Medical Image Analysis 82 (2022) 102643

Contents lists available at ScienceDirect

Medical Image Analysis

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

Multi-modal sequence learning for Alzheimer's disease progression prediction with incomplete variable-length longitudinal data

Lei Xu [a,b], Hui Wu [b], Chunming He [c], Jun Wang [d], Changqing Zhang [e], Feiping Nie [a], Lei Chen [b,*]

[a] School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, PR China
[b] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, PR China
[c] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, PR China
[d] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, PR China
[e] College of Intelligence and Computing, Tianjin University, Tianjin 300350, PR China

ARTICLE INFO

Keywords:
Alzheimer's disease
Disease progression prediction
Missing modality
Multi-modal learning
Sequence learning
Latent representation learning

ABSTRACT

Alzheimer's disease (AD) is a neurodegenerative disorder with a long prodromal phase. Predicting AD progression will clinically help improve diagnosis and empower sufferers in taking proactive care. However, most existing methods only target individuals with a fixed number of historical visits, and only predict the cognitive scores once at a fixed time horizon in the future, which cannot meet practical requirements. In this study, we consider a flexible yet more challenging scenario in which individuals may suffer from the (arbitrary) modality-missing issue, as well as the number of individuals' historical visits and the length of target score trajectories being not prespecified. To address this problem, a multi-modal sequence learning framework, highlighted by deep latent representation collaborated sequence learning strategy, is proposed to flexibly handle the incomplete variable-length longitudinal multi-modal data. Specifically, the proposed framework first employs a deep multi-modality fusion module that automatically captures complementary information for each individual with incomplete multi-modality data. A comprehensive representation is thus learned and fed into a sequence learning module to model AD progression. In addition, both the multi-modality fusion module and sequence learning module are collaboratively trained to further promote the performance of AD progression prediction. Experimental results on Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset validate the superiority of our method.

1. Introduction

Alzheimer's disease (AD) is an irreversible and progressive neurodegenerative disease that gradually impairs patients' memory and other cognitive functions. Currently, AD affects over 55 million people all over the world, and the number is predicted to reach 78 million by 2030 (Gauthier et al., 2021). AD not only causes patients endless psychological and emotional burdens but also imposes a substantial financial burden on the whole health care system (Association, 2019). Unfortunately, AD can only be controlled but not cured, which typically progresses slowly and lasts over a long period (Wang et al., 2014). Therefore, it is of practical significance to develop accurate disease progression models for early AD detection during the presymptomatic phases, so as to carry out timely therapeutic intervention and avoid disease deterioration (Marinescu et al., 2019).

To date, a definitive AD diagnosis can only be made by an analysis of brain tissues during a biopsy or autopsy, which may cause severe brain injury and is not applicable for early detection. Consequently, many alternative measures have been used for evaluating AD progression. For example, Mini-Mental State Examination (MMSE) measures cognitive impairment and is associated with progressive deterioration of functional ability (Folstein et al., 1975; Petrella et al., 2003). AD Assessment Scale-Cognitive Subscale (ADAS-Cog) measures the severity of the most critical symptoms of AD, and it is the gold standard in AD drug trials for cognitive function assessment (Rosen et al., 1984). Besides, both Clinical Dementia Rating-Global (CDR-Global) and Clinical Dementia Rating-Sum Of Boxes (CDR-SOB) are used to evaluate the cognitive and functional impairment of AD (Yang et al., 2019). These cognitive scores serve as critical criteria for diagnosing AD.

There have been extensive studies on the disease progression modeling (DPM) problem. However, most previous methods utilize statistical models to process medical data, such as regression models (McDonnell et al., 2012) and risk prediction models (Green et al., 2011). In recent years, due to the rapid development of machine learning, people attempt to employ machine learning techniques for DPM. Compared with

* Corresponding author.
E-mail address: chenlei@njupt.edu.cn (L. Chen).

https://doi.org/10.1016/j.media.2022.102643
Received 31 December 2021; Received in revised form 27 August 2022; Accepted 23 September 2022
Available online 28 September 2022

traditional statistical methods, machine learning models do not require too many assumptions. The mainstream machine learning methods for DPM fall into three categories: time-series methods, multi-task learning methods, and deep learning methods.

Time-series methods (Brookmeyer and Abdalla, 2018; Sukkar et al., 2012) assume that disease progression is driven by clinical manifestations at different time points, for which the temporal correlation at different time points is introduced in the data modeling, and the longitudinal trajectories are parameterized into linear or sigmoidal curves (Ito et al., 2010; Samtani et al., 2012; Sabuncu et al., 2014; Vemuri et al., 2009). However, these approaches require prior knowledge on score trajectories. Furthermore, due to the heterogeneity of AD that the pathological characteristic are different for each individual (Rahimi and Kovacs, 2014), the individual progression trajectory may deviate from the assumed parametric form.

Instead of assuming score trajectories to follow a specific function form, multi-task learning models regard the DPM as a multi-task learning problem (i.e. considering each time point as a prediction task) and consider the correlations between different tasks by employing regularization terms such as temporal smooth constraint and low-rank constraint (Nie et al., 2017; Thung et al., 2018; Zhou et al., 2013; Zhu et al., 2017). However, the length of historical visits and target score trajectories are fixed in most multi-task learning studies (Wang et al., 2019; Xie et al., 2016), which cannot satisfy the practical requirement. Traditional multi-task learning methods cannot handle individuals that do not meet the requirement for the length of historical visits, which further exacerbates data scarcity because individual data is quite limited in practical longitudinal AD studies. Therefore, it is urgent to develop a flexible framework that does not restrict the length of historical input and the length of target score trajectories.

More recently, people have witnessed the potentiality of deep learning methods to be powerfully expressive in capturing the intrinsic data patterns. One of the most competitive deep learning models in DPM is the recurrent neural network (RNN), which captures long-term temporal dependencies with its ability to memorize historical information in longitudinal data (Marinescu et al., 2019). Compared with the other two paradigms (i.e. time-series methods and multi-task learning methods), RNN-based methods require neither the prior knowledge of score trajectories nor sophisticated regularization terms (El-Sappagh et al., 2020; Mehdipour-Ghazi et al., 2019; Nguyen et al., 2020). Moreover, the characteristic of the recurrent unit in RNN makes it feasible to process variable-length data and predict the cognitive scores over an arbitrarily long period of time. We therefore propose an RNN-based DPM method to meet the flexibility requirements of the length of historical visits and target trajectories in practical applications.

In addition to the variable length problem for longitudinal data, effectively excavating the correlations among different modalities should also be carefully considered. AD clinical data usually comprise multiple heterogeneous yet complementary modalities, such as magnetic resonance imaging (MRI), positron emission tomography (PET), and demographics. Most existing studies investigated DPM problem with single modality (Zhou et al., 2013), while the complementary information across multiple modalities is not excavated. Compared with single-modality methods, combining multiple modalities can synthetically characterize AD individuals and yield more comprehensive insight into AD progression. To be specific, MRI and PET measure nerve cell injuries and the individual's beta-amyloid level, respectively. While the demographic data (e.g., site, gender, age, education level, ApoE4 gene, etc.) can also help to diagnose AD. The ApoE4 gene is known as the most risk factor for Alzheimer's disease (Kanekiyo and Bu, 2016; Spasov et al., 2019), and age, education level as well as other biomarkers are also verified as critical influencing factors of AD (Fleet et al., 2016; Kim et al., 2020; Williams et al., 2010). Hence, the combination of multi-modal data augments the prospects for a more accurate prediction. Moreover, these modalities are often complementary since they depict the same individuals from different aspects, whereas most existing works merely concatenate the multi-modal data on each historical visit (Nguyen et al., 2018, 2020; Thung et al., 2018) and do not take the complementary information into account, which leads to sub-optimal results.

Last but not least, some DPM models are developed based on the assumption of "data completeness", whereas in practical situations, missing data is a prevalent and severe problem which always exists in both imaging records and clinical scores. For example, the elderly patients may not show up at pre-agreed time points or even drop out from the study (Tabarestani et al., 2020), which is referred as "visit missing" where no brain imaging data (e.g., MRI and PET) are recorded. Besides, due to the high cost of the clinical examination, some individuals only have partial records on some visits (Liu et al., 2021). Likewise, we refer to this issue as "partial modality missing" where only one imaging modality is available. In addition, the clinical scores for many patients are missing at some time points. Directly dropping these individuals will inevitably result in information loss. To address this issue, most conventional methods consider handling missing data and modeling progression as two separate steps, addressing the missing issue before the training process (Stekhoven and Bühlmann, 2012; White et al., 2011; Zhou et al., 2013). In the "preprocessing" step, they first impute the missing data with mean or other statistical values based on the observed visits. Then they predict cognitive scores based on imputed data in the "prediction" step. The performance of this separate approach heavily depends on the imputation strategy that is irrelevant to the prediction task. Recently, some integrative approaches are proposed to address the missing issue during the training process (Mehdipour-Ghazi et al., 2019; Nguyen et al., 2018, 2020). Among them, the indicator matrix is one of the most commonly used strategies to indicate the missing data and alleviate the effect of these incomplete values in the training process (Zhou et al., 2013; Zhang et al., 2021). Instead of filtering out missing values with the indicator matrix, some imputation-based methods are proposed to further predict and impute the missing values on the current visit based on the previous visit (Jung et al., 2021; Nguyen et al., 2020).

To address the above challenges, we devise a unified framework that is (1) flexible enough to handle variable-length historical data and predict arbitrary-length score trajectories, (2) able to effectively exploit the intrinsic correlation between multiple modalities, and (3) capable of processing and imputing arbitrary-missing data. Specifically, we propose a deep latent representation collaborated sequence learning framework that integrates a deep multi-modality fusion module and an RNN-based sequence learning module for AD progression modeling. We first devise a multi-modality fusion module to exploit the underlying complementary information from different modalities and learn modality-shared latent representations at each historical visit. All individuals and modalities can be jointly exploited regardless of incomplete modality data, which provides the framework with flexibility to handle "partial-modality-missing" and "visit-missing" data. We assume that multiple modalities originate from the common latent representation, which essentially describes the data and reveals the underlying latent structure shared by different modalities. Thus, the observations of different modalities can be reconstructed through their respective degradation network with shared latent representations in the multi-modality fusion module. Based on the learned longitudinal representations, we utilize the RNN-based sequence learning module to flexibly process variable-length longitudinal data and model AD progression by predicting future cognitive score trajectories. Furthermore, the "Model Filling" strategy based on RNN is employed to handle the "visit-missing" issue, which predicts and imputes the missing data of the current visit based on the estimated values from the previous visit. We integrate the multi-modality fusion module and the sequence learning module into a unified framework for collaborative training to learn task-oriented representations and optimal network parameters.

The contributions of this paper are summarized as follows:

Table 1
Statistical information for demographics of 805 individuals in the original ADNI dataset. The first two lines describe the mean value, standard deviation, minimal value and maximal value of Age and Education in terms of different patient groups. The last line describes the individual numbers on different ApoE-ϵ 4 values in terms of different patient groups.

