Computers in Biology and Medicine 138 (2021) 104935

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Rethinking modeling Alzheimer's disease progression from a multi-task learning perspective with deep recurrent neural network

Wei Liang [a], Kai Zhang [a], Peng Cao [a,b,*], Xiaoli Liu [c], Jinzhu Yang [a,b], Osmar Zaiane [d]

[a] Computer Science and Engineering, Northeastern University, Shenyang, China
[b] Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang, China
[c] Alibaba A.I. Labs, Hangzhou, China
[d] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

Keywords:
Alzheimer's disease
Multitask learning
Time_LSTM
Disease progression
Missing value

ABSTRACT

Alzheimer's disease (AD) is a severe neurodegenerative disorder that usually starts slowly and progressively worsens. Predicting the progression of Alzheimer's disease with longitudinal analysis on the time series data has recently received increasing attention. However, training an accurate progression model for brain network faces two major challenges: missing features, and the small sample size during the follow-up study. According to our analysis on the AD progression task, we thoroughly analyze the correlation among the multiple predictive tasks of AD progression at multiple time points. Thus, we propose a multi-task learning framework that can adaptively impute missing values and predict future progression over time from a subject's historical measurements. Progression is measured in terms of MRI volumetric measurements, trajectories of a cognitive score and clinical status. To this end, we propose a new perspective of predicting the AD progression with a multi-task learning paradigm. In our multi-task learning paradigm, we hypothesize that the inherent correlations exist among: ***(i)***. the prediction tasks of clinical diagnosis, cognition and ventricular volume at each time point; ***(ii)***. the tasks of imputation and prediction; and ***(iii)***. the prediction tasks at multiple future time points. According to our findings of the task correlation, we develop an end-to-end deep multi-task learning method to jointly improve the performance of assigning missing value and prediction. We design a balanced multi-task dynamic weight optimization. With in-depth analysis and empirical evidence on Alzheimer's Disease Neuroimaging Initiative (ADNI), we show the benefits and flexibility of the proposed multi-task learning model, especially for the prediction at the M60 time point. The proposed approach achieves 5.6%, 5.7%, 4.0% and 11.8% improvement with respect to mAUC, BCA and MAE (ADAS-Cog13 and Ventricles), respectively.

1. Introduction

Alzheimer's disease (AD) is the most common neurological disorder disease that begins with severe memory impairment and a continuous decline in conversation, orientation and cognitive abilities [1]. An estimated 5.5 million people aged 65 and older are living with AD in the USA. The loss of cognitive abilities is irreversible for AD, hence the therapeutic intervention is critical at pre-symptomatic stages. Neurodegeneration of AD begins years before the onset of the disease and understanding the disease progression is important for patients in designing a treatment plan, predicting prognosis, and evaluate the effects of treatments [2–4]. This is a disease progression modeling (DPM) problem. Developing data-driven methods for DPM on the longitudinal data is necessary to exploit the relationship between neuroimaging and the development of the disease progress for better diagnosis, monitoring and prognosis [5]. In our work, the aim is to build a predictive model for the Alzheimer's Disease progression with the relationship between the multiple predictive tasks (cognitive scores, clinical diagnosis, and ventricular volume) and historical Magnetic resonance imaging (MRI) markers over time.

Recently, predictive models have been successfully proposed for predicting the future cognition of ventricular volume, ADAS-Cog13, and clinical diagnosis of the participants. Although promising progress has been made by researchers in studies related to disease progression modeling, the challenging issues still remain.

* Corresponding author. Computer Science and Engineering, Northeastern University, Shenyang, China.
E-mail address: caopeng@cse.neu.edu.cn (P. Cao).

https://doi.org/10.1016/j.compbiomed.2021.104935
Received 6 September 2021; Received in revised form 7 October 2021; Accepted 8 October 2021
Available online 13 October 2021

1. The missing data is prevalent in most medical applications, which prevents the analysis of multivariate time series data. The longitudinal cohorts often include missing biomarker values due to patients dropping out of the studies or unsuccessful measurements. Most previous works on modeling AD progression often ignore missing data issues due to the assumption of traditional statistical analysis models that the progression data is feature-complete. However, ignoring the missing data results in the loss of information for the machine learning models. Existing imputation approaches try to deal with missing values by a statistical imputation scheme. However, these methods are not capable of modeling both the feature relation and the temporal relation. Moreover, the tasks of imputation and prediction are studied independently. However, they can benefit each other. Recent studies have shown that the model filling methods achieve a better performance than the previous interpolation methods [6,7], such as forward filling and linear interpolation. In Ref. [6], MinimalRNN model is proposed to impute the missing data both in the training and testing stages with a model filling scheme for AD progression. The imputed missing values are used to extrapolate features as input, which is then used to forecast the MRI biomarker as well as cognitive test scores and clinical status simultaneously. MinimalRNN is trained to predict the observations in future months given the observations for the previous 3 months. The errors between the predicted outputs and the ground truth outputs are used to update the model parameters. The major limitation is that it generates many inaccurate values at the unobserved time points since it takes one month as a time step, which transgresses the fact that the visit interval is over 6 months. The inaccurate value results in decreasing performance of the following imputation and prediction. Moreover, the current model filling methods cannot handle the imputation for the unobserved value at the farther time points such as 60 months later, due to the challenge of the dynamic temporal relation.
2. Current works focus on the prediction of the target at multiple time points with a single model. To improve the performance, we employ a joint analysis strategy for multiple tasks at multiple time points, which is especially effective for the task with a limited number of patients at some certain times. The difficulty of prediction of farther time points and nearer time points is different. The prediction task of the nearest time points, e.g. 6 months later, is relatively easier, whilst the relationship between the MRI features and the farther time points prediction, for example, 5 years later, is not obvious, resulting in a challenging task. A single model is not capable of achieving an accurate prediction performance at different time points. Therefore, it is important to build multiple longitudinal prediction models, each of which focuses on the prediction at different time points. Moreover, we can exploit the inherent correlation among the predictive tasks to improve the progression modeling performance.

To cope with the above challenges, we propose a unified framework that can adaptively impute missing values and predict future MRI volumetric measurements, trajectories of a cognitive score and clinical status over time in the longitudinal data. Jointly modeling the inherent relations within both the multivariate and the temporal tasks facilitates the estimation of missing values and AD progression. To collaboratively construct a better model for each task, several multi-task learning model have been proposed by incorporating the potentially inherent correlations among multiple clinical cognitive measures [8–14]. Multi-task learning benefits from its ability to learn a shared representation across related tasks and to improve the generalization performance of each task. Identifying how the tasks are related and how to build predictive models to capture such relatedness are critical issues in multi-task learning.

Accordingly, we thoroughly analyze the correlation among the multiple predictive tasks of AD progression at multiple time points. According to our analysis on the AD progression task, we find that: 1) The multiple predictive tasks for the multiple future time points are highly relevant, given the same subject's historical features. 2) there exists temporal smoothness among the multiple tasks.

Thus, we propose a new perspective on predicting the AD progression with a multi-task learning paradigm. In our multi-task learning paradigm, we hypothesize that the inherent correlations exist among: ***(1)*** the prediction tasks of clinical diagnosis, cognition and ventricular volume at each time point; ***(2)*** the tasks of imputation and prediction; and ***(3)*** the prediction tasks at multiple future time points. With the assumption, we design an objective function such that the network parameters of all the tasks can be trained jointly in an end-to-end manner, which allows simultaneous joint imputation and prediction tasks at multiple time points. All the relations in the longitudinal data could be informative to the AD progression. More specifically, we develop an end-to-end deep multi-task learning method for AD progression, called MTL-ATM, according to our findings of the task correlation with respect to the model parameters and the temporal smoothness. Furthermore, the intervals of multiple time points are different and are important factors for temporal patterns. The previous works with RNN or LSTM fail to properly model the varied temporal relation when modeling the time relation. To exploit time intervals, we choose Time_LSTM as our module to implicitly capture the interval information. Finally, the appropriate loss weights are important for the final prediction performance. In order to achieve optimal weights and reduce the occurrences of negative transfer, we design a balanced multi-task dynamic weight optimization. Our multi-task model outperforms most state-of-the-art baselines, especially for the prediction at farther time points. Extensive experiments verify that our method not only achieves excellent performance in missing data imputation, but also obtains competitive results in the progression prediction performance.

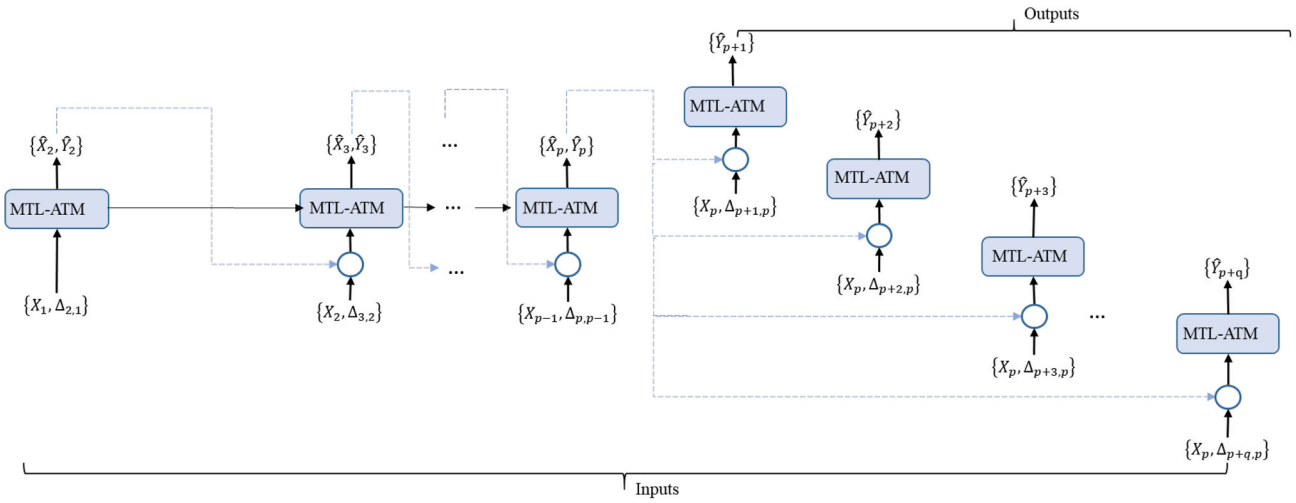
The rest of the paper is organized as follows. The related works for predicting Alzheimer's disease progression are discussed in Section 2. A thorough problem formulation and task correlation analysis are presented in Section 3. In Section 4, we provide the overall network architectures of the proposed MTL-ATS. In Section 5, we conduct comprehensive experiments to demonstrate the advantage of our method for the progression prediction of Alzheimer's disease. The conclusion is drawn in Section 6.

