, which included sex, age, MRI, APOE, and baseline MMSE, might boost the prediction of cognitive scores, such as MMSE and clinical dementia rating (CDR). However, PET is far less commonly used and more expensive than MRI [17], and obtaining CSF is highly invasive.

Accordingly, an interesting approach to biomarkers is the use of the only standard, T1-weighted MRI, to predict the progression rate of dementia. MRI enables the study of various noninvasive aspects of the human brain to detect biomarkers associated with AD [18], and it is a widely available imaging modality. Indeed, several studies have investigated the association between cognitive scores and MRI biomarkers, such as gray matter volume [19], [20] and cortical and subcortical volumes [21], [22]. Lei *et al.* [23] studied the relationship between MRI data and cognitive scores by introducing a framework that includes correntropy regularized joint learning and a deep polynomial network for feature construction, as well as ensemble learning based on support vector regression for the prediction of cognitive scores. Bhagwat *et al.* [24] proposed an artificial neural network model for predicting cognitive scores from the cortical thickness and hippocampal subfield volumes. Jiang *et al.* [25] proposed a novel MTL

formulation that considers a correlation-aware sparse and low-rank constrained regularization in order to explore the relationship between the MRI features and cognitive scores. Huang *et al.* [26] presented a random forest (RF) with sparse regression and soft-split technique, which adopted probabilistic paths during the testing stage in RF to predict cognitive scores at multiple time points. Zhou *et al.* [27] proposed two MTL formulations based on a temporal group Lasso regularizer and the convex fused sparse group Lasso, which utilize the common temporal patterns of MRI biomarkers to predict disease progression measured by the cognitive scores. To date, most existing progression models focus on predicting cognitive scores derived from the entire AD continuum, from healthy elderly to moderate AD, using a single model, e.g., [28]–[31], with the exception of Duchesne *et al.* [32], who studied the relationship between MRI and one-year MMSE changes in the MCI population. However, there is little evidence that this one-size-fits-all strategy would be optimal. Moreover, individuals at different stages of the continuum can be expected to regress differently (for example, most cognitively normal individuals are likely to be cognitively normal after three years, while most AD patients are expected to have regressed during that time period). Therefore, evaluating the prediction models using the whole-continuum data leads to results that are perhaps hard to interpret, and we argue that the prediction models should be evaluated while stratifying the subject population based on the baseline diagnosis.

The evaluation of the prediction models while stratifying for the baseline diagnosis leads to a question of whether the other diagnostic groups are still useful when training the predictive models. This subsequently leads to the consideration of MTL approaches to improve the generalization performance by simultaneously solving multiple learning tasks while exploiting commonalities and differences across tasks [33]. One of the critical issues in MTL is to identify the essential relatedness between the tasks and to build learning models in order to obtain this task-relatedness. MTL approaches with sparsity-inducing regularization have been studied to investigate the prediction of cognitive measures. For example, Tabarestani *et al.* [34] applied ℓ_1-norm regularization to introduce sparsity among all features that could select a small subset of features to predict MMSE at six-time points. Zhou *et al.* [27] and Lei *et al.* [35] employed joint sparsity regularization ($\ell_{2,1}$-norm) in order to share a common subset of features for all tasks simultaneously, where each task refers to AD progression prediction at a single time point. Wang *et al.* [36] formulated the progression of AD as a weakly supervised temporal multitask matrix regression framework that considers the prediction of cognitive scores at each time point as a regression task. However, MTL has not been studied in cases where different tasks correspond to cognitive score prediction of different diagnostic groups within the AD continuum. In addition, MTL has not been studied for adapting predictive modeling for differences in MRI acquisition.

In this study, we explore whether MRI at the baseline can potentially predict changes in ADAS-Cog scores while stratifying the population based on the baseline diagnosis. To our knowledge, no such analysis has been applied to study the progression of different stages of the AD continuum. We frame this as a single-task or multitask prediction problem, where different tasks correspond to three diagnostic groups of subjects at baseline. We will also address multitask learning in the presence of differences in MRI acquisition; in this case, MRIs have been acquired using two different magnetic field strengths (MFSs). We compare MTL to two heterogeneity reduction approaches, partial least squares (PLS) domain adaptation [37] and ComBat [38].

II. MATERIALS AND METHODS
A. ADNI DATASET
Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).[1] The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early AD.

B. SUBJECTS AND MRI
The data in this study included baseline MRIs from 1376 ADNI subjects (430 NC, 662 MCI, 284 AD), aged 54 to 91 years old. Detailed characteristics of the subjects are presented in Table 1. For these subjects, the baseline MRI data were obtained with a T1-weighted MP-RAGE sequence at 1.5 T (with a $256 \times 256 \times 170$ acquisition matrix and a voxel size of $1.25 \times 1.25 \times 1.2 \ mm^3$) and 3.0 T (with a $256 \times 256 \times 170$ acquisition matrix and a voxel size of $1.0 \times 1.0 \times 1.2 \ mm^3$). Specifically, 808 subjects were from the ADNI-1 cohort with the MRI acquired at 1.5 T, and 571 subjects were from the ADNI-2 cohort with the MRI acquired at 3.0 T. As seen in Table 1, the number of subjects with ADAS-Cog-13 scores decreased during the follow-up due to subject drop-out and missing data. The roster identification numbers (RIDs) of the subjects employed in this study are provided in the Supplementary Material.

C. IMAGE PREPROCESSING
The preprocessing of the T1-weighted images was performed using the fully automated CAT12 package [39] running under MATLAB[2]. T1-weighted images were first denoised by using adaptive nonlocal means filtering [40], then they were corrected for bias field inhomogeneities and segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) [41]. After segmentation, partial volume estimation (PVE) with a simplified mixed model with a maximum of two tissue types was performed, resulting in maps of the tissue type densities [42]. Furthermore, the segmented images

were spatially normalized by utilizing the high-dimensional DARTEL normalization algorithm into the standard MNI space [43]. This procedure resulted in spatially aligned maps for tissue fractions of WM and GM. We only utilized the GM images in this study. Finally, we averaged the gray matter density values according to the brain regions defined by the AAL atlas, resulting in 122 regional GM density values.

D. ADAS-COG SCORE
The ADAS-Cog was developed as an outcome measure to assess the severity of cognitive dysfunction in AD. We consider ADAS-Cog-13, which yields a measure of cognitive performance by combining the original tasks of the ADAS-Cog-11 [44] (subject-completed tests and observer-based assessments) as well as a test of delayed word recall and a number cancellation or maze task [9]. The ADAS-Cog-13, later referred to as ADAS-Cog, scores are in the range of 0 to 85, with higher scores indicating more severe impairment [45]. The ADAS-Cog scores used in this study were acquired once a year as described in the ADNI General Procedures Manuals.[3]

E. OVERVIEW OF THE METHODS
We aim to predict the change in the ADAS-Cog score ($\Delta_t ADAS_i$) for each subject i and each time point ($t = 12, 24, 36$ months) based on regional, MRI-derived gray matter density values at the baseline ($t = 0$ months). The change in the ADAS-Cog score is defined as $\Delta_t ADAS_i = ADAS_i(t) - ADAS_i(0)$, where $ADAS_i(t)$ is the ADAS score of subject i at time t. We frame this as a single-task (Fig. 1(A)) or multitask (Fig. 1(B)) prediction problem, where different tasks correspond to the groups of different, but related, subjects (three diagnostic groups and/or two magnetic field strengths used to acquire the data). More formally, we build S prediction models:

$$\Delta_t ADAS_i = f_d(x_i), \qquad (1)$$

where x_i denotes the MRI (122 regional gray matter density values) of subject i at baseline and f_d is the prediction model for subject group d. We consider the cases where $S = 3$, where the groups are determined based on the baseline diagnosis (NC, MCI, and AD), and $S = 6$, where the groups are determined based on the baseline diagnosis and MFS (NC at 1.5 T, NC at 3 T, MCI at 1.5 T, MCI at 3 T, AD at 1.5 T and AD at 3 T). multitask learning aimed at enhancing the precision of learning algorithms by jointly learning independent variables for multiple tasks. The learning approach works well, especially if these tasks have some commonalities, as we expect f_d to have.

We compare multitask learning to harmonize MRI acquired with different field strengths. We do this by adopting a recent domain adaptation technique [37] and using ComBat data harmonization [46]. The algorithms compared in

---

[1]For up-to-date information, see https://www.adni-info.org.

[2]For more information, see http://www.neuro.uni-jena.de/cat.
[3]http://adni.loni.usc.edu/wp-content/uploads/2010/09/ADNI_General ProceduresManual.pdf

TABLE 1. Demographic information of subjects at different time points. Under the ranges, the mean and standard deviation are provided.

Time point	Baseline diagnosis	No. of Subjects (1.5 T, 3.0 T)	Age	Male/Female	ADAS-Cog
Baseline	All	1376 (808, 568)	Range: [54-91] 74.05(7.10)	759/617	Range: [0-55] 17.34(9.59)
	NC	430 (227, 203)	Range: [56-90] 74.39(5.77)	214/216	Range: [0-24] 9.38(4.27)
	MCI	662 (394, 268)	Range: [54-91] 73.43(7.60)	394/268	Range: [3-40] 17.19(6.67)
	AD	284 (187, 97)	Range: [55-91] 74.95(7.59)	151/133	Range: [13-55] 29.77(8.03)
12-Months	All	1160 (711, 449)	Range: [55-91] 74.24(7.04)	647/513	Range: [0-73] 18.34(11.97)
	NC	347 (206, 141)	Range: [56-90] 74.89(5.61)	177/170	Range: [0-23] 8.75(4.59)
	MCI	594 (351, 243)	Range: [55-91] 73.47(7.57)	352/242	Range: [0-73] 18.17(8.89)
	AD	219 (154, 65)	Range: [56-91] 75.30(7.39)	118/101	Range: [9-71] 33.99(11.02)
24-Months	All	1011 (617, 394)	Range: [55-91] 73.98(6.97)	559/452	Range: [0-71] 18.55(13.16)
	NC	367 (200, 167)	Range: [56-90] 74.41(5.83)	189/178	Range: [0-26] 9.06(4.86)
	MCI	501 (296, 205)	Range: [55-91] 73.30(7.46)	297/204	Range: [0-71] 19.85(10.46)
	AD	143 (121, 22)	Range: [56-89] 75.25(7.63)	73/70	Range: [12-68] 38.40(12.25)
36-Months	All	629 (437, 192)	Range: [55-90] 73.75(6.99)	353/276	Range: [0-74] 17.09(11.96)
	NC	208 (182, 26)	Range: [56-90] 75.30(5.46)	104/104	Range: [0-33] 9.32(5.24)
	MCI	411 (245, 166)	Range: [55-87] 72.86(7.52)	245/166	Range: [0-74] 20.49(12.09)
	AD	10 (10, 0)	Range: [66-85] 77.95(5.51)	4/6	Range: [22-73] 39.33(15.14)

this work are summarized in Table 2. Supplementary Figure 1 delineates the workflow of the applied methods for predicting the change of ADAS-Cog scores.