	CN (Total=226; 108 F/118 M)	MCI (Total=393; 140 F/253 M)	AD (Total=186; 87 F/99 M)
Age (mean±std/[min, max])	75.8 ± 5.0/[59.9-89.6]	74.9 ± 7.3/[54.4-89.3]	75.3 ± 7.6/[55.1-90.9]
Education (mean±std/[min–max])	16.0 ± 2.9/[6-20]	15.6 ± 3.0/[4-20]	14.7 ± 3.1/[4-20]
ApoE-ϵ 4 (individual number of 0, 1, 2)	(166, 55, 5)	(182, 166, 45)	(63, 87, 36)

CN = Cognitively Normal, MCI = Mild Cognitive Impairment, AD = Alzheimer's Disease, F = female, M = male.

Table 2
Number of observed individuals for different data sources at different visits in the original ADNI dataset.

Type	Baseline	M06	M12	M18	M24	M36	M48	M60	M72	M84	M96
MRI	805	725	675	282	479	50	0	0	0	0	0
PET	396	360	329	152	278	175	0	0	0	0	0
Demographics	805	0	0	0	0	0	0	0	0	0	0
MMSE	805	769	721	322	636	449	267	222	222	175	89
ADAS-Cog	805	769	721	322	636	450	267	222	222	173	86
CDR-Global	805	768	721	322	637	450	270	238	240	192	94
CDR-SOB	805	768	721	322	637	450	270	238	240	192	94

- The AD progression is investigated from the perspective of incomplete and variable-length longitudinal multi-modal data. This is quite different from existing works that only focus on complete or fixed-length data. We consider a flexible yet more challenging scenario in which individuals may suffer from (arbitrary) modality-missing issue, as well as the number of individual's historical visits and the length of target score trajectories being not prespecified.
- To address the above problem, we first design a deep multi-modality fusion module to handle the challenge of individuals with possible (arbitrary) modality-missing patterns in practical scenarios. The proposed module can automatically capture complementary information from incomplete multiple modalities, and thus obtain a common comprehensive representation to characterize individuals.
- Based on the designed multi-modality fusion module, we further propose a deep latent representation collaborated sequence learning framework to model AD progression of individuals with incomplete variable-length longitudinal multi-modal data. The proposed framework can predict the progression of AD measured by cognitive scores at each time point (indefinitely) in the future for individuals with even arbitrary modality-missing patterns.

2. Materials

2.1. Subjects

Data utilized in this paper are obtained from the ADNI dataset[1] (Jack Jr. et al., 2008), the goal of which is to explore whether the combination of longitudinal MRI, PET, and other biological biomarkers can be used to measure the progression of Cognitively Normal (CN) controls, Mild Cognitive Impairment (MCI) and early AD.

Our research is based on three modality data (i.e., MRI, PET, and demographics) and four cognitive scores (i.e., MMSE, ADAS-Cog, CDR-Global, and CDR-SOB) of 805 individuals from the ADNI dataset. More specifically, original MRI and PET are imaging data that will be respectively extracted as 93-dimension feature vectors, demographics consist of five demographic features (i.e., site, age, gender, education, and ApoE-ϵ 4). The demographics only contain single time point data, while the other three data sources (i.e., MRI, PET, and cognitive scores) are longitudinal data from multiple time points.

[1] The dataset is available in http://www.adni-info.org.

Based on the description of the original dataset, the date when the individual takes the first examination is called "baseline", and follow-up visits are named by the duration from the baseline time point. For example, M06 denotes the visit when the individual was examined six months after the first visit. In this paper, we utilize 805 individuals with multi-modal data from six visits (i.e., baseline, M06, M12, M18, M24, and M36) and clinical scores recorded at 11 visits (i.e., baseline, M06, M12, M18, M24, M36, M48, M60, M72, M84, and M96), including 186 AD individuals, 393 MCI individuals, and 226 CN individuals. The summary of demographic information of all these individuals is outlined in Table 1.

Each individual has complete baseline demographic information, yet there are missing modalities at some visits in longitudinal data. Table 2 shows the number of observed individuals for MRI, PET, demographics, and four cognitive scores at different time points. It can be seen that each individual has complete MRI, demographic, and score data at baseline. However, the number of observed individuals decreases over time due to many practical reasons such as the death of the individuals or dropout individuals.

2.2. Image processing

We first download 1.5T MR images from the ADNI website and follow the work in Zhou et al. (2019) to process the original data and extract Region Of Interest (ROI) based features. Specifically, these images are processed with the following steps: anterior commissure–posterior commissure (AC–PC) correction, intensity inhomogeneity correction, brain extraction, cerebellum removal, tissues segmentation, registration to a 93-ROI template (Kabani et al., 1998) and ROI labels projection. For each ROI, we finally use the gray matter tissue volume normalized with the intracranial volume as the feature representation. For PET data, the PET images are aligned to their corresponding T1 MR images with affine registration, then the average PET intensity value of each ROI is computed as the feature representation. Based on the 93-ROI template, we extract a 93-dimensional ROI-based feature vector from the specific modality (i.e., MRI or PET).

3. Proposed framework

3.1. Notations

Throughout this paper, we denote **MATRICES** as boldface uppercase letters, **vectors** as boldface lowercase letters, and *scalars* as normal italic letters. For an arbitrary matrix $\mathbf{A}$, $a_{i,j}$ denotes the (i, j)-th element in $\mathbf{A}$. For a better understanding, we summarize the main notations used in this paper including their respective meanings in Table 3.

Table 3
Summation of main notations used in this paper.

Notation	Meaning	Notation	Meaning
Notations for modality features		*Notations for RNN-based sequence learning network*	
N	total number of individuals	$\mathbf{s}_t$	RNN input at tth time point
T	maximal number of historical visits	$\mathbf{z}_t$	hidden state at tth time point
n_t	number of available individuals at tth time point	$\mathbf{c}_t$	cell state at tth time point
d_v	dimension of feature vector for vth modality	$\tilde{\mathbf{c}}_t$	candidate state at tth time point
$\mathbf{X}_t^{(v)}$	feature matrix of vth modality at tth time point	$\mathbf{f}_t$	forget gate
$\mathcal{X}^{(v)} = \left\{\mathbf{X}_1^{(v)}, \mathbf{X}_2^{(v)}, \ldots, \mathbf{X}_T^{(v)}\right\}$	longitudinal features of vth modality	$\mathbf{i}_t$	input gate
$\mathbf{x}_t^v$	feature vector for vth modality of one individual at tth time point	$\mathbf{o}_t$	output gate
$\boldsymbol{x}_t = \{\mathbf{x}_t^{(1)}, \mathbf{x}_t^{(2)}, \mathbf{x}_t^{(3)}\}$	clinical records of one individual at tth time point	$\mathbf{W}_f, \mathbf{W}_i, \mathbf{W}_o, \mathbf{W}_c$	weight matrices for LSTM
Notations for cognitive scores		$\mathbf{b}_f, \mathbf{b}_i, \mathbf{b}_o, \mathbf{b}_c$	bias vectors for LSTM
F	number of target time points	$\boldsymbol{\delta}_{i,t}$	masking vector indicating missing values in $\mathbf{s}_{i,t}$
$\mathbf{Y}_t$	cognitive score matrix at t-th time point	$\odot$	dot product operator
$\mathcal{Y} = \{\mathbf{Y}_1, \ldots \mathbf{Y}_T, \mathbf{Y}_{T+1}, \ldots, \mathbf{Y}_{T+F}\}$	overall cognitive score data	$\tanh(\cdot)$	hyperbolic tangent function
$\mathcal{Y}_p = \{\mathbf{Y}_1, \ldots \mathbf{Y}_T\}$	cognitive score matrices at historical time points	$\sigma(\cdot)$	sigmoid activation function
$\mathcal{Y}_f = \{\mathbf{Y}_{T+1}, \ldots \mathbf{Y}_{T+F}\}$	cognitive score matrices at future time points	*Notations for objective function*	
$\mathbf{y}_t$	ground truth score vector of one individual at tth time point	$o_{i,t}^{(v)}$	indicator variable for v-th modality of i-th individual at t-th time point
$\tilde{\mathbf{y}}_t$	predicted score vector of one individual at tth time point	$p_{(i,j),t}, q_{(i,j),t}$	indicator variables for j-th feature of i-th individual at t-th time point
Notations for multi-modality fusion network		$\boldsymbol{\Theta}_v$	parameters for the degradation network
$\mathbf{H}_t$	latent representation matrix at tth time point	$\boldsymbol{\Omega}$	parameters for the sequence learning network
$\mathbf{h}_t$	latent representation vector of one individual at tth time point	$f_v\left(\cdot; \boldsymbol{\Theta}_v\right)$	degradation network for vth modality
Notations for metrics		$g(\cdot; \boldsymbol{\Omega})$	sequence learning network
$m_{k,t}$	number of observed individuals for score type k at tth time point	α_1, α_2	hyperparameters for objective function
$\mathbf{Y}_{(:,k),t}$	ground truth vector of all observed individuals for score type k at at tth time point	$\mathcal{F}(\cdot)$	slice operator
$\tilde{\mathbf{Y}}_{(:,k),t}$	predicted score vector of all observed individuals for score type k at tth time point	$\lvert\cdot\rvert$	sum of absolute values of entries

Assuming the total number of individuals to be N and each individual contains modality records from up to T historical visits, we denote the v-th modality data as $\mathcal{X}^{(v)} = \left\{\mathbf{X}_1^{(v)}, \mathbf{X}_2^{(v)}, \ldots, \mathbf{X}_T^{(v)}\right\}$, where $\mathbf{X}_t^{(v)} \in \mathbb{R}^{n_t \times d_v}$ represents the feature matrix of vth modality at tth visit, which consists of n_t individuals and d_v features (for each individual). Since our study is based on MRI, PET and demographics, we denote them respectively as $\mathcal{X}^{(1)}$, $\mathcal{X}^{(2)}$ and $\mathcal{X}^{(3)}$, where $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ are longitudinal data containing MRI and PET records from T historical visits, respectively. $\mathcal{X}^{(3)} \in \mathbb{R}^{N \times d_3}$ is complete and only contains baseline data. Although $\mathcal{X}^{(3)}$ is composed of mostly static biomarkers, features such as age, education, and ApoE4 gene play an important role in disease diagnosis, which contribute to the comprehensive characterization of AD individuals (Kanekiyo and Bu, 2016; Kim et al., 2020; Williams et al., 2010). Therefore, we expand $\mathcal{X}^{(3)}$ to each time point for the learning of latent representations, i.e. $\mathbf{X}_t^{(3)} = \mathcal{X}^{(3)}$ with $t = 1, 2, \ldots, T$. Note that we are focusing on processing variable-length longitudinal data, the number of available[2] individuals n_t varies at each time point because not all individuals contain data at the tth time point. There exist severe missing issues in $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$, which has been explicated in Section 2 and will be handled hereinafter.

