Pattern Recognition 102 (2020) 107247

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

# Deep and joint learning of longitudinal data for Alzheimer's disease prediction

Baiying Lei[a], Mengya Yang[a], Peng Yang[a], Feng Zhou[b], Wen Hou[c], Wenbin Zou[c], Xia Li[c], Tianfu Wang[a], Xiaohua Xiao[d,*], Shuqiang Wang[e,*]

[a] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China
[b] Department of Industrial and Manufacturing, Systems Engineering, The University of Michigan, Dearborn, MI, USA
[c] Shenzhen Key Lab of Advanced Telecommunication and Information Processing, College of Electronic and Information Engineering, Shenzhen University, Shenzhen 518060, China
[d] First Affiliated Hospital of Shenzhen University, Health Science Center, Shenzhen University, Shenzhen 518060, China
[e] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Shenzhen 518000, China

ARTICLE INFO

*Article history:*
Received 15 February 2019
Revised 14 January 2020
Accepted 26 January 2020
Available online 28 January 2020

*Keywords:*
Alzheimer's disease
Longitudinal scores prediction
Joint learning
Correntropy
Deep polynomial network

ABSTRACT

Alzheimer's disease (AD) is an irreversible and progressive neurodegenerative disease. The close AD monitoring of this disease is essential for the patient treatment plan adjustment. For AD monitoring, clinical score prediction via neuroimaging data is highly desirable since it is able to reveal the disease status, adequately. For this task, most previous studies are focused on a single time point without considering relationship between neuroimaging data (e.g., Magnetic Resonance Imaging (MRI)) and clinical scores at multiple time points. Differing from these studies, we propose to build a framework based on longitudinal multiple time points data to predict clinical scores. Specifically, the proposed framework consists of three parts, feature selection based on correntropy regularized joint learning, feature encoding based on deep polynomial network, and ensemble learning for regression via the support vector regression method. Two scenarios are designed for scores prediction. Namely, scenario 1 uses the baseline data to achieve the longitudinal scores prediction, while scenario 2 utilizes all the previous time points data to obtain the predicted scores at the next time point, which can improve the score prediction's accuracy. Meanwhile, the missing clinical scores at longitudinal multiple time points are imputated to solve the incompleteness of the data. Extensive experiments on the public database of Alzheimer's Disease Neuroimaging Initiative (ADNI) demonstrate that our proposed framework can effectively reveal the relationship between clinical score and MRI data and outperforms the state-of-the-art methods in scores prediction.



## 1. Introduction

Alzheimer's disease (AD) is an irreversible neurodegenerative disease that results in progressive loss of memory and other mental functions. It is a major cause of dementia and is the sixth principal cause of death in the United States [1]. AD possess a mass of population and has given rise to a huge economic burden to the society. In 2018, there is about 50 million AD patients in the world and has caused about 1 trillion US dollars economic cost, and this will be doubled by 2030 [2]. Therefore, AD has garnered growing attention in the last several years. Several previous studies have revealed that AD may affect both functions and structures of the brain [3,4]. It is characterized by a decline in cognitive competence such as memory and problem-solving, which seriously affects person's activity of daily living. The accurate diagnosis of AD is highly important for patients to receive timely treatment and can delay the progression of the disease as much as possible. During the past few decades, neuroimaging techniques such as positron emission topography (PET) [5,6] and magnetic resonance imaging (MRI) [7,8] are dominant tools to expedite AD diagnoses and treatments [9–11].

Most of the existing researches focus on developing novel classification frameworks at a single time point to predict class labels such as normal control (NC) and AD, by the patterns in neuroimaging data [4,5]. Lately, regression analysis models have also been utilized to predict longitudinal clinical scores, like the AD assessment scale-cognitive subscale (ADAS-Cog) [12], the clinical demen-

* Corresponding authors.
*E-mail addresses:* leiby@szu.edu.cn (B. Lei), 2172243161@email.szu.edu.cn (M. Yang), fenzhou@umich.edu (F. Zhou), wzou@szu.edu.cn (W. Zou), lixia@szu.edu.cn (X. Li), tu_xi8888@163.com (X. Xiao), sq.wang@siat.ac.cn (S. Wang).

https://doi.org/10.1016/j.patcog.2020.107247
0031-3203/

tia rating-global and the sum of boxes (CDR-GLOB and CDR-SOB) [13], and mini-mental state examination (MMSE) using MRI and/or PET data [14,15]. There are many studies on longitudinal clinical score prediction since it can help evaluate the status of patients and predict disease progression. For example, Wang et al. [15] proposed a high-dimensional kernel regression framework to achieve the prediction of MMSE and ADAS-Cog. Duchesne et al. [16] employed the linear regression method from MRI to evaluate one year MMSE changes. However, these regression framework mainly focus on the traditional machine learning methods, and still face the challenge of missing data [17], which is caused by a variety of subjects dropping out for various reasons at different time points.

Different from existing studies, we design a novel combined machine learning and deep learning framework to realize the longitudinal clinical score prediction. The regression process is devised in two scenarios. In scenario 1, only the baseline data is utilized for training to achieve longitudinal clinical score prediction, which is similar to the traditional methods used in previous studies [18,19]. Then, we design a novel training framework in scenario 2, where all the previous time points data is combined for clinical score prediction at next time point. As for data missing problem, one common approach to deal with the incompleteness is to remove the subjects with missing scores, which results in reduction of available subjects and limits our investigation. Another approach is to imputate missing scores via interpolation methods, which is heavily dependent on temporal smoothness and can be biased if the subject's situation exacerbates. Considering these shortcomings, we propose to combine all the available data at previous and current time points to achieve the clinical scores prediction. Meanwhile, the missing clinical scores are imputated by our proposed regression framework, and thus the problem of data incompleteness that presents at multiple time points is solved.

With the multiple time points data, one difficulty faced by the prediction task is the overfitting. This problem can be addressed by two ways. One way is by feature selection [18], and the other way is by subspace learning. The first category includes statistical Chi-squared methods, *t*-test, and sparse models [20], which is advantageous in identifying biomarkers. The second category projects features into a low-dimensional space [21], which demonstrates preferable performance in indicating disease status. Therefore, it is advantageous for clinical scores prediction by combining the strengths of the above categories [22]. Inspired by this, we explore the temporally constrained group LASSO method [23] to realize the feature selection. In addition, the correntropy [24] is incorporated to eliminate outliers and improve prediction performance.

A great success has been witnessed by the recent development of deep learning (DL) method [25] especially for AD diagnosis and prognosis [26,27]. As a particular type in the renaissance of DL, deep polynomial network (DPN) [28] is an effective and supervised method with solid theoretical foundation. DPN attempts to build a network that offers a superior approximation basis for the values achieved by all polynomials of bounded degree on the training cases. By integrating features between different dimensions and samples through the network in a hierarchical way, feature representation performance can be greatly boosted. It has similar or even better performance compared with the sparse autoencoder and deep belief networks algorithms on some image datasets [29]. Hence, it would be beneficial to combine the temporally constrained group LASSO model with deep feature representation via DPN for accurate prediction performance.

In this paper, the proposed regression framework develops a feature selection method based on the correntropy [30], temporally constrained group LASSO (CT), and feature encoding in DPN to realize the longitudinal clinical score prediction. The ensemble learning technique is also utilized via the support vector regression (SVR) [31] method. We name the proposed method as CTDE, where 'D' stands for DPN based feature encoding model and 'E' is for ensemble. Extensive experiments demonstrate that, the overall CTDE model achieves superior performance than partial models, such as the model with only feature encoding or only feature selection. The extensive experiments have been performed to validate the proposed method and details are described in the following sections. Our main contributions are summarized as below:

(1) The proposed framework addresses both longitudinal characteristics and data incompleteness presented in the ADNI [32] database.
(2) An effective feature selection model is developed by incorporating the temporal constraint and the correntropy, followed by the implementation of an efficient optimization algorithm. Hereby, the most relevant features can be found.
(3) The combination of correntropy regularized joint learning and DPN promotes the advantages of each other in the sense of increasing the prediction accuracy and discovering AD biomarkers.
(4) Instead of concatenating the output of all DPN layers, a weighted ensemble method is proposed to further explore the benefits of feature encoding via DPN.

## 2. Materials

### 2.1. Data acquisition

In this work, we obtain the dataset from the public Alzheimer's disease neuroimaging initiative (ADNI) database as it is the most popular database to analyze development process of AD. A total of 805 subjects including the baseline MRI T1-weighted (T1w) data and various clinical scores (ADAS-Cog, CDR-GLOB, CDR-SOB, and MMSE) at baseline are utilized in this study. A return visit was expected at the 06th month (M06), 12th month (M12), 18th month (M18), 24th month (M24), and 36th month (M36) from baseline. As shown in Fig. 1, the proposed longitudinal scores prediction model is conducted for two scenarios. Specifically, scenario 1 is the traditional training model, where only baseline data is employed to realize the scores prediction at future time points. Scenario 2 is a novel method to improve the scores prediction performance, where all the previous time points data are combined to get the predicted scores at next time point, which not only increases the quantity but also considers the correlation between subjects at multiple time points.

In particular, the splits of the training and testing set in two scenarios are different. In scenario 1, we employ the baseline data (e.g., MRI data and four type of clinical scores) to be the training set, and the testing set will change with time. For example, we put the MRI data at M06 as the testing set into the proposed regression framework to predict the scores at M06, and the MRI data at M12 as the testing set into the framework to predict scores at M12. But there is a big difference between the two scenarios. In scenario 2, both the training and testing set will change when we predict scores at different time points. We take the clinical scores prediction at M06 for example. The training set is MRI data and scores at baseline, and the testing set is the MRI data and scores at M06. Afterwards, we predict the clinical scores at M12 by the proposed regression framework. The training set is the MRI data and scores at all previous time points (i.e., baseline, M06). The testing set is the MRI data and scores of subjects at M12. Ultimately, we can obtain the predicted clinical scores at M12. The other time points can be obtained in the same manner too. The details of our dataset information are presented in Table 1. The first part of the category is the number of available subjects with MRI data, and other one is the subjects both MRI and clinical scores.