2. Related work

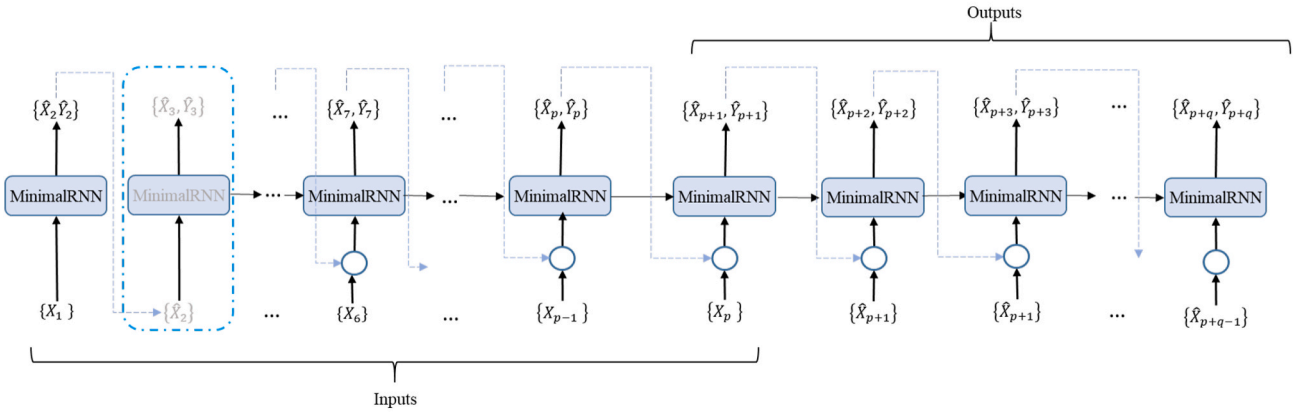
Recently, using machine learning approaches to improve the early AD diagnosis and prognosis has attracted increasing attention [12,15–20]. Due to the advances of deep learning techniques, many methods are proposed to model the AD diagnosis as a classification and regression formulation [14,19,21–25].

On the one hand, many of the existing studies focus on cross-section data analysis for AD diagnosis [14,26–29]. Among them, Dolph [27] utilizes features extracted from structural MRI and proposes a multi-class deep learning model. Liu [14] first identifies the discriminative anatomical landmarks from MR images in a data-driven manner and then extracts multiple image patches around these detected landmarks. Cheng [28] proposes to make classification for AD diagnosis by constructing multiple deep convolutional neural networks with various features from local brain images. Although these research demonstrates promising classification performances, all of them do not consider the progressive deterioration of AD, which is an important characteristic of AD.

On the other hand, in order to solve the above problems, longitudinal data is utilized by some researchers [6,7,22,25,30]. Sappagh [25] predicts multiple variables jointly based on stacked convolutional neural network and bidirectional long short-term networks. Lee [22] proposes an integrative framework that combines both cross-sectional neuroimaging biomarkers at baseline and longitudinal cognitive performance biomarker obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI). Nguyen [6] explores three different strategies to solve the issue of missing data and impute the missing value for subsequent modeling.



(a) Problem formulation of the proposed method.



(b) Problem formulation of MinimalRNN.

Fig. 1. Illustration of the problem formulation for imputing missing values and the progression in our method and MinimalRNN.

Jung [7] takes into account inherent temporal and multivariate relations in time series data, and estimates missing values in a systematic way.

Based on the above work, the technologies of early diagnosis of AD have gradually attracted attention and have been studied to predict the disease severity of AD. However, few works focus on multi-task joint learning of multiple categorical and continuous variables on multiple time points, which could perform better than predicting each time point separately. Thus, this paper constructs a multi-task learning model by taking into account the relationship between prediction tasks on different time points. In the next section, we detail the methodology of our proposed approach for the disease progression modeling.

3. Problem formulation and task correlation analysis

We formulate progression prediction for multiple time points as a multi-task learning problem. We explore two strategies to control the correlation among the multiple tasks, involving both parameter sharing and temporal smoothness constraints. The formulation consists of q prediction tasks (Fig. 1). In this section, we first formulate the problems, and then analyze the task correlation and introduce the proposed MTL-ATM approach.

3.1. Problem formulation

We start by giving the formulation of the problem of AD progression prediction. The number of the training samples is denoted as N, and each subject has its corresponding data at T different time points, represented as $X_t = [X_{t,1}, X_{t,2}, \ldots, X_{t,N}] \in \mathbb{R}^{N\times D}$, where $X_{t,n} \in \mathbb{R}^{1\times D}$ is a D-dimensional vector at time point t. $Y_t = [Y_{t,1}, Y_{t,2}, \ldots, Y_{t,N}] \in \mathbb{R}^{N\times 1}$ is the corresponding target at time point t for all subjects, where $Y_{t,n} = [y_{t,n}^{Diag}, y_{t,n}^{Ven}, y_{t,n}^{Cog}]$ represents clinical diagnosis, Ventricles and ADAS-Cog13 of the n-th subject at time point t, respectively. To enable the model to know which input features or labels are observed, we utilize mask vectors M_x and $M_y^{Cog}, M_y^{Diag}, M_y^{Ven}$ to indicate the input mask vector and the label mask vectors, respectively. Given observations and time intervals of the previous p time points.

$\{(X_1, \Delta_{2,1}), (X_2, \Delta_{3,2}), \ldots, (X_p, \Delta_{p+1,p})\}$, the proposed MTL-ATM approach aims to learn a non-linear mapping for predicting the future status:

$$f_{pred} : \{(X_1, \Delta_{2,1}), (X_2, \Delta_{3,2}), \ldots, (X_p, \Delta_{p+1,p})\} \rightarrow \{\widehat{Y}_{p+1}, \widehat{Y}_{p+2}, \ldots, \widehat{Y}_{p+q}\} \quad (1)$$

where each prediction of $\widehat{Y}_{p+k}, k \in [1, q]$ indicates the corresponding task at the $(p + k)$-th time point. In our work, the tasks are indicated as $T24$, $T36$, $T48$ and $T60$. Fig. 1 shows the problem formulation of the proposed

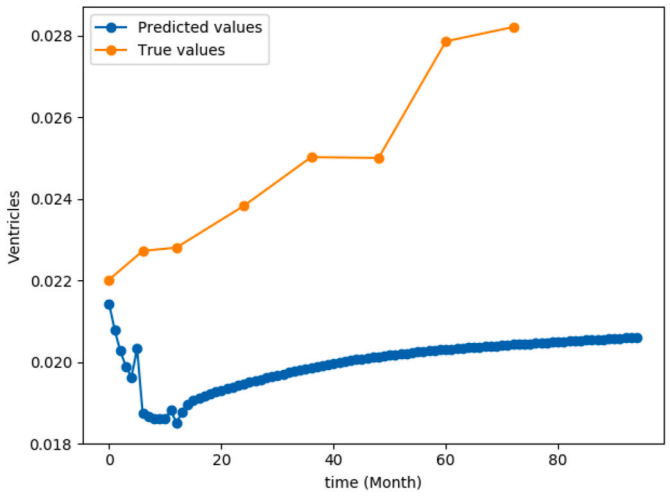


Fig. 2. Predicted Ventricles of MinimalRNN during 94 months. The blue circles are the predicted values of MinimalRNN and the orange circles are the true values at multiple time points. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

method and MinimalRNN method [6], respectively.

During the training stage, the previous temporal context information up to $t-1$ is encoded in the hidden state h_{t-1} of the MTL-ATM module. When $t \leqslant p$, the information is used to impute the missing observations in X_t and predict $\widehat{Y}_t$ for loss calculation simultaneously. The first t tasks are formulated as Eq. (2):

$$\widehat{X}_t, \widehat{Y}_t = f_{pred\&imp}((X_1, \Delta_{2,1}), (X_2, \Delta_{3,2}), \ldots, (X_{t-1}, \Delta_{t,t-1}), \widehat{X}_{t-1}) \quad (1 < t \leqslant p) \tag{2}$$

where $\widehat{X}_t$ is the imputed value of the corresponding variables in X_t. The complete value $\widetilde{X}_t$ after imputation is formulated as:

$$\widetilde{X}_t = X_t \odot M_x + \widehat{X}_t \odot (1 - M_x) \quad (1 < t \leqslant p) \tag{3}$$

When $t > p$, the previous temporal context information up to p is used to predict Y_t.

$$\widehat{Y}_t = f_{pred}((\widetilde{X}_1, \Delta_{2,1}), (\widetilde{X}_2, \Delta_{3,2}), \ldots, (\widetilde{X}_p, \Delta_{p+1,p})) \quad (t > p) \tag{4}$$

It is important to note that for each task k, $k \in [1, q]$, the training loss is calculated by $\{(\widetilde{X}_2, \widehat{X}_2), \ldots, (\widetilde{X}_p, \widehat{X}_p), (\widetilde{X}_{p+k}, \widehat{X}_{p+k})\}$ and $\{(\widetilde{Y}_2, \widehat{Y}_2), \ldots, (\widetilde{Y}_p, \widehat{Y}_p), (\widetilde{Y}_{p+k}, \widehat{Y}_{p+k})\}$. However, at the test stage, performance is only evaluated by $(\widetilde{Y}_{p+k}, \widehat{Y}_{p+k})$.

The main differences between our proposed approach and MinimalRNN are as follows:

1 For MinimalRNN, prediction at each time point is conducted based on all the historical observed values and the predicted values(if unobserved or missing). As shown in Fig. 3(b), for the prediction task at time $p+3$, MinimalRNN has to impute all the values at each month before $p+3$ including the unobserved time point. In contrast, our proposed approach considers the variable time interval between two observed time points, without imputing the values at the completely unobserved time points.
2 For prediction tasks at time $p+1$, $p+2$, …, $p+q$, MinimalRNN predicts each time point in sequence, which means the predicted values at time $p+2$ are used to predict Y_{P+3}. In this way, the prediction at time $p+3$ inevitably produces a poorer result if the prediction at time $p+2$ is inaccurate. This infringes on the fact that data were collected at a minimum interval of 6 months, resulting in generating inconsistent predictions. Fig. 2 shows the predicted Ventricles value by MinimalRNN during 94 months. Circles in blue are the predicted values of MinimalRNN and circles in orange are the true values at some time points. As we analyzed above, MinimalRNN produces an inaccurate prediction at M6, which results in poorer prediction in the following months, a cascading error effect.