Let $\mathcal{Y} = \left\{\mathbf{Y}_1, \ldots \mathbf{Y}_T, \mathbf{Y}_{T+1}, \ldots, \mathbf{Y}_{T+F}\right\}$ represents the cognitive score set at $(T + F)$ time points, where we have at most T historical visits for each individual and aim to predict F subsequent time points. $\mathbf{Y}_t \in \mathbb{R}^{n_t \times d_y}$ denotes the score matrix on the t-th visit with n_t individuals and d_y scores (in this paper $d_y = 4$, i.e. MMSE, CDR-Global, CDR-SOB, and ADAS-Cog). Likewise, $\mathcal{Y}$ has also missing scores on some visits. We split $\mathcal{Y}$ into two parts, one is the historical cognitive scores $\mathcal{Y}_p = \left\{\mathbf{Y}_1, \ldots, \mathbf{Y}_T\right\}$ at history time points, and the other is the target scores $\mathcal{Y}_f = \left\{\mathbf{Y}_{T+1}, \ldots, \mathbf{Y}_{T+F}\right\}$ at F future time points. It is noteworthy that the fixed number F is just for facilitating the experimental verification, our model can theoretically predict any number of time points in the future.

3.2. Problem definition

Fig. 1 illustrates the problem considered in this paper, with the blue part representing the training set and the green part representing the testing set. We train our model with the blue part and predict the green target scores marked with "?" based on green history input. Each row represents one individual, each column represents one time point, and each layer represents one data source. The block in the ith row and the tth column of the v-th layer represents the feature vector of vth modality of individual i at tth time point, which is denoted as $\mathbf{x}_{i,t}^{(v)}$. For each modality in Fig. 1, the number of blocks (including white ones marked with "?") in column t in one layer indicates the number of available individuals n_t at time point t.

As can be seen from Fig. 1, we focus on modeling AD progression based on the longitudinal multi-modal data that contains missing records at some time points. Furthermore, in practical situations, the number of visits varies in individuals, for which we focus on processing variable-length data and making predictions on the score trajectories over an arbitrary period of time, which provides more flexibility in practical applications.

[2] In this paper, "*observed*" means individuals have complete modalities on a certain visit, while "*available*" means the prespecified visits even without modality records. Namely, the semantic range of "*available*" is broader than "*observed*".

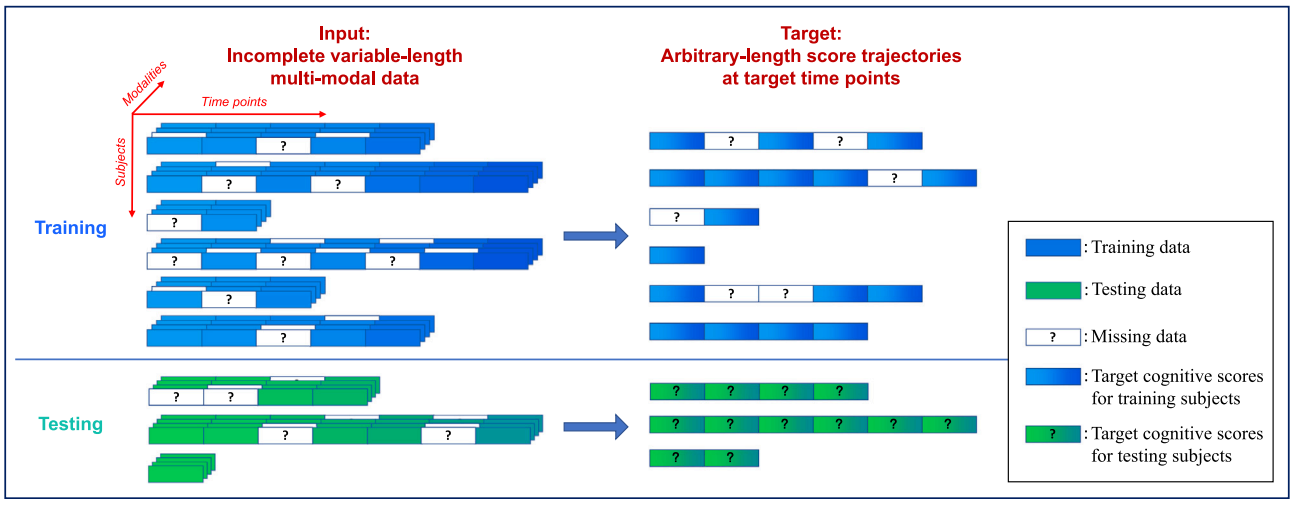

Fig. 1. Illustration of DPM problem based on incomplete variable-length longitudinal multi-modal data, where each block represents the feature vector of a certain modality.

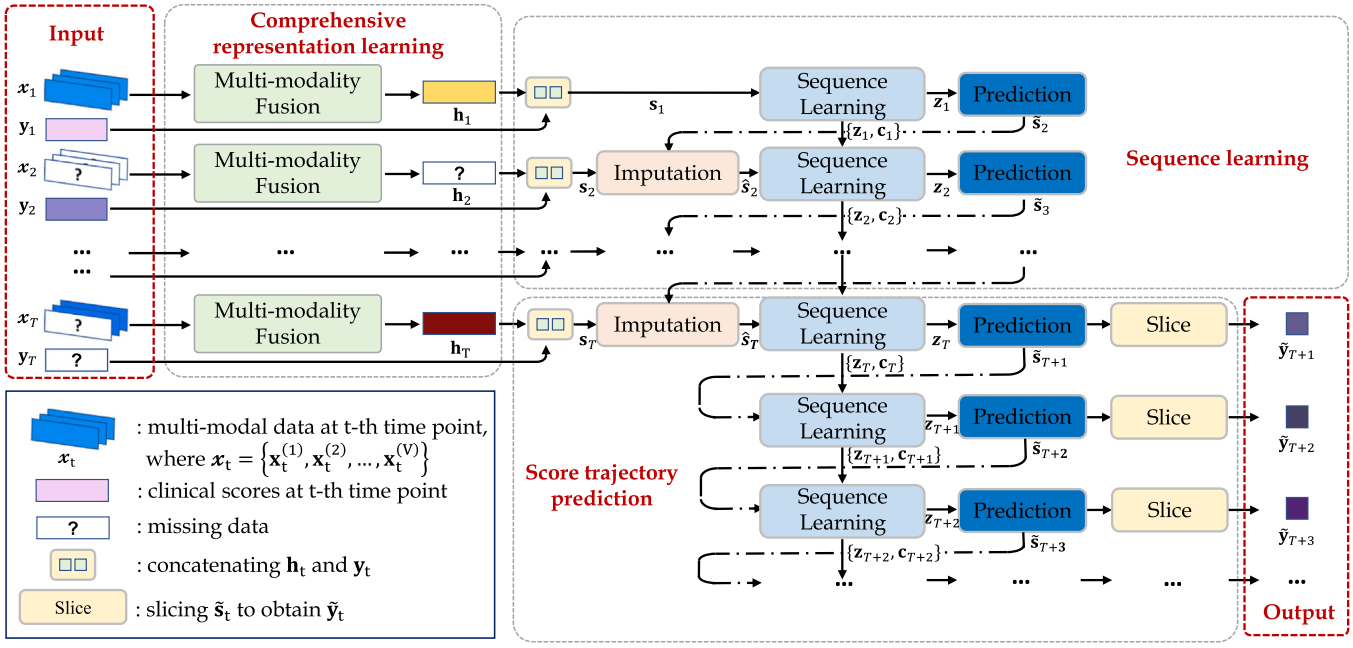

Fig. 2. Illustration of the proposed deep latent representation collaborated sequence learning framework.

3.3. Proposed model

To address the above challenge, we propose a unified DPM framework composed of the deep multi-modality fusion module and the RNN-based sequence learning module, as illustrated in Fig. 2.

3.3.1. Model overview

The deep multi-modality fusion module learns latent representations on each historical visit based on longitudinal multi-modal data. During this process, the degradation networks (Zhang et al., 2019) in the fusion module will explore the complementary information between different modalities, and latent representations are learned based on the observed modalities.

After that, we concatenate the longitudinal latent representations with cognitive scores on corresponding visits and feed the concatenated vectors into RNN for sequence learning. The missing part will be replaced by estimated values from the dense layer (also known as the fully connected layer) based on the hidden state on the previous visit.

Since RNN encodes the underlying temporal characteristic of each individual, when predicting future scores, the estimated values at the previous visit will be used as input despite that there is no available longitudinal input at future time points.

We employ the collaborative training strategy for the multi-modality fusion module and the sequence learning module to learn task-oriented representations and optimal parameters.

3.3.2. Deep multi-modality fusion module

For each time point t, we combine MRI, PET, and demographics at the current visit to learn the latent representation matrix $\mathbf{H}_t \in \mathbb{R}^{n_t \times d_h}$, where d_h denotes the size of latent representations. $\mathbf{H}_t$ comprehensively characterizes AD individuals and contributes to AD progression modeling.

Conventional multi-modal methods learn modality-shared representations by mapping multi-modal data into a common space and require a certain degree of similarity between the learned representation and each original modality. Different from traditional methods, considering that different modalities depict the sample individual from different

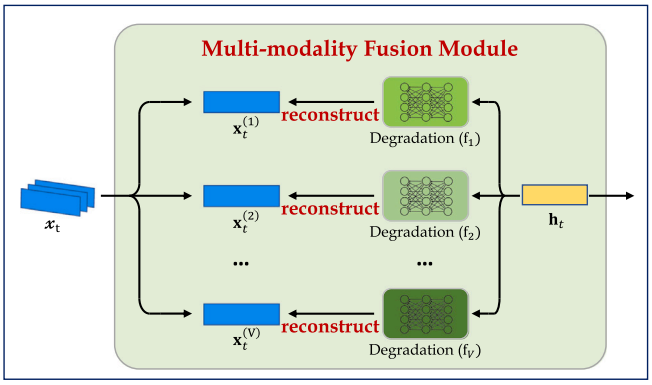

Fig. 3. Illustration of the multi-modality fusion module, where $\mathbf{x}_t^{(v)}$ represents the feature of v-th modality and $\mathbf{h}_t$ represents the learned representation at t-th visit for individual i. "Degradation (f_v)" denotes the degradation layer for v-th modality.

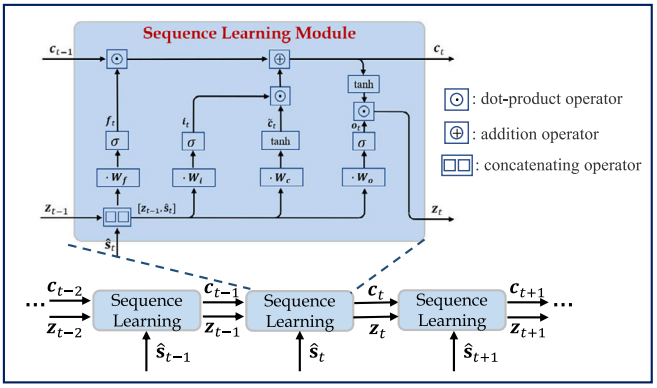

Fig. 4. Illustration of sequence learning module, where $\hat{\mathbf{s}}_t, \mathbf{c}_t, \mathbf{z}_t$, represent the input, the cell state and the hidden state at t-th visit, respectively.

aspects, we explore the complementary information between different modalities and regard the desired representation as the comprehensive characterization of individuals. Concretely, we focus on finding the complete representation that contains information from all original modalities. To this end, we construct the degradation functions for each modality and require the learned representation to reconstruct original modal data through degradation functions. In this way, we ensure the representation to encode complete information from all observed modalities.