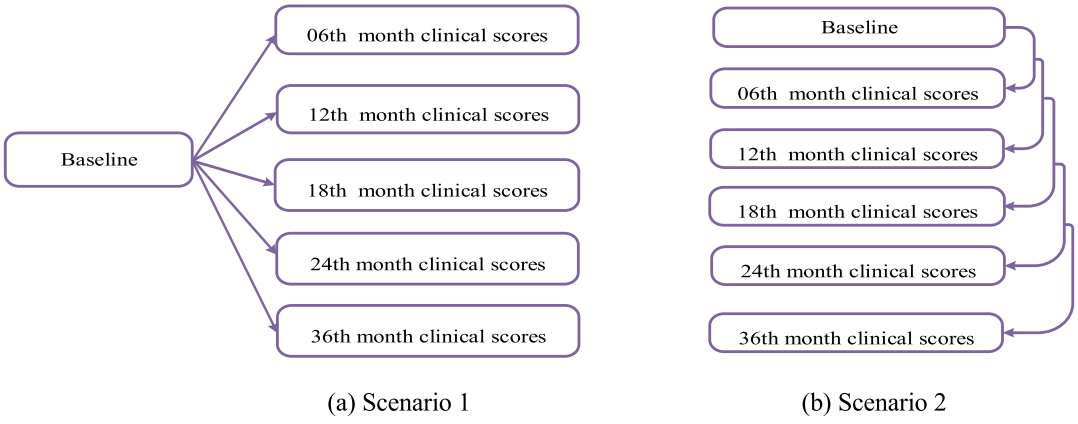



**Fig. 1.** Two different scenarios for clinical scores prediction at multiple time points.

**Table 1**
Number of subjects used in our experiments at different time points.

| Category | Clinical score | Baseline | M06 | M12 | M18 | M24 | M36 |
|---|---|---|---|---|---|---|---|
| MRI data | MMSE | 805 | 725 | 675 | 282 | 479 | 50 |
| | CDR-SOB | 805 | 725 | 675 | 282 | 479 | 50 |
| | CDR-GLOB | 805 | 725 | 675 | 282 | 479 | 50 |
| | ADAS-Cog | 805 | 725 | 675 | 282 | 479 | 50 |
| MRI data and clinical score | MMSE | 805 | 705 | 637 | 247 | 430 | 50 |
| | CDR-SOB | 805 | 725 | 675 | 280 | 473 | 50 |
| | CDR-GLOB | 805 | 722 | 667 | 280 | 473 | 50 |
| | ADAS-Cog | 805 | 725 | 674 | 281 | 477 | 50 |

From Table 1, we note that clinical scores of some subjects are missing at different time points (i.e., M24). For example, there are 725, 675, 282, 479, and 50 subjects with MRI data at the five time points from baseline, respectively. Among these subjects, only 705, 637, 247,430 and 50 subjects have MMSE clinical scores. Hence, we propose to imputate the missing scores by the regression framework. We take MMSE clinical scores prediction at M06 for example. In scenario 2, we put the MRI data of the subject who has MRI data but without the clinical scores at M06 into the proposed regression framework and get the predicted missing clinical scores. Then we put the predicted missing clinical scores into the incomplete clinical scores obtained by return visit at M06. Therefore, we obtain the complete clinical scores at M06. Moreover, we put the data into the proposed regression framework to predict the clinical scores at M06. In the experiment, the baseline data is defined as the training set and the complete data at M06 is defined as the testing set. Namely, all the previous time points data are utilized to achieve clinical scores prediction at next time point. The prediction at other time points can be achieved in the same manner.

### 2.2. Feature extraction

We preprocess MRI data with the same method in [33], which is divided into four steps as below:

(1) The intensity homogeneity is corrected through N3 method [34], and skull of brain is removed.
(2) The FAST in the open source software package FSL (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) is employed to segment the brain tissue into cerebrospinal, gray matter (GM), and white matter (WM) [35].
(3) The HAMMER tool [36] is applied to register the segmented images to a commonly used Jacob atlas. It is a T1w GM parcellation with 93 cerebral regions of interest (ROIs) using the corresponding anatomic definitions.

**Table 2**
List of the used notations in this paper.

| | |
|---|---|
| $\mathbf{X}_t$ | The feature matrix of $N$ subjects at $t$ time point |
| $\mathbf{Y}_t$ | The clinical scores of $N$ subjects at $t$ time point |
| $\mathbf{x}_t^i$ | The $i$th row vetor of $\mathbf{X}_t$ |
| $\mathbf{x}_{t,j}$ | The $j$th column vetor of $\mathbf{X}_t$ |
| $x_{t,j}^i$ | The element in $i$th row and $j$th column of $\mathbf{X}_t$ |
| $\mathbf{W}$ | The weight matrix across all time points |
| $\mathbf{w}$ | a vector of $\mathbf{W}$ |
| $\|\mathbf{X}_t\|_{2,1}$ | The $\ell_{2,1}$norm of $\mathbf{X}_t$, i.e., $\|\mathbf{X}_t\|_{2,1} = \sum_i \sqrt{\sum_j (x_{t,j}^i)^2}$ |

(4) The volumes of the 93 ROIs for each subject are calculated and normalized through the total intracranial volume, and utilized as features in this paper.

## 3. Methods

### 3.1. System overview

The overall architecture of our proposed longitudinal scores prediction framework is illustrated in Fig. 2. The procedures of the regression framework include feature extraction, feature selection, feature encoding and final SVR regression. The details of the proposed method are described in the following subsections.

### 3.2. Feature selection via joint learning

In this paper, there is $N$ number of all subjects and the MRI data with clinical scores of each subject are derived from $T$ different time points. Let the $\mathbf{X}_t = [\mathbf{x}_t^1; \ldots; \mathbf{x}_t^i; \ldots; \mathbf{x}_t^N] \in \mathbb{R}^{N\times D}$ and $\mathbf{Y}_t \in \mathbb{R}^{N\times 1}$ stand for the $D$-dimension MRI data and scores of all subjects at time point $t$, respectively, $\mathbf{x}_t^i \in \mathbb{R}^{1\times D}$ denotes the $D$-dimensional row vector at time point $t$, $t = 1 \ldots T$. Note that, uppercase boldface letters represent matrices while small bold letters are vectors. We summarize all notations in Table 2.

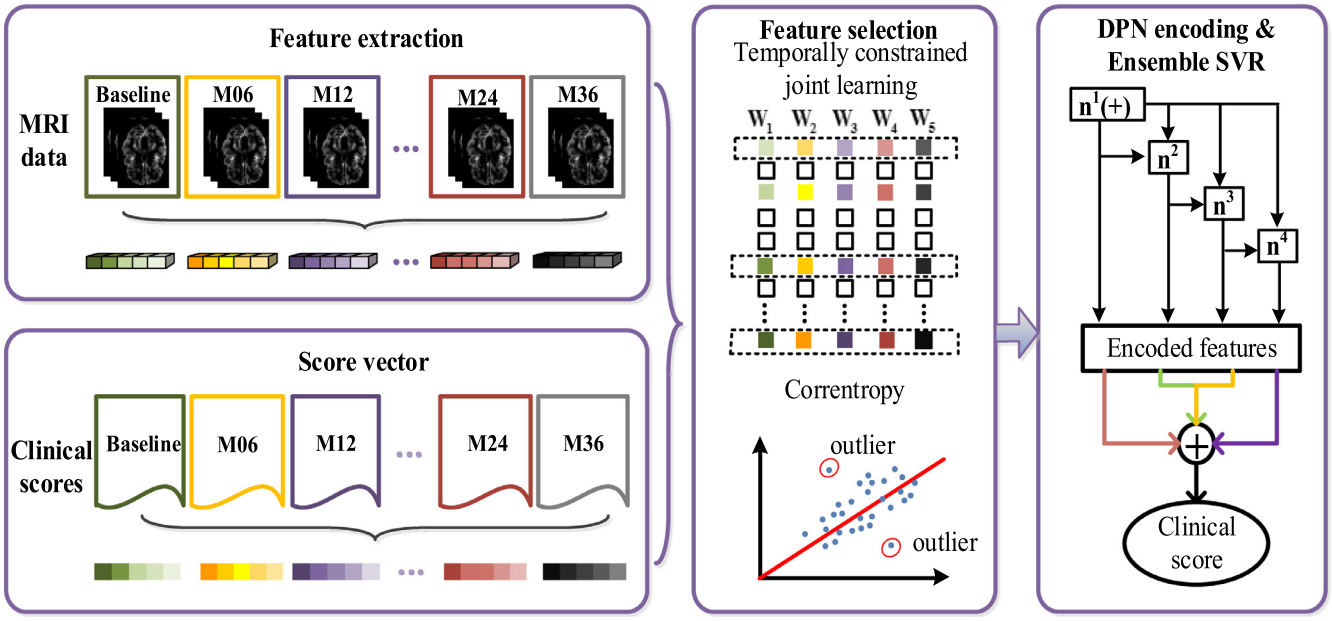



**Fig. 2.** Flowchart of the proposed method with deep and joint learning.

Due to the high dimensionality of MRI features and small sample size, it is easy to lead to the problem of overfitting, which can affect the analysis of experimental results. Hence, the sparsity inducing regularization method has been proposed to reduce the high dimensionality of MRI features [37], such as LASSO [38]. The standard group LASSO method is formulated as

$$\underset{\mathbf{W}}{\operatorname{argmin}} \ \frac{1}{2}\sum_{t=1}^{T}\mathbf{Y}_t - \mathbf{X}_t\mathbf{w}_t{}_2^2 + \rho_0\|\mathbf{W}\|_{2,1}, \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_t\ , \cdots, \mathbf{w}_T] \in \mathbb{R}^{D\times T}$ is weight coefficient matrix of all time points and $\mathbf{X}_t\mathbf{w}_t = \mathbf{Y}_t$ is utilized to evaluate the scores at $t$ time point, where $\mathbf{w}_t \in \mathbb{R}^{D\times 1}$ holds different weights of each feature. The $\|\mathbf{W}\|_{2,1}$ denotes the regularization term which can improve the generalization ability, and $\rho_0$ is regularization parameter. The $\ell_{2,1}$ norm [39,40] $\mathbf{W}_{2,1} = \sum_{t=1}^{T}\mathbf{w}_{t2}$ is used for joint feature selection by penalizing weight coefficients in the same row of $\mathbf{W}$, and then features corresponding to the small weight coefficients are discarded. In addition, a fused smoothness term [41] is incorporated into the group LASSO model to explore the temporal characteristics of longitudinal data. The fused smoothness term can minimize the weight difference over time and fully explore the longitudinal similarity. The Eq. (1) can be rewritten as

$$\underset{\mathbf{W}}{\operatorname{argmin}} \ \frac{1}{2}\sum_{t=1}^{T}\mathbf{Y}_t - \mathbf{X}_t\mathbf{w}_t{}_2^2 + \rho_0\mathbf{W}_{2,1} + \ \rho_1\sum_{t=1}^{T-1}\mathbf{w}_t - \mathbf{w}_{t+1}{}_2^2. \tag{2}$$

Furthermore, we add correntropy into the model to remove the potential outliers such as non-Gaussian noise and impulsive noise. The joint feature selection model with correntropy is finally defined as

$$\underset{\mathbf{W}}{\operatorname{argmin}} \ 1 - \frac{1}{2}\sum_{t=1}^{T}\exp\left(-\frac{\mathbf{Y}_t - \mathbf{X}_t\mathbf{w}_t{}_2^2}{\rho_2}\right) + \rho_0\mathbf{W}_{2,1} + \rho_1\sum_{t=1}^{T-1}\mathbf{w}_t - \mathbf{w}_{t+1}{}_2^2, \tag{3}$$

where $\rho_1$and $\rho_2$ are positive tuning parameters, $\rho_2$ refers to the kernel size and dominates the properties of the correntropy. Finally, Eq. (3) can obtain the informative features which are favorable for the following scores prediction.