F. PENALIZED LINEAR REGRESSION

In the simplest cases, we assume that either 1) f_d are independent or 2) f_d are equal and can proceed with single-task learning. As a single-task learning algorithm, we use least-squares linear regression with an elastic net penalty to predict the ADAS-Cog changes. More formally, the ΔADAS scores are predicted by solving the following linear regression problems:

$$\Delta_t ADAS_i = a_d^T x_i + b_d + \epsilon_i, \qquad (2)$$

where i refers to a subject, a_d and b_d are the model parameters for the task d and ϵ_i is the error term. Adding the elastic net penalty, the model is solved by minimizing the following elastic net cost function:

$$\frac{1}{2N_d} \sum_{i=1}^{N_d} (\Delta_t ADAS_i - b_d - x_i^T a_d)^2$$
$$+ \lambda[(1-\alpha)||a_d||_2^2/2 + \alpha||a_d||_1], \qquad (3)$$

where N_d is the number of training samples, λ is the complexity parameter found by cross-validation, $\alpha \in [0, 1]$ defines the compromise between ridge $||a||_2^2/2$ and lasso penalties $||a||_1$, and $||.||_1$ denotes the L1-norm. Here, we selected $\alpha = 0.5$ to give equal weights for the ridge and lasso penalties.
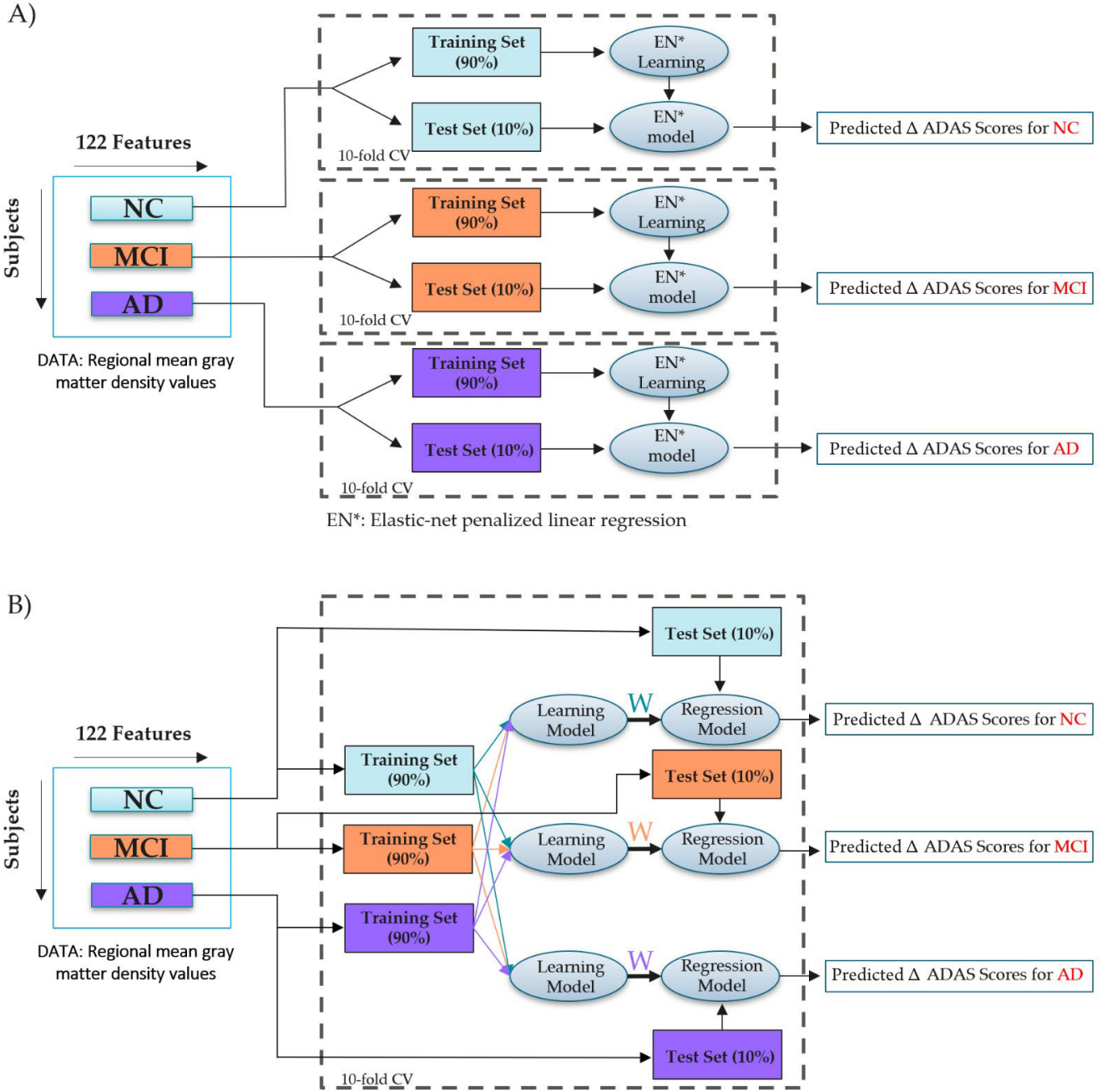
FIGURE 1. The framework of two different learning models to predict the future change of ADAS-Cog scores. A) Single-task learning, B) Multitask learning.

We modeled each time point ($t = 12, 24, 36$ months) separately.

G. MULTITASK LINEAR REGRESSION

Multitask regression incorporates the shared information among different tasks (task relatedness) into the regression model. To introduce the multitask regression approaches, we need to introduce additional notation. Let X_d be an $N_d \times M$ matrix of the input MRI data at the baseline corresponding to

task d, where N_d is the number of subjects in a group d, and M is the dimensionality of the feature space. We formulate the multitask learning as minimization of the penalized empirical loss:

$$\min_W \mathcal{L}(W) + E(W), \qquad (4)$$

where $W \in \mathbb{R}^{M \times d}$ is the weight matrix, which is estimated from the feature matrices X_d and Δ ADAS-Cog scores in the training set, $\mathcal{L}(W)$ is the empirical loss on the training set, and

TABLE 2. Summary of the compared algorithms. $\mathcal{L}$(W) and E(W) represent the empirical loss and the regularization term, respectively.

Purpose	Method	Algorithm
MFS harmonization	ComBat [38]	ComBat harmonization.
	ComBat$_{Age}$	ComBat harmonization; Age as a covariate.
	ComBat$_{Reg_{Age}}$	Eliminate age as a confound from harmonized image data.
	PLS [37]	PLS based domain adaptation; MFSs as the response variable.
	PLS_{Age}	PLS based domain adaptation; MFSs along with age as the response variable.
	Dirty Model$_{6-Tasks}$	6-task Dirty learning model
Learning models	SEP-EN	Elastic-net penalized linear regression; Training the model separately for each group.
	ALL-EN	Elastic-net penalized linear regression; Training the model using all groups
	Least Lasso [47]	Multitask learning; $\mathcal{L}$(W) = Least Squares; E(W) = $\rho_1||W||_1 + \rho_{L2}$.
	JFS [48]	Multitask learning; $\mathcal{L}$(W) = Least Squares; E(W) = $\lambda||W||_{1,2}$.
	Dirty Model [49]	Multitask learning; $\mathcal{L}$(W) = Least Squares; E(W) = $\rho_1||R||_{1,\infty} + \rho_2||S||_1$, W = R+S.
	LRA [50]	Multitask learning; $\mathcal{L}$(W) = Least Squares; E(W) = $\lambda||W||_*$.

$E(W)$ is the regularization term that encodes the relatedness of different tasks. Moreover, we use w_d to denote the weights related to the group d, i.e., $W = [w_1, \ldots, w_S]^T$. We then use the least squares loss as the empirical loss function for the regression tasks:

$$\mathcal{L}(W_t) = \sum_{d \in \{AD,MCI,NC\}} ||w_{d,t}^T X_d + b_{d,t}1 - y_{d,t}||^2, \quad (5)$$

where $y_{d,t}$ is a shorthand of the $\Delta_t ADAS$ scores of N_d subjects in group d at the time point t (12, 24, or 36 months), $b_{d,t}$ are the bias terms, and 1 is an M-element vector of ones. Different regularization terms were used to determine different assumptions on task relatedness [51]–[54]. The regularization terms are described as in the following, where we drop the subindex t as we separately consider the prediction at three time points.

1) MULTITASK LASSO
Multitask Lasso is a simple generalization of the elastic net penalty [47] to multitask regression. The regularization term is defined as:

$$E(W) = \rho_1||W||_1 + \rho_{L2}||W||_F^2, \quad (6)$$

where ρ_1 is a regularization parameter for controlling the sparsity among all tasks and ρ_{L2} is an optional regularization parameter that controls the ℓ2-norm penalty.

2) JOINT FEATURE SELECTION
Joint feature selection is used to constrain all models to share a common set of features [48], [51]. This goal is achieved by setting $E(W)$ and minimizing the following ℓ2,1-norm regularized learning [55], [56]:

$$E(W) = \rho_1||W||_{2,1} + \rho_{L2}||W||_F^2, \quad (7)$$

where $||W||_{1,2} = \sum_{D=1}^3 ||W_D||_2$, with $W_D \in \mathbb{R}^{1 \times D}$, is the group sparse penalty, ρ_1 is a regularization parameter for controlling the sparsity among all tasks and ρ_{L2} is an optional regularization parameter that controls the ℓ2-norm penalty.

3) DIRTY MODEL FOR MULTITASK LEARNING
The Dirty model estimates a superposition of two sets of parameters and regularizes them differently [49]. In more detail, each w_d is written as a sum $w_d = s_d + r_d$, where the corresponding matrices S and R are encouraged to have elementwise sparsity and block-structured row sparsity. This leads to the minimization problem:

$$\min_{S,R} \sum_d ||(s_d + r_d)^T X_d + b_d 1 - y_{d,t}||^2$$
$$+ \rho_1||R||_{1,\infty} + \rho_2||S||_1, \quad (8)$$

where $||R||_{1,\infty} = \sum_j \max_d |r_{d_j}|$. The final output is $W = \hat{S} + \hat{R}$, where $\hat{S}, \hat{R}$ are the minimizers of (8).

4) LOW RANK ASSUMPTION

The task commonalities can be utilized to constrain the prediction models from different tasks to share a low-dimensional subspace, i.e., constrain W to be of low rank [50]. Since rank optimization is, in general, NP-hard, the rank function [57] is replaced by the trace norm [52], [53], which is given by the sum of the singular values of W: $||W||_* = \sum_d \sigma_d(W)$, and the regularization term is

$$E(W) = \rho_1 ||W||_*, \qquad (9)$$

where the regularization parameter ρ_1 controls the rank of W.

H. MAGNETIC FIELD STRENGTH HARMONIZATION

To correct for differences in features caused by imaging at two MFSs in ADNI1 and ADNI2, we studied three different techniques: 1) PLS-based domain adaptation introduced in [37], 2) ComBat harmonization originating from genetics [38], which has become widely used in brain imaging [46], [58], and 3) multitask learning.