Concretely, in this paper, we construct the deep multi-modality fusion module based on degradation networks, as shown in Fig. 3. The degradation layer for each modality comprises several dense layers. To reduce parameter numbers and avoid network overfitting, the same modality among different visits are passed through the same degradation layer. On each visit, the latent representations will be passed through degradation layers to reconstruct original modalities.

For partial-modality-missing cases, the learned representation will be required to reconstruct the available modalities to guarantee the sufficiency in utilizing multi-modal data. Note that for each individual, those visit-missing time points will also take part in the fusion process and their corresponding representations can be obtained based on complete $\mathbf{x}_t^{(3)}$ (i.e. demographics). However, from our perspective, these representations lack temporal information and may interfere with the sequence learning, for which we still regard these visits as missing visits in the following sequence learning module.

To reconstruct the corresponding modality via the v-th degradation network using the latent representation, we design the loss function $\mathcal{L}_v$ for the v-th degradation network as following Eq. (1):

$$\mathcal{L}_v = \sum_{i=1}^{N}\sum_{t=1}^{T}\left\| o_{i,t}^{(v)}\left(f_v\left(\mathbf{h}_{i,t};\boldsymbol{\Theta}_v\right)-\mathbf{x}_{i,t}^{(v)}\right)\right\|_2^2 \tag{1}$$

where $f_v\left(\cdot;\boldsymbol{\Theta}_v\right)$ is the degradation layer for the vth modality with parameters $\boldsymbol{\Theta}_v$ including multiple dense layers, $\mathbf{x}_{i,t}^{(v)}$ denotes the v-th modality data of i-th individual at the t-th time point, and $o_{i,t}^{(v)}$ indicates the missing data; if the ith individual has data in the vth modality on tth visit, $o_{i,t}^{(v)}=1$, otherwise $o_{i,t}^{(v)}=0$.

Note that the multi-modality fusion module includes V degradation networks, we develop the reconstruction loss $\mathcal{L}_{rec}$ as follows:

$$\mathcal{L}_{rec} = \sum_{v=1}^{V}\mathcal{L}_v = \sum_{v=1}^{V}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\| o_{i,t}^{(v)}\left(f_v\left(\mathbf{h}_{i,t};\boldsymbol{\Theta}_v\right)-\mathbf{x}_{i,t}^{(v)}\right)\right\|_2^2 \tag{2}$$

3.3.3. RNN based sequence learning module

Based on the learned longitudinal representation, we encode and capture the temporal correlations of these representations with RNN-based sequence learning module. In this paper, we employ long short-term memory (LSTM) network (Gers et al., 2000) as the basic RNN model for sequence learning. Note that LSTM is not the only choice in our framework, it can be replaced by any other RNN models such as gated recurrent unit (GRU) (Cho et al., 2014) and minimal gated unit (MGU) (Zhou et al., 2016).

(a) Long short-term memory network

The structure of LSTM is shown in Fig. 4 and update formulations of LSTM are as follows:

$$\mathbf{f}_t = \sigma\left(\left[\mathbf{z}_{t-1},\hat{\mathbf{s}}_t\right]\mathbf{W}_f+\mathbf{b}_f\right) \tag{3}$$

$$\mathbf{i}_t = \sigma\left(\left[\mathbf{z}_{t-1},\hat{\mathbf{s}}_t\right]\mathbf{W}_i+\mathbf{b}_i\right) \tag{4}$$

$$\mathbf{o}_t = \sigma\left(\left[\mathbf{z}_{t-1},\hat{\mathbf{s}}_t\right]\mathbf{W}_o+\mathbf{b}_o\right) \tag{5}$$

$$\tilde{\mathbf{c}}_t = \tanh\left(\left[\mathbf{z}_{t-1},\hat{\mathbf{s}}_t\right]\mathbf{W}_c+\mathbf{b}_c\right) \tag{6}$$

$$\mathbf{c}_t = \mathbf{f}_t\odot\mathbf{c}_{t-1}+\mathbf{i}_t\odot\tilde{\mathbf{c}}_t \tag{7}$$

$$\mathbf{z}_t = \mathbf{o}_t\odot\tanh\left(\mathbf{c}_t\right) \tag{8}$$

where $\hat{\mathbf{s}}_t$, $\mathbf{c}_t$, $\tilde{\mathbf{c}}_t$, and $\mathbf{z}_t$ represent the current input, the cell state, the candidate state, and the hidden state at the tth visit, respectively. $\{\mathbf{W}_f,\mathbf{W}_i,\mathbf{W}_o,\mathbf{W}_c,\mathbf{b}_f,\mathbf{b}_i,\mathbf{b}_o,\mathbf{b}_c\}$ are network parameters. $\odot$ denotes the dot product operator, $\tanh(\cdot)$ denotes the hyperbolic tangent function, and $\sigma(\cdot)$ denotes the sigmoid activation function. On each visit t, the current input $\mathbf{s}_t$, the cell state $\mathbf{c}_{t-1}$ and the hidden state $\mathbf{z}_{t-1}$ of the previous visit are simultaneously fed into LSTM to update $\mathbf{c}_t$ and $\mathbf{z}_t$. The hidden state $\mathbf{z}_{t-1}$ can be interpreted as fusing history information of all past visits up to the current visit, for which $\mathbf{z}_{t-1}$ is used to predict the input $\mathbf{s}_t$.

For different time points, LSTM always processes data with the same parameters. In other words, each time point will be processed by the same LSTM unit regardless of the sequence length of input data, for which LSTM can process variable-length input.

(b) Processing longitudinal data with sequence learning module

Next, we are going to explicate how to process longitudinal data with the sequence learning module. Assuming that the longitudinal latent representations have been obtained from the multi-modality fusion module, we take the ith individual for example and denote its longitudinal latent representations as $\{\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_T\}$.[3]

As can be seen from Fig. 2, for each visit t, we concatenate the hidden representations $\mathbf{h}_t$ and cognitive score vector $\mathbf{y}_t$ as the longitudinal input $\mathbf{s}_t$, i.e., $\mathbf{s}_t=\left[\mathbf{h}_t,\mathbf{y}_t\right]\in\mathbb{R}^{d_h+d_y}$. To address the data missing issue (i.e., visit missing or score missing or both) in $\mathbf{s}_t$, we utilize the "Model Filling" method, imputing $\mathbf{s}_t$ with predicted input $\tilde{\mathbf{s}}_t$ from the

[3] For notational simplicity, we leave out the subscript i throughout this paper unless explicitly required.

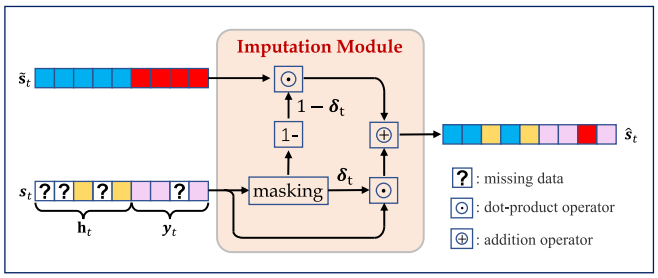

Fig. 5. Illustration of the imputation module, where $\mathbf{s}_t = [\mathbf{h}_t, \mathbf{y}_t]$. The masking vector $\boldsymbol{\delta}_t$ indicates missing values in $\mathbf{s}_t$, which will be imputed with corresponding entries in $\tilde{\mathbf{s}}_t$.

dense layer based on the hidden state $\mathbf{z}_{t-1}$. To be specific, we first feed $\mathbf{z}_{t-1}$ into the dense layer to predict $\tilde{\mathbf{s}}_t$:

$$\tilde{\mathbf{s}}_t = \mathbf{z}_{t-1}\mathbf{W}_d + \mathbf{s}_{t-1} \tag{9}$$

where $\mathbf{W}_d$ denotes the weight parameter of the dense layer. Note that the estimated value $\tilde{\mathbf{s}}_t$ is dependent on the hidden state $\mathbf{z}_{t-1}$, which is not available at the first time point. If data are missing at the first time point, the mean value across all time points of all the training individuals will be used for imputation.

Based on the estimated value $\tilde{\mathbf{s}}_t$, we perform element-wise imputation on $\mathbf{s}_t$ in the imputation layer:

$$\hat{\mathbf{s}}_t = \boldsymbol{\delta}_t \odot \mathbf{s}_t + (1 - \boldsymbol{\delta}_t) \odot \tilde{\mathbf{s}}_t \tag{10}$$

where $\boldsymbol{\delta}_t \in \mathbb{R}^{d_h+d_y}$ is a masking vector that indicates missing values in $\mathbf{s}_t$, which is defined as follows:

$$\delta_{t,d} = \begin{cases} 1, & s_{t,d} \text{ is complete} \\ 0, & s_{t,d} \text{ is missing} \end{cases} \tag{11}$$

where $\delta_{t,d}$ and $s_{t,d}$ denote the d-th component in $\boldsymbol{\delta}_t$ and $\mathbf{s}_t$, respectively. The structure of the imputation module is shown in Fig. 5.

Finally, we feed the imputed data $\tilde{\mathbf{s}}_t$, cell state $\mathbf{c}_{t-1}$, and hidden state $\mathbf{z}_{t-1}$ into the LSTM unit to update the current states $\mathbf{c}_t$ and $\mathbf{z}_t$.

During the longitudinal data processing stage, we minimize the fitting loss $\mathcal{L}_{fit}$ as follows:

$$\mathcal{L}_{fit} = \sum_{i=1}^{N}\sum_{t=1}^{T-1}\left|\mathbf{p}_{i,t} \odot \left(g\left(\mathbf{s}_{i,t};\boldsymbol{\Omega}\right) - \mathbf{s}_{i,t+1}\right)\right| \tag{12}$$

where $|\cdot|$ denotes the sum of absolute values of entries, $\mathbf{s}_{i,t} \in \mathbb{R}^{1\times(d_h+d_y)}$ denotes the input vector of individual i at the tth time point and $g(\cdot;\boldsymbol{\Omega})$ denotes the network including the imputation layer, the sequence learning module, and the dense layer with the parameters $\boldsymbol{\Omega}$. $\mathbf{p}_{i,t}$ is an indicator vector with $p_{(i,j),t} = 1$ if the j-th feature of the i-th individual is available at t-th time point and $p_{(i,j),t} = 0$ otherwise.

(c) Modeling future AD progression with sequence learning module

Finally, we elaborate on how to predict the cognitive scores at target time points, as shown in Fig. 2. Likewise, we take the ith individual for example, when it comes to future progression modeling, there is no longer available longitudinal input from the $(T+1)$-th time point onward. Since LSTM encodes the underlying temporal characteristic of individuals, the estimated value based on the hidden state at the previous time point is directly fed into LSTM for prediction. Recall that $\tilde{\mathbf{s}}_t = \left[\tilde{\mathbf{h}}_t, \tilde{\mathbf{y}}_t\right]$, hence the predicted scores $\tilde{\mathbf{y}}_{T+1}$ can be easily obtained by taking the last d_y dimensions of $\tilde{\mathbf{s}}_{T+1}$ through the slice layer. The same procedure is adopted to subsequent time points.