To solve Eq. (3), we use the accelerated gradient method (AGM) [42] to optimize the objective function. Specifically, the objective function $f(\mathbf{W})$ described in Eq. (3) is divided into two parts, one is a smooth function $f_s(\mathbf{W})$, and the other is non-smooth function $f_{ns}(\mathbf{W})$, where

$$f_s(\mathbf{W}) = 1 - \frac{1}{2}\sum_{t=1}^{T}\exp\left(-\frac{\mathbf{Y}_t - \mathbf{X}_t\mathbf{w}_t{}_2^2}{\rho_2}\right) + \ \rho_1\sum_{t=1}^{T-1}\mathbf{w}_t - \mathbf{w}_{t+1}{}_2^2, \tag{4}$$

$$f_{ns}(\mathbf{W}) = \rho_0\mathbf{W}_{2,1}\ . \tag{5}$$

Supposing $k$ is the iteration index and $\mathbf{W}^k$ is weight matrix at the iteration index $k$, the first-order Taylor expansion is employed at $\mathbf{W}^k$ for $f_s(\mathbf{W})$ and obtain an proximate composite function of $f(\mathbf{W})$ [43] as the model g( · ) using

$$g_{L,\mathbf{W}^k} = f_s\left(\mathbf{W}^k\right) + \left\langle f'\left(\mathbf{W}^k\right), \mathbf{W} - \mathbf{W}^k\right\rangle \ + \frac{L}{2}\mathbf{W} - \mathbf{W}^k{}_F^2 + f_{ns}\left(\mathbf{W}^k\right), \tag{6}$$

where $<\boldsymbol{A},\boldsymbol{B}>=\mathrm{Tr}(\mathbf{A}^{\mathbf{T}}\mathbf{B})$ denotes the matrix inner product. The $\|\cdot\|_F$ denotes the Frobenious norm [44] and the regularization term $L/2\mathbf{W} - \mathbf{W}^k{}_F^2$ keeps $\mathbf{W}^k$ in the neighborhood of $\mathbf{W}$. $L$ is a regularization parameter and $L>0$. At the $k$-th iteration, $\mathbf{U}^k$ is the affine combination of $\mathbf{W}^k$ and $\mathbf{W}^{k-1}$, which is defined as

$$\mathbf{U}^k = \mathbf{W}^k + \beta_k\left(\mathbf{W}^k - \mathbf{W}^{k-1}\right), \tag{7}$$

with the carefully selected parameter $\beta_k$. Hence, the approximate solution of $\mathbf{W}^{k+1}$ is given by minimizing $g_{L_k,\mathbf{U}^k}(\mathbf{W})$, and $\mathbf{W}^{k+1}$ is used to solve the $\ell_{2,1}$-norm regularized Euclidean projection problem, where $L_k$ is found by line search on the basis of the Armijo-Goldstein rule. After ignoring constant terms, Eq. (6) can be rewritten as

$$\mathbf{W}^{k+1} = \arg\min_{\mathbf{W}}\frac{1}{2}\mathbf{W} - \mathbf{V}_F^2 + \frac{1}{L_k}f_{ns}(\mathbf{W}) \ = \underset{\mathbf{w}_1\cdots\mathbf{w}_D}{\operatorname{argmin}} \ \frac{1}{2}\sum_{j=1}^{D}\mathbf{w}_j - \mathbf{v}_j{}_2^2 + \frac{\rho_0}{L_k}f_{ns}\mathbf{w}_j{}_2, \tag{8}$$

where $\mathbf{V} = \mathbf{U}^k - 1/\boldsymbol{L}_k\ f'(\mathbf{U}^k)$, and $\mathbf{w}_j$ and $\mathbf{v}_j$represent $j$-th row of $\mathbf{W}$ and $\mathbf{V}$, respectively. Therefore, the problem changes into $D$ separate subproblems, and we use the algorithm [45] to calculate the

**Algorithm 1**
Optimization algorithm for correntropy based joint group learning.

**Input:** $\rho_0 > 0, \rho_1 > 0, \rho_2 > 0\, L_0 > 0, \mathbf{W}^0$, $K$
**Output:** $\mathbf{W}^{K+1}$
1: Initialize $\mathbf{W}^1 = \mathbf{W}^0$, $\alpha_{-1} = 0$, $\alpha_0 = 1$, and $L = L_0$
2: **for** $k = 1$ to $K$ **do**
3: $\beta_k = \frac{\alpha_{k-2}-1}{\alpha_{k-1}}$, $\mathbf{U}^k = \mathbf{W}^k + \beta_k(\mathbf{W}^k - \mathbf{W}^{k-1})$
4: Update $\mathbf{W}^{k+1}$ in Eq. (8)
5: Find the minimum value of $L$ among $\{L_{k-1}, 2L_{k-1}, \cdots\}$, ...,such that $f(\mathbf{W}^{k+1}) \le g_{L_k,\mathbf{U}^k}(\mathbf{W}^{k+1})$
6: Update $L_k = L$
7: $\alpha_{k+1} = \frac{1+\sqrt{1+4\alpha_k^2}}{2}$
8: **end for**

optimal solution $\mathbf{w}^*_j$ as

$$\mathbf{w}^*_j = \begin{cases} \left(1 - \frac{\rho_0}{L\mathbf{v}_{j2}}\right) & if\left(\mathbf{v}_{j2} > \frac{\rho_0}{L_k}\right) \\ 0 & otherwise. \end{cases} \tag{9}$$

The convergence rate is $O(1/K^2)$ as the optimization algorithm is an exact AGM and $K$ is the maximum iteration number. The overall procedure is described in Algorithm 1.

### 3.3. Feature encoding via deep polynomial network

The DPN is able to describe any function on a certain number of dataset succinctly due to its network architecture. The learned predictors are polynomial functions over input space since each of DPN node calculates a linear or quadratic function of the input. In order to present complex predictions compactly, DPN builds a deep architecture, which relies on its multi-layered structure. The width of DPN is denoted as the maximum number of nodes in each layer, and the depth is denoted as the number of layers. The predictors of DPN are the polynomial function on the input space. There is a degree-4 network architecture as an example in Fig. 2. Following the feature selection in Section 3.2, it is assumed that each subject has $M$-dimension features with $M < D$. Here, we take the DPN application at $t$ time point as an example. Starting from constructing the first degree in DPN, the approximate basis is given by

$$\left\{\left(\mathbf{z}, \left[1\ \mathbf{x}_t^1\right], \mathbf{z}, \left[1\ \mathbf{x}_t^2\right], \cdots, \mathbf{z}, \left[1\ \mathbf{x}_t^N\right]\right) : \mathbf{z} \in \mathbb{R}^{M+1}\right\}, \tag{10}$$

which is the $(M+1)$-dimensional linear subspace of $\mathbb{R}^N$. To conduct a basis for it, the singular value decomposition (SVD) is applied to find $M+1$ vectors $\mathbf{z}_1, \mathbf{z}_2, \cdots \mathbf{z}_{M+1}$, so that $\{(\mathbf{z}_i, [1\ \mathbf{x}_t^1], \mathbf{z}_i, [1\ \mathbf{x}_t^2], \cdots, \mathbf{z}_i, [1\ \mathbf{x}_t^N])\}_{i=1}^{M+1}$ are linearly independent. We denote a linear transformation matrix $\mathbf{Z}$ which is used to map $[\ \mathbf{1}\ \mathbf{X}_t]$ into the constructed basis, where $\mathbf{1}$ is the all-ones vector. The columns in $\mathbf{Z}$ represent the $M+1$ linear functions, which form the 1-st layer network in DPN. Let the $\mathbf{F}^1 \in \mathbb{R}^{N\times(M+1)}$ denotes the output of the first layer with the independent vectors as columns.

According to the decomposition theorem of polynomials in [46], any polynomial of degree $P$ can be obtained by the degree- $(P-1)$ polynomials and the degree-1 polynomials, and the more network layers are built in the same way. Taking the construction of the $P$th layer network as an example, we define the new matrix

$$\tilde{\mathbf{F}}^P = \left[(\mathbf{F}_1^{P-1} \circ \mathbf{F}_1^1)(\mathbf{F}_1^{P-1} \circ \mathbf{F}_2^1) \cdots (\mathbf{F}_1^{P-1} \circ \mathbf{F}_{|\mathbf{F}_1|}^1) \cdots (\mathbf{F}_{|\mathbf{F}^{P-1}|}^{-1} \circ \mathbf{F}_{|\mathbf{F}_1|}^1)\right], \tag{11}$$

where $\mathbf{F}_i$ denotes the $i$-th column of the output matrix of DPN layer, $|\cdot|$ stands for the number of columns, and the $\circ$ operation represents the Hadamard product. Let $\mathbf{F}^P$ be a subset of the columns of $\tilde{\mathbf{F}}^P$. $\mathbf{F}^P$generates the basis of degree-$P$ polynomial, and it can be obtained by SVD to select the linear independent columns from $\tilde{\mathbf{F}}^P$. Finally, the output of all DPN layers creates the matrix of encoded features.