1) PARTIAL LEAST SQUARES DOMAIN ADAPTATION

We next briefly explain the PLS-based domain adaptation method introduced in [37] for correcting the site dependency of cortical thickness measurements in predicting the severity of autism spectrum disorder. PLS is a linear feature transformation method for modeling relations between two sets of observed variables. The idea of PLS-based domain adaptation is to project the input features into a new (lower-dimensional) space, which is a product of two orthogonal subspaces: a subspace that is dependent on MFSs and its (orthogonal) complement. This can be achieved by applying a PLS algorithm that ensures the orthogonality of the resulting latent features so that the response variable is a codification of the scanner MFS and the predictor variables are the original features. However, as demonstrated in [37], this works the best, and different brain regions are separately corrected. Therefore, [37] suggested a two-stage strategy where in the first stage a PLS-based domain adaptation is performed for each brain region separately, then for each brain region the prediction task is performed. These predictions are then combined in the stacking framework. We applied an AAL-atlas to decompose the gray matter density values into 122 distinct regions. For the prediction $\Delta ADAS$ for each brain region, we utilized support vector regression (with a radial basis function kernel), as suggested in [37]. Finally, the per-region predictions were combined using the elastic-net penalized linear regression described in Section II-F.

2) ComBat HARMONIZATION

We applied ComBat harmonization to reduce any potential biases induced by different MFSs. This method was initially proposed to correct for the site effects in genomics [38]. Later, ComBat was applied to correct for site effects in imaging applications, including diffusion tensor imaging data [46], cortical thickness [58], positron emission tomography [59]

and functional connectivity [60]. In this study, ComBat utilizes a multivariate linear mixed-effects regression to model MFS-adapted feature measurements. Let x_{ij} denote a regional gray matter density value for subject j with MRI acquired at MFS $i \in \{1, 2\}$ (where 1 refers to $1.5T$ and 2 refers to $3T$). Then, x_{ij} can be written as:

$$x_{ij} = \alpha + C_{ij}\beta + \gamma_i + \delta_i \epsilon_{ij}, \qquad (10)$$

where α is the overall GM density, C_{ij} is a design matrix for the covariates of interest (age), and β is the regression coefficient corresponding to the covariate C. The terms γ_i represent the location parameter effect of MFS i, δ_i describes the multiplicative effect of MFS i, and ϵ_{ij} is an error term from a normal distribution with a zero mean [46]. The ComBat-harmonized GM densities (MFSs-adjusted) are then defined as:

$$x_{ij}^{ComBat} = \frac{x_{ij} - \widehat{\alpha} - C_{ij}\widehat{\beta} - \gamma_i^*}{\delta_i^*} + \widehat{\alpha} + C_{ij}\widehat{\beta}, \qquad (11)$$

in which $\widehat{\alpha}$ and $\widehat{\beta}$ are estimators of parameters α and β, and γ_i^* and δ_i^* are the empirical Bayes estimators of γ_i and δ_i. To correct the difference among MFSs, we considered two strategies: 1) using ComBat without adjusting any biological covariates (i.e., setting $C = 0$) and 2) using ComBat while adjusting the age as a biological covariate.

I. IMPLEMENTATION AND EVALUATION

To tune the parameters and evaluate the performance of the models, we utilized repeated, nested 10-fold cross-validation (CV). The number of repeats was 10. In nested CV, we have a double loop where an inner CV loop is performed inside the outer CV loop [61]. The whole data set is divided into 10-folds (outer loop), and one at a time, it is used to test the model trained with the remaining folds. To select the optimal parameters for each of the outer ten folds, a 10-fold CV is applied to each of the 10 training sets. For each outer fold, these are then used to train the model with the (outer loop) training sets. We used two evaluation metrics: (1) R is the correlation coefficient between the predicted and observed Δ ADAS-Cog scores, averaged over ten repeats of the CV; and (2) MAE is the mean absolute error between the observed and predicted Δ ADAS-Cog values, averaged over the subjects and ten CV repeats. We then used ten repeats of the 10-fold CV and averaged the metrics to reduce the random variation due to the sampling of subjects to different folds. We computed 95% confidence intervals for cross-validated, averaged correlations R and MAEs using a bootstrap method [62], [63]. Confidence intervals represent an approximation of the overall performance of the prediction model, and the specific bootstrap method used is adapted to be used in repeated CV (see [63] for details).

The values of the hyperparameters were selected in the inner CV loop, and predictions of the Δ ADAS-Cog scores were evaluated in the outer CV loop to avoid the problem of training on the testing data. The distribution of the

ADAS-Cog score changes in each of the CV folds was similar since we used stratified cross-validation folds[4] [64].

The implementation of the elastic-net penalized linear regression model was performed by using the glmnet library[5] [65]. The optimal value of the regularization parameter (λ) was selected in the 10-fold inner CV loop by minimizing the mean squared error (MSE). The implementation of multitask learning techniques was performed using the MALSAR package running in MATLAB [54]. The functions implemented in MALSAR have many parameters to tune. Since fine-tuning all parameters with a grid search was impractical, we only considered ρ_1 (the regularization parameter for controlling the sparsity among all tasks) and ρ_2 (an optional regularization parameter that controls the $\ell2$-norm penalty) as the most crucial parameters for the grid search. The ρ_1 parameter was selected among the candidate set $\{10^{-3}, 10^{-2.5}, \ldots, 10^2, 2 \cdot 10^2, 2.5 \cdot 10^2, \ldots, 5 \cdot 10^2)\}$, where the parameter ρ_2 was chosen among the candidate set $\{10^{-3}, 10^{-2.5}, \ldots, 10^2, 2 \cdot 10^2, 2.5 \cdot 10^2, \ldots, 10 \cdot 10^2)\}$ by minimizing the root-mean-square error (rmse). For the tuning parameters, default values were used for the optional optimization parameters (starting points, termination criterion, endurance, and a maximum number of repetitions).

The implementation of ComBat was performed using a publicly available MATLAB package.[6] PLS was performed by the PLSREGRESS function in MATLAB with a constant number of components for all groups of diagnoses at each time point. The number of components for PLS was selected at each time point by initial experiments between the candidate set {5, 10, 15, 17, 20, 25}.

The PLS-based domain adaption was performed as instructed in [37]. The implementation of SVR, required by the PLS domain adaptation, was performed using LIBSVM[7] [66]. The SVR model parameters were set to their default values ($C = 1$, $\nu = 0.5$, $\lambda = 1/K$, where K refers to the dimensionality of the feature space). The implementation of elastic-net penalized linear regression, required by the PLS domain adaptation, was performed as described above.

III. RESULTS

A. PREDICTION PERFORMANCE OF SINGLE- AND MULTITASK LEARNING

We evaluated the performance of single- and multitask learning approaches by predicting the future change in ADAS-Cog scores using baseline MRI features. The experimental results presented in this subsection ignore the variation in the magnetic field strengths to acquire the MRIs.

Table 3 indicates the comparison results between different multitask learning methods based on the least-squares loss

function, including multitask Lasso (least lasso), joint feature selection (JFS), dirty model (least dirty), and trace-norm regularization (least trace), with two single-task learning strategies (SEP-EN and ALL-EN) based on elastic-net penalized linear regression (EN). Note that the low number of subjects in the AD group at 36 months (only ten subjects) cannot provide a reliable validation of predictive models; however, the results for this group and time point are shown in Table 3. As shown in Table 3, the average correlation coefficients between the predicted and actual Δ ADAS-Cog scores were positive for all baseline diagnoses and time points. This demonstrates that MRI-based predictive models were able to predict the disease progression. A comparison of two single-task learning strategies, SEP-EN and ALL-EN, indicated that ALL-EN, which utilized all diagnostic groups for training, performed better in the NC and MCI groups. For instance, R of the NC subjects increased from 0.09 to 0.12 at 12 months, from 0.09 to 0.17 at 24 months, and from 0.04 to 0.19 at 36 months. Moreover, the ALL-EN prediction model achieved the best performance among all methods in terms of a correlation score for NC and MCI groups (e.g., R for MCI group at 12, 24, and 36 months were 0.22, 0.41, and 0.39). However, R of the other methods were typically within the 95% confidence intervals of R of ALL-EN, indicating that the improvement was not large. All multitask learning algorithms performed highly similarly to ALL-EN. However, especially in the predictions concerning the AD group, these performed slightly better than ALL-EN: For instance, R of the AD subjects increased from 0.22 to 0.27 at 12 months and from 0.26 to 0.27 at 24 months when comparing the Dirty Model and ALL-EN. SEP-EN was clearly the method of choice for predicting 24-month changes in the ADAS-Cog in AD patients, with a R value of 0.40 compared to the maximum of 0.31 of others. Supplementary Figures 2-3 provide scatter plots of the observed vs. predicted change in ADAS scores in the CV run with the median correlation (R). These scatter plots imply that the Δ ADAS-Cog scores with very high values for all time points were the most difficult to predict, since the number of individuals with observed ADAS-Cog score changes over 20 was small (e.g., the numbers of MCI subjects at 12, 24, and 36 months were 2, 8, and 19, respectively). In summary, these results support the notion that auxiliary data were useful in predicting the ADAS-Cog change in the NC and MCI groups, but in the AD group. Complex multitask learning algorithms did not demonstrate benefits over simpler single-task learning methods.

Interestingly, in the majority of methods, R scores typically increased with the length of the follow-up (e.g., R in Δ ADAS-24 was higher than R of Δ ADAS-12). The potential reason for this higher correlation is that as the changes in ADAS-Cog become more prominent, they are easier to predict based on MRI.

Table 3 the ALL column lists the evaluation results while agglomerating all subject groups in the validation. For all methods, the R values were higher when all the subjects were combined than stratified based on the baseline

---

[4]The stratified cross-validation code is available at https://github.com/jussitohka/general_matlab.

[5]The glmnet library is available at https://web.stanford.edu/~hastie/glmnet_matlab.

[6]The ComBat harmonization package is available at https://github.com/Jfortin1/ComBatHarmonization.

[7]https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

TABLE 3. Comparison of single and multitask learning on predicting the change in ADAS-Cog. The methods are given in Table 2. SEP-EN refers to the single-task method trained for each diagnostic group separately. ALL-EN refers to the single-task method trained with the data from all diagnostic groups. R is the cross-validated correlation between the actual and predicted Δ ADAS-Cog scores averaged over 10 CV runs and *MAE* is the mean absolute error averaged over 10 CV runs. Values in parentheses give the bootstrapped 95% confidence intervals. The asterisk (*) implies that the validation result is not trustworthy due to the low number of samples.