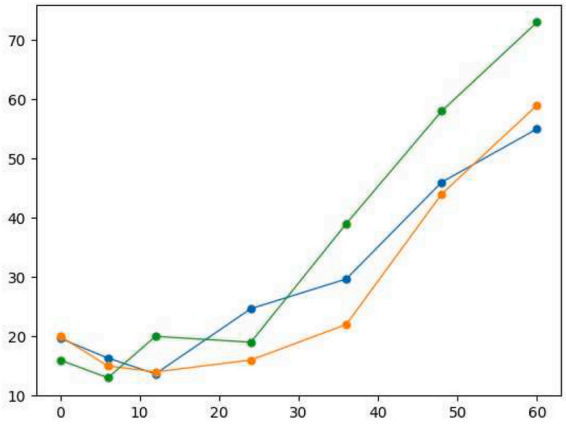
3.2. Task correlation analysis

In this section, we thoroughly analyze the correlation among the tasks of multiple longitudinal predictions. According to the analysis, two findings are investigated as follows.

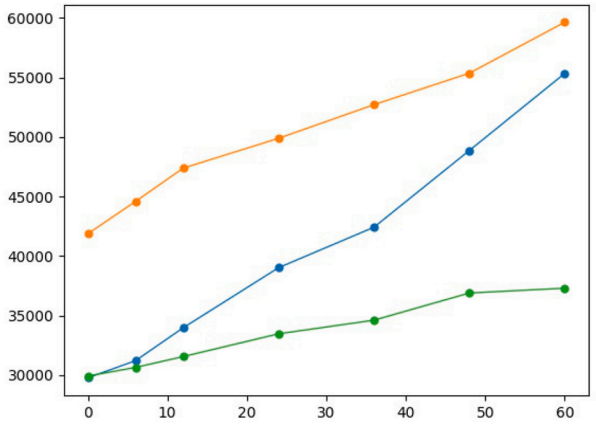
Finding 1: The multiple predictive tasks for the multiple future time points are highly relevant, given the same subject's historical features. Therefore the models at different time points should be collectively learned.

Analysis: We describe the prediction task proposed in Section 3.1 in the following way. First, $task_{24}$ aims to predict values at time M24, $task_{36}$ aims to predict values at time M36, $task_{48}$ aims to predict values at time point M48, and $task_{60}$ aims to predict values at time M60.

The main challenge tackled in this work is to perform an accurate prediction at a further time point M60, giving that only a small amount of samples are provided. Therefore we regard $T60$ as a reference, and $T24$, $T36$, $T48$ as auxiliary. Let $y_m^{Ven} = (y_m^{Ven,1}, y_m^{Ven,2}, \ldots, y_m^{Ven,N})$ be a vector



(a) Variation patterns of ADAS-Cog13.



(b) Variation patterns of Ventricles.

Fig. 3. Variation patterns of ADAS-Cog13 and Ventricles over multiple time points.

Table 1
Correlation coefficients between *T*60 and the other tasks.

Corr(*T*60, *T*24)	*Corr*(*T*60, *T*36)	*Corr*(*T*60, *T*48)
0.67	0.83	0.89

that represents true Ventricles of reference task, $y_a^{Ven} = (y_a^{Ven,1}, y_a^{Ven,2}, \ldots, y_a^{Ven,N})$ is a vector that represents true Ventricles of the auxiliary task. N is the number of subjects. Then the correlation coefficients $r_{m,a}$ between reference task and each auxiliary task is calculated as Formulation (5):

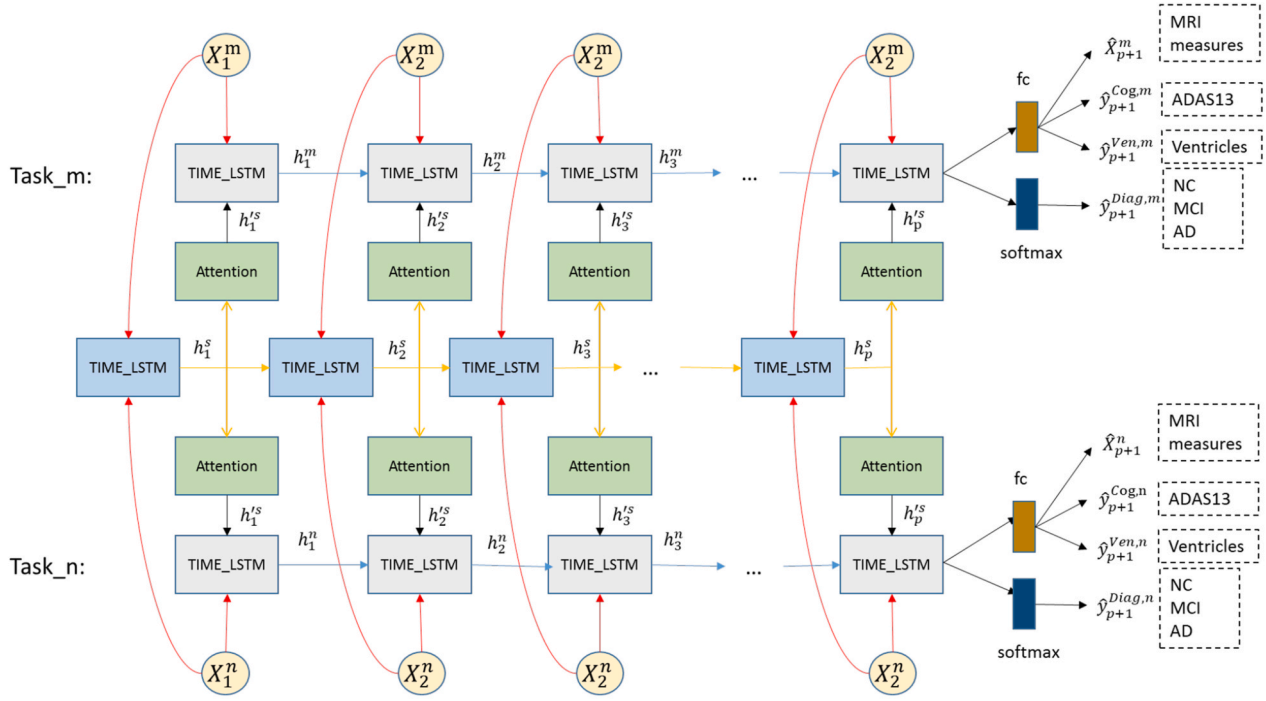
$$Corr(m,a) = \frac{\sum_{j=1}^{N}(y_m^{Ven,j} - \bar{y}_m^{Ven})(y_a^{Ven,j} - \bar{y}_a^{Ven})}{\sqrt{\sum_{j=1}^{N}(y_m^{Ven,j} - \bar{y}_m^{Ven})}\sqrt{\sum_{j=1}^{N}(y_a^{Ven,j} - \bar{y}_a^{Ven})}} \quad (5)$$

where $\bar{y}_m^{Ven}$ and $\bar{y}_a^{Ven}$ are average of $y_m^{Ven,j}$ and $y_a^{Ven,j}$. The larger the $|r_{m,a}|$ is, the stronger correlation between the reference task and the auxiliary task exists. We obtain the correlation coefficients between $task_{60}$ and the other tasks as follows:

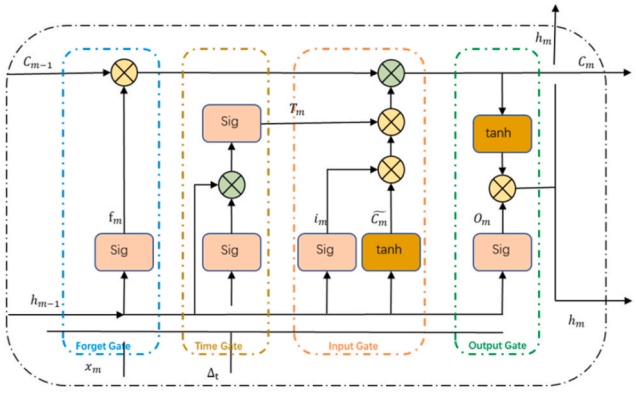
Table 1 shows that the correlation coefficients between *T*60 and *T*24, *T*36, *T*48 are all larger than 0.5, which demonstrates that strong correlations exist among these tasks. These observations subsequently motivate us to propose a multi-task learning method for modeling the related prediction tasks.

Finding 2: There exists temporal smoothness among the multiple tasks.

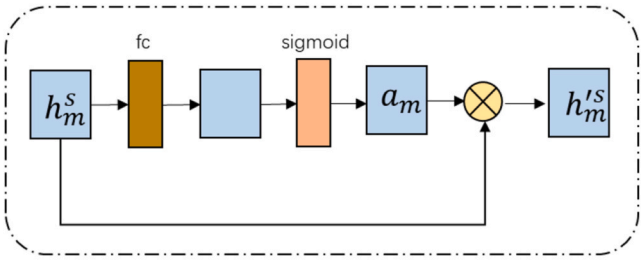
Analysis: Here, we further study the correlation among the longitudinal tasks. We assume that the difference in the cognitive scores between two successive time points is relatively small. To evaluate it, we explore the variation pattern at multiple time points. Fig. 3 shows the varying pattern of ADAS-Cog13 and Ventricles over seven time points, respectively. In Fig. 3(a) and (b), We can clearly observe the temporal smoothness between two adjacent time points. To guarantee the prediction value is appropriate, a temporal smoothness regularization is incorporated to penalize the large difference between the predictive



(a) Architecture of the proposed MTL-ATM approach.



(b) The unit of Time_LSTM.



(c) Attention module.

Fig. 4. (a) Overview of our proposal MTL-ATM. (b)The unit of Time_LSTM. (c) Attention module.

score at the consecutive time points. Without the temporal smoothness regularization, the scores obtained by a predictive model are prone to fluctuation. In my experimental results, the model performs much better than the single task learning approaches. This is discussed later.