### 3.4. Weighted ensemble prediction of clinical scores

After the feature selection and feature encoding, the SVR method is used for the prediction of clinical scores. Similar to the support vector machine (SVM) [47,48], SVR tries for minimizing error via individualizing the hyperplane which maximizes the margin. Features of testing data are then mapped to the same space so that the corresponding classes of the examples can be recognized. Regression is a challenging task because it has infinite probabilities. However, SVR can solve this task proficiently by a given margin of tolerance. Since the features obtained from DPN are from different layers and possess their own characteristics, a simple concatenation may not fully explore their benefits. Therefore, it is more reasonable to deal with them separately. Here, SVR deals with matrix of encoded features which is the output of all DPN layers, then outputs $\mathbf{R} = [\mathbf{R}^1\ \mathbf{R}^2 \cdots \mathbf{R}^P\ ]$ as the final prediction results, respectively. For the purpose of ensembling the prediction results to attain the best prediction performance, we propose the following optimization problem:

$$min \frac{1}{2}\mathbf{RG} - \mathbf{Y}_2^2\ s.t\ 0 \le \boldsymbol{G}_i \le 1\ \&\ \sum_{i=1}^{P} \boldsymbol{G}_i = 1. \tag{12}$$

Supposing the number of the training samples is $N$, the dimension of $\mathbf{R}$ is $N \times P$, and $\mathbf{G}$ denotes the weight vector for predictions of $P$ layers. $\mathbf{Y}$ is ground truth vector and $\mathbf{G}_i$ is the $i$ th element of $\mathbf{G}$. After resolving the constrained linear least-squares problem in Eq. (12), we can obtain the optimal weights for ensemble predictions based on features from different layers.

## 4. Results

### 4.1. Experiment setup

We employ the proposed method in two scenarios to predict scores (ADAS-Cog, CDR-GLOB, CDR-SOB, and MMSE) at five time points (M06, M12, M18, M24, M36) with the dataset obtained from ADNI database. The Pearson correlation coefficient (R) and mean absolute error (MAE) between the ground truth and the predicted scores are applied to assess the overall prediction performance, respectively. The definitions of R and MAE at single time point are given as:

$$R = \frac{\mathrm{cov}\left(\mathbf{Y}, \hat{\mathbf{Y}}\right)}{\sigma(\mathbf{Y})\sigma\left(\hat{\mathbf{Y}}\right)},\ \mathrm{MAE} = \mathrm{mean}\left(\left|\mathbf{Y} - \hat{\mathbf{Y}}\right|\right), \tag{13}$$

where $\mathbf{Y}$ stands for the ground truth and $\hat{\mathbf{Y}}$ denotes the predicted scores, $\sigma(\cdot)$ is the standard deviation, and $\mathrm{cov}(\cdot)$ is the convariance.

### 4.2. Scenario 1: predictions with baseline dataset

In scenario 1, the proposed correntropy and temporally constrained group LASSO model is downgraded to the correntropy regularized LASSO model since the training dataset are only baseline

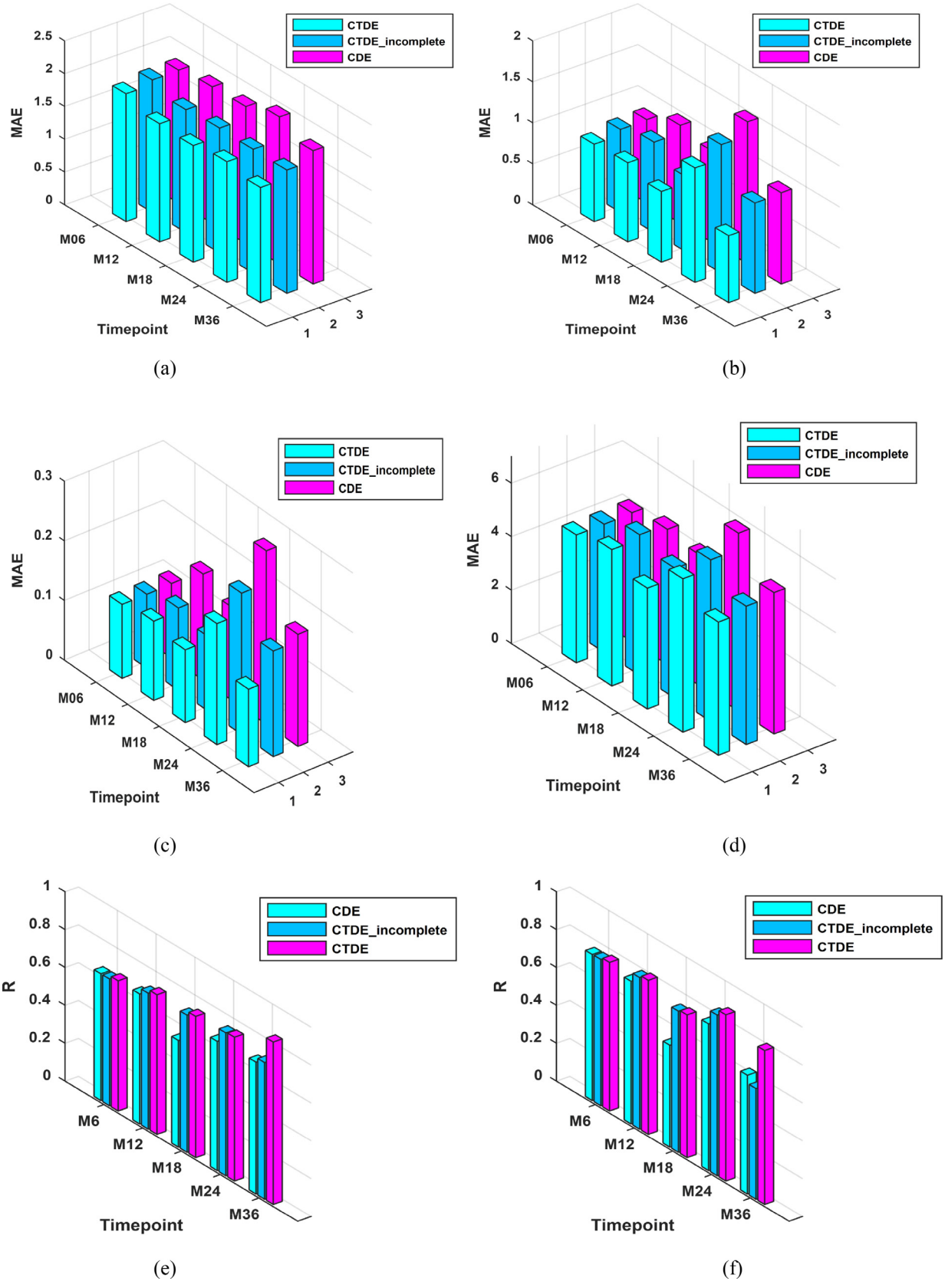
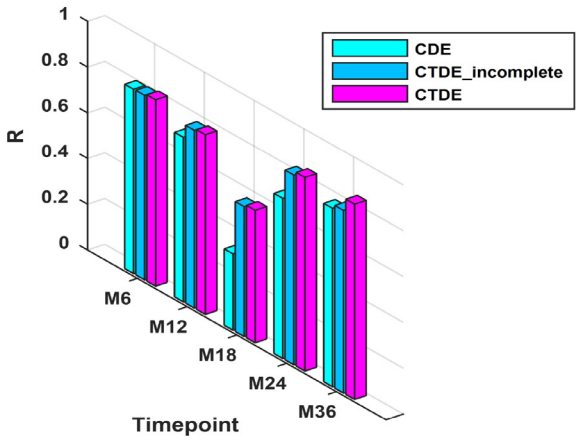
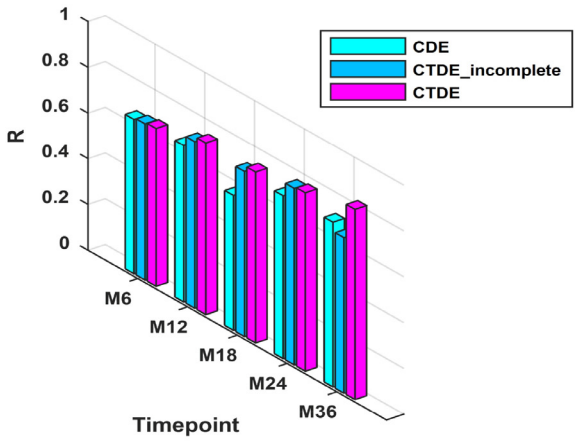



**Fig. 3.** Comparisons between CDE (in scenario 1), CTDE_incomplete (in scenario 2, without data filling), and CTDE in terms of the MAE of (a) MMSE; (b) CDR-SOB; (c) CDR-GLOB; (d) ADAS-Cog, and R of (e) MMSE; (f) CDR-SOB; (g) CDR-GLOB; (h) ADAS-Cog.




(g)




(h)

**Fig. 3.** Continued

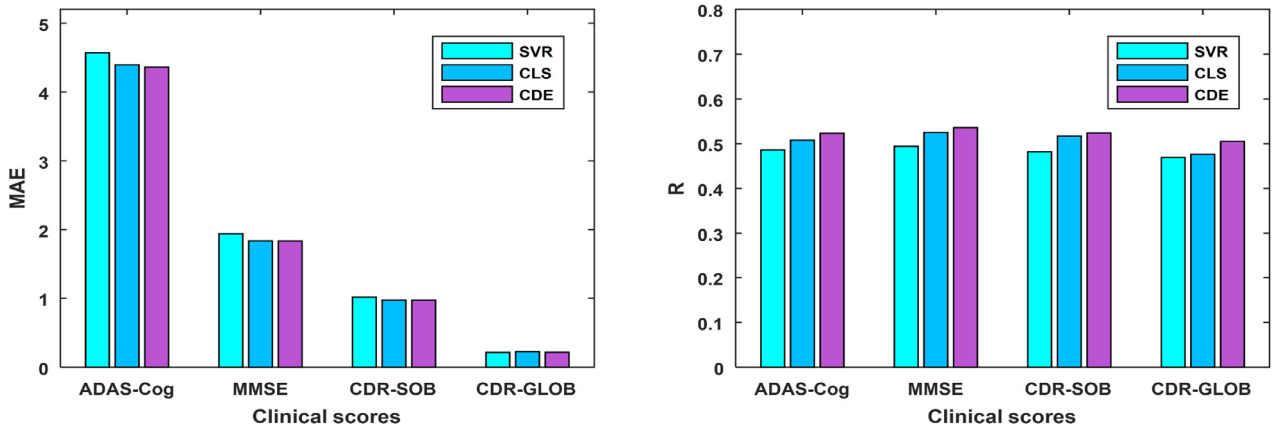



**Fig. 4.** Comparison of baseline performances for the competing methods.

data, and the proposed model in scenario 1 is denoted as CDE. For further evaluation of the composite CDE model, we compare its performance with that of partial model to investigate the role of each part and prove the overall utility of the proposed model. Specifically, the proposed CDE model is mainly divided into the three portions, such as SVR, CLS, DPNS. Specifically, CL is short for feature selection incorporating correntropy. CLS is the combination of the CL and SVR. DPNS is the combination of the DPN and SVR. The results are shown in Table 3. We observe that the composite model CDE gets the largest R and the smallest MAE, and proves advantages of the concatenation of feature selection and encoding.