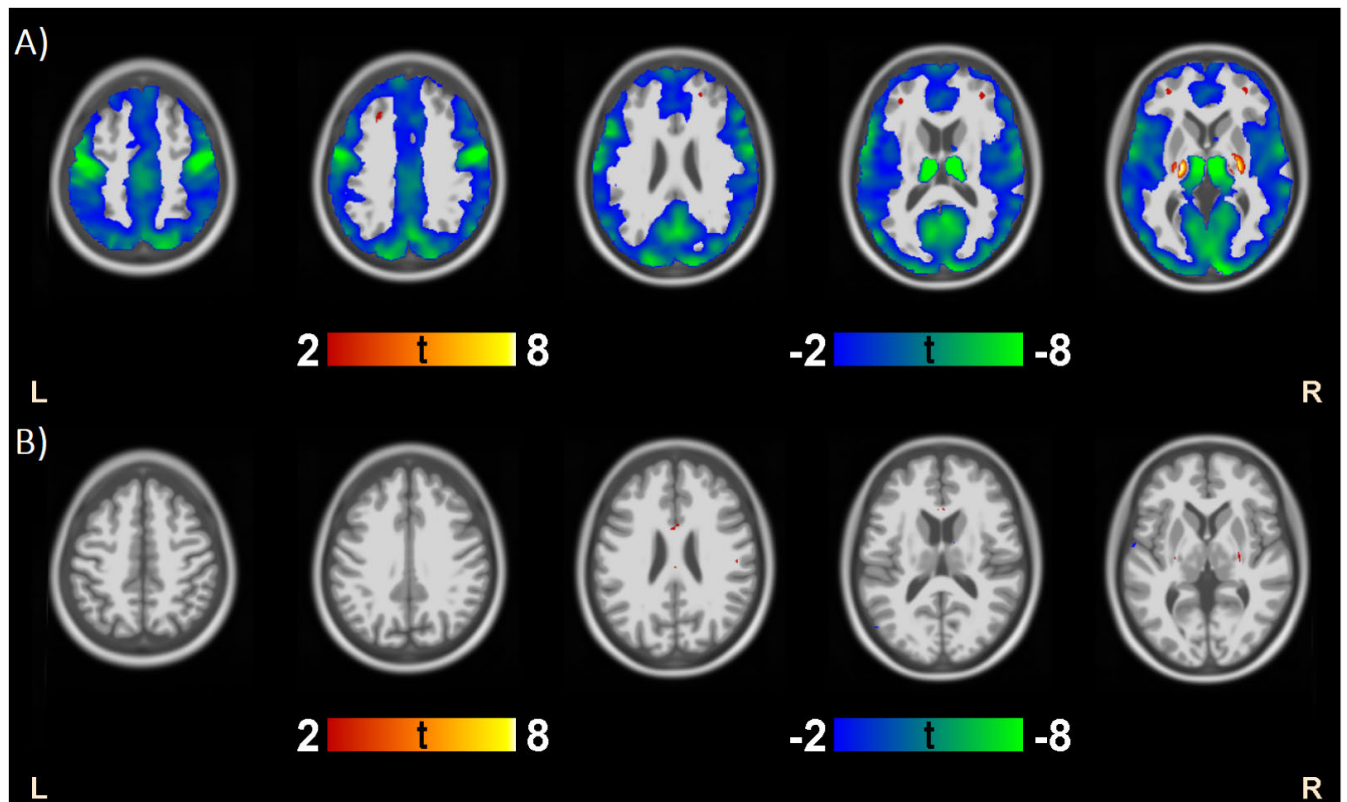|	NC		MCI		AD		ALL	
	R	MAE	R	MAE	R	MAE	R	MAE
SEP-EN								
Δ ADAS-12	0.09	3.07	0.19	3.84	0.24	4.86	0.38	3.81
	(0.00 to 0.19)	(2.79 to 3.32)	(0.11 to 0.27)	(3.58 to 4.17)	(0.12 to 0.35)	(4.33 to 5.49)	(0.32 to 0.43)	(3.60 to 4.01)
Δ ADAS-24	0.09	3.16	0.36	4.74	0.40	6.46	0.55	4.43
	(0.00 to 0.19)	(2.93 to 3.41)	(0.29 to 0.44)	(4.37 to 5.13)	(0.23 to 0.52)	(5.74 to 7.37)	(0.49 to 0.60)	(4.18 to 4.65)
Δ ADAS-36	0.04	3.25	0.37	5.95	-0.28^*	7.06^*	0.43	5.07
	(-0.02 to 0.11)	(2.90 to 3.69)	(0.29 to 0.45)	(5.43 to 6.50)	(-0.67 to 0.33)	(3.81 to 10.35)	(0.37 to 0.50)	(4.73 to 5.52)
ALL-EN								
Δ ADAS-12	0.12	3.23	0.22	3.84	0.22	4.74	0.32	3.83
	(0.03 to 0.23)	(2.96 to 3.49)	(0.15 to 0.30)	(3.56 to 4.12)	(0.06 to 0.34)	(4.10 to 5.39)	(0.27 to 0.38)	(3.63 to 4.04)
Δ ADAS-24	0.17	3.59	0.41	4.65	0.26	6.55	0.48	4.54
	(0.07 to 0.28)	(3.36 to 3.89)	(0.34 to 0.48)	(4.28 to 5.05)	(0.11 to 0.40)	(5.46 to 7.66)	(0.44 to 0.53)	(4.29 to4.82)
Δ ADAS-36	0.19	3.86	0.39	5.83	0.71^*	5.07^*	0.42	5.17
	(0.09 to 0.30)	(3.47 to 4.27)	(0.32 to 0.46)	(5.28 to 6.39)	(-0.07 to 0.90)	(2.19 to 8.58)	(0.35 to 0.48)	(4.81 to 5.56)
Least Lasso								
Δ ADAS-12	0.10	3.07	0.17	3.87	0.26	5.01	0.33	3.85
	(0.01 to 0.21)	(2.81 to 3.33)	(0.10 to 0.26)	(3.56 to 4.21)	(0.15 to 0.37)	(4.35 to 5.60)	(0.26 to 0.40)	(3.63 to 4.07)
Δ ADAS-24	0.09	3.15	0.37	5.10	0.29	7.36	0.47	4.70
	(0.02 to 0.22)	(2.92 to 3.42)	(0.29 to 0.45)	(4.69 to 5.46)	(0.11 to 0.43)	(6.30 to 8.59)	(0.40 to 0.52)	(4.41 to 4.97)
Δ ADAS-36	0.15	3.24	0.37	6.46	-0.17^*	8.16^*	0.38	5.34
	(0.07 to 0.30)	(2.88 to 3.65)	(0.30 to 0.46)	(5.72 to 6.95)	(-0.53 to 0.14)	(3.85 to 13.47)	(0.31 to 0.45)	(4.92 to 5.80)
JFS								
Δ ADAS-12	0.11	3.07	0.20	3.86	0.25	4.99	0.34	3.84
	(0.01 to 0.21)	(2.81 to 3.34)	(0.13 to 0.29)	(3.56 to 4.14)	(0.13 to 0.37)	(4.31 to 5.59)	(0.28 to 0.40)	(3.62 to 4.05)
Δ ADAS-24	0.10	3.17	0.37	5.03	0.30	7.24	0.48	4.67
	(0.02 to 0.22)	(2.93 to 3.39)	(0.29 to 0.45)	(4.63 to 5.47)	(0.14 to 0.42)	(6.23 to 8.35)	(0.42 to 0.53)	(4.40 to 4.96)
Δ ADAS-36	0.09	3.25	0.38	6.33	0.38^*	6.77^*	0.40	5.33
	(0.01 to 0.21)	(2.98 to 3.68)	(0.30 to 0.46)	(5.76 to 6.91)	(-0.13 to 0.79)	(3.62 to 10.35)	(0.32 to 0.46)	(4.91 to 5.80)
Dirty Model								
Δ ADAS-12	0.09	3.09	0.21	3.84	0.27	4.96	0.34	3.82
	(0.01 to 0.18)	(2.81 to 3.35)	(0.13 to 0.28)	(3.55 to 4.15)	(0.15 to 0.38)	(4.28 to 5.61)	(0.28 to 0.40)	(3.62 to 4.06)
Δ ADAS-24	0.08	3.18	0.37	5.06	0.27	7.18	0.47	4.70
	(-0.01 to 0.16)	(2.96 to 3.46)	(0.29 to 0.45)	(4.63 to 5.48)	(0.12 to 0.41)	(6.20 to 8.53)	(0.41 to 0.53)	(4.43 to 4.99)
Δ ADAS-36	0.18	3.27	0.39	6.29	0.38^*	5.87^*	0.41	5.27
	(0.05 to 0.30)	(2.92 to 3.70)	(0.31 to 0.45)	(5.71 to 6.94)	(-0.30 to 0.76)	(3.32 to 8.55)	(0.33 to 0.48)	(4.90 to 5.76)
LRA								
Δ ADAS-12	0.09	3.08	0.21	3.84	0.25	4.96	0.33	3.82
	(-0.01 to 0.20)	(2.79 to 3.36)	(0.13 to 0.29)	(3.54 to 4.16)	(0.12 to 0.37)	(4.34 to 5.66)	(0.27 to 0.40)	(3.61 to 4.04)
Δ ADAS-24	0.13	3.16	0.39	5.00	0.31	7.11	0.49	4.65
	(0.03 to 0.23)	(2.94 to 3.41)	(0.31 to 0.46)	(4.64 to 5.42)	(0.16 to 0.45)	(6.11 to 8.32)	(0.43 to 0.55)	(4.38 to 4.94)
Δ ADAS-36	0.01	3.48	0.34	6.62	0.28^*	6.66^*	0.35	5.56
	(-0.08 to 0.10)	(3.13 to 3.94)	(0.27 to 0.43)	(6.00 to 7.19)	(-0.44 to 0.82)	(3.82 to 9.12)	(0.28 to 0.43)	(5.16 to 6.03)

FIGURE 2. t-scores of voxel-wise differences in the gray matter density in cognitively normal subjects due to differences in magnetic field strength. The t-score values have a threshold at $|t| > 2$, corresponding to uncorrected $p < 0.05$. Panel (A) before using the ComBat harmonization and (B) after using the ComBat harmonization approach. The differences in the gray matter density were strong in panel (A), but ComBat harmonization was successful in removing the contribution of MFS at the group level.

diagnosis. Since the prediction models were the same in both cases, this inflation in the prediction performance can be seen as artificial and one to avoid. It is likely a product of the interaction between the heterogeneity of subject groups and the particular evaluation measure (correlation) that scales according to this heterogeneity.

B. PREDICTION PERFORMANCE WHILE ACCOUNTING FOR DIFFERENCES IN MFS

This subsection focuses on the situation associated with heterogeneity reduction between data from two MFSs and explores the prediction models by including MFS correction approaches to reduce unwanted variance across features. In addition, based on the results of the previous subsection, we selected ALL-EN as a single-task learning method and a Dirty model as a multitask learning method for further analysis.

To demonstrate the differences between GM density values of images acquired at 1.5 T and 3.0 T, we applied a standard voxel-based morphometry approach to compare GM densities of NC subjects acquired at 1.5 T and 3.0 T. Voxelwise t-statistics in Fig. 2(A) demonstrate considerable differences in the GM density values between 1.5 T and 3.0 T. We repeated the analysis after using the ComBat harmonization method. Fig. 2(B) delineates that, at the group level,

the ComBat harmonization performed exceptionally well in removing nuisance variability associated with two different MFSs.