These observations are important for our later analysis of jointly modeling the multiple tasks in a unified framework. So there is a natural question, how to capture such task relatedness. Here after we introduce our multi-task learning architecture.

4. Methods: multi-task learning for disease progression modeling

Multi-task learning (MTL) improves the performance of each task by optimizing multiple tasks jointly and utilizing the correlation between related tasks. Despite the steady growth of multi-task learning research for healthcare and computer-aided diagnosis, there are still two problems with multi-task learning:

1. How to design a shared network architecture: It is important to develop a network of multi-task learning that exploits the shared information among the multiple tasks while maximally preserving the specific information of specific tasks.
2. How to optimize the learning of multi-task learning: Different tasks are required to be properly balanced to enable the model converge to a state that is beneficial for all tasks. It is desirable to learn a balanced global task weight to avoid some tasks dominating the training process.

4.1. Architecture design

To enable both shared and task-specific features to be learned automatically, we integrate Time_LSTM into the multi-task learning framework. Fig. 4(a) shows the architecture of the proposed MTL-ATM approach. It consists of a shared network and multiple task-specific networks. The shared network is trained for modeling the shared information among the multiple tasks, whilst each task-specific network consists of an attention module to capture the task-specific information from the shared information. The network can be seen as an end-to-end architecture with the shared and task-specific parameters. By learning those parameters jointly, we arrive at a collaborative learning method to jointly improve the performance of the prediction tasks at multiple time points. In the proposed architecture, the shared layer contains:

- The shared information between classification task and regression task at each time point.
- The shared information between the multiple prediction tasks on multiple time points.

Considering the existence of missing visits as we mentioned above, we adopt Time_LSTM as the shared layer and task-specific layer network. The architecture of Time_LSTM is shown in Fig. 4(b). Time_LSTM is a variant of LSTM. More specifically, LSTM, as a variant of RNN, has an advanced ability to model short and long-term temporal dependencies and has become an effective and scalable model for sequential prediction problems. The basic functions of an LSTM unit are defined as follows:

$$i_t = \sigma_i(X_t W_{xi} + h_{t-1} W_{hi} + b_i) \tag{6}$$

$$f_t = \sigma_f(X_t W_{xf} + h_{t-1} W_{hf} + b_f) \tag{7}$$

$$c_t = i_t \odot \sigma_c(X_t W_{xc} + h_{t-1} W_{hc} + b_c) \tag{8}$$

$$o_t = \sigma_o(X_t W_{xo} + h_{t-1} W_{ho} + b_o) \tag{9}$$

$$h_t = o_t \odot \sigma_h(c_t) \tag{10}$$

Where $\{W_{xi}, W_{xf}, W_{xc}, W_{xo}\}$ and $\{W_{hi}, W_{hf}, W_{hc}, W_{ho}\}$ are learned weights of LSTM. $\{b_i, b_f, b_c, b_o\}$ represents corresponding biases. Symbol $\odot$ denotes element-wise multiplication. $\{\sigma_i, \sigma_f, \sigma_o\}$ are sigmoidal nonlinearities and $\{\sigma_c, \sigma_h\}$ are tanh nonlinearities.

During the disease progression, time intervals between subjects' visits are of significant importance in capturing the relations of subjects' progression. However, the traditional LSTM architectures are incapable of modeling them. In order to solve this problem, Time_LSTM is proposed to model the time information by adding a time gate to Long Short Term Memory (LSTM) [31]. Hence, it allows Time_LSTM to model the temporal relation by capturing the variable interval information. Based on the update equations from Eq. (6) to Eq. (10), Time_LSTM add one update equation for T_m as:

$$T_m = \sigma_t(X_m W_{xt} + \sigma_{\triangle t}(\triangle t_m W_{tt}) + b_t) \tag{11}$$

Then Eq. (8) and Eq. (9) are modified as:

$$c_m = i_m \odot T_m \odot \sigma_c(X_m W_{xc} + h_{m-1} W_{hc} + b_c) \tag{12}$$

$$o_m = \sigma_o(X_m W_{xo} + \triangle t_m W^s_{to} + h_{m-1} W_{ho} + b_o) \tag{13}$$

Where $\triangle t_m$ is the time interval and $\sigma_{\triangle t}$ is a sigmoid function. The processing of the shared layer at time t is updated as:

$$h^s_t = Time\_LSTM(X_t, h^s_{t-1}, \theta_s) \tag{14}$$

Where $X_t \in \mathbb{R}^{1\times D}$ is the D dimensional input feature at time t, and θ_s denotes the parameters of the shared layer. Concretely, the internal operations in the shared layer Time_LSTM are as Eq. (6) to Eq. (13)

To enhance the interaction between task-specific layers and the shared layer, an attention mechanism is developed to enable the task-specific module to learn task-related information. Through assigning one attention mask for each task, it is able to automatically determine the relevant information for each specific task from the shared knowledge. The collaborate learning strategy allows learning of both task-shared and task-specific information at the same time.

We denote the learned attention mask of task q at time t as a^q_t:

$$a^q_t = \sigma(h^s_t W^q_{att}) \tag{15}$$

Where W^q_{att} is the learnable parameter in the attention layer for task q. a^q_t is utilized to control how much information is transformed from the shared layer to each task-specific layer. If the learned a^q_t is close to 1, that means other tasks contribute more to task q. On the contrary, the learned a^q_t close to 0 means that task q gets less information from other tasks.

The information $h^{'s}_t$ transformed into the task-specific layer is then computed by element-wise multiplication of the attention mask with the hidden output of the shared layer:

$$h^{'s}_t = a^q_t \odot h^s_t \tag{16}$$

Fig. 4(c) shows the attention module, where all the attention modules for different tasks have the same design, although their weights are individually learned.

For the task-specific layers, the input features are concatenated with the output $h^{'s}_t$ of the attention module. This enables each task to share the information from all tasks. The output of task q at time t can be denoted as:

$$h^q_t = Time\_LSTM(\begin{bmatrix} X_t \\ h^{'s}_t \end{bmatrix}, h^q_{t-1}, \theta_q) \tag{17}$$

where θ_q denotes the parameters of a task-specific layer.

4.2. Temporal smoothness

We propose a regularization for capturing the temporal correlation in disease progression. We assume that there is a relatively small difference between the cognitive scores of two successive time points, which is also in line with the fact (as shown in Fig. 3). The temporal smoothness is modeled by penalizing the large difference between the predictions of multiple tasks at the consecutive time points. .

4.3. Loss function

The MTL-ATM is trained to predict the future observation given three initial observations (e.g., predicting the disease progression at M24, M36, M48 and M60 given the MRI features at M0, M6 and M12). Errors between the actual observations and predictions are used to update the model parameters. It is worth noting that the loss function was only calculated with available observations. The overall loss $L_{overall}$ is defined as follows:

$$L_{overall} = \sum_{k=1}^{q} \lambda_k (\lambda_c L_c + \lambda_r L_r + \lambda_s L_s) \tag{18}$$

Where λ_k is a parameter that determines the weight of the *k-th* task, q is the number of tasks. λ_c, λ_r and λ_s are the weights for classification loss, regression loss and temporal smoothing loss, respectively. L_c and L_r are prediction losses for the categorical variable (Diagnosis status) and continuous variables (Ventricles, ADAS-Cog13, MRI measurements) at all time points except time M0, respectively. Here, we choose cross-entropy loss and mean absolute error(MAE) for L_c and L_r:

$$\begin{aligned} L_c &= \sum_{i>1} (CrossEntropy(y_i^{Diag} \odot M_{y,i}^{Diag}, \hat{y}_i^{Diag} \odot M_{y,i}^{Diag})) \\ &= \sum_{i>1} \sum_{j=1}^{3} (-(y_i^{Diag,j} \odot M_{y,i}^{Diag,j}) \log(\hat{y}_i^{Diag,j} \odot M_{y,i}^{Diag,j})) \end{aligned} \tag{19}$$

$$\begin{aligned} L_r &= \sum_{i>1} (MAE(y_i^{Ven} \odot M_{y,i}^{Ven}, \hat{y}_i^{Ven} \odot M_{y,i}^{Ven}) \\ &\quad + MAE(y_i^{Cog} \odot M_{y,i}^{Cog}, \hat{y}_i^{Cog} \odot M_{y,i}^{Cog}) \\ &\quad + MAE(X \odot M_x^i, \hat{X} \odot M_x^i)) \end{aligned} \tag{20}$$

In order to capture the temporal smoothness on MRI biomarker and cognitive scores, we incorporate a time smooth regularization term [32] into the loss function to capture the smoothness of outputs from adjacent time points. The temporal smoothness can be enforced by penalizing the difference between models of consecutive time points, as shown in Eq. (21):

$$\begin{aligned} L_s = \sum_{i>2} &((\hat{y}_i^{Ven} \odot M_{y,i}^{Ven} - \hat{y}_{i-1}^{Ven} \odot M_{y,i-1}^{Ven})^2 \\ &+ (\hat{y}_i^{Cog} \odot M_{y,i}^{Cog} - \hat{y}_{i-1}^{Cog} \odot M_{y,i-1}^{Cog})^2) \end{aligned} \tag{21}$$

Based on the above loss function, our prediction module products several outcomes, i.e., MRI measurements $\hat{X}$, cognitive tests $\hat{y}^{Cog}$, clinical state $\hat{y}^{Diag}$ and Ventricles $\hat{y}^{Ven}$. The outputs at each time point t are formulated as:

$$\hat{y}_t^{Diag} = softmax(W_{cla} h_t + b_{cla}) \tag{22}$$

$$\hat{X}_t, \hat{y}_t^{Ven}, \hat{y}_t^{Cog} = W_{reg} h_t + b_{reg} \tag{23}$$

where $\hat{y}_t^{Diag}$ is prediction probabilities of clinical diagnosis, $\hat{y}_t^{Ven}$ and $\hat{y}_t^{Cog}$ are the prediction value of Ventricles and ADAS-Cog13, respectively. $\hat{X}_t$ is the predicted MRI measurement that is utilized to impute the missing X_t value at time t. W_{cla} and W_{reg} are weight matrices that need to be learned. b_{cla} and b_{reg} are bias terms.