### 4.3. Scenario 2: predictions with longitudinal dataset

In scenario 2, MRI data and clinical scores at previous time points are utilized to predict scores at the next time point. The proposed model consists of the feature selection incorporating correntropy and temporal constraints (CT), DPN based feature encoding (DE), and ensemble SVR, which is denoted as CTDE. Since CTDE is a composite model, we compare its performance with that of partial models to evaluate the role of each part and prove the overall utility of the proposed composite model. Here, CTS is the combination of CT and SVR while DPNS is the combination of DPN with SVR. The comparison results are summarized in Table 4. From these results, it is found that the proposed composite model CTDE attains the largest R and smallest MAE, which proves the advantages of the concatenation of feature selection and encoding. Meanwhile, the missing clinical scores of many subjects at different time points are filled and then be utilized for future scores prediction.

Furthermore, we conduct a comparative experiment to verify the effectiveness of scores completion. In the CTDE model, we conduct the scores completion experiment to imputate the missing clinical scores of many subjects at different time points and then these scores are utilized for future scores prediction. Simultaneously, the model without this imputating process is defined as CTDE_incomplete model and the experimental results are illustrated in Fig. 3. From the experimental results, it is found that CTDE model has an obvious improvement of scores prediction accuracy. The main reason is that, there are an increase in the number of training samples and thus a larger range of scores is covered.

### 4.4. Baseline predictions performance

In addition to predicting the longitudinal scores at multiple time points, we also use the cross-validation method to evaluate baseline prediction performance and indicate the versatility of the proposed model. The CDE model in scenario 1 is compared with previous methods such as SVR, CLS model. The experiments are conducted using a 10-fold cross-validation. Specifically, the baseline MRI data and corresponding clinical scores are randomly divided into ten subsets. Ten percent of the dataset is utilized for testing and the remaining is utilized for training. We repeat this process ten times to obtain the generalization results. The experimental results are obtained by averaging the repeated experimental results. The barplot of MAE and R are shown in Fig. 4. By con-

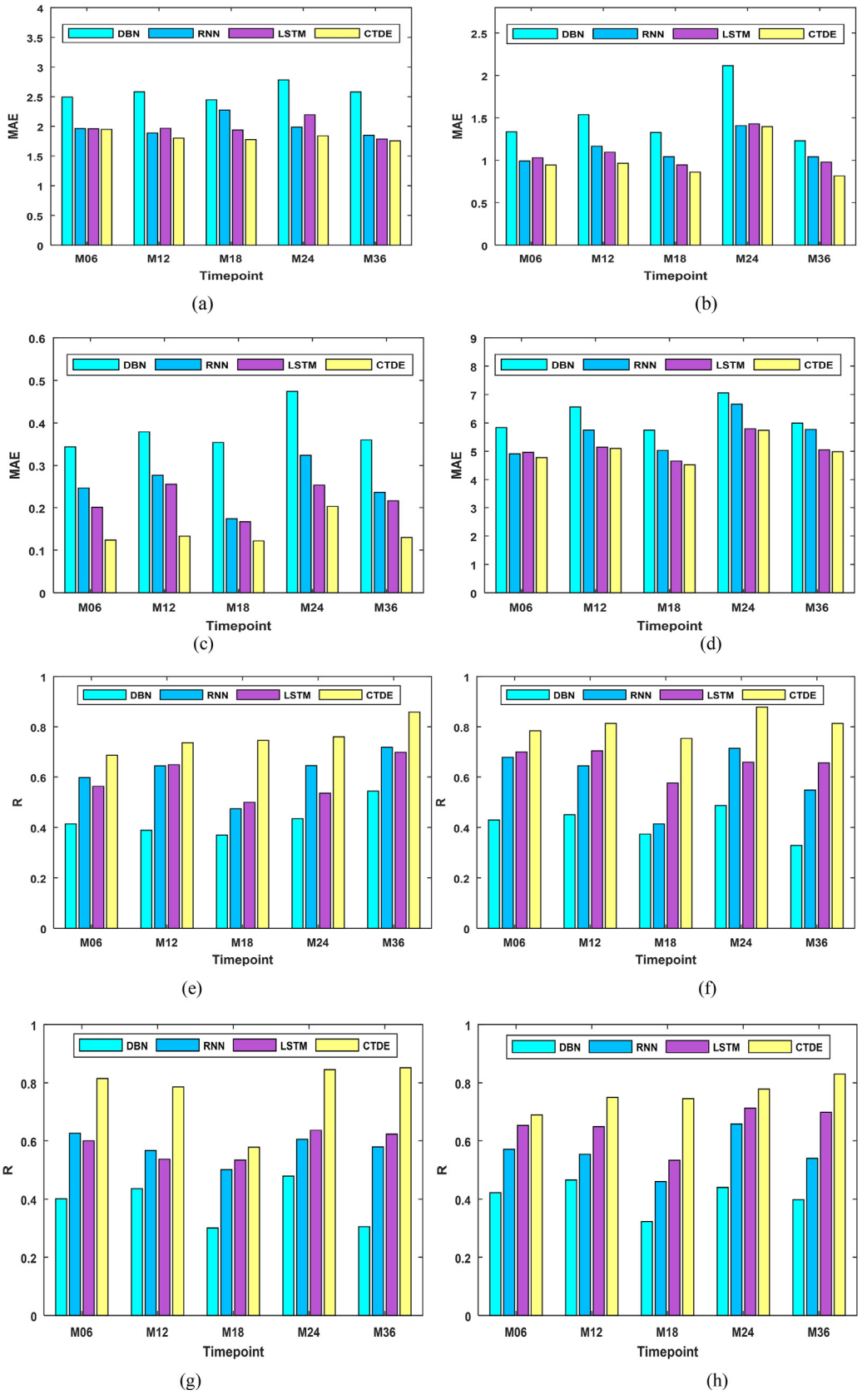



**Fig. 5.** Comparisons between the proposed model and state-of-the-art methods. The MAE of (a) MMSE; (b) CDR-SOB; (c) CDR-GLOB; (d) ADAS-Cog, and the R of (e) MMSE; (f) CDR-SOB; (g) CDR-GLOB; (h) ADAS-Cog.

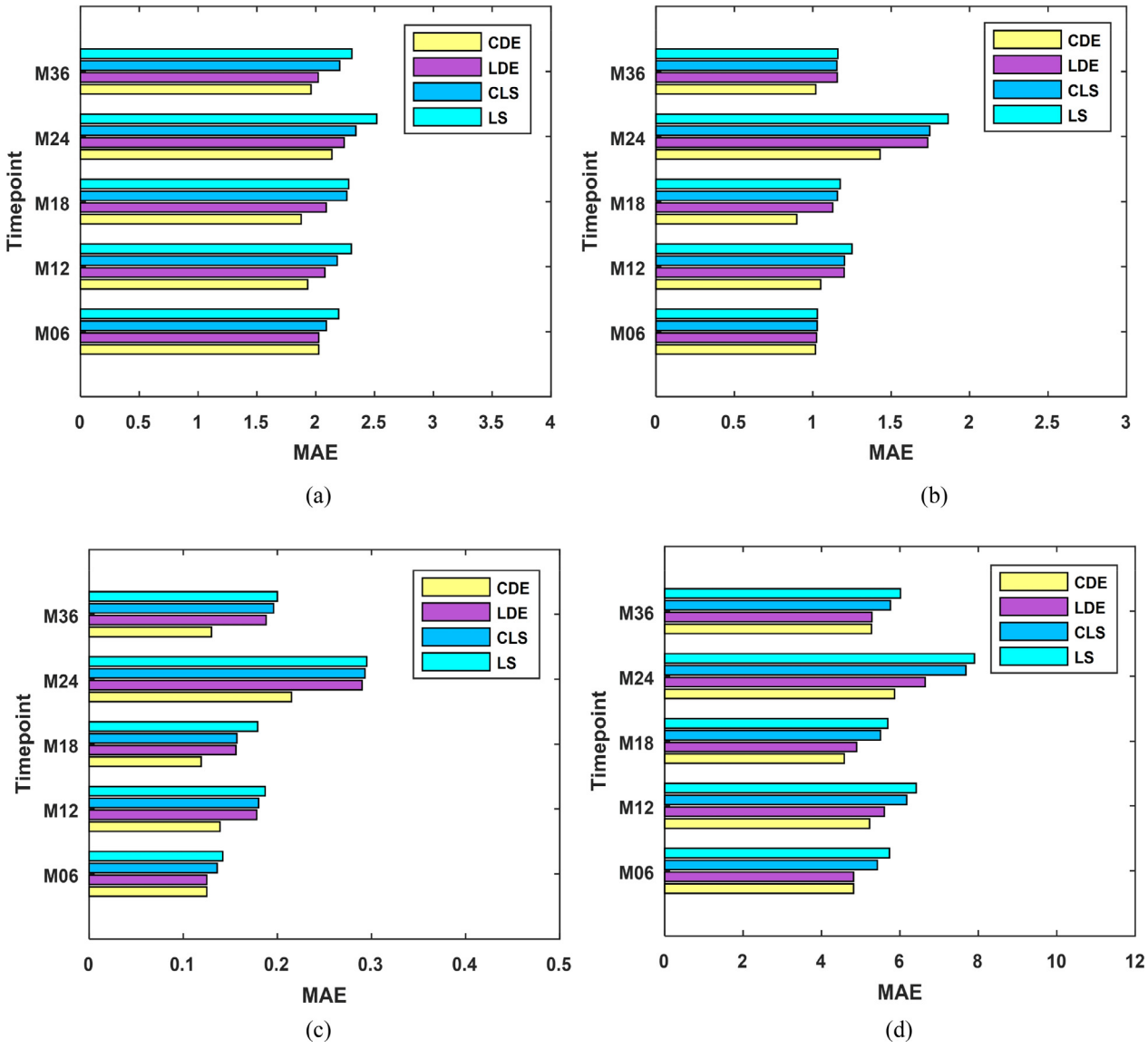



**Fig. 6.** MAE comparisons between LS (CLS without correntropy), CLS, LDE, CDE. (a) MMSE; (b) CDR-SOB; (c) CDR-GOLB; (d) ADAS-Cog.