We adopted two strategies to harmonize the MRI data for MFS differences and studied whether harmonization can improve the performance of ADAS-Cog prediction. First, we performed the ComBat approach on 122 regional GM density measurements and then used ALL-EN to predict Δ ADAS-Cog scores. Second, we applied the PLS-based domain adaptation method, as described in Section 2.8.1. Table 4 presents the R and MAE scores. Combined with ALL-EN, PLS-based domain adaptation methods performed slightly better than the ComBat method in terms of the average correlation. For example, in the PLS approach, R for the NC, MCI, and AD groups at 24 months were 0.16, 0.40, and 0.28, respectively. In the ComBat approach, R for NC, MCI, and AD groups at 24 months were 0.11, 0.38, and 0.24, respectively.

Table 4 delineates that the performance of the Dirty model was similar to or worse than the performance of ALL-EN after the correction for the MFS differences. For example, the prediction performance of the PLS method for the AD group at 12 months, when the Dirty model substituted the ALL-EN, the average R score dropped from 0.30 to 0.24. Table 4 also shows the results of 6-task

TABLE 4. Comparison of the predictive model performance with and without MFS harmonization. ALL-EN refers to the baseline method without any harmonization. $ComBat_{ALL-EN}$ (PLS_{ALL-EN}) refers to Combat (PLS) harmonization followed by ALL-EN. $ComBat_{DirtyModel}$ ($PLS_{DirtyModel}$) refers to Combat (PLS) harmonization followed by Dirty Model. $DirtyModel_{6-Tasks}$ ($LRA_{6-Tasks}$) refers to the 6-task learning with Dirty Model (LRA). The asterisk (*) implies that the validation result is not trustworthy due to the low number of samples.

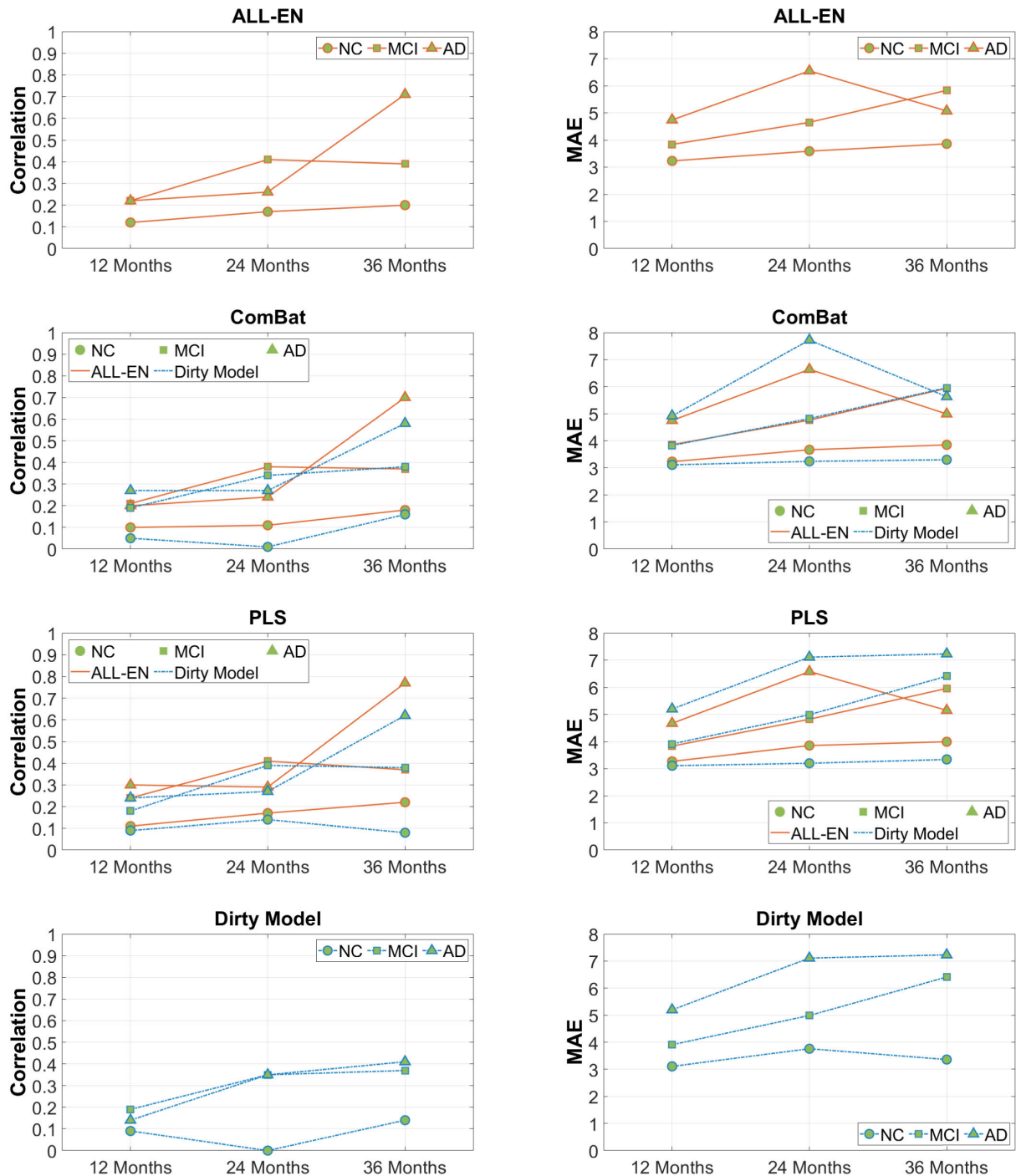|	NC		MCI		AD		ALL	
	R	MAE	R	MAE	R	MAE	R	MAE
ALL-EN								
Δ ADAS-12	0.12	3.23	0.22	3.84	0.22	4.74	0.32	3.83
	(0.03 to 0.23)	(2.96 to 3.49)	(0.15 to 0.30)	(3.56 to 4.12)	(0.06 to 0.34)	(4.10 to 5.39)	(0.27 to 0.38)	(3.63 to 4.04)
Δ ADAS-24	0.17	3.59	0.41	4.65	0.26	6.55	0.48	4.54
	(0.07 to 0.28)	(3.36 to 3.89)	(0.34 to 0.48)	(4.28 to 5.05)	(0.11 to 0.40)	(5.46 to 7.66)	(0.44 to 0.53)	(4.29 to 4.82)
Δ ADAS-36	0.19	3.86	0.39	5.83	0.71*	5.07*	0.42	5.17
	(0.09 to 0.30)	(3.47 to 4.27)	(0.32 to 0.46)	(5.28 to 6.39)	(-0.07 to 0.90)	(2.19 to 8.58)	(0.35 to 0.48)	(4.81 to 5.56)
$ComBat_{ALL-EN}$								
Δ ADAS-12	0.10	3.23	0.21	3.86	0.20	4.75	0.31	3.84
	(-0.01 to 0.19)	(2.98 to 3.51)	(0.13 to 0.29)	(3.57 to 4.17)	(0.05 to 0.34)	(4.06 to 5.40)	(0.26 to 0.36)	(3.63 to 4.06)
Δ ADAS-24	0.11	3.67	0.38	4.76	0.27	6.64	0.45	4.64
	(0.02 to 0.21)	(3.41 to 3.95)	(0.29 to 0.45)	(4.40 to 5.16)	(0.12 to 0.42)	(5.69 to 7.78)	(0.40 to 0.49)	(4.39 to 4.91)
Δ ADAS-36	0.18	3.85	0.37	5.94	0.70*	4.99*	0.41	5.22
	(0.06 to 0.30)	(3.42 to 4.23)	(0.29 to 0.44)	(5.40 to 6.49)	(0.03 to 0.91)	(2.60 to 7.87)	(0.34 to 0.47)	(4.85 to 5.61)
PLS_{ALL-EN}								
Δ ADAS-12	0.11	3.27	0.23	3.83	0.30	4.67	0.34	3.83
	(0.01 to 0.21)	(3.00 to 3.56)	(0.16 to 0.32)	(3.56 to 4.12)	(0.16 to 0.43)	(4.05 to 5.27)	(0.27 to 0.39)	(3.62 to 4.05)
Δ ADAS-24	0.16	3.85	0.40	4.82	0.28	6.58	0.48	4.70
	(0.06 to 0.25)	(3.53 to 4.09)	(0.34 to 0.47)	(4.48 to 5.21)	(0.13 to 0.43)	(5.55 to 7.59)	(0.43 to 0.52)	(4.45 to 4.95)
Δ ADAS-36	0.22	3.99	0.36	5.96	0.77*	5.15*	0.40	5.26
	(0.10 to 0.34)	(3.56 to 4.35)	(0.30 to 0.44)	(5.40 to 6.49)	(0.34 to 0.91)	(2.47 to 7.97)	(0.34 to 0.46)	(4.92 to 5.65)
$ComBat_{DirtyModel}$								
Δ ADAS-12	0.05	3.11	0.19	3.86	0.27	4.92	0.33	3.84
	(-0.04 to 0.14)	(2.83 to 3.36)	(0.11 to 0.27)	(3.59 to 4.19)	(0.13 to 0.38)	(4.30 to 5.58)	(0.27 to 0.38)	(3.63 to 4.07)
Δ ADAS-24	0.01	3.24	0.34	5.12	0.27	7.72	0.42	4.81
	(-0.08 to 0.07)	(3.02 to 3.52)	(0.26 to 0.42)	(4.71 to 5.52)	(0.12 to 0.40)	(6.56 to 8.91)	(0.36 to 0.48)	(4.52 to 5.13)
Δ ADAS-36	0.16	3.30	0.38	6.33	0.58*	5.63*	0.40	5.31
	(0.02 to 0.26)	(2.93 to 3.71)	(0.30 to 0.45)	(5.72 to 6.94)	(0.01 to 0.89)	(3.45 to 7.25)	(0.33 to 0.47)	(4.88 to 5.74)
$PLS_{DirtyModel}$								
Δ ADAS-12	0.09	3.11	0.18	3.91	0.24	5.20	0.30	3.93
	(-0.01 to 0.20)	(2.86 to 3.40)	(0.12 to 0.26)	(3.64 to 4.26)	(0.13 to 0.35)	(4.60 to 5.94)	(0.24 to 0.36)	(3.70 to 4.15)
Δ ADAS-24	0.15	3.20	0.39	4.99	0.27	7.11	0.48	4.62
	(0.02 to 0.23)	(2.95 to 3.44)	(0.31 to 0.47)	(4.59 to 5.45)	(0.10 to 0.43)	(5.97 to 8.30)	(0.43 to 0.54)	(4.35 to 4.93)
Δ ADAS-36	0.08	3.34	0.38	6.41	0.62*	7.23*	0.39	5.40
	(-0.04 to 0.18)	(2.97 to 3.73)	(0.31 to 0.46)	(5.79 to 7.05)	(-0.42 to 0.90)	(3.97 to 10.52)	(0.32 to 0.46)	(4.98 to 5.88)
$DirtyModel_{6-Tasks}$								
Δ ADAS-12	0.09	3.11	0.19	3.91	0.14	4.98	0.34	3.88
	(0.00 to 0.18)	(2.87 to 3.40)	(0.12 to 0.25)	(3.66 to 4.23)	(0.02 to 0.26)	(4.40 to 5.62)	(0.29 to 0.39)	(3.68 to 4.09)
Δ ADAS-24	−0.005	3.76	0.35	4.93	0.35	6.59	0.50	4.74
	(-0.09 to 0.08)	(3.50 to 4.06)	(0.27 to 0.43)	(4.56 to 5.28)	(0.20 to 0.48)	(5.73 to 7.45)	(0.44 to 0.55)	(4.49 to 5.02)
Δ ADAS-36	0.14	3.36	0.37	6.00	0.41*	5.34*	0.44	5.12
	(0.00 to 0.25)	(2.98 to 3.78)	(0.29 to 0.45)	(5.51 to 6.55)	(-0.08 to 0.78)	(2.43 to 8.72)	(0.38 to 0.50)	(4.76 to 5.52)
$LRA_{6-Tasks}$								
Δ ADAS-12	0.02	3.15	0.18	3.92	0.16	5.02	0.33	3.90
	(-0.06 to 0.11)	(2.88 to 3.43)	(0.11 to 0.25)	(3.66 to 4.23)	(0.04 to 0.28)	(4.44 to 5.64)	(0.27 to 0.38)	(3.70 to 4.12)
Δ ADAS-24	0.02	3.60	0.35	4.82	0.25	6.77	0.49	4.65
	(-0.06 to 0.10)	(3.32 to 3.87)	(0.27 to 0.42)	(4.46 to 5.23)	(0.10 to 0.41)	(5.83 to 7.69)	(0.44 to 0.54)	(4.40 to 4.94)
Δ ADAS-36	0.03	3.43	0.40	5.98	−0.40*	6.90*	0.44	5.15
	(-0.07 to 0.14)	(3.02 to 3.83)	(0.31 to 0.48)	(5.50 to 6.53)	(-0.74 to -0.03)	(3.57 to 10.78)	(0.37 to 0.52)	(4.77 to 5.54)