The prediction error $\mathcal{L}_{error}$ for the sequence learning module is defined as follows:

$$\mathcal{L}_{error} = \sum_{i=1}^{N}\sum_{t=T+1}^{T+F}\left|\mathbf{q}_{i,t} \odot \left(\mathcal{F}(\tilde{\mathbf{s}}_{i,t}) - \mathbf{y}_{i,t}\right)\right| \tag{13}$$

where $\mathcal{F}(\cdot)$ denotes the slice operator of on $\tilde{\mathbf{s}}_{i,t}$. $\mathbf{q}_{i,t}$ is an indicator matrix with $q_{(i,j),t} = 1$ if the j-th score of the i-th individual is available at the t-th time point, and $p_{(i,j),t} = 0$ otherwise.

3.3.4. Collaborative model learning

The most straightforward way to address the DPM problem is to separately train the above two modules by first preprocessing the multi-modal data with the multi-modality fusion module based on the objection function $\mathcal{L}_{rec}$, and then feeding the learned representation into sequence learning module for AD progression modeling based on $\mathcal{L}_{fit}$ and $\mathcal{L}_{error}$. In contrast to this two-step processing manner, we assume that the collaborative training of these two modules will facilitate learning the task-oriented representation and promote parameter learning. In other words, the classification results of LSTM will be fed back to the fusion module, making the learned representations better fit for specific prediction tasks, and the optimal representations will assist the network in learning optimal parameters.

Therefore, we combine two modules and train the unified framework with the following composite objective function:

$$\begin{aligned}\mathcal{L} &= \alpha_1\mathcal{L}_{rec} + \alpha_2\mathcal{L}_{fit} + \mathcal{L}_{error} \\ &= \alpha_1\sum_{v=1}^{V}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|o_{i,t}^{(v)}\left(f_v\left(\mathbf{h}_{i,t};\boldsymbol{\Theta}_v\right) - \mathbf{x}_{i,t}^{(v)}\right)\right\|_2^2 \\ &+ \alpha_2\sum_{i=1}^{N}\sum_{t=1}^{T-1}\left|\mathbf{p}_{i,t}\odot\left(g\left(\mathbf{s}_{i,t};\boldsymbol{\Omega}\right) - \mathbf{s}_{i,t+1}\right)\right| + \sum_{i=1}^{N}\sum_{t=T+1}^{T+F}\left|\mathbf{q}_{i,t}\odot\left(\mathcal{F}(\tilde{\mathbf{s}}_{i,t}) - \mathbf{y}_{i,t}\right)\right|\end{aligned} \tag{14}$$

where α_1, α_2 are hyperparameters. It is important to emphasize that this loss function is only calculated on the observed values of the original data, missing data is not taken into account when computing the loss. We will verify in the following experiment that this organic combination does contribute to better results.

3.3.5. Testing scenario

Suppose that we obtain a trained model, for given individuals, we aim to predict their score trajectories at K target time points. If it is the first visit of the individuals, we assume that they have the complete baseline multi-modal data $(\{\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_1^{(V)}\}, \mathbf{y}_1)$, then the baseline multi-modal data $\{\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_1^{(V)}\}$ is first fed into the multi-modality fusion module to obtain a baseline latent representation $\mathbf{h}_1$. Then $\mathbf{h}_1$ is concatenated with the baseline scores $\mathbf{y}_1$ as baseline input $\mathbf{s}_1$ and fed into LSTM for sequence learning. Next, LSTM will process $\mathbf{s}_1$ and update the current state $\mathbf{z}_1$. $\mathbf{z}_1$ will be passed through the dense layer to predict input $\tilde{\mathbf{s}}_2$. Since the individuals only have baseline data, the estimated input $\tilde{\mathbf{s}}_2$ will be directly used as the LSTM input of the next time point to update $\mathbf{z}_2$ and predict $\tilde{\mathbf{s}}_3$. The whole sequence learning process will be repeated until $\tilde{\mathbf{s}}_{K+1}$ is obtained. Finally, $\{\tilde{\mathbf{s}}_2, \tilde{\mathbf{s}}_3, \ldots, \tilde{\mathbf{s}}_{K+1}\}$ are fed into the slice layer to obtain the predicted scores $\{\tilde{\mathbf{y}}_2, \tilde{\mathbf{y}}_3, \ldots, \tilde{\mathbf{y}}_{K+1}\}$.

Likewise, when individuals have multiple historical visits, we first feed the longitudinal multi-modal data into the multi-modality fusion module to learn longitudinal representations. However, in fusion module, those visits that contain neither the MRI record nor the PET record will be marked as missing visits in the sequence learning module. In the sequence learning module, the missing values will be imputed with the estimated values from the previous visit. Then the sequence learning module employs the same procedure as baseline individuals to predict the target score trajectories.

4. Experiment

4.1. Experiment settings

In the experiment, we predict the progression of given individuals by predicting their clinical scores (MMSE, ADAS-Cog, CDR-Global, and CDR-SOB) at up to 15 target time points (with a six-month interval between consecutive time points). To verify the performance of the proposed method, we compare it with one multi-task learning method,

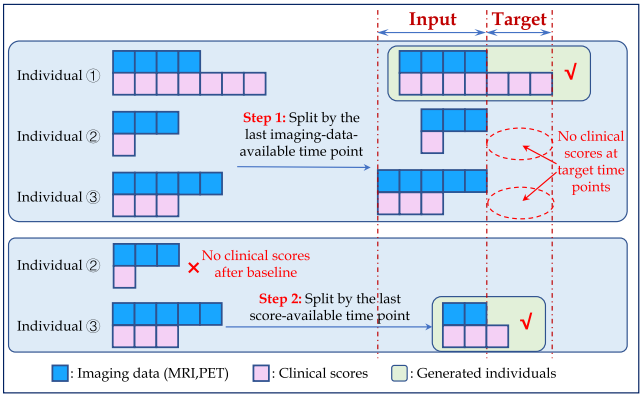

Fig. 6. Illustration of generating target time points from the original data.

three state-of-the-art RNN-based methods, and the variants of three benchmark models with respect to different missing-imputation strategies. Furthermore, we conduct several ablation experiments to verify the hypotheses employed in our method: the effectiveness of utilizing multi-modal data, the effectiveness of the multi-modality fusion module, and the effectiveness of the collaborative training strategy. The source code used in this paper can be found at https://github.com/solerxl/Code_For_MIA_2022.

In the following section, we successively elaborate on data preprocessing steps, performance metrics, comparison methods, regression results, and further analysis in our experiments.

4.1.1. Data preprocessing

1. *Interval alignment*: There are two kinds of intervals in the original dataset, i.e. 6-month and 12-month. Since RNN requires data with a fixed time interval, we first unify the time interval to 6-month for model learning and use demographics for multi-modality fusion for the additional visits where no MRI and PET data are available (i.e. M30, M42, M54, M66, M78, M90).
2. *Age processing*: As in practical situations, the age of the individuals will grow over time. To this end, we first regard the age in the original demographics as baseline age and increase the age at subsequent visits based on the duration from the baseline. For example, if a individual is 65 years old at baseline, then the age will be 65.5 at M06.
3. *Generating target time points*: Here we elaborate on how to generate historical visits and target time points from the original longitudinal data. The whole generation process is shown in Fig. 6. Recall that our model is able to predict the score trajectory of arbitrary length. To reflect such a scenario, for each individual, we regard the last available time point for imaging data (MRI/PET) as the end of historical visits and the subsequent time points as the target time points to be predicted (Individual 1 in Fig. 6). In this way, some individuals will contain no clinical scores at target time points (Individual 2 and 3 in Fig. 6). To maximize data utilization, we remove those individuals that do not contain any score data after baseline (Individual 2 in Fig. 6). For the remaining individuals, we predict the last time point containing clinical scores and treat the previous time points as historical visits (Individual 3 in Fig. 6). Finally, we obtain a dataset of 773 individuals, of which the length of target time points ranges from 1 to 15. The distributions of the length of historical visits and target time points of the obtained dataset are shown in Fig. 7.

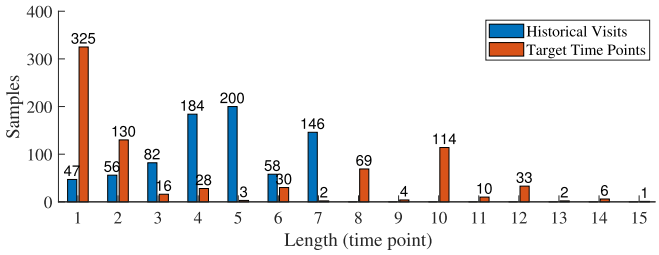

Fig. 7. Distribution of the length of historical visits and the length of target time points.

4.1.2. Evaluation metrics

We evaluate the prediction performance of the model with mean absolute error (MAE) (Mehdipour-Ghazi et al., 2019) and weighted correlation coefficient (wR) (Duchesne et al., 2009; Ito et al., 2011; Stonnington et al., 2010), which are wildly used in DPM-related literature. The two metrics are defined as

$$\mathrm{MAE}(k) = \sum_{t=1}^{F} \frac{1}{m_{k,t} F} \left| \mathbf{Y}_{(:,k),t} - \tilde{\mathbf{Y}}_{(:,k),t} \right| \tag{15}$$

$$\mathrm{wR}(k) = \frac{\sum_{t=1}^{F} \mathrm{Corr}\left(\mathbf{Y}_{(:,k),t}, \tilde{\mathbf{Y}}_{(:,k),t}\right) m_{k,t}}{\sum_{t=1}^{F} m_{k,t}} \tag{16}$$

For a certain score type k (e.g. MMSE/CDR-Global/CDR-SOB/ADAS-Cog), $\mathbf{Y}_{(:,k),t}$ and $\tilde{\mathbf{Y}}_{(:,k),t}$ denote the corresponding ground truth score vector and predicted score vector of all observed individuals at tth time point. $m_{k,t}$ denotes the number of observed individuals for score type k at t-th time point. $\mathrm{Corr}(\mathbf{y}, \tilde{\mathbf{y}})$ represents the correlation coefficient between the ground truth $\mathbf{y}$ and the predicted scores $\tilde{\mathbf{y}}$, which is calculated as:

$$\mathrm{Corr}(\mathbf{y}, \tilde{\mathbf{y}}) = \frac{\sum_i \left(y_i - \bar{y}\right)\left(\tilde{y}_i - \bar{\tilde{y}}\right)}{\sqrt{\sum_i \left(y_i - \bar{y}\right)^2} \sqrt{\sum_i \left(\tilde{y}_i - \bar{\tilde{y}}\right)^2}} \tag{17}$$

where y_i and $\tilde{y}_i$ represent the i-th element of $\mathbf{y}$ and $\tilde{\mathbf{y}}$, $\bar{y}$ and $\bar{\tilde{y}}$ represent the mean value of $\mathbf{y}$ and $\tilde{\mathbf{y}}$.