**Table 3**
The MAE and R of various models for comparison in Scenario 1.

| Month | Method | MAE | | | | R | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MMSE | CDR-SOB | CDR-GLOB | ADAS-Cog | MMSE | CDR-SOB | CDR-GLOB | ADAS-Cog |
| M06 | SVR | 2.430 | 1.277 | 0.291 | 5.694 | 0.592 | 0.705 | 0.607 | 0.635 |
| | CLS | 2.056 | 1.114 | 0.212 | 5.022 | 0.656 | 0.760 | 0.749 | 0.654 |
| | DPNS | 2.117 | 1.192 | 0.248 | 5.316 | 0.628 | 0.721 | 0.742 | 0.636 |
| | **CDE** | **2.026** | **1.017** | **0.125** | **4.817** | **0.669** | **0.771** | **0.806** | **0.675** |
| M12 | SVR | 2.518 | 1.460 | 0.322 | 6.425 | 0.632 | 0.702 | 0.596 | 0.636 |
| | CLS | 2.111 | 1.267 | 0.246 | 5.757 | 0.673 | 0.706 | 0.679 | 0.652 |
| | DPNS | 2.172 | 1.362 | 0.279 | 6.276 | 0.630 | 0.673 | 0.702 | 0.633 |
| | **CDE** | **2.080** | **1.194** | **0.178** | **5.602** | **0.686** | **0.753** | **0.720** | **0.684** |
| M18 | SVR | 2.408 | 1.238 | 0.277 | 5.664 | 0.483 | 0.500 | 0.306 | 0.501 |
| | CLS | 2.134 | 1.096 | 0.218 | 5.161 | 0.513 | 0.523 | 0.322 | 0.516 |
| | DPNS | 2.103 | 1.120 | 0.227 | 5.302 | 0.524 | 0.499 | 0.239 | 0.516 |
| | **CDE** | **2.088** | **1.116** | **0.156** | **4.896** | **0.564** | **0.539** | **0.333** | **0.589** |
| M24 | SVR | 2.719 | 2.013 | 0.414 | 7.895 | 0.651 | 0.746 | 0.669 | 0.600 |
| | CLS | 2.235 | 1.746 | 0.332 | 6.956 | 0.668 | 0.664 | 0.699 | 0.592 |
| | DPNS | 2.363 | 1.909 | 0.358 | 7.727 | 0.653 | 0.724 | 0.696 | 0.710 |
| | **CDE** | **2.239** | **1.731** | **0.291** | **6.645** | **0.683** | **0.775** | **0.699** | **0.711** |
| M36 | SVR | 2.496 | 1.178 | 0.326 | 5.922 | 0.631 | 0.514 | 0.567 | 0.488 |
| | CLS | 2.039 | 1.122 | 0.253 | 5.318 | 0.640 | 0.541 | 0.636 | 0.511 |
| | DPNS | 2.185 | 1.182 | 0.290 | 5.829 | 0.628 | 0.527 | 0.551 | 0.594 |
| | **CDE** | **2.032** | **1.113** | **0.188** | **5.287** | **0.697** | **0.628** | **0.779** | **0.719** |

**Table 4**
The MAE and R of various models for comparison in Scenario 2.

| Month | Method | MAE | | | | R | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MMSE | CDR-SOB | CDR-GLOB | ADAS-Cog | MMSE | CDR-SOB | CDR-GLOB | ADAS-Cog |
| M06 | SVR | 2.096 | 1.124 | 0.232 | 5.060 | 0.592 | 0.706 | 0.715 | 0.620 |
| | CTS | 1.964 | 1.025 | 0.142 | 4.861 | 0.666 | 0.760 | 0.792 | 0.654 |
| | DPNS | 2.036 | 1.039 | 0.169 | 5.008 | 0.631 | 0.743 | 0.751 | 0.643 |
| | **CTDE** | **1.949** | **0.944** | **0.124** | **4.781** | **0.687** | **0.784** | **0.815** | **0.690** |
| M12 | SVR | 2.030 | 1.221 | 0.253 | 5.575 | 0.670 | 0.743 | 0.708 | 0.646 |
| | CTS | 1.890 | 1.009 | 0.149 | 5.356 | 0.734 | 0.799 | 0.769 | 0.712 |
| | DPNS | 1.907 | 1.112 | 0.170 | 5.462 | 0.722 | 0.785 | 0.767 | 0.726 |
| | **CTDE** | **1.801** | **0.965** | **0.133** | **5.099** | **0.737** | **0.813** | **0.786** | **0.750** |
| M18 | SVR | 2.046 | 1.029 | 0.213 | 4.937 | 0.577 | 0.615 | 0.576 | 0.538 |
| | CTS | 1.939 | 0.883 | 0.134 | 4.803 | 0.723 | 0.733 | 0.621 | 0.685 |
| | DPNS | 1.944 | 0.948 | 0.144 | 4.762 | 0.707 | 0.705 | 0.504 | 0.683 |
| | **CTDE** | **1.777** | **0.861** | **0.122** | **4.521** | **0.746** | **0.754** | **0.578** | **0.746** |
| M24 | SVR | 1.982 | 1.590 | 0.318 | 6.340 | 0.747 | 0.829 | 0.672 | 0.742 |
| | CTS | 1.847 | 1.416 | 0.204 | 6.010 | 0.772 | 0.862 | 0.783 | 0.713 |
| | DPNS | 2.085 | 1.521 | 0.257 | 6.860 | 0.750 | 0.825 | 0.780 | 0.732 |
| | **CTDE** | **1.840** | **1.396** | **0.203** | **5.736** | **0.760** | **0.878** | **0.845** | **0.779** |
| M36 | SVR | 1.886 | 1.086 | 0.268 | 5.003 | 0.735 | 0.658 | 0.760 | 0.658 |
| | CTS | 1.786 | 0.842 | 0.136 | 4.996 | 0.798 | 0.776 | 0.837 | 0.717 |
| | DPNS | 1.815 | 1.126 | 0.259 | 5.336 | 0.740 | 0.552 | 0.685 | 0.681 |
| | **CTDE** | **1.756** | **0.816** | **0.130** | **4.981** | **0.858** | **0.813** | **0.852** | **0.830** |

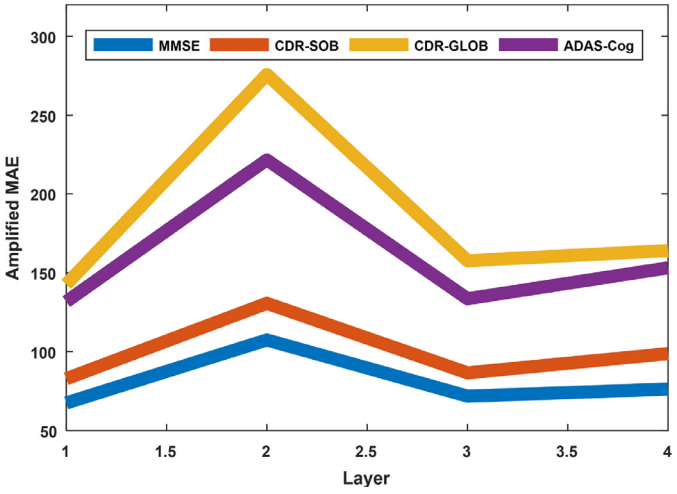



**Fig. 7.** Investigation of each layer.

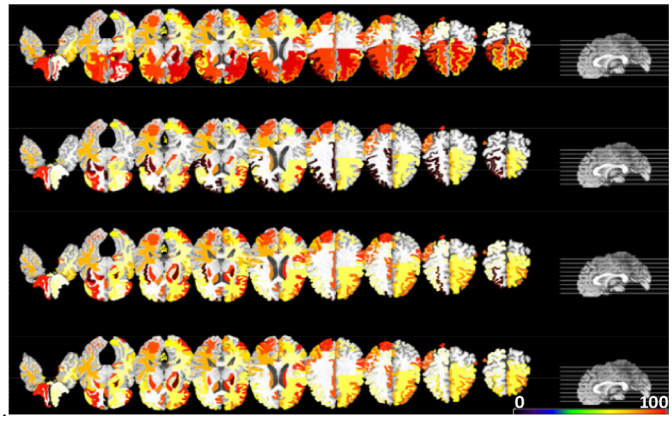



**Fig. 8.** The most discriminative ROIs for the prediction of four type of clinical scores. From top to bottom: MMSE, CDR-SOB, CDR-GLOB and ADAS-Cog, respectively. Here, different colors denote different ROIs.

trast, the proposed CDE model is effective and outperforms the other competing methods.

## 5. Discussion

### 5.1. Advantages over state-of-art methods

To reveal the effect of our proposed CTDE prediction network, we compare the performance with the state-of-the-art methods [46,49,50]. Specifically, Cui *et.al* put forward a classification framework based on Recurrent Neural Networks (RNN) for longitudinal analysis for AD diagnosis [49]. Hong et.al. put forward a novel framework based on Long short-term memory (LSTM) to analyze longitudinal dataset [50]. The longitudinal prediction experimental results at multiple points are shown in Fig. 5. Obviously, the proposed CTDE model achieves the best performance.

### 5.2. Advantages of correntropy

The related experiment is conducted to evaluate the effectiveness of the correntropy. In scenario 1, the group LASSO incorporating correntropy framework is downgraded to the group LASSO model (L) model when we remove the correntropy. Hence, the proposed CDE model also becomes the LDE model. We compare the performance of CDE model with LDE model, CLS model, and group LASSO model with SVR (LS). To assess the scores prediction performance, we compute the MAE values between the ground truth and the predicted scores. The results are plotted in Fig. 6. It is observed that CL model obtains more accurate predictions than the LS model. On the other hand, the CDE model obtains lower MAE values than LDE and other models. The results demonstrate that the using of correntropy method is quite effective.

### 5.3. Effect of DPN

Simulations are conducted in order to deepen the understanding of DPN. The output of each layer is passed to SVR and results of MAE are shown in Fig. 7. It is found that the features of the first layer carry more discriminative information than other layers, and thus the corresponding prediction is consistently more accurate than that of other layers. Hence, we choose to ensemble the results of concatenated layers rather than the results of each layer.