Through this process, our proposed MTL-ATM encodes the longitudinal data and captures underlying temporal characteristics from it. Lastly, the encoded representations are further transformed to predict the clinical status, volumetric measurements and cognitive scores at the next time point.

4.4. Optimization

The proposed formulation is challenging to solve due to a large number of tasks, resulting in the involved task weight parameters to be difficult to optimize. The appropriate task weight is important for the prediction performance. How to obtain the best weight parameter for each task for balancing the corresponding contribution at each training step is a challenge.

In our work, we assign different weights for each term in Eq. (18) by developing a dynamic task weighting scheme into the optimization, which enables the model to achieve a balanced training automatically by dynamically tuning gradient magnitudes. We define the weighting $\lambda_{k,c} = \lambda_k \lambda_c$, $\lambda_{k,r} = \lambda_k \lambda_r$, $\lambda_{k,s} = \lambda_k \lambda_s$ of task k at the *i-th* batch as:

$$\lambda_{k,c}^i = \left(\frac{L_{k,c}^i}{L_{k,c}^1}\right)^{\alpha} \qquad \lambda_{k,r}^i = \left(\frac{L_{k,r}^i}{L_{k,r}^1}\right)^{\alpha} \qquad \lambda_{k,s}^i = \left(\frac{L_{k,s}^i}{L_{k,s}^1}\right)^{\alpha} \tag{24}$$

where α is a hyperparameter, and the weighting of each loss item considers the loss ratio between the current loss and the initial loss, which can measure how well the model has been trained for that loss. When a loss is poorly trained, the ratio is close to 1 and takes a larger proportion in the overall loss and gradient, and vice versa. The procedure of the network optimization is shown in Algorithm 1:

Algorithm 1. Training MTL-ATM

Algorithm 1 Training MTL-ATM

Input: Data from multiple tasks
Output: Calibrated network weights sets Θ

repeat
  Select a random task k
  Select a batch of examples from datasets of the task k
  Calculate the loss according to Eq.(18) to Eq.(24)
  Update $\lambda_{k,c}$, $\lambda_{k,r}$, $\lambda_{k,s}$ according to Eq.(24)
  Update all parameters Θ by taking a gradient step with respect to this batch
until Convergence

Table 2
Description of different tasks in the experiments.

Tasks	Observed time points			Predicted time points	Number of subjects
Task1	M0	M6	M12	M24	676
Task2	M0	M6	M12	M36	425
Task3	M0	M6	M12	M48	274
Task4	M0	M6	M12	M60	140

Table 3
The grid search space for the hyperparameters.

Hyperparameter	Range
Hidden size	32,64,128,256,512
Dropout rate	0.1,0.2,0.3,0.4
Learning rate	$1e^{-2}$, $1e^{-3}$, $5e^{-3}$, $5e^{-4}$
Number of hidden layer	1, 2, 3
Weight decay	$1e^{-3}$, $1e^{-4}$, $1e^{-5}$

5. Experiments

5.1. Datasets and experiment setting

5.1.1. Datasets

The data utilized in this work is provided by the TADPOLE challenge [33], consisting of 1737 subjects from the ADNI database [34]. Although the TADPOLE dataset offers numerous kinds of biomarkers to forecast the AD progression, in this paper, we consider 22 variables recommended by the TADPOLE challenge, which include:

- Cognitive tests: MMSE, CDRSB, ADAS-Cog11, RAVLT_perc_forgetting, FAQ, MOCA RAVLT_immediate,

  RAVLT_learning, ADAS-Cog13, RAVLT_forgetting.

- MRI measures: Hippocampus, WholeBrain, Fusiform, ICV, Ventricles, Entorhinal, MidTemp.
- PET measures: FDG, AV45.
- CSF measures: ABETA_UPENNBIOMK9_04_19_17,

  TAU_UPENNBIOMK9_04_19_17, PTAU_UPENNBIOMK9_04_19_17.

The clinical groups are labeled as AD, CN, EMCI, LMCI and SMC, which mean early MCI, late MCI and significant memory concern respectively. Similar to the previous work [6,7,21], CN and SMC are merged into the CN group, EMCI and LMIC are merged into the MCI group, thus resulting in three categories: CN(523), MCI(872) and AD (342). The number of subjects in different tasks is summarized in Table 2. The first time at the hospital is indicated as M0, and the follow-up visit is denoted by the duration starting from the baseline. For example, M6 refers to the time point at the 6th month.

5.1.2. Experiment setting

We conduct experiments to study how the changes in the brain are associated with different clinical biomarkers at four time points (i.e. 24-month, 36-month, 48-month, 60-month), which can be taken as four tasks. This yields a total of n = 676 subjects for Task1 (Baseline, M6, M12, M24) and for Task2, Task3 and Task4, the sample size are 425, 274 and 140, respectively. The sample amount of each task is different in Table 2 due to the dropout from the study of some patients for various reasons.

We report the average test results from 10-fold cross-validation. Specifically, the dataset is partitioned into three non-overlapping subsets (training set, validation set and testing set). The radio of subjects in each subset is 8:1:1. All variables are z-normalized except clinical status. The z-normalization is performed on the training set. The mean and standard deviation from the training set are then utilized to z-normalize the validation and test sets.

We use the Adam optimizer, early stopping criterion and hyperparameters selection strategy for training. The validation set is used to select the hyperparameters and we stop the training if the validation loss does not improve for 50 epochs. The optimal hyperparameters are obtained by grid search. The ranges of grid search are summarized in Table 3.

For quantitative evaluation, we choose the metrics of multi-class area under the receiver operating characteristic curve (mAUC) as well as balanced class accuracy (BCA) for clinical status prediction, and MAE for the MRI biomarker prediction as well as cognitive scores forecasting. Higher values of both mAUC and BCA metrics indicate better performance. Whilst lower MAE indicates better performance.

The proposed method is compared with four methods that deal with clinical time series prediction:

- LSTM-F [35]: LSTM-F focuses on data gathered from the Children's Hospital Los Angeles pediatric intensive care unit (PICU). In LSTM-F, the missing values are imputed by the most recent recorded measurement, and then an LSTM is trained to addressing the task of multilabel classification of diagnosis giving clinical time series.
- LSTM-M [36]: LSTM-M is an AD disease progression model based on a deep RNN with an LSTM module. For the missing data, LSMTM-M applies a median imputation for the ordinal variables and a mode imputation for the nominal variables.
- MinimalRNN [6]: MinimalRNN is an AD disease progression model that utilizes the data provided by the TADPOLE challenge [33]. The predictions of the MinimalRNN are used as the inputs for the next time point.
- DeepRNN [7]: DeepRNN is also a model for modeling the AD with incomplete longitudinal data from the TADPOLE challenge. In Ref. [7], missing value imputation, forecasting of future MRI biomarker, cognitive score and clinical status prognosis over multiple time points are jointly learned in a unified framework.

5.2. Results

To verify the effectiveness of our proposed method for multiple time points prediction, we compare the proposed MTL-ATM approach with LSTM-F, LSTM-M, MinimalRNN, and DeepRNN. For the four time points from M24 to M60, we predict their ADAS-cog13, Ventricles and clinical diagnosis with the features of MRI, PET and CSF from the first three time points(M0, M6 and M12). Table 4 and Table 5 show the performance on

Table 4
Performance on a multi-class (AD vs. MCI vs. CN) classification task in terms of mAUC. The best results are bold, and superscript symbol indicates that our proposed approach significantly outperformed that method on that score. p represent Student's t-test level (†: p = 0.05, *:p = 0.01, ▲:p = 0.001).

Methods	M24	M36	M48	M60
LSTM-F	0.882 ± 0.046*	0.858 ± 0.028	0.783 ± 0.050▲	0.779 ± 0.048*
LSTM-M	0.870 ± 0.044*	0.836 ± 0.040▲	0.790 ± 0.052▲	0.780 ± 0.034*
MinimalRNN	0.921 ± 0.039†	0.883 ± 0.026	0.833 ± 0.091†	0.830 ± 0.051†
DeepRNN	**0.943 ± 0.021**	0.899 ± 0.035	0.876 ± 0.027†	0.822 ± 0.083
MTL-ATM (Ours)	0.935 ± 0.023	**0.920 ± 0.039**	**0.905 ± 0.034**	**0.897 ± 0.085**

Table 5
Performance on a multi-class (AD vs. MCI vs. CN) classification task in terms of BCA. The best results are bold, and superscript symbol * indicates that our proposed approach significantly outperformed that method on that score. p represent Student's t-test level (†: p = 0.05, *:p = 0.01, ▲:p = 0.001).

Methods	M24	M36	M48	M60
LSTM-F	0.774 ± 0.075†	0.703 ± 0.044	0.635 ± 0.071▲	0.650 ± 0.033*
LSTM-M	0.771 ± 0.038▲	0.696 ± 0.041▲	0.691 ± 0.085*	0.646 ± 0.052†
MinimalRNN	0.833 ± 0.045†	0.750 ± 0.051*	0.746 ± 0.070	0.717 ± 0.070
DeepRNN	**0.858 ± 0.036**	0.782 ± 0.047	0.764 ± 0.047	0.731 ± 0.084
MTL-ATM (Ours)	0.842 ± 0.037	**0.804 ± 0.032**	**0.818 ± 0.062**	**0.752 ± 0.096**

Table 6
Performance of forecasting Cog13 at time M24, M36, M48, M60 based on data in M0, M6, M12 in terms of MAE. The best results are bold, and superscript symbol * indicates that our proposed approach significantly outperformed that method on that score. p represent Student's t-test level (†: p = 0.05, *:p = 0.01, ▲:p = 0.001).

Methods	M24	M36	M48	M60
LSTM-F	4.40 ± 0.348*	5.270 ± 0.802▲	5.112 ± 0.836	5.866 ± 1.763
LSTM-M	3.892 ± 0.278	4.672 ± 0.737	5.163 ± 0.867	6.197 ± 1.734
MinimalRNN	**3.801** ± 0.238	4.198 ± 0.721	5.070 ± 1.125	5.802 ± 2.303
DeepRNN	4.067 ± 0.311†	4.698 ± 0.779	5.564 ± 1.031†	6.411 ± 1.444
MTL-ATM (Ours)	3.826 ± 0.316	**4.196 ± 0.754**	**4.712 ± 0.913**	**5.238 ± 1.691**

Table 7
Performance of forecasting Ventricles at time M24, M36, M48, M60 based on data in M0, M6, M12 in terms of MAE. The best results are bold, and superscript symbol * indicates that our proposed approach significantly outperformed that method on that score. p represent Student's t-test level (†: p = 0.05, *:p = 0.01, ▲:p = 0.001).