For MAE, the lower value indicates the better performance. In contrast, the higher wR indicates the better performance. It is important to note that the above metrics only measure the prediction results on the complete data. In the experiments, we use a 10-fold cross-validation strategy with 10 repetitions to evaluate the average performance of different methods. To determine the optimal hyperparameters for the model, in each fold, we further partition the training data into two non-overlapping subsets using an 8:2 ratio, with one subset used for model training and the other used for model evaluation. The optimal hyperparameters are selected based on the minimum MAE value on predicting MMSE. After determining the optimal hyperparameters, we train the model on the whole training set and report the prediction results on the testing set.

Moreover, to check the statistical significance of our method, we performed the t-test based on the results in terms of MAE and wR on four cognitive scores at the 95% confidence level (Dietterich, 1998).

4.2. Competing methods

First of all, we compare our model with three benchmark models: support vector regression(SVR), Lasso, and GRU. We design their variants according to different missing filling strategies. Taking the ith individual as an example, assuming that its jth feature of the vth modality at the tth visit (i.e., $x^{(v)}_{(i,j),t}$) is missing, we consider the following missing imputation strategies:

Table 4
Hyperparameter search space for different methods.

Model	Hyperparameter	Range
Lasso variants	λ	$(10^{-3} - 10^{3})$
SVR variants	Kernel	RBF
	C	$(10^{-2} - 10^{2})$
	γ	$(10^{-2} - 10^{0})$
	ϵ	$(10^{-2} - 10^{2})$
GRU variants, MinimalRNN, Our model	Size of hidden parameters (d_h)	$\{64, 128, 256\}$
	Learning rate	$(10^{-3} - 10^{-1})$
	Weight decay	$(10^{-3} - 10^{-1})$
	α_1 (if any)	$(10^{-2} - 10^{0})$
	α_2 (if any)	$(10^{-1} - 10^{1})$
LSTM-P,LSTM-T	Size of hidden parameters (d_h)	$\{16, 32, 48, 64, 80, 96\}$
	Learning rate	$\{5\times10^{-5}, 5\times10^{-4}, 5\times10^{-3}, 5\times10^{-2}\}$
	Weight decay	$\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$
	α_2	$(10^{-2} - 10^{0})$

1. *Mean Filling*: The mean value of the corresponding feature of $\mathbf{x}_i$ at other observed visits is utilized to impute the missing data:

$$x^{(v)}_{(i,j),t} = \frac{1}{\left|t^{(v)}_{i,j}\right|} \sum_{m \in t^{(v)}_{i,j}} x^{(v)}_{(i,j),m} \tag{18}$$

where $t^{(v)}_{i,j}$ is the set of observed visits defined as $t^{(v)}_{i,j} = \left\{ m \mid x^{(v)}_{(i,j),m} is\ not\ missing \right\}$, $\left|t^{(v)}_{i,j}\right|$ denotes the cardinality of $t^{(v)}_{i,j}$. If $\left|t^{(v)}_{i,j}\right| = 0$, then value 0 is used for imputation.

2. *Forward Filling*: The corresponding feature at the previous observed visit of $\mathbf{x}_i$ is used for imputation, that is:

$$x^{(v)}_{(i,j),t} = x^{(v)}_{(i,j),t-1} \tag{19}$$

If there is no observed data before current visit t, value 0 is used for imputation

3. *Linear Filling*: In the "Linear Filling" strategy, a linear function is established based on the corresponding features of the two closest observed visits before and after the current visit. The corresponding imputation value is calculated by substituting the current visit into the established linear function. Assuming that the closest forward observed visit is t_1 and the closest backward observed visit is t_2, it is easy to establish the linear function for $x^{(v)}_{(i,j),t}$ as follows:

$$x^{(v)}_{(i,j),t} = \frac{x^{(v)}_{(i,j),t_2} - x^{(v)}_{(i,j),t_1}}{t_2 - t_1}\left(t - t_1\right) + x^{(v)}_{(i,j),t_1} \tag{20}$$

4. *Model Filling*: As we mentioned before, for RNN models, the estimated value calculated on the previous visit can be used to impute the missing data on the current visit.

We denote "Mean Filling" strategy as MeanF, "Forward Filling" strategy as FF, "Linear Filling" strategy as LF, and "Model Filling" strategy as ModelF to design variants for SVR, Lasso, and GRU, that is, SVR-MeanF, SVR-LF, SVR-FF, Lasso-MeanF, Lasso-LF, Lasso-FF, GRU-MeanF, GRU-LF, GRU-FF, and GRU-ModelF.

Moreover, we compare our method with the following DPM methods:

- **Convex Fused Sparse Group Lasso** (cFSGL)[4] (Zhou et al., 2013) is a multi-task learning method for the DPM problem. It simultaneously selects task-shared and task-specific features using the sparse group Lasso penalty. This method introduces the indicator matrix to deal with missing data.

[4] https://github.com/jiayuzhou/MALSAR

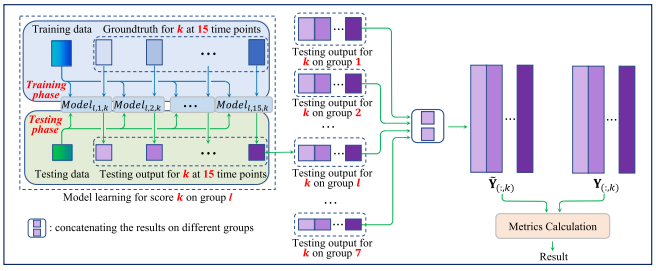

Fig. 8. Illustration of how SVR/Lasso predict score k with variable-length data, where $15\times7 = 105$ models (15 time points, 7 groups) are constructed for predicting score k at different time points of different groups. The prediction results at different time points of different groups will be ultimately concatenated for metrics calculation.

- **Peephole LSTM** (LSTM-P)[5] (Mehdipour-Ghazi et al., 2019) is an LSTM-based method that tackles the missing issue by adding peephole connections to the LSTM network.
- **MinimalRNN**[6] (Nguyen et al., 2020) is a MinimalRNN-based method that utilizes "Model Filling" strategy to impute missing data with MinimalRNN for AD progression modeling.
- **Temporal LSTM** (LSTM-T)[7] (Jung et al., 2021) is an LSTM-based method that considers the temporal correlation between different time points and utilizes the "Model Filling" strategy based on LSTM.

4.3. Group training strategy for SVR and Lasso

Since SVR and Lasso can only process individuals with fixed-length input, we divide the dataset into seven groups according to their sequence length. Individuals in the sample group have the same sequence length.

For each group, we concatenate the multi-modal longitudinal data in the order of "modalities first, sequences follow". We construct different SVR and Lasso models for predicting each score type k at each time point t based on each group l, which consequently leads to a total of $4 \times 7 \times 15 = 420$ (4 score types, 7 groups, and up to 15 target time points) models for SVR and Lasso, respectively. Fig. 8 illustrates how SVR/Lasso models AD progression with variable-length longitudinal data.

[5] We use the code provided in https://github.com/ssikjeong1/Deep_Recurrent_AD/tree/master/Competing_methods.

[6] https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/predict_phenotypes/Nguyen2020_RNNAD

[7] https://github.com/ssikjeong1/Deep_Recurrent_AD

Table 5
Overall regression results of proposed method and competing methods.