FIGURE 3. Comparison of single and multitask learning with MFS correction (ComBat and PLS), without MFS correction (ALL-EN), and 6-task Dirty model. Combat and PLS panels give the results with both the ALL-EN and 3-task Dirty model.

learning approaches for MFS adaptation. These methods were performed on par with other correction approaches, but failed to consistently improve the prediction of the baseline model (ALL-EN). Fig. 3 illustrates the performance comparison between single and multitask learning strategies before and after utilizing correction approaches. The performance comparison shows that ComBat did not improve the

prediction performance at the individual level, although it worked well at the group level, as demonstrated in Fig. 2. Fig. 3 indicates that the PLS domain adaptation based on the single-task learning model performed consistently better than the other methods. Moreover, combining multitask learning with ComBat slightly improved the performance in the AD group.

TABLE 5. Comparison of ComBat, PLS with Age as a covariate. In *ComBat*$_{Age}$ Age was used as a covariate to preserve its effect while removing the variability associated with MFS. In *ComBat*$_{Reg_{Age}}$, age was regressed out from MRI after the MFS correction with *ComBat*$_{Age}$. In *PLS*$_{age}$, Age and MFS were used as response variables. The asterisk (*) implies that the validation result is not trustworthy due to the low number of samples.

	NC		MCI		AD		ALL	
	R	MAE	R	MAE	R	MAE	R	MAE
$ComBat_{Age}$								
Δ ADAS-12	0.10	3.23	0.20	3.89	0.19	4.75	0.31	3.87
	(0.01 to 0.21)	(2.97 to 3.50)	(0.13 to 0.28)	(3.60 to 4.18)	(0.05 to 0.34)	(4.08 to 5.39)	(0.26 to 0.37)	(3.66 to 4.07)
Δ ADAS-24	0.08	3.79	0.36	4.83	0.24	6.77	0.44	4.73
	(-0.01 to 0.17)	(3.54 to 4.08)	(0.29 to 0.44)	(4.48 to 5.20)	(0.09 to 0.38)	(5.66 to 7.90)	(0.39 to 0.48)	(4.47 to 5.01)
Δ ADAS-36	0.22	3.81	0.36	6.00	0.65*	5.43*	0.41	5.25
	(0.10 to 0.33)	(3.44 to 4.24)	(0.29 to 0.44)	(5.52 to 6.51)	(-0.12 to 0.90)	(2.45 to 8.75)	(0.35 to 0.47)	(4.89 to 5.66)
$ComBat_{Reg_{Age}}$								
Δ ADAS-12	0.11	3.18	0.19	3.91	0.24	4.75	0.31	3.85
	(0.02 to 0.21)	(2.94 to 3.47)	(0.11 to 0.27)	(3.61 to 4.20)	(0.07 to 0.39)	(4.12 to 5.42)	(0.26 to 0.37)	(3.63 to 4.07)
Δ ADAS-24	0.11	3.73	0.39	4.95	0.37	6.54	0.47	4.75
	(0.02 to 0.20)	(3.51 to 4.02)	(0.30 to 0.45)	(4.63 to 5.34)	(0.21 to 0.52)	(5.50 to 7.55)	(0.42 to 0.52)	(4.49 to 5.01)
Δ ADAS-36	0.16	3.80	0.38	6.07	0.34*	5.57*	0.40	5.32
	(0.08 to 0.33)	(3.47 to 4.27)	(0.31 to 0.46)	(5.40 to 6.42)	(0.15 to 0.83)	(2.26 to 8.56)	(0.35 to 0.47)	(4.81 to 5.63)
PLS_{Age}								
Δ ADAS-12	0.12	3.27	0.25	3.79	0.31	4.66	0.35	3.80
	(0.03 to 0.22)	(3.02 to 3.54)	(0.18 to 0.34)	(3.52 to 4.11)	(0.20 to 0.42)	(4.08 to 5.23)	(0.29 to 0.41)	(3.60 to 4.03)
Δ ADAS-24	0.16	3.80	0.44	4.78	0.34	6.41	0.51	4.84
	(0.06 to 0.26)	(3.48 to 4.06)	(0.38 to 0.50)	(4.44 to 5.10)	(0.19 to 0.49)	(5.41 to 7.38)	(0.46 to 0.54)	(4.39 to 4.88)
Δ ADAS-36	0.22	3.92	0.38	5.92	0.75*	5.16*	0.41	5.27
	(0.08 to 0.33)	(3.60 to 4.36)	(0.31 to 0.45)	(5.38 to 6.48)	(0.17 to 0.92)	(2.88 to 8.39)	(0.35 to 0.47)	(4.91 to 5.63)

C. AGE AS A COVARIATE

In the MRI-based predictive modeling of AD, age plays an essential role; for example, regressing age out of MRI has been shown to improve MCI-to-AD conversion prediction [67]. Therefore, we studied whether removing or preserving age as a biological variable can improve the ADAS-Cog prediction, focusing on the ALL-EN model. We considered different data harmonization methods: (1) we applied ComBat by considering age as a covariate to preserve its effect while removing the variability associated with MFS (*ComBat*$_{Age}$) and (2) additionally regressed age out of the MRI data before estimating the predictive model (*ComBat*$_{Reg_{Age}}$). (3) We applied PLS-based domain adaptation using age and MFS as response variables in PLS, effectively removing the effects of both MFS and age (*PLS*$_{Age}$). The prediction performances are listed in Table 5. The comparison between *ComBat* and *ComBat*$_{Age}$ (see Tables 4 and 5) shows that the retaining age as a covariate did not improve R or MAE. For instance, R at 24 months dropped from 0.11 to 0.08 for NC, 0.38 to 0.36 for MCI, and 0.27 to 0.24 for AD. However, *ComBat*$_{Reg_{Age}}$ slightly improved the prediction performance for the AD group at 12 and 24 months: R for the AD group at 12 and 24 months increased from 0.19 to 0.24 and from 0.24 to 0.37, respectively. *PLS*$_{Age}$ slightly improved

the performance for each diagnosis group at the time points of 12 and 24 months compared to PLS without age as a covariate (see Table 4). For example, R increased from 0.40 to 0.44 for MCI and 0.28 to 0.34 for AD at 24 months. *PLS*$_{Age}$ featured slightly increased R values for all diagnosis groups at all time points compared to Combat-based approaches, e.g., R improved from 0.08 to 0.16 for NC, from 0.36 to 0.44 for MCI, and from 0.24 to 0.34 for AD at 24 months. The scatter plots are provided in Supplementary Figure 4. In addition, we illustrate the cross-validated accuracy of ComBat and PLS with and without considering age as a covariate in Fig. 4.

IV. DISCUSSION

We predicted the changes in the ADAS-Cog scores (Δ ADAS-Cog) in three distinct subject groups (NC, MCI, and AD) based on MRI for up to 36 months. We explored this problem by comparing various formulations of single- and multitask learning algorithms and scrutinizing whether multitask learning can help to cope with differences in the MRI data caused by different MFSs. MTL models aim to enhance generalization performance by utilizing relatedness among various tasks; here, predicting Δ ADAS-Cog in different subject groups. We predicted Δ ADAS-Cog scores from regional GM density values by single-task learning via
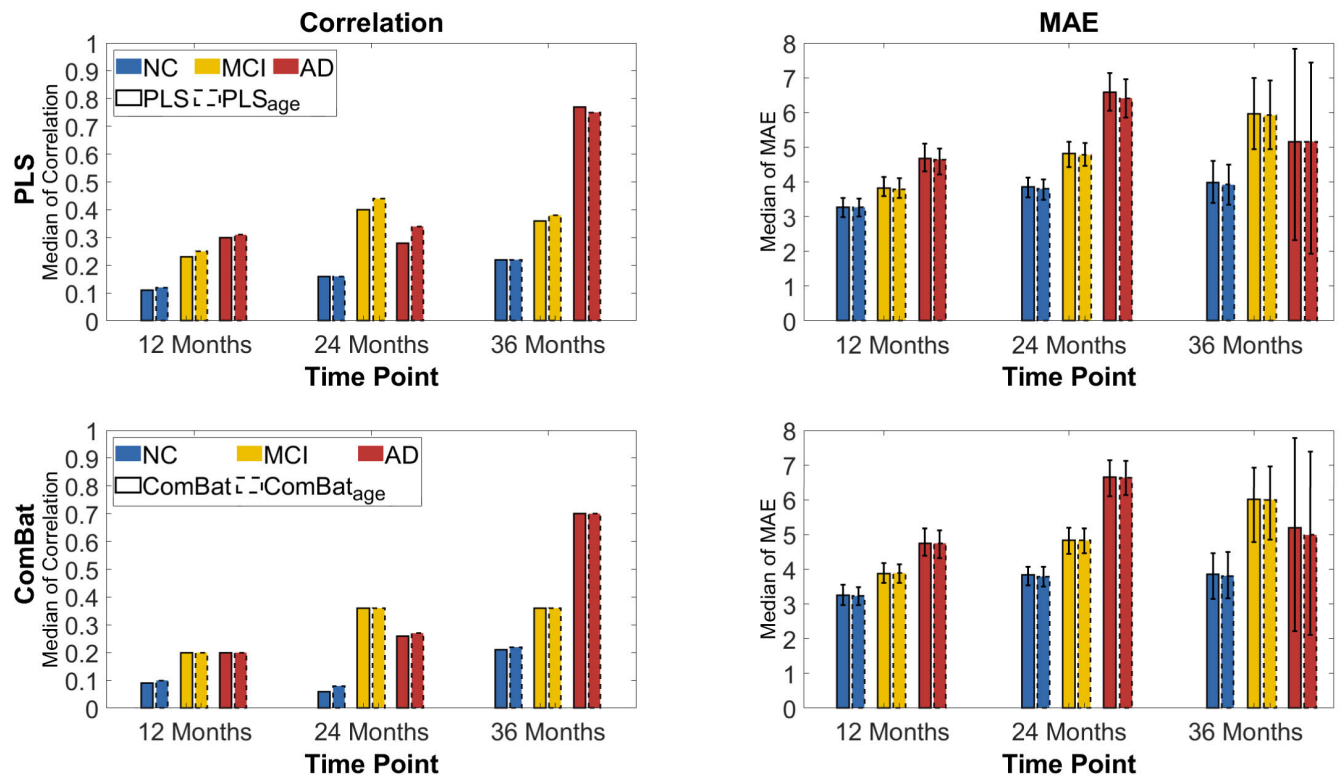
FIGURE 4. Comparison of the accuracy of PLS and ComBat harmonization with age (dashed line) and without age (solid line) as a biological covariate. The figure provides the median MAE and R among 10 CV runs along with the 95% confidence intervals obtained by the bootstrap method.