Methods	M24($\times 10^{-3}$)	M36($\times 10^{-3}$)	M48($\times 10^{-3}$)	M60($\times 10^{-3}$)
LSTM-F	8.42(±0.67)▲	8.61(±1.09)▲	8.29(±1.75)	9.26(±1.83)
LSTM-M	8.73(±0.64)▲	8.71(±1.15)*	8.57(±1.75†)	9.65 ± 2.13†
MinimalRNN	7.55(±0.47)	7.70(±0.98)	8.50(±2.65)	9.63(±2.37)
DeepRNN	**6.45(± 1.65)**	7.92(±2.17)	8.00(±3.21)	9.00(±3.23)
MTL-ATM (Ours)	7.35(±0.51)	**6.78(± 1.13)**	**7.32(± 2.07)**	**7.81(± 2.57)**

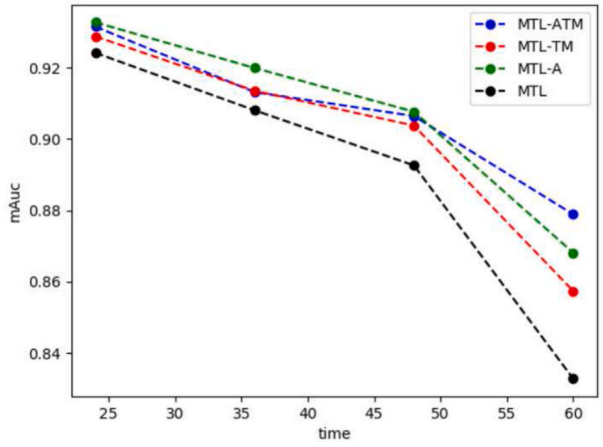
Table 8
Performance at different time points of comparable filling methods.

Filling strategies	mAUC ↑ ($\times 10^{-1}$)				BCA ↑ ($\times 10^{-1}$)				MAE (Cog13) ↓ ($\times 10^{0}$)				MAE (Ventricles) ↓ ($\times 10^{-3}$)			
	M24	M36	M48	M60	M24	M36	M48	M60	M24	M36	M48	M60	M24	M36	M48	M60
LSTM-F	8.82	8.58	7.83	7.79	7.74	7.03	6.35	6.50	4.40	5.27	5.11	5.87	8.42	8.61	8.29	9.26
LSTM-M	8.70	8.36	7.90	7.80	7.71	6.96	6.91	6.46	3.89	4.67	5.16	6.20	8.73	8.71	8.57	9.65
Mean filling	8.82	8.70	8.59	8.45	8.16	7.72	7.84	7.41	3.83	4.23	**4.55**	5.26	7.51	7.02	6.92	8.31
Linear filling	9.20	8.72	8.96	8.45	8.32	7.67	7.87	7.03	3.85	4.32	4.71	5.25	7.38	7.03	7.48	7.95
Forward filling	9.25	9.11	9.04	8.49	8.37	7.96	7.90	7.49	**3.80**	4.23	4.79	5.47	7.63	7.03	6.89	8.45
Model filling (Ours)	**9.35**	**9.20**	**9.05**	**8.97**	**8.42**	**8.04**	**8.18**	**7.52**	3.83	**4.20**	4.71	**5.23**	**7.35**	**6.85**	**7.32**	**7.81**

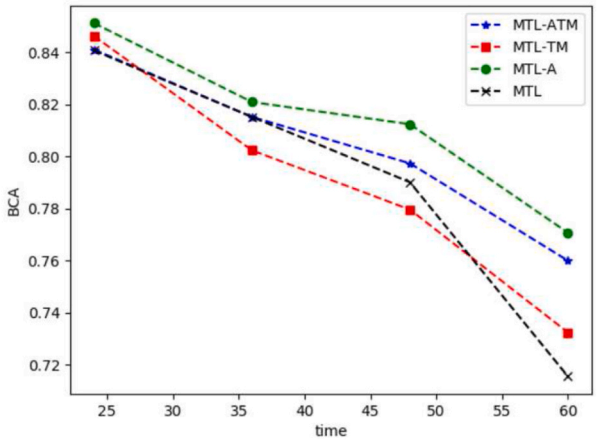
classification tasks in terms of mAUC and BCA over the 4 time points. Table 6 and Table 7 show the prediction errors in MAE over ADAS-cog13 and Ventricles.

According to Tables 4–7, we can draw the following conclusions:

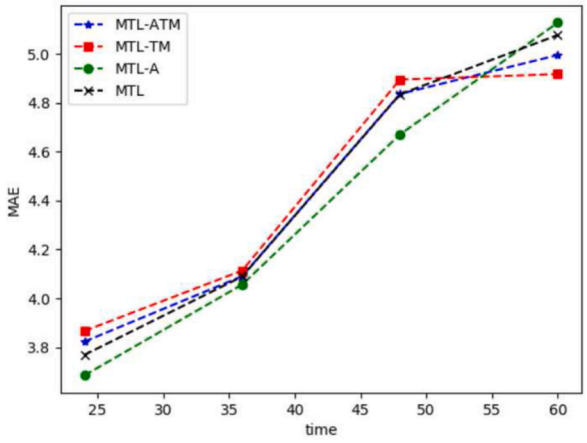
1. The proposed method achieves decent and comparable performance when compared to state-of-the-art methods, resulting in 5.6%, 5.7%, 4.0%, 11.8% and 3.5%, 2.7%, 12.4%, 5.5% increase in mAUC, BCA, MAE (ADAS-cog13) and MAE (Ventricles) compared with the previous MinimalRNN model and DeepRNN model, respectively. This result indicates that learning multiple tasks jointly and utilizing the correlation among related tasks greatly improve the performance of each task.
2. From the perspective of the filling approach, the proposed method, MinimalRNN and DeepRNN belong to the model filling whilst LSTM-F and LSTM-M belong to linear filling and mean filling. It is clear that the three model filling methods outperform the linear filling and mean filling methods, which is a collaborative learning method for the imputation and prediction tasks while both the LSTM-F and LSTM-M conduct the imputation and prediction independently. This again validates that accurate imputation can introduce contributions to the prediction, while appropriate prediction can also benefit imputation performance. The observation is consistent with the result in Refs. [6,7].
3. Our model is more effective at the farther time points. For example, at the prediction at time point M60, the proposed method improves MinimalRNN by 8.1% and DeepRNN by 9.1% in terms of mAUC. The improved performance can be attributed to the correlation of the tasks, which helps to learn a better prediction ability for the prediction at the farther time points of AD diagnosis and hence yield better results compared to other works. Our results shed new light on the importance of multi-task learning in the AD progression model for improving prediction performance.



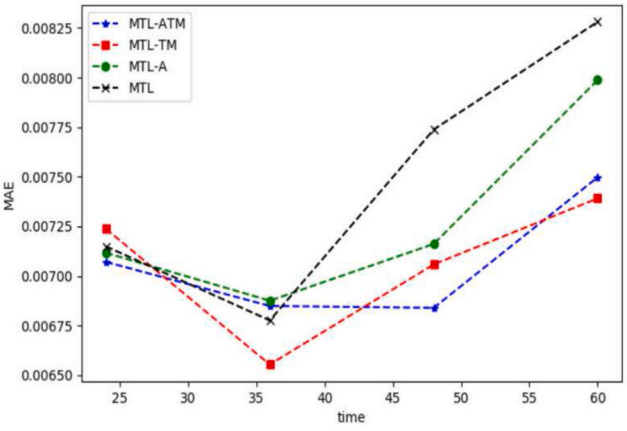
(a) Performance on mAUC.



(b) Performance on BCA.



(c) Performance on MAE of ADAS-Cog13.



(d) Performance on MAE of Ventricles.

Fig. 5. Result of ablation studies on the effect of time smooth loss and attention component.

Table 9
ranking_score at four time points.

Methods	M24	M36	M48	M60
MTL	13	12	14	15
MTL-TM	13	11	13	8
MTL-A	**5**	**7**	7	9
MTL-ATM	9	8	**6**	**7**

6. Discussion

6.1. Effect of imputation

In order to investigate the influence of different interpolation methods on the following prediction performance, we conduct four experiments with four missing values interpolation strategies: Mean filling, Linear filling, Forward filling and Model filling. It is worth noting that the experimental settings are the same for the four experiments except for the filling strategies.

We present the experimental results in Table 8. The results, shown in Table 8, once again validate that model filling exhibits an impressive improvement over the competing filling methods in most cases at all time points, except the MAE of ADAS-Cog13 at time M24 and time M48, which further indicates the values imputed by the model are more accurate than other filling strategies. Particularly, our model achieved a promising prediction performance at the task of M60, which is vital for the accurate progression estimation of AD at its prodromal stage.

In addition, it can be seen that the proposed MTL-ATM method with Mean filling and Forward filling achieves 0.816 and 0.837 in terms of BCA, resulting in a 5.5% and 7.5% increase compared with the LSTM-M and LSTM-F, respectively. This demonstrates that our proposed MTL-ATM method provides a flexible framework into which different filling strategies can be incorporated.

6.2. Ablation studies

In addition to the above-mentioned results, we are also interested in the effectiveness of each component in the proposed MTL model. To examine the effectiveness of the component of attention and temporal smoothness regularization, we conduct experiments on the three categories of variants, denoted as MTL-A, MTL-TM and MTL:

- MTL-A: MTL-ATM without temporal smoothness regularization.

Table 10
Performance at different time points with 1, 2, 3 and 4 tasks.