Metric	Method	Clinical Score (mean ± std (p-value))			
		MMSE	CDR-Global	CDR-SOB	ADAS-Cog
MAE (smaller = better)	Lasso-MeanF	2.2171 ± 0.0238 (3.5×10^{-4})*	0.3223 ± 0.0010 (1.2×10^{-12})*	1.6372 ± 0.0065 (5.7×10^{-12})*	6.6206 ± 0.0536 (1.2×10^{-10})*
	Lasso-FF	2.9151 ± 0.0277 (5.6×10^{-13})*	0.3268 ± 0.0018 (1.6×10^{-12})*	1.6794 ± 0.0078 (9×10^{-12})*	7.0454 ± 0.0561 (3.6×10^{-13})*
	Lasso-LF	2.2819 ± 0.0268 (5×10^{-6})*	0.3350 ± 0.0025 (1.9×10^{-12})*	1.6860 ± 0.0097 (4.8×10^{-12})*	6.7834 ± 0.0482 (1.7×10^{-11})*
	SVR-MeanF	3.0900 ± 0.0331 (3.7×10^{-13})*	0.3724 ± 0.0031 (1.5×10^{-14})*	2.0465 ± 0.0150 (1.1×10^{-14})*	7.8284 ± 0.0710 (1.4×10^{-13})*
	SVR-FF	3.0720 ± 0.0385 (8.2×10^{-13})*	0.3729 ± 0.0029 (4.9×10^{-15})*	2.0569 ± 0.0219 (2.8×10^{-14})*	7.8026 ± 0.0626 (2.3×10^{-13})*
	SVR-LF	2.8494 ± 0.0233 (4×10^{-11})*	0.3655 ± 0.0046 (9.9×10^{-14})*	2.0338 ± 0.0193 (6.6×10^{-14})*	7.4960 ± 0.0863 (5.1×10^{-12})*
	GRU-MeanF	2.2153 ± 0.0160 (1.1×10^{-3})*	0.2882 ± 0.0050 (7.1×10^{-5})*	1.4917 ± 0.0171 (1.3×10^{-8})*	6.1111 ± 0.0721 (1×10^{-4})*
	GRU-FF	2.2028 ± 0.0239 (1.3×10^{-3})*	0.2882 ± 0.0038 (6.4×10^{-6})*	1.4969 ± 0.0231 (4.5×10^{-7})*	6.1786 ± 0.1101 (8.7×10^{-5})*
	GRU-LF	2.2602 ± 0.0158 (1.9×10^{-5})*	0.3024 ± 0.0090 (5.8×10^{-6})*	1.5221 ± 0.0203 (2×10^{-8})*	6.1002 ± 0.1119 (1.2×10^{-3})*
	GRU-ModelF	2.1786 ± 0.0450 (4.5×10^{-2})*	0.2848 ± 0.0053 (1×10^{-3})*	1.3929 ± 0.0312 (2.8×10^{-2})*	6.1422 ± 0.1292 (3.1×10^{-4})*
	cFSGL	2.7392 ± 0.0127 (5.3×10^{-11})*	0.3882 ± 0.0019 (6.5×10^{-15})*	2.1472 ± 0.0120 (5.9×10^{-16})*	7.6065 ± 0.0478 (9.2×10^{-14})*
	LSTM-P	2.5492 ± 0.1563 (1.4×10^{-5})*	0.3457 ± 0.0110 (4.3×10^{-9})*	1.8175 ± 0.0860 (5.6×10^{-8})*	7.5018 ± 0.2994 (2.8×10^{-8})*
	LSTM-T	4.5031 ± 1.1185 (1×10^{-4})*	0.3576 ± 0.0166 (1.1×10^{-7})*	1.9734 ± 0.1221 (5.7×10^{-8})*	17.810 ± 0.1199 (3.2×10^{-19})*
	MinimalRNN	4.5600 ± 0.0164 (1.3×10^{-17})*	0.3556 ± 0.0063 (4.2×10^{-11})*	2.1074 ± 0.0069 (1.2×10^{-15})*	9.2416 ± 0.0175 (7.2×10^{-17})*
	Our	**2.1373 ± 0.0442**	**0.2753 ± 0.0029**	**1.3560 ± 0.0190**	**5.8689 ± 0.0623**
wR (larger = better)	Lasso-MeanF	**0.7795 ± 0.0164** (3.2×10^{-2})*	0.7076 ± 0.0075 (8×10^{-7})*	0.7613 ± 0.0082 (2.6×10^{-6})*	0.6796 ± 0.0136 (1.2×10^{-5})*
	Lasso-FF	0.6858 ± 0.0168 (2.4×10^{-6})*	0.6979 ± 0.0072 (3.3×10^{-8})*	0.7459 ± 0.0104 (5.5×10^{-7})*	0.6500 ± 0.0157 (9.8×10^{-8})*
	Lasso-LF	0.7674 ± 0.0161 (4.3×10^{-1})	0.6778 ± 0.0087 (4.6×10^{-9})*	0.7540 ± 0.0118 (2.6×10^{-6})*	0.6685 ± 0.0131 (1.2×10^{-6})*
	SVR-MeanF	0.5776 ± 0.0139 (1.4×10^{-9})*	0.5755 ± 0.0119 (2×10^{-11})*	0.6528 ± 0.0079 (2.2×10^{-11})*	0.5941 ± 0.0151 (1.2×10^{-8})*
	SVR-FF	0.5847 ± 0.0219 (2.1×10^{-8})*	0.5732 ± 0.0100 (1×10^{-12})*	0.6455 ± 0.0102 (1.3×10^{-10})*	0.5944 ± 0.0131 (7×10^{-9})*
	SVR-LF	0.6320 ± 0.0147 (5.1×10^{-8})*	0.5908 ± 0.0156 (2.2×10^{-10})*	0.6536 ± 0.0164 (1.9×10^{-9})*	0.6233 ± 0.0105 (7.6×10^{-8})*
	GRU-MeanF	0.7612 ± 0.0114 (8.5×10^{-1})	0.7160 ± 0.0095 (3.3×10^{-5})*	0.7879 ± 0.0107 (5.2×10^{-2})	0.7060 ± 0.0184 (5.2×10^{-2})
	GRU-FF	0.7633 ± 0.0084 (8.3×10^{-1})	0.7182 ± 0.0106 (8.5×10^{-6})*	0.7901 ± 0.0163 (1.8×10^{-1})	0.6985 ± 0.0154 (5.6×10^{-3})*
	GRU-LF	0.7615 ± 0.0127 (8.6×10^{-1})	0.7025 ± 0.0186 (1×10^{-4})*	0.7871 ± 0.0094 (1.5×10^{-2})*	0.7117 ± 0.0136 (7.5×10^{-2})
	GRU-ModelF	0.7559 ± 0.0151 (3×10^{-1})	0.7264 ± 0.0072 (4.4×10^{-4})*	0.7884 ± 0.0069 (3.9×10^{-2})*	0.6825 ± 0.0157 (3.7×10^{-5})*
	cFSGL	0.6969 ± 0.0137 (2.1×10^{-6})*	0.5526 ± 0.0128 (1.3×10^{-11})*	0.6207 ± 0.0111 (1.3×10^{-11})*	0.6209 ± 0.0135 (2.6×10^{-8})*
	LSTM-P	0.7449 ± 0.0228 (5.5×10^{-2})	0.6862 ± 0.0210 (2.5×10^{-5})*	0.7309 ± 0.0183 (3.8×10^{-6})*	0.6364 ± 0.0380 (2.2×10^{-4})*
	LSTM-T	0.6599 ± 0.0753 (3.3×10^{-3})*	0.6527 ± 0.0436 (1.2×10^{-4})*	0.6841 ± 0.0506 (4.8×10^{-5})*	0.6654 ± 0.0615 (1.4×10^{-2})*
	MinimalRNN	0.5963 ± 0.0086 (3.9×10^{-11})*	0.5112 ± 0.0087 (3.2×10^{-14})*	0.5909 ± 0.0082 (1.3×10^{-13})*	0.5224 ± 0.0122 (3.7×10^{-14})*
	Our	0.7620 ± 0.0132	**0.7415 ± 0.0050**	**0.7979 ± 0.0083**	**0.7249 ± 0.0165**

For a certain group l, in the training phase, we construct 4 models for predicting score k with each model corresponding to one target time point, and then we input features and scores for model learning. In the testing phase, we feed the testing data into four models to obtain the testing output. We further concatenate the testing output of each group to get the final predicted score matrix $\tilde{\mathbf{Y}}_{(:,k)}$, which will be compared with $\mathbf{Y}_{(:,k)}$ in metrics calculation.

4.4. Experimental results

In this paper, we use two dense layers to construct the degradation layer in our network. The "ReLU" was used as the activation function. For GRU-ModelF, LSTM-P, LSTM-T, and MinimalRNN, we employ $\alpha_2\mathcal{L}_{fit}+\mathcal{L}_{error}$ as the objective function. For other GRU variants, we employ $\mathcal{L}_{error}$ as the objective function. All network parameters of deep learning methods are randomly initialized before training. To select optimal hyperparameters, we further divide the training data and use 80% for training and 20% for validating. To simplify the tuning process for deep learning methods, we set the dimension of all hidden parameters (i.e., degradation layer parameters, dense layer parameters, and the states in LSTM) as d_h, which is the dimension of the latent representation. The Hyperopt toolbox (Bergstra et al., 2013) is utilized to find the best hyperparameter combinations by minimizing the MAE value on the validation set, and the hyperparameter search space for different methods are presented in Table 4.

Table 5 exhibits the performance of all methods in predicting four cognitive scores in terms of MAE and wR. The **bold** font represents the best result and (*) indicates that the performance of the competing method is significantly different from ours. It can be seen that our algorithm performs significantly better than all other methods in most cases.

Furthermore, in Figs. 9 and 10, we illustrate the specific performance of all methods at each single time point, from which we can see that our algorithm maintains a stable and competitive performance at most time points. It is noteworthy that, in Fig. 10, all methods have very low wR at the last three time points. In particular, there are even no wR results at the last time point. The reason is that the available data at the last three time points are extremely limited, and we only calculate wR on the observed data, which results in very low wR values. In addition, there is only one individual available at the 15th time point (which can be seen from Fig. 7) and the wR value cannot be calculated, for which no wR results at this time point are presented.

4.5. Further analysis

In this part, we conduct three ablation experiments to verify the hypothesis employed in our method, that is, the effectiveness of utilizing multi-modal data, the effectiveness of the multi-modality fusion module, and the effectiveness of the collaborative training strategy. Furthermore, we also explore the importance of different modalities and apply our method to predict the progression of modalities.

4.5.1. Effect of multi-modal data

In this experiment, we train and test our framework with the single-modal data to validate the effectiveness of combining multi-modal data. First, considering the "partial-modality-missing" and the "visit-missing" issues in the dataset, we discard those individuals containing no MRI or no PET data at historical visits and finally obtain a subset of 384 individuals. Then we train and test our model with the single-modal and multi-modal data of the selected individuals, respectively, and compare their results in Table 6. We can find that multi-modal data exhibits the most competitive performance in Table 6, which suggests that the combination of multi-modal data contributes to the prediction of AD score trajectories. Besides, we also notice that PET modality performs better in the single-modality case than MRI and demographics.

4.5.2. Effect of multi-modality fusion

To verify the efficacy of the fusion module, we remove the multi-modality fusion module from the proposed framework and train the

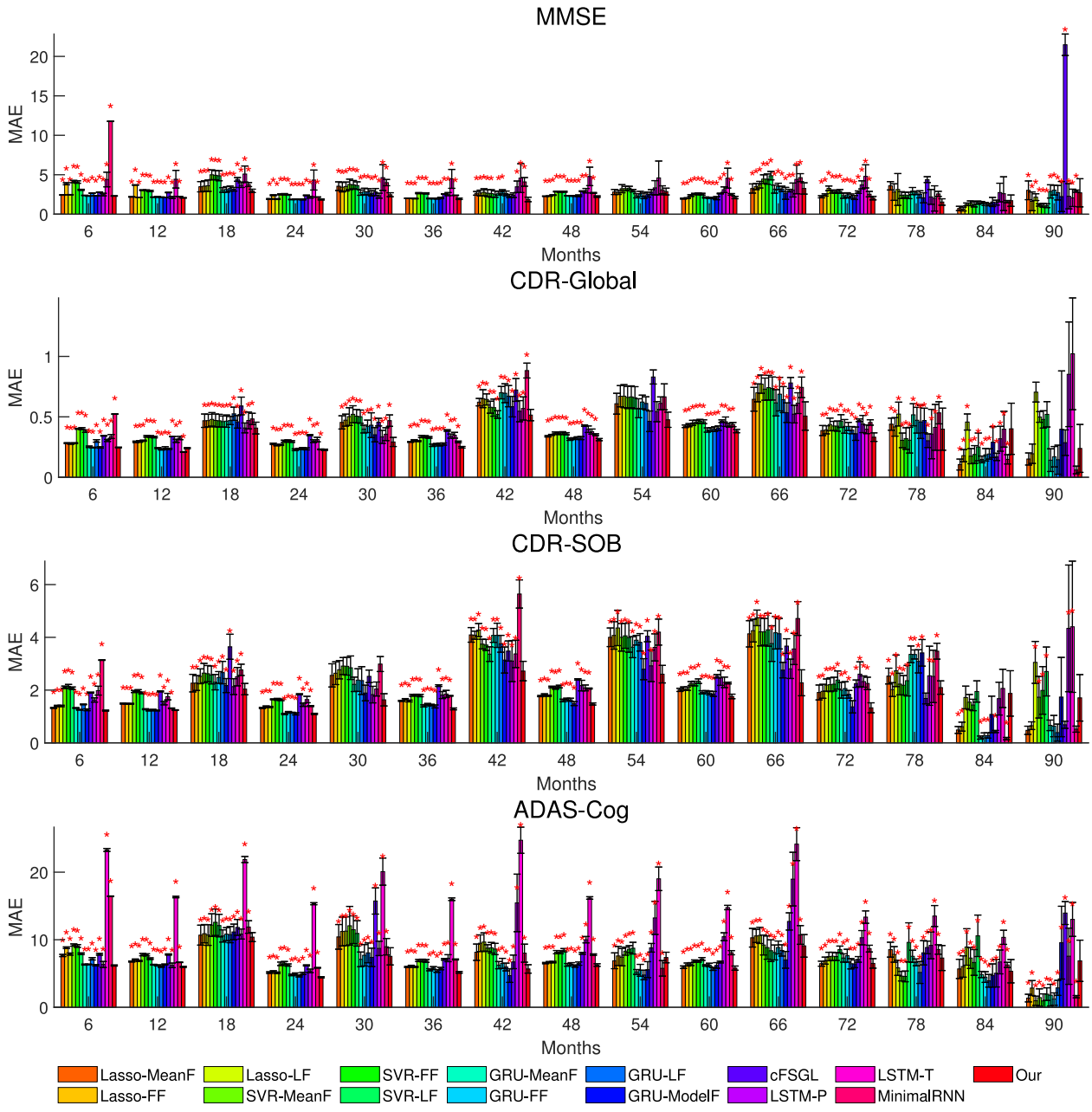

Fig. 9. Performance comparison between competing methods and our method in terms of MAE at each time point.