elastic net penalized linear regression as a baseline learning method. The single-task learning was applied based on two distinct strategies: 1) training the model by pooling together the data (ALL-EN) across the diagnostic groups, and 2) training a separate model for each diagnostic group (SEP-EN). We compared single-task models to multitask learning approaches, where we treated the prediction of cognitive scores in different baseline diagnoses as separate tasks.

The experiments revealed a positive correlation between observed and predicted Δ ADAS-Cog scores in all diagnostic groups at all time points. This indicates that MRI has predictive value for changes in ADAS-Cog scores across all subject groups. As shown in Table 3), the SEP-EN and MTL methods performed similarly; however, the ALL-EN method performed slightly better than the other methods regarding the average correlation score. In addition, a comparison of average correlation scores obtained from two single-task learning strategies (SEP-EN vs. ALL-EN) showed that simultaneous prediction in all diagnostic groups was beneficial for predicting disease progression in the NC and MCI groups. More complex multitask learning approaches were unable to provide benefits over single-task learning in our experiments.

Considering different MFS used in ADNI1 and ADNI2 cohorts, we studied 1) whether correcting for this difference affects the ADAS-Cog prediction and 2) whether multitask learning would be useful for such a correction. We used

two heterogeneity reduction approaches, typically applied for correcting for the site differences: PLS-based domain adaptation [37] and ComBat [38]. Correcting the MRI with the help of PLS-based domain adaptation marginally improved the ADAS-Cog change prediction, but the improvement was typically not statistically significant, as seen by comparing the confidence intervals in Table 4. Multitask learning with six tasks corresponding to the three baseline diagnoses and two MFSs of MRI did not bring any improvements over single-task learning.

We investigated the role of age as a covariate in the prediction models. The evaluation demonstrated that the accuracy of predicted Δ ADAS-Cog scores improved by regressing out the age from the MRI data. This agrees with previous studies, indicating that age has a significant effect on the accuracy of cognitive score prediction [28], [45].

Several studies have analyzed the role of ADAS-Cog scores in the evaluation of AD, as well as the relationship between ADAS-Cog and MRI ([25]–[30], [68], [69]). For instance, Wang *et al.* [68] proposed a multitask exclusive relationship learning model to automatically capture the intrinsic relationship among tasks at different time points for estimating clinical measures based on longitudinal imaging data. Yan *et al.* [30] proposed a new group-sparse multitask regression model for predicting ADAS, MMSE, and RAVLT cognitive scores at the baseline using cortical thickness measurements. Duchesne *et al.* [32] applied a linear regression

model to predict one-year MMSE changes using baseline MRI features and revealed that the baseline MRI features moderately predict one-year MMSE changes in the general MCI population. However, to the best of our knowledge, this work is the first to stratify the estimation of the prediction performance between diagnostic groups and utilize the relatedness between the diagnostic groups to boost the prediction performance. In addition, we demonstrated the necessity of stratifying subjects based on a baseline diagnosis to evaluate the predictive modeling of the change in ADAS-Cog.

V. CONCLUSION

We explored single and multitask learning to predict the changes in ADAS-Cog scores based on T1-weighted anatomical MRI. We stratified the subjects based on their baseline diagnoses and evaluated the prediction performances in each group. Our results indicated a positive relationship between the predicted and observed ADAS-Cog score changes in each diagnostic group, suggesting that standard T1-weighted MRI has a predictive value for evaluating the cognitive decline in the AD continuum. We further studied whether correction of the differences in MFS of MRI would improve the ADAS-Cog score prediction. The PLS-based domain adaptation slightly improved the prediction performance, but the improvement was marginal. In summary, this study demonstrated that ADAS-Cog changes could be, to some extent, predicted based on anatomical MRI. Based on this study, the recommended method for learning the predictive models is ALL-EN, due to its simplicity and good performance.

AVAILABILITY OF MATERIALS AND METHODS

The source codes of the single and multitask algorithms, as well as MRI pre-processing, are available at https://github.com/vandadim/ADAS_MRI. The dataset used in this paper was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and is available at http://adni.loni.usc.edu/data-samples/access-data/. The roster identification numbers (RIDs) of the subjects employed in this study are provided as a supplementary csv-file.

AUTHOR CONTRIBUTIONS

Vandad Imani: Methodology, software, formal analysis, writing (original draft), and visualization. Mithilesh Prakash: Writing (original draft), and visualization. Marzieh Zare: Software, data curation, and writing (review and editing). Jussi Tohka: Conceptualization, methodology, writing (original draft), supervision, and funding acquisition.

REFERENCES

[1] C. Patterson, World Alzheimer Report 2018: The State of the Art of Dementia Research: New Frontiers. London, U.K.: Alzheimer's Disease International, 2018.

[2] P. S. Aisen, J. Cummings, C. R. Jack, J. C. Morris, R. Sperling, L. Frölich, R. W. Jones, S. A. Dowsett, B. R. Matthews, J. Raskin, P. Scheltens, and B. Dubois, "On the path to 2025: Understanding the Alzheimer's disease continuum," Alzheimer's Res. Therapy, vol. 9, no. 1, pp. 1–10, Dec. 2017.

[3] K. M. Langa and D. A. Levine, "The diagnosis and management of mild cognitive impairment: A clinical review," J. Amer. Med. Assoc., vol. 312, no. 23, pp. 2551–2561, 2014.

[4] T. D. Koepsell and S. E. Monsell, "Reversion from mild cognitive impairment to normal or near-normal cognition: Risk factors and prognosis," Neurology, vol. 79, no. 15, pp. 1591–1598, Oct. 2012.

[5] J. Cummings, G. Lee, A. Ritter, M. Sabbagh, and K. Zhong, "Alzheimer's disease drug development pipeline: 2020," Alzheimer's Dementia, Transl. Res. Clin. Intervent., vol. 6, no. 1, 2020, Art. no. e12050.

[6] M. E. de Vugt and F. R. J. Verhey, "The impact of early dementia diagnosis and intervention on informal caregivers," Prog. Neurobiol., vol. 110, pp. 54–62, Nov. 2013.

[7] J. Rasmussen and H. Langerman, "Alzheimer's disease—Why we need early diagnosis," Degenerative Neurol. Neuromuscular Disease, vol. 9, p. 123, Dec. 2019.

[8] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "'Mini-mental state': A practical method for grading the cognitive state of patients for the clinician," J. Psychiatric Res., vol. 12, no. 3, pp. 189–198, 1975.

[9] R. C. Mohs, D. Knopman, R. C. Petersen, S. H. Ferris, C. Ernesto, M. Grundman, M. Sano, L. Bieliauskas, D. Geldmacher, C. Clark, and L. J. Thai, "Development of cognitive instruments for use in clinical trials of antidementia drugs: Additions to the Alzheimer's disease assessment scale that broaden its scope," Alzheimer Disease Associated Disorders, vol. 11, pp. 13–21, Nov. 1997.

[10] D. M. Kaufman, H. L. Geyer, and M. J. Milstein, "Dementia," in *Kaufman's Clinical Neurology for Psychiatrists*, D. M. Kaufman, H. L. Geyer, and M. J. Milstein, Eds., 8th ed. Amsterdam, The Netherlands: Elsevier, 2017, ch. 7, pp. 105–149. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323415590000071

[11] J. Skinner, J. O. Carvalho, G. G. Potter, A. Thames, E. Zelinski, P. K. Crane, and L. E. Gibbons, "The Alzheimer's disease assessment scale-cognitive-plus (ADAS-Cog-Plus): An expansion of the ADAS-Cog to improve responsiveness in MCI," *Brain Imag. Behav.*, vol. 6, no. 4, pp. 489–501, Dec. 2012.

[12] M. Ansart, S. Epelbaum, G. Bassignana, A. Bêne, S. Bottani, T. Cattai, R. Couronné, J. Faouzi, I. Koval, M. Louis, E. Thibeau-Sutre, J. Wen, A. Wild, N. Burgos, D. Dormont, O. Colliot, and S. Durrleman, "Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101848.

[13] Y. Utsumil, O. O. Rudovicl, K. Petersonl, R. Guerrero, and R. W. Picardl, "Personalized Gaussian processes for forecasting of Alzheimer's disease assessment scale-cognition sub-scale (ADAS-Cog13)," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 4007–4011.

[14] X. Zhu, H.-I. Suk, S.-W. Lee, and D. Shen, "Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis," *Brain Imag. Behav.*, vol. 10, no. 3, pp. 818–828, Sep. 2016.

[15] M. Prakash, M. Abdelaziz, L. Zhang, B. A. Strange, and J. Tohka, "Quantitative longitudinal predictions of Alzheimer's disease by multi-modal predictive learning," *J. Alzheimer's Disease*, vol. 79, pp. 1–14, Dec. 2020.

[16] S. Tsao, N. Gajawelli, J. Zhou, J. Shi, J. Ye, Y. Wang, and N. Leporé, "Feature selective temporal prediction of Alzheimer's disease progression using hippocampus surface morphometry," *Brain Behav.*, vol. 7, no. 7, Jul. 2017, Art. no. e00733.

[17] R. Wittenberg, M. Knapp, M. Karagiannidou, J. Dickson, and J. M. Schott, "Economic impacts of introducing diagnostics for mild cognitive impairment Alzheimer's disease patients," *Alzheimer's Dementia, Transl. Res. Clin. Intervent.*, vol. 5, no. 1, pp. 382–387, Jan. 2019.