Task num	mAUC ↑ ($\times 10^{-1}$)				BCA ↑ ($\times 10^{-1}$)				MAE (ADAS-Cog13) ↓ ($\times 10^{0}$)				MAE (Ventricles) ↓ ($\times 10^{-3}$)			
	M24	M36	M48	M60	M24	M36	M48	M60	M24	M36	M48	M60	M24	M36	M48	M60
1 task	9.24	8.30	7.20	7.20	8.33	7.22	6.63	6.29	**3.783**	4.731	5.655	6.983	7.49	7.83	8.99	9.93
2 tasks	9.28	**9.25**	–	–	**8.44**	**8.20**	–	–	3.884	4.253	–	–	7.69	6.95	–	–
3 tasks	9.34	9.18	**9.16**	–	8.42	8.05	**8.24**	–	3.795	4.220	4.758	–	7.41	**6.61**	**6.92**	–
4 tasks	**9.35**	9.20	9.05	**8.97**	8.42	8.04	8.18	**7.52**	3.826	**4.196**	**4.712**	**5.238**	**7.35**	6.78	7.32	**7.81**

Table 11
Effect of the number of input time points. Content in brackets is the total number of samples used in each task.

Num of the input time points	mAUC ↑ ($\times 10^{-1}$)			BCA ↑ ($\times 10^{-1}$)			MAE (ADAS-Cog13) ↓ ($\times 10^{0}$)			MAE (Ventricles) ↓ ($\times 10^{-3}$)		
	M36	M48	M60	M36	M48	M60	M36	M48	M60	M36	M48	M60
2 input points (499,326,167 subjects)	8.61	8.51	**8.86**	**7.72**	**7.61**	**7.69**	**4.20**	4.91	6.17	7.65	8.13	**8.28**
3 input points (425,274,140 subjects)	8.58	8.38	8.53	7.22	7.28	6.67	4.32	**4.69**	**5.17**	7.49	7.83	8.37
4 input points (360, 234, 117 subjects)	**8.80**	**8.56**	8.61	7.19	7.27	7.48	4.42	4.77	5.65	**6.94**	**7.35**	9.63

Table 12
Comparison among the methods with the same task number but a different combination.

Task combination	mAUC ↑ ($\times 10^{-1}$)				BCA ↑($\times 10^{-1}$)				MAE (ADAS-Cog13) ↓ ($\times 10^{0}$)				MAE (Ventricles) ↓ ($\times 10^{-3}$)			
	M24	M36	M48	M60	M24	M36	M48	M60	M24	M36	M48	M60	M24	M36	M48	M60
24,36,48	9.34	**9.18**	**9.16**	–	8.42	**8.05**	**8.24**	–	3.80	**4.22**	4.76	–	7.41	**6.61**	**6.92**	–
36,48,60	–	8.58	8.38	8.53	–	7.22	7.28	6.67	–	4.32	**4.69**	5.17	–	7.49	7.83	8.37

- MTL-TM: MTL-ATM without the attention component. The shared information is transformed to the task-specific layer directly.
- MTL: MTL-ATM without both temporal smoothness regularization and attention component.

Fig. 5 shows the performance of MTL-ATM and its three variants in terms of mAUC, BCA and MAE. From Fig. 5(a,b,c,d), it can be seen that MTL-ATM, MTL-A, MTL-TM are better than MTL in most cases. The proposed multi-task learning schemes involving the attention mechanism and temporal regularization not only improve the performance of multi-task learning, but also benefit single-task learning. An interesting observation is that the performance improvement of MTL-TM on Ventricles is more obvious than on ADAS-Cog13, which is also consistent with the fact that with the progress of AD disease, the Ventricular volume becomes gradually higher. However, ADAS-Cog13 is a neuropsychological test administered by a clinical expert, although it may increase in a direct and quantifiable manner, due to some practice effects, the test results may be unstable.

In order to show the performance of the proposed approach and the three variants intuitively, we propose a measurement *ranking_score* to evaluate the overall performance. *ranking_score* is the sum of ranking score on four indicators. *ranking_score* of each approach is calculated on four indicators and the results are shown in Table 9. The lower the *ranking_score* is, the better.

In Table 9, it is obvious to see that MTL-ATM, MTL-A and MTL-TM generally obtain better performance than MTL, verifying the usefulness of the attention component and temporal smoothness regularization. Exploiting two complementary multiple task correlations is important. More specifically, on the one hand, due to the task correlation, the attention component can adaptively determine how much information can be transformed from the shared layer to the task-specific layer. On the other hand, the temporal smoothness regularization guides the model to predict more appropriate values by penalizing the large difference between two consecutive prediction values.

It is more challenging for the predictive tasks at the farther time points with less available data such as M60. MTL-ATM performs better than MTL-A and MTL-TM at M48 and M60, which verifies the predictive ability of MTL-ATM at farther time points. The fact implies the necessity to develop multi-task learning scheme to tackle disease progression problems, which also validates our motivation to build multi-task learning for multiple prediction tasks over time points. The ablation study further demonstrates the effectiveness of the proposed attention and time smooth mechanism for the prediction at the farther time points.

6.3. Effect of multi-task learning

It is noteworthy that in the AD progress prediction task, the major difference between our work and other works is the joint prediction of multiple tasks at different time points. Thus we conduct experiments to discuss the effect of multi-task learning in our proposed approach by varying the task number. It is worth noting that by removing the shared layer and the attention layer, our model degenerates to a single task learning model.

Table 10 shows the results of the proposed model predicting at multiple time points with different task numbers. First, in the case of a single task, we can see that the performance of various indicators gradually decreases over time. It reflects that the difficulty of the prediction task increases over time since the relationship to be modeled between the features and the prediction tasks becomes weaker. Such a single learning model is incapable of jointly predicting multiple tasks over time. However, as the number of tasks increases, we observe consistent improvements in all metrics. As shown in Table 10, jointly training of multiple tasks greatly improves the performance at each time point compared to the single learning task. Even though MAE performance of ADAS-Cog13 at time M24 drops by 2.67%, 3.17%, 1.14% with the jointly learning the other tasks, its performance at time M36, M48, M60 improves by 1.12%, 16.7%, 25.0% when trained with 4 tasks. These observations suggest that collaboratively learning of multi-task at multiple time points has merit. And such a finding is crucial for the early

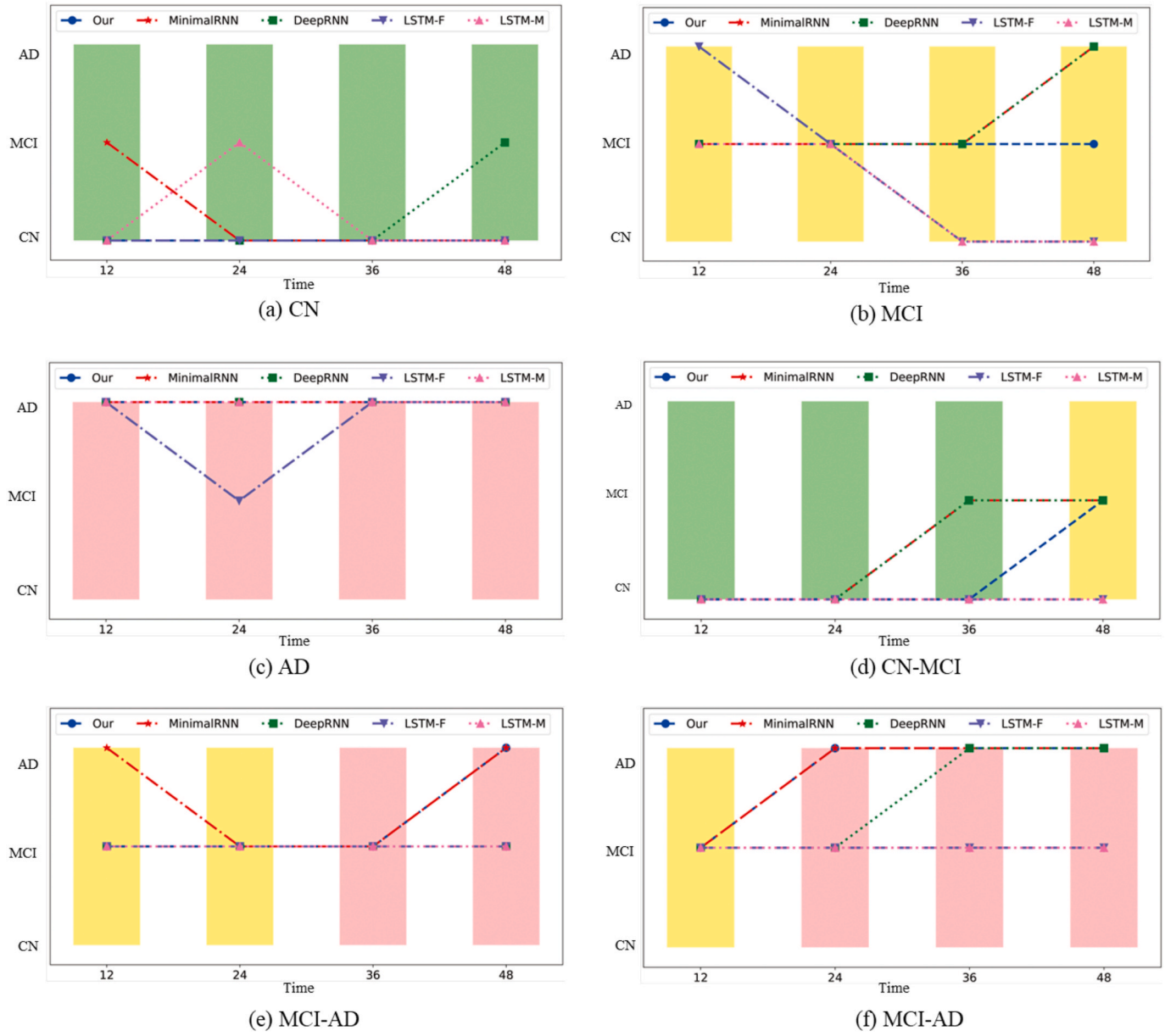


Fig. 6. Results of longitudinal status prediction between MTL-ATM our the proposed method, MinimalRNN, DeepRNN, LSTM-F and LSTM-M.(Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Alzheimer's disease prediction especially for the prediction at the farther time points.

6.4. *Effect of the number of the input time points*

We evaluate our method with three previous inputs at time points M0, M6 and M12 in the previous experiments. In fact, our model is also effective with different numbers of input time points. In this section, we validate the proposed approach on 2, 3 and 4 input time points. For a fair comparison, we fix the predictive future tasks at the M36, M48 and M60 with varying the input time point numbers, respectively.