### 5.4. Most discriminative regions

In order to find the frequency distribution of discriminative brain regions, we perform four experiments to discover potential relevant biomarkers. It is meaningful for us to find useful information from the biomarkers to discover the underlying relationships for clinical scores prediction. We calculate the frequency of occurrences of each feature and sort them in an ascending way. Lastly, we find the most discriminative regions and map them in the template space. In the feature selection model, $\rho_0$ is the most influential parameter over the number of selected features. In order to predict the four type scores (ADAS-Cog, CDR-GLOB, CDR-SOB, and MMSE), $\rho_0$ is set to 20, 10, 1, and 20, respectively, to select the corresponding features. The feature maps for the prediction of different scores are shown in Fig. 8. It is sensible to observe certain consistence across different scores prediction as these scores are highly correlated. Specifically, the selected brain regions, such as amygdala, hippocampus areas, temporal pole, inferior temporal gyrus, and uncus are considered as AD biomarkers in previous literatures [51,52]. The corresponding brain regions are listed in Appendix A.

## 6. Conclusion

In this work, we introduce a deep and joint learning along with a two scenarios regression model for AD scores prediction. Different from the commonly used scores prediction methods which focus on the machine learning or deep learning based on baseline dataset, we utilize all the previous time points dataset to obtain the predicted scores at the next time point. Specifically, we integrate the feature selection with fused smoothness term, and employ the correntropy to construct the joint learning model. Meanwhile, the DPN algorithm is proposed to further improve feature representation, and then SVR is applied to predict four type of clinical scores. The extensive experiments on ADNI dataset demonstrate that our method presents superior performance than its counterparts. In the meantime, the proposed composite model outperforms its partial models in scores prediction accuracy. Furthermore, we generate relevant ROIs according to their weighing values to demonstrate the important brain regions for further studies.

Despite the good performance of the proposed model, there are also several limitations of our current study that should be explored in future study. First, as the experimental data, we just collect the longitudinal MRI data from ADNI. It will be more interesting to learn more knowledge from multiple modalities, such as functional MRI (fMRI), PET and diffusion tensor imaging (DTI). Using biomarkers of multiple modalities may reveal hidden information that may be overlooked by using a single modality. Second, the relevant clinical details (e.g., age, gender, education level) and other physiological factors of AD were not taken into account in the experiments. Considering a variety of clinical details and psychosocial factors, we can further identify the impact of related information and study the progression of Alzheimer's disease.

In the future, we will try to conduct more sophisticated feature selection and encoding methods, for further improving clinical scores prediction performance. Especially, we plan to investigate feature selection methods using both features similarity and adaptive sparseness learning [53]. Besides, we try to collect other imaging modalities and employ the fusion method [54] to construct an effective prediction model. Finally, we are testing the performance of the proposed method on other patient groups.

## Acknowledgments

This work was supported partly by National Natural Science Foundation of China (Nos. 61871274, 61872351 and 61801305), International Science and Technology Cooperation Projects of Guangdong (No. 2019A050510030), Key Laboratory of Medical Image Processing of Guangdong Province (No. K217300003), Guangdong Pearl River Talents Plan (2016ZT06S220), Shenzhen Peacock Plan (Nos. KQTD2016053112051497 and KQTD20150330161049 26), Shenzhen Key Basic Research Project (Nos. JCYJ20180507184647636 and JCYJ20170818094109846), and SZU Medical Young Scientists Program (No. 71201-000001).

## Appendix A: List of 93 ROIs

| Num | Name | Num | Name | Num | Name | Num | Name |
|---|---|---|---|---|---|---|---|
| 1 | medial front-orbital gyrus right | 25 | frontal lobe WM left | 47 | middle occipital gyrus right | 71 | parietal lobe WM right |
| 2 | middle frontal gyrus right | 26 | precuneus right | 48 | middle temporal gyrus left | 72 | insula left |
| 3 | lateral ventricle left | 27 | subthalamic nucleus left | 49 | lingual gyrus left | 73 | postcentral gyrus right |
| 4 | insula right | 28 | posterior limb of internal capsule | 50 | superior frontal gyrus left | 74 | lingual gyrus right |
| 5 | precentral gyrus right | | inc. cerebral peduncle left | 51 | nucleus accumbens left | 75 | medial frontal gyrus right |
| 6 | lateral front-orbital gyrus right | 29 | posterior limb of internal capsule | 52 | occipital lobe WM left | 76 | amygdala left |
| 7 | cingulate region right | | inc. cerebral peduncle right | 53 | postcentral gyrus left | 77 | medial occipitotemporal gyrus left |
| 8 | lateral ventricle right | 30 | hippocampal formation right | 54 | inferior frontal gyrus right | 78 | parahippocampal gyrus right |
| 9 | medial frontal gyrus left | 31 | inferior occipital gyrus left | 55 | precentral gyrus left | 79 | anterior limb of internal capsule right |
| 10 | superior frontal gyrus right | 32 | superior occipital gyrus right | 56 | temporal lobe WM left | 80 | middle temporal gyrus right |
| 11 | globus palladus right | 33 | caudate nucleus left | 57 | medial front-orbital gyrus left | 81 | occipital pole right |
| 12 | globus palladus left | 34 | supramarginal gyrus left | 58 | perirhinal cortex right | 82 | corpus callosum |
| 13 | putamen left | 35 | anterior limb of internal capsule left | 59 | superior parietal lobule right | 83 | amygdala right |
| 14 | inferior frontal gyrus left | 36 | occipital lobe WM right | 60 | lateral front-orbital gyrus left | 84 | inferior temporal gyrus right |
| 15 | putamen right | 37 | middle frontal gyrus left | 61 | perirhinal cortex left | 85 | superior temporal gyrus right |
| 16 | frontal lobe WM right | 38 | superior parietal lobule left | 62 | inferior temporal gyrus left | 86 | middle occipital gyrus left |
| 17 | parahippocampal gyrus left | 39 | caudate nucleus right | 63 | temporal pole left | 87 | angular gyrus left |
| 18 | angular gyrus right | 40 | cuneus left | 64 | entorhinal cortex left | 88 | medial occipitotemporal gyrus right |
| 19 | temporal pole right | 41 | precuneus left | 65 | inferior occipital gyrus right | 89 | cuneus right |
| 20 | subthalamic nucleus right | 42 | parietal lobe WM left | 66 | superior occipital gyrus left | 90 | lateral occipitotemporal gyrus left |
| 21 | nucleus accumbens right | 43 | temporal lobe WM right | 67 | lateral occipitotemporal gyrus right | 91 | thalamus right |
| 22 | uncus right | 44 | supramarginal gyrus right | 68 | entorhinal cortex right | 92 | occipital pole left |
| 23 | cingulate region left | 45 | superior temporal gyrus left | 69 | hippocampal formation left | 93 | fornix right |
| 24 | fornix left | 46 | uncus left | 70 | thalamus left | | |

## References

[1] A.S. Association, 2018 Alzheimer's disease facts and figures, Alzheimer's Dementia 14 (2018) 367–429.
[2] C. Patterson, World Alzheimer Report 2018—The State of the Art of Dementia Research: New Frontiers, Alzheimer's Disease International (ADI), London, UK, 2018.
[3] T. Zhou, M. Liu, K.-H. Thung, D. Shen, Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data, IEEE Trans. Med. Imaging 38 (2019) 2411–2422.
[4] T. Tong, K. Gray, Q. Gao, L. Chen, D. Rueckert, A. s. D. N. Initiative, Multi-modal classification of Alzheimer's disease using nonlinear graph fusion, Pattern Recognit. 63 (2017) 171–181.