[18] P. Vemuri and C. R. Jack, "Role of structural MRI in Alzheimer's disease," *Alzheimer's Res. Therapy*, vol. 2, no. 4, pp. 1–10, 2010.

[19] Y. Wang, Y. Fan, P. Bhatt, and C. Davatzikos, "High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables," *NeuroImage*, vol. 50, no. 4, pp. 1519–1535, May 2010.

[20] B. Jie, M. Liu, J. Liu, D. Zhang, and D. Shen, "Temporally constrained group sparse learning for longitudinal data analysis in Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 238–249, Jan. 2017.

[21] P. Cao, X. Liu, J. Yang, D. Zhao, M. Huang, and O. Zaiane, "$\ell_{2,1}-\ell_1$ regularized nonlinear multi-task representation learning based cognitive performance prediction of Alzheimer's disease," *Pattern Recognit.*, vol. 79, pp. 195–215, Jul. 2018.

[22] P. Cao, S. Tang, M. Huang, J. Yang, D. Zhao, A. Trabelsi, and O. Zaiane, "Feature-aware multi-task feature learning for predicting cognitive outcomes in Alzheimer's disease," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 1–5.

[23] B. Lei, M. Yang, P. Yang, F. Zhou, W. Hou, W. Zou, X. Li, T. Wang, X. Xiao, and S. Wang, "Deep and joint learning of longitudinal data for Alzheimer's disease prediction," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107247.

[24] N. Bhagwat, J. Pipitone, A. N. Voineskos, and M. M. Chakravarty, "An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures," *J. Psychiatry Neurosci.*, vol. 44, no. 4, p. 246, 2019.

[25] P. Jiang, X. Wang, Q. Li, L. Jin, and S. Li, "Correlation-aware sparse and low-rank constrained multi-task learning for longitudinal analysis of Alzheimer's disease," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1450–1456, Jul. 2019.

[26] L. Huang, Y. Jin, Y. Gao, and K.-H. Thung, "Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest," *Neurobiol. Aging*, vol. 46, pp. 180–191, Oct. 2016.

[27] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via multi-task learning," *NeuroImage*, vol. 78, pp. 233–248, Jun. 2013.

[28] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack, Jr., J. Ashburner, and R. S. Frackowiak, "Predicting clinical scores from magnetic resonance scans in Alzheimer's disease," *NeuroImage*, vol. 51, no. 4, pp. 1405–1413, 2010.

[29] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 814–822.

[30] J. Yan, T. Li, H. Wang, H. Huang, J. Wan, K. Nho, S. L. Risacher, A. J. Saykin, and L. Shen, "Cortical surface biomarkers for predicting cognitive outcomes using group $\ell_{2,1}$ norm," *Neurobiol. Aging*, vol. 36, pp. S185–S193, Jan. 2015.

[31] L. Lu, H. Wang, S. Elbeleidy, and F. Nie, "Predicting cognitive declines using longitudinally enriched representations for imaging biomarkers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4827–4836.

[32] S. Duchesne, A. Caroli, C. Geroldi, D. L. Collins, and G. B. Frisoni, "Relating one-year cognitive change in mild cognitive impairment to baseline MRI features," *NeuroImage*, vol. 47, no. 4, pp. 1363–1370, 2009.

[33] A. Jalali, P. Ravikumar, and S. Sanghavi, "A dirty model for multiple sparse regression," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7947–7968, Dec. 2013.

[34] S. Tabarestani, M. Aghili, M. Eslami, M. Cabrerizo, A. Barreto, N. Rishe, R. E. Curiel, D. Loewenstein, R. Duara, and M. Adjouadi, "A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study," *NeuroImage*, vol. 206, Feb. 2020, Art. no. 116317.

[35] B. Lei, F. Jiang, S. Chen, D. Ni, and T. Wang, "Longitudinal analysis for disease progression via simultaneous multi-relational temporal-fused learning," *Frontiers Aging Neurosci.*, vol. 9, p. 6, Mar. 2017.

[36] L. Wang, L. Xu, P. Li, S. Zha, and L. Chen, "Modeling disease progression via weakly supervised temporal multitask matrix completion," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 1141–1148.

[37] E. Moradi, B. Khundrakpam, J. D. Lewis, A. C. Evans, and J. Tohka, "Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data," *NeuroImage*, vol. 144, pp. 128–141, Jan. 2017.

[38] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[39] C. Gaser and R. Dahnke, "CAT—A computational anatomy toolbox for the analysis of structural MRI data," in *Proc. OHBM Annu. Meeting*, 2016, pp. 336–348.

[40] J. V. Manjón, P. Coupé, L. Martí-Bonmatí, D. L. Collins, and M. Robles, "Adaptive non-local means denoising of MR images with spatially varying noise levels," *J. Magn. Reson. Imag.*, vol. 31, no. 1, pp. 192–203, 2010.

[41] J. C. Rajapakse, J. N. Giedd, and J. L. Rapoport, "Statistical approach to segmentation of single-channel cerebral MR images," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 176–186, Apr. 1997.

[42] J. Tohka, A. Zijdenbos, and A. Evans, "Fast and robust parameter estimation for statistical partial volume models in brain MRI," *NeuroImage*, vol. 23, no. 1, pp. 84–97, Sep. 2004.

[43] J. Ashburner, "A fast diffeomorphic image registration algorithm," *Neuroimage*, vol. 38, no. 1, pp. 95–113, 2007.

[44] W. G. Rosen, R. C. Mohs, and K. L. Davis, "A new rating scale for Alzheimer's disease," *Amer. J. Psychiatry*, vol. 111, pp. 1356–1364, Nov. 1984.

[45] J. K. Kueper, M. Speechley, and M. Montero-Odasso, "The Alzheimer's disease assessment scale–cognitive subscale (ADAS-Cog): Modifications and responsiveness in pre-dementia populations. A narrative review," *J. Alzheimer's Disease*, vol. 63, no. 2, pp. 423–444, Apr. 2018.

[46] J.-P. Fortin, D. Parker, B. Tunç, T. Watanabe, M. A. Elliott, K. Ruparel, D. R. Roalf, T. D. Satterthwaite, R. C. Gur, R. T. Schultz, R. Verma, and R. T. Shinohara, "Harmonization of multi-site diffusion tensor imaging data," *NeuroImage*, vol. 161, pp. 149–170, Nov. 2017.

[47] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[48] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 41–48.

[49] A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar, "A dirty model for multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 964–972.

[50] D. He, D. Kuhn, and L. Parida, "Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction," *Bioinformatics*, vol. 32, no. 12, pp. i37–i43, Jun. 2016.

[51] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.

[52] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 457–464.

[53] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statist. Comput.*, vol. 20, no. 2, pp. 231–252, 2010.

[54] J. Zhou, J. Chen, and J. Ye, "MALSAR: Multi-task learning via structural regularization," Dept. Comput. Sci. Eng., Arizona State Univ., Tempe, AZ, USA, Tech. Rep., 2011, vol. 21.

[55] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization," 2012, *arXiv:1205.2631*.

[56] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[57] S. M. Fazel, "Matrix rank minimization with applications," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2003.

[58] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, M. McInnis, M. L. Phillips, M. H. Trivedi, M. M. Weissman, and R. T. Shinohara, "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage*, vol. 167, pp. 104–120, Feb. 2018.

[59] F. Orlhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, and I. Buvat, "A postre-construction harmonization method for multicenter radiomic studies in PET," *J. Nucl. Med.*, vol. 59, no. 8, pp. 1321–1328, Aug. 2018.

[60] M. Yu, K. A. Linn, P. A. Cook, M. L. Phillips, M. McInnis, M. Fava, M. H. Trivedi, M. M. Weissman, R. T. Shinohara, and Y. I. Sheline, "Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data," *Hum. Brain Mapping*, vol. 39, no. 11, pp. 4213–4227, 2018.

[61] J. Tohka and M. van Gils, "Evaluation of machine learning algorithms for health and wellness applications: A tutorial," *Comput. Biol. Med.*, vol. 132, May 2021, Art. no. 104324.

[62] P. I. Good, *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. U.K.: Springer, 2006.

[63] J. D. Lewis, A. C. Evans, and J. Tohka, "T1 white/gray contrast as a predictor of chronological age, and an index of cognitive performance," *NeuroImage*, vol. 173, pp. 341–350, Jun. 2018.

[64] H. Huttunen, T. Manninen, and J. Tohka, "Meg mind reading: Strategies for feature selection," in *Proc. Fed. Comput. Sci. Event*, 2012, pp. 42–49.

[65] J. Qian, T. Hastie, J. Friedman, R. Tibshirani, and N. Simon, "Glmnet for MATLAB," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2013, pp. 2–4. Accessed: Nov. 13, 2017.

[66] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[67] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," *NeuroImage*, vol. 104, pp. 398–412, Jan. 2015.

[68] M. Wang, D. Zhang, D. Shen, and M. Liu, "Multi-task exclusive relationship learning for Alzheimer's disease progression prediction with longitudinal data," *Med. Image Anal.*, vol. 53, pp. 111–122, Feb. 2019.

[69] A. Zandifar, V. S. Fonov, S. Ducharme, S. Belleville, and D. L. Collins, "MRI and cognitive scores complement each other to accurately predict Alzheimer's dementia 2 to 7 years before clinical onset," *NeuroImage, Clin.*, vol. 25, Jan. 2020, Art. no. 102121.

VANDAD IMANI received the master's degree in artificial intelligence from the Central University of Hyderabad, India, in 2014. He is currently pursuing the Ph.D. degree with the A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland. His research interests include machine learning, pattern recognition, and their applications to predict brain disorders.

MITHILESH PRAKASH was born in Bellary, India, in 1988. He received the B.E. degree in medical electronics from V.T.U, India, in 2009, the M.Sc. (Tech) degree in biomedical engineering from the Tampere University of Technology, Finland, in 2015, and the Ph.D. degree in applied physics from the University of Eastern Finland, Finland, in 2019. He has over four years of software programming industry experience. He is currently working as a Postdoctoral Researcher with the A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland. His research interests include machine learning, algorithm designing, and programming for their applications in biomedical engineering.

MARZIEH ZARE received the M.S. degree from the Signal Processing Department, Tampere University, Tampere, Finland, in 2017, where she is currently pursuing the Ph.D. degree in computing sciences. She was also a Research Assistant with the A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland with supervision of Prof. Jussi Tohka, from 2017 to 2018. Her research interests include data analysis, signal processing, machine learning, and pattern recognition and their applications to industrial solutions.

JUSSI TOHKA received the Ph.D. degree in signal processing from the Tampere University of Technology, Finland, in 2003. He was a Postdoctoral Fellow with the University of California at Los Angeles, Los Angeles, USA. He held an academy research fellow position with the Department of Signal Processing, Tampere University of Technology. He was a CONEX Professor with the Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Spain. He is currently with the A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland. His research interests include machine learning, image analysis, and pattern recognition and their applications to brain imaging.

● ● ●