As shown in Table 11, when the number of input time points increases, the amount of data gradually decreases. However, the performance of the proposed method does not drop dramatically, even improves in some cases. With the increase of the input time points, more available historical information can be obtained, and then the correlation between multiple tasks can be better captured. The results demonstrate that the proposed MTL-ATM method is effective regardless of the number of input time points.

In addition, we make a comparison among the methods with the same task number (3) but a different combination. As shown in Table 12, the result demonstrates that the [24,36,48] achieves a better performance than the [36,48,60] with respect to mAUC, BCA and MAE(Ventricles) at the 36 and 48 time points. The reason is that the task of 24 is more beneficial for the 36 and 48. Moreover, it also indicates that the correlation between 24 and 36/48 is stronger than the one between 60 and 36/48.

6.5. *Irreversibility analysis*

Alzheimer's disease is a progressive neurodegenerative disease, so the method should also reflect the character of irreversible neurodegeneration. That is, the clinical status can only be converted from CN to MCI or AD, and there is no reverse conversion. To explore the irreversible characteristics in different methods, we compare the performance of different approaches on subjects with different clinical states.

In Fig. 6, the plots in each panel represent clinical status conversion from different subjects. Green, yellow and red represent CN, MCI and AD, respectively.

It is clear that all of the MinimalRNN, LSTM-M and LSTM-F methods lead to some reversible conversions. Among them, LSTM-M and LSTM-F have the worst performance in Fig. 6(d,e,f), they cannot catch the change of state in clinical status. As for DeepRNN, although its predictions do not appear reversible conversions, its predictions are not as accurate as our proposed method, and there are some misclassifications in Fig. 6(a,d,f). There is no inverse transformation made by our proposed method, which reflects the irreversible characteristics of our method. For three examples without status changing (Fig. 6(a,b,c)), our model still achieve a stable prediction whereas the comparable methods yield poor results. The results verify that our model has ability of accurate and consistent prediction. That contributes to the time smoothing mechanism in our proposed approach, which considers that the predicting results of adjacent time points are similar and there is a certain trend, which could help our method capture changes in time and ensure the irreversibility. Therefore, our proposed method not only predicts accurately, but also reflects the irreversible characteristics in all situations.

7. Conclusion

In this paper, we have proposed a general deep multi-task learning model (MTL-ATM) to jointly consider missing imputation and future prediction. We propose a new perspective of predictive progression with a multi-task learning paradigm. Different from the previous work, we introduce a multi-task learning architecture with an attention mechanism and smoothness regularization to better model the correlation across the different tasks. It can simultaneously achieve the optima of both the imputation and prediction, both of which benefit each other. Moreover, we develop an optimization strategy for jointly learning the multi-task over time points. The ATM-MTL not only improves the performance of progression, but also benefits the missing value imputation. Experimental results on the ADNI dataset demonstrate the effectiveness of the proposed model.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (No.62076059) and the Fundamental Research Funds for the Central Universities (No. N2016001).

References

[1] M. M. Ghazi, M. Nielsen, A. Pai, M. J. Cardoso, M. Modat, S. Ourselin, L. Sørensen, Robust Training of Recurrent Neural Networks to Handle Missing Data for Disease Progression Modeling, arXiv preprint arXiv:1808.05500(2018).
[2] B. Dubois, H. Hampel, H.H. Feldman, P. Scheltens, P. Aisen, S. Andrieu, H. Bakardjian, H. Benali, L. Bertram, K. Blennow, et al., Preclinical alzheimer's disease: definition, natural history, and diagnostic criteria, Alzheimer's Dementia 12 (3) (2016) 292–323.
[3] M.W. Bondi, E.C. Edmonds, D.P. Salmon, Alzheimer's disease: past, present, and future, J. Int. Neuropsychol. Soc. 23 (9–10) (2017) 818–831.
[4] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, A.D.N. Initiative, et al., Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects, Neuroimage 104 (2015) 398–412.
[5] N.P. Oxtoby, D.C. Alexander, Imaging plus x: multimodal models of neurodegenerative disease, Curr. Opin. Neurol. 30 (4) (2017) 371.
[6] M. Nguyen, T. He, L. An, D.C. Alexander, J. Feng, B.T. Yeo, A.D.N. Initiative, et al., Predicting alzheimer's disease progression using deep recurrent neural networks, Neuroimage 222 (2020) 117203.
[7] W. Jung, E. Jun, H.-I. Suk, A.D.N. Initiative, et al., Deep recurrent model for individualized prediction of alzheimer's disease progression, Neuroimage 237 (2021) 118143.
[8] A. Søgaard, Y. Goldberg, Deep multi-task learning with low level tasks supervised at lower layers, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ume 2, 2016, pp. 231–235. Short Papers.
[9] S. Liu, E. Johns, A.J. Davison, End-to-end multi-task learning with attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1871–1880.
[10] J. Chen, X. Qiu, P. Liu, X. Huang, Meta multi-task learning for sequence modeling, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.
[11] Z. Chen, E. Jiaze, X. Zhang, H. Sheng, X. Cheng, Multi-task time series forecasting with shared attention, in: 2020 International Conference on Data Mining Workshops (ICDMW), IEEE, 2020, pp. 917–925.
[12] M. Wang, D. Zhang, D. Shen, M. Liu, Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data, Med. Image Anal. 53 (2019) 111–122.
[13] J. Zhou, L. Yuan, J. Liu, J. Ye, A multi-task learning formulation for predicting disease progression, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, pp. 814–822.
[14] M. Liu, J. Zhang, E. Adeli, D. Shen, Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 66 (5) (2018) 1195–1206.
[15] L. Ferrarini, W.M. Palm, H. Olofsen, R. van der Landen, G.J. Blauw, R. G. Westendorp, E.L. Bollen, H.A. Middelkoop, J.H. Reiber, M.A. van Buchem, et al., Mmse scores correlate with local ventricular enlargement in the spectrum from cognitively normal to alzheimer disease, Neuroimage 39 (4) (2008) 1832–1838.
[16] D. Zhang, J. Liu, D. Shen, Temporally-constrained group sparse learning for longitudinal data analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2012, pp. 264–271.
[17] B. Jie, M. Liu, J. Liu, D. Zhang, D. Shen, Temporally constrained group sparse learning for longitudinal data analysis in alzheimer's disease, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 64 (1) (2016) 238–249.
[18] X. Zhang, Y. Yang, T. Li, Y. Zhang, H. Wang, H. Fujita, Cmc: a consensus multi-view clustering model for predicting alzheimer's disease progression, Comput. Methods Progr. Biomed. 199 (2021) 105895.
[19] H.M. Tavakoli, T. Xie, J. Shi, M. Hadzikadic, Y. Ge, Predicting neural deterioration in patients with alzheimer's disease using a convolutional neural network, in: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2020, pp. 1951–1958.
[20] N. Bhagwat, J.D. Viviano, A.N. Voineskos, M.M. Chakravarty, A.D.N. Initiative, et al., Modeling and prediction of clinical symptom trajectories in alzheimer's disease using longitudinal data, PLoS Comput. Biol. 14 (9) (2018), e1006376.
[21] W. Jung, A.W. Mulyadi, H.-I. Suk, Unified modeling of imputation, forecasting, and prediction for ad progression, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 168–176.
[22] G. Lee, K. Nho, B. Kang, K.-A. Sohn, D. Kim, Predicting alzheimer's disease progression using multi-modal deep learning approach, Sci. Rep. 9 (1) (2019) 1–12.
[23] W. Zhu, L. Sun, J. Huang, L. Han, D. Zhang, Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural mri, IEEE Trans. Med. Imaging PP.99 (2021), 1-1.
[24] B. Lim, M. van der Schaar, Forecasting Disease Trajectories in Alzheimer's Disease Using Deep Learning, arXiv preprint arXiv:1807.03159(2018).
[25] S. El-Sappagh, T. Abuhmed, S.R. Islam, K.S. Kwak, Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data, Neurocomputing 412 (2020) 197–215.
[26] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, A. D.N. Initiative, et al., Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks, Neuroimage: Clinic 21 (2019) 101645.
[27] C.V. Dolph, M. Alam, Z. Shboul, M.D. Samad, K.M. Iftekharuddin, Deep learning of texture and structural features for multiclass alzheimer's disease classification, in: 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, 2017, pp. 2259–2266.
[28] D. Cheng, M. Liu, J. Fu, Y. Wang, Classification of mr brain images by combination of multi-cnns for ad diagnosis, in: Ninth International Conference on Digital Image Processing (ICDIP 2017), vol. 10420, International Society for Optics and Photonics, 2017, p. 1042042.
[29] H. Wang, Y. Shen, S. Wang, T. Xiao, L. Deng, X. Wang, X. Zhao, Ensemble of 3d densely connected convolutional network for diagnosis of mild cognitive impairment and alzheimer's disease, Neurocomputing 333 (2019) 145–156.
[30] M.M. Ghazi, M. Nielsen, A. Pai, M.J. Cardoso, M. Modat, S. Ourselin, L. Sørensen, A.D.N. Initiative, et al., Training recurrent neural networks robust to incomplete data: application to alzheimer's disease progression modeling, Med. Image Anal. 53 (2019) 39–46.
[31] Y. Zhu, H. Li, Y. Liao, B. Wang, Z. Guan, H. Liu, D. Cai, What to do next: modeling user behaviors by time-lstm, in: IJCAI, 17, 2017, pp. 3602–3608.
[32] J. Zhou, J. Liu, V.A. Narayan, J. Ye, A.D.N. Initiative, et al., Modeling disease progression via multi-task learning, Neuroimage 78 (2013) 233–248.
[33] R. V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, S. Klein, D. C. Alexander, et al., Tadpole Challenge: Prediction of Longitudinal Evolution in Alzheimer's Disease, arXiv preprint arXiv: 1805.03909(2018).
[34] C.R. Jack Jr., M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, J.L. Whitwell, C. Ward, et al., The alzheimer's disease neuroimaging initiative (adni): mri methods, J. Magn. Reson. Imag.: Off. J. Int. Soc. Magn. Reson. Med. 27 (4) (2008) 685–691.
[35] Z.C. Lipton, D.C. Kale, R. Wetzel, et al., Modeling missing data in clinical time series with rnns, Mach. Learn. Healthcare 56 (2016).
[36] T. Wang, R.G. Qiu, M. Yu, Predictive modeling of the progression of alzheimer's disease with recurrent neural networks, Sci. Rep. 8 (1) (2018) 1–12.