[5] B. Shi, Y. Chen, P. Zhang, C.D. Smith, J. Liu, A. s. D. N. Initiative, Nonlinear feature transformation and deep fusion for Alzheimer's disease staging analysis, Pattern Recognit. 63 (2017) 487–498.
[6] A. Nordberg, J.O. Rinne, A. Kadir, B. Långström, The use of PET in Alzheimer disease, Nat. Rev. Neurol. 6 (2010) 78–87.
[7] L.K. McEvoy, C. Fennema-Notestine, J.C. Roddey, D.J. Hagler Jr, D. Holland, D.S. Karow, et al., Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment, Radiology 251 (2009) 195–205.
[8] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, et al., Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database, Neuroimage 56 (2011) 766–781.
[9] J. Peng, X. Zhu, Y. Wang, L. An, D. Shen, Structured sparsity regularized multiple kernel learning for Alzheimer's disease diagnosis, Pattern Recognit. 88 (2019) 370–382.
[10] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J.V. Hajnal, D. Rueckert, et al., Multiple instance learning for classification of dementia in brain MRI, Med. Image Anal. 18 (2014) 808–818.
[11] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, et al., Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease, IEEE Trans. Biomed. Eng. 62 (2014) 1132–1140.
[12] C.M. Stonnington, C. Chu, S. Klöppel, C.R. Jack Jr, J. Ashburner, R.S. Frackowiak, et al., Predicting clinical scores from magnetic resonance scans in Alzheimer's disease, Neuroimage 51 (2010) 1405–1413.
[13] J. Morris, Current vision and scoring rules The Clinical Dementia Rating (CDR), Neurology 43 (1993) 2412–2414.
[14] L. Huang, Y. Jin, Y. Gao, K.-H. Thung, D. Shen, A. s. D. N. Initiative, Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest, Neurobiol. Aging 46 (2016) 180–191.
[15] Y. Wang, Y. Fan, P. Bhatt, C. Davatzikos, High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables, Neuroimage 50 (2010) 1519–1535.
[16] S. Duchesne, A. Caroli, C. Geroldi, D.L. Collins, G.B. Frisoni, Relating one-year cognitive change in mild cognitive impairment to baseline MRI features, Neuroimage 47 (2009) 1363–1370.
[17] S.E. Hardy, H. Allore, S.A. Studenski, Missing data: a special challenge in aging research, J. Am. Geriatr. Soc. 57 (2009) 722–729.
[18] W. Zheng, X. Zhu, G. Wen, Y. Zhu, H. Yu, J. Gan, Unsupervised feature selection by self-paced learning regularization, Pattern Recognit. Lett. (2018).
[19] X. Zhu, H.-I. Suk, S.-W. Lee, D. Shen, Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification, IEEE Trans. Biomed. Eng. 63 (2015) 607–618.
[20] B. Jie, M. Liu, J. Liu, D. Zhang, D. Shen, Temporally constrained group sparse learning for longitudinal data analysis in Alzheimer's disease, IEEE Trans. Biomed. Eng. 64 (2017) 238–249.
[21] L. Zhang, L. Wang, W. Lin, Conjunctive patches subspace learning with side information for collaborative image retrieval, IEEE Trans. Image Process. 21 (2012) 3707–3720.
[22] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 2010–2023.
[23] Y. Shi, H.-I. Suk, Y. Gao, D. Shen, Joint coupled-feature representation and coupled boosting for AD diagnosis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2014, pp. 2721–2728.
[24] W. Liu, P.P. Pokharel, J.C. Príncipe, Correntropy: Properties and applications in non-Gaussian signal processing, IEEE Trans. Signal Process. 55 (2007) 5286–5298.
[25] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, Q. Sun, Deep learning for image-based cancer detection and diagnosis− a survey, Pattern Recognit. 83 (2018) 134–149.
[26] H.-I. Suk, S.-W. Lee, D. Shen, A. s. D. N. Initiative, Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, Neuroimage 101 (2014) 569–582.
[27] A. Ortiz, J. Munilla, J.M. Gorriz, J. Ramirez, Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease, Int. J. Neural Syst. 26 (2016) 1650025.
[28] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, T. Wang, Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset, Neurocomputing 194 (2016) 87–94.
[29] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, et al., Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans, Sci. Rep. 6 (2016) 24454.
[30] I. Santamaría, P.P. Pokharel, J.C. Principe, Generalized correlation function: definition, properties, and application to blind equalization, IEEE Trans. Signal Process. 54 (2006) 2187–2197.
[31] E. Yildizer, A.M. Balci, M. Hassan, R. Alhajj, Efficient content-based image retrieval using multiple support vector machines ensemble, Expert Syst. Appl. 39 (2012) 2385–2396.
[32] K. Ito, B. Corrigan, Q. Zhao, J. French, R. Miller, H. Soares, et al., Disease progression model for cognitive deterioration from Alzheimer's Disease Neuroimaging Initiative database, Alzheimer's Dementia 7 (2011) 151–160.
[33] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A. s. D. N. Initiative, Multimodal classification of Alzheimer's disease and mild cognitive impairment, Neuroimage 55 (2011) 856–867.
[34] J.G. Sled, A.P. Zijdenbos, A.C. Evans, A nonparametric method for automatic correction of intensity nonuniformity in MRI data, IEEE Trans. Med. Imaging 17 (1998) 87–97.
[35] Y. Zhang, M. Brady, S. Smith, Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm, IEEE Trans. Med. Imaging 20 (2001) 45–57.
[36] D. Shen, C. Davatzikos, HAMMER: hierarchical attribute matching mechanism for elastic registration, IEEE Trans. Med. Imaging 21 (2002) 1421–1439.
[37] P. Cao, X. Liu, J. Yang, D. Zhao, M. Huang, O. Zaiane, $l_{2,1}$-$l_1$ regularized nonlinear multi-task representation learning based cognitive performance prediction of Alzheimer's disease, Pattern Recognit. 79 (2018) 195–215.
[38] J. Liu, J. Ye, Efficient Euclidean projections in linear time, in: Proceedings of the 26th Annual International Conference on Machine Learning (ICML), 2009, pp. 657–664.
[39] J. Yan, T. Li, H. Wang, H. Huang, J. Wan, K. Nho, et al., Cortical surface biomarkers for predicting cognitive outcomes using group $l_{2,1}$ norm, Neurobiol. Aging 36 (2015) S185–S193.
[40] X. Zhu, X. Li, S. Zhang, C. Ju, X. Wu, Robust joint graph sparse coding for unsupervised spectral feature selection, IEEE Trans. Neural Netw. Learn. Syst. 28 (2017) 1263–1275.
[41] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, J. R. Stat. Soc. 67 (2005) 91–108.
[42] Z. Zhao, L. Zhang, X. He, W. Ng, Expert finding for question answering via graph regularized matrix completion, IEEE Trans. Knowl. Data Eng. 27 (2014) 993–1004.
[43] Y. Nesterov, Gradient methods for minimizing composite functions, Math. Program. 140 (2013) 125–161.
[44] Z. Zha, X. Zhang, Q. Wang, Y. Bai, Y. Chen, L. Tang, et al., Group sparsity residual constraint for image denoising with external nonlocal self-similarity prior, Neurocomputing 275 (2018) 2294–2306.
[45] X. Chen, W. Pan, J.T. Kwok, J.G. Carbonell, Accelerated gradient method for multi-task sparse learning problem, in: 2009 Ninth IEEE International Conference on Data Mining (ICDM), 2009, pp. 746–751.
[46] J. Shi, X. Zheng, Y. Li, Q. Zhang, S. Ying, Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease, IEEE J. Biomed. Health Inf. 22 (2017) 173–183.
[47] B. Gaonkar, R.T. Shinohara, C. Davatzikos, A.D.N. Initiative, Interpreting support vector machine models for multivariate group wise analysis in neuroimaging, Med. Image Anal. 24 (2015) 190–204.
[48] J. Shi, J. Wu, Y. Li, Q. Zhang, S. Ying, Histopathological image classification with color pattern random binary hashing-based pcanet and matrix-form classifier, IEEE J. Biomed. Health Inf. 21 (2017) 1327–1337.
[49] R. Cui, M. Liu, A. s. D. N. Initiative, RNN-based longitudinal analysis for diagnosis of Alzheimer's disease, Comput. Med. Imaging Graph. 73 (2019) 1–10.
[50] X. Hong, R. Lin, C. Yang, N. Zeng, C. Cai, J. Gou, et al., Predicting Alzheimer's disease using LSTM, IEEE Access 7 (2019) 80893–80901.
[51] C. Misra, Y. Fan, C. Davatzikos, Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI, Neuroimage 44 (2009) 1415–1422.
[52] A. Convit, J. De Asis, M. De Leon, C. Tarshish, S. De Santi, H. Rusinek, Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease, Neurobiol. Aging 21 (2000) 19–26.
[53] H. Lei, Z. Huang, F. Zhou, A. Elazab, E.-L. Tan, H. Li, et al., Parkinson's disease diagnosis via joint learning from multiple modalities and relations, IEEE J. Biomed. Health Inf. 23 (2018) 1437–1449.
[54] F. Liu, C.-Y. Wee, H. Chen, D. Shen, Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification, Neuroimage 84 (2014) 466–475.

**Baiying Lei** received her M. Eng degree in electronics science and technology from Zhejiang University, China in 2007, and Ph.D degree from Nanyang Technological University (NTU), Singapore in 2013. She is currently with School of Biomedical Engineering, Health Science Center, Shenzhen University, China. Her current research interests include medical image analysis, machine learning, and pattern recognition. Dr. Lei has coauthored more than 130 scientific articles, e.g., IEEE TCYB, IEEE TMI, IEEE TBME, IEEE JBHI. Pattern Recognition and Information Sciences. She is an IEEE senior member and serves as the editorial board member of Scientific Reports, Frontiers in Neuroinformatics, Frontiers in Aging Neuroscience, and Academic Editor of Plos One.

**Mengya Yang** received the B.Sc degree in Biomedical Engineering from Northeast University, Qin Huangdao, China, in 2016. She is currently working toward her M.Eng degree in the School of Biomedical Engineering, Health Science Center, Shenzhen University, Guangdong, China. His research interests include medical image analysis and deep learning.

**Peng Yang** received the B.Sc degree in Biomedical Engineering from Northeast University, Qin Huangdao, China, in 2016. He is currently working toward his Ph.D degree in the School of Biomedical Engineering, Health Science Center, Shenzhen University, Guangdong, China. His research interests include medical image analysis and machine learning.

**Feng Zhou** received the Bachelor's degree from Ningbo University, Ningbo, China, in 2005, and the M.S. degree from Zhejiang University, Hangzhou, China, in 2007,

both in electronic engineering, and the Ph.D degrees in human factors engineering from Nanyang Technological University, Singapore, in 2011, and in mechanical engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2014. He is currently an Assistant Professor at the Department of Industrial and Manufacturing Systems Engineering, The University of Michigan, Dearborn, MI, USA. His current research interests include engineering design, human factors engineering, human–computer interaction, and user research.

**Wen Hou** received her Ph.D degree in Signal Processing from Swinburne University in 2015. She was with College of Information Engineering, Shenzhen University, Shenzhen, China. Her research interests include medical image analysis, signal processing.

**Wenbin Zou** received the M.E. degree in software engineering with a specialization in multimedia technology from Peking University, China, in 2010, and the Ph.D degree from the National Institute of Applied Sciences, Rennes, France, in 2014. From 2014 to 2015, he was a Researcher with the UMR Laboratoire d'informatique Gaspard-Monge, CNRS, and the École des Ponts ParisTech, France. Since 2015, he has been with the Faculty of the College of Information Engineering, Shenzhen University, China. His current research interests include saliency detection, object segmentation, and semantic segmentation.

**Xia Li** received the B.S. and M.S. degrees in electronic engineering and signal and information processing from Xidian University in 1989 and 1992, respectively, and the Ph.D degree from the Department of Information Engineering Philosophy, The Chinese University of Hong Kong, in 1997. She was the former Dean of the College of Information Engineering, Shenzhen University, and the Director of the Shenzhen Key Laboratory of Advanced Communication and Information Processing. She is currently an Associate Vice-President with The Chinese University of Hong Kong at Shenzhen. Her research interests include intelligent computing and its applications, image processing, and pattern recognition.

**Tianfu Wang** received his Ph.D degree in biomedical engineering from Sichuan University in 1997. He is currently a Professor in School of Biomedical Engineering, and the Associate Chair of Health Science Center, Shenzhen University, China. His research interests include ultrasound image analysis, medical image processing, pattern recognition and medical imaging.

**Xiaohua Xiao** received MD. degree from Zhongshan Medical University, in 2000. He has been engaged in clinical work of Neurology for more than 20 years and is good at diagnosis and treatment of epilepsy, vertigo, and cerebrovascular diseases. He is proficient in neuroelectrophysiological examination and is good at diagnosis and treatment of epilepsy, motor neuron disease, and other neuromuscular diseases by using neuroelectrophysiological examination technology. He has deep experience in diagnosis and treatment of benign paroxysmal positional vertigo and is good at using manual reduction to treat benign paroxysmal positional vertigo.

**Shuqiang Wang** received the Ph.D. degree from the City University of Hong Kong, Hong Kong, in 2012. From 2012 to 2013, he was a Research Scientist at Huawei Technologies Noah's Ark Lab. From 2013 to 2014, he held a Post-doctoral fellowship in the University of Hong Kong. He is currently an Associate Professor with Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. His current research interests include machine learning, medical image computing and optimization theory.