Table 6
Comparison results of training with single-modal data and training with multi-modal data.

Metric	Method	Clinical Score (mean ± std (p-value))			
		MMSE	CDR-Global	CDR-SOB	ADAS-Cog
MAE (smaller = better)	Single MRI	2.4007 ± 0.0600 (1×10^{-2})*	0.3018 ± 0.0113 (1×10^{-2})*	1.5171 ± 0.0568 (6.7×10^{-2})	6.6059 ± 0.3854 (2.2×10^{-2})*
	Single PET	2.3347 ± 0.0580 (4.8×10^{-1})	0.2966 ± 0.0085 (5×10^{-2})*	1.4970 ± 0.0294 (2.4×10^{-1})	6.3513 ± 0.1472 (9.6×10^{-2})
	Single Demographics	2.3703 ± 0.0899 (2×10^{-1})	0.2938 ± 0.0133 (3.5×10^{-1})	1.4855 ± 0.0408 (8.4×10^{-1})	6.4087 ± 0.4012 (2.3×10^{-1})
	Complete multi view (Our)	**2.3172 ± 0.0473**	**0.2895 ± 0.0047**	**1.4828 ± 0.0280**	**6.2342 ± 0.1098**
wR (larger = better)	Single MRI	0.7602 ± 0.0119 (4.6×10^{-1})	0.7111 ± 0.0104 (2.4×10^{-1})	0.7841 ± 0.0123 (1×10^{0})	0.7073 ± 0.0255 (7.3×10^{-2})
	Single PET	**0.7690 ± 0.0109** (3×10^{-1})	0.7147 ± 0.0167 (7.9×10^{-1})	**0.7871 ± 0.0134** (4.6×10^{-1})	0.7224 ± 0.0148 (4.9×10^{-1})
	Single Demographics	0.7628 ± 0.0121 (7.9×10^{-1})	0.7092 ± 0.0118 (2.9×10^{-1})	0.7811 ± 0.0093 (3.2×10^{-1})	0.7040 ± 0.0272 (4.9×10^{-2})*
	Complete multi view (Our)	0.7641 ± 0.0123	**0.7166 ± 0.0169**	0.7841 ± 0.0110	**0.7259 ± 0.0124**

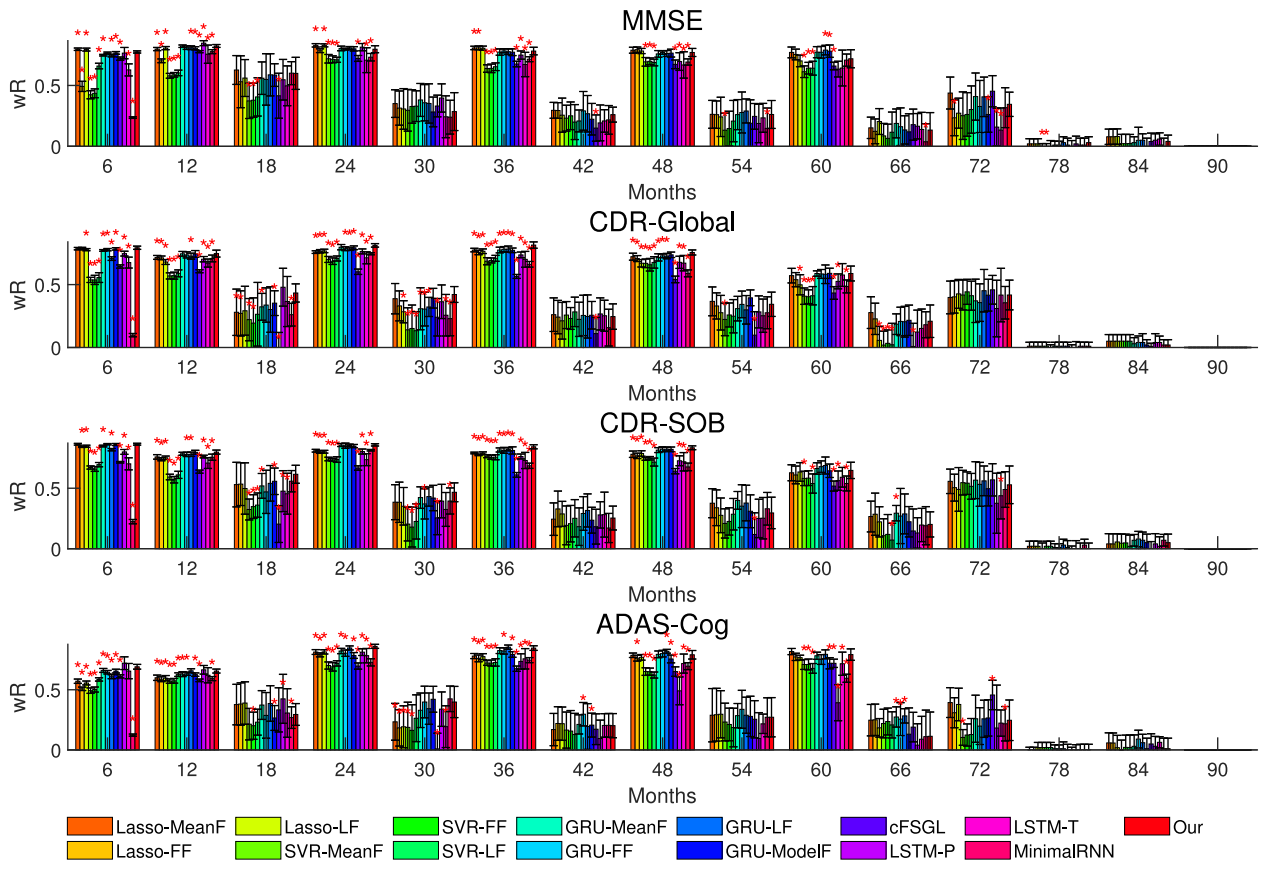

Fig. 10. Performance comparison between competing methods and our method in terms of wR at each time point.

Table 7
Comparison results of training with modality-concatenated data and training with modality-fused data.

Metric	Method	Clinical Score (mean ± std (p-value))			
		MMSE	CDR-Global	CDR-SOB	ADAS-Cog
MAE (smaller = better)	Concatenation Fusion (Our)	2.1993 ± 0.0496 (1.8×10^{-2})* **2.1630 ± 0.0287**	0.2795 ± 0.0058 (8×10^{-2}) **0.2749 ± 0.0065**	1.4059 ± 0.0234 (2.3×10^{-4})* **1.3706 ± 0.0142**	6.1796 ± 0.0801 (2.3×10^{-5})* **6.0873 ± 0.1150**
wR (larger = better)	Concatenation Fusion (Our)	**0.7650 ± 0.0128** (6.5×10^{-1}) 0.7613 ± 0.0099	0.7212 ± 0.0131 (7.3×10^{-4})* **0.7346 ± 0.0119**	0.7890 ± 0.0071 (1.7×10^{-2})* **0.7944 ± 0.0068**	0.6857 ± 0.0164 (1.8×10^{-3})* **0.7010 ± 0.0162**

Table 8
Comparison results of different training strategies.

Metric	Method	Clinical Score (mean ± std (p-value))			
		MMSE	CDR-Global	CDR-SOB	ADAS-Cog
MAE (smaller = better)	Separate Training Collaborative Training (Our)	2.9476 ± 0.2156 (1.1×10^{-6})* **2.1630 ± 0.0287**	1.4312 ± 0.1659 (3.4×10^{-9})* **0.2749 ± 0.0065**	2.4735 ± 0.1617 (4.1×10^{-9})* **1.3706 ± 0.0142**	7.2200 ± 0.2308 (2.6×10^{-8})* **6.0873 ± 0.1150**
wR (klarger = better)	Separate Training Collaborative Training (Our)	0.6994 ± 0.0150 (5×10^{-6})* **0.7613 ± 0.0099**	0.3907 ± 0.1204 (6×10^{-6})* **0.7346 ± 0.0119**	0.7013 ± 0.0318 (1.5×10^{-5})* **0.7944 ± 0.0068**	0.6347 ± 0.0127 (3.6×10^{-11})* **0.7010 ± 0.0162**

remaining sequence learning module with modality-concatenated data based on the objective function $\alpha_2\mathcal{L}_{fit}+\mathcal{L}_{error}$. The results are exhibited in Table 7, from which we can see that the proposed framework with multi-modality fusion significantly outperforms concatenating modalities.

4.5.3. Effect of collaborative training

The collaborative training of the multi-modality fusion module and sequence learning module contributes to model learning. In this experiment, we compare this collaborative training manner with the separate training manner that the longitudinal multi-modal data are first used to train multi-modality fusion module based on the objection function $\mathcal{L}_{rec}$, then the module output (i.e., the longitudinal latent representations) are utilized for training the sequence learning module based on the objection function $\alpha_2\mathcal{L}_{fit}+\mathcal{L}_{error}$. The results are shown in Table 8. It is evident that the collaboratively training manner outperforms the separate training manner. Moreover, note that in Section 4.5.2 we present the learning performance of concatenating modalities, the separate training manner is even worse than concatenating modalities despite the utilization of multi-modality fusion module, which indicates that if the fusion module and the prediction module are not trained collaboratively, the representation learned by the fusion

Table 9
Comparison results of dropping different modalities.

Metric	Method	Clinical Score (mean ± std (*p*-value))			
		MMSE	CDR-Global	CDR-SOB	ADAS-Cog
MAE (smaller = better)	Dropping MRI	2.3225 ± 0.0550 (5.2×10^{-1})	0.3013 ± 0.0086 (1.4×10^{-1})	1.5118 ± 0.0261 (6.9×10^{-2})	6.3460 ± 0.1839 (3.3×10^{-1})
	Dropping PET	2.3138 ± 0.0575 (2.2×10^{-1})	0.3019 ± 0.0113 (1.8×10^{-1})	1.5245 ± 0.0360 (8.1×10^{-2})	6.2814 ± 0.1417 (9.5×10^{-1})
	Dropping Demographics	**2.3062 ± 0.0588** (1.9×10^{-1})	0.2992 ± 0.0069 (1.6×10^{-1})	1.5111 ± 0.0342 (2.2×10^{-1})	6.2960 ± 0.1351 (7×10^{-1})
	Complete Modalities (Our)	2.3375 ± 0.0841	**0.2942 ± 0.0105**	**1.4952 ± 0.0228**	**6.2765 ± 0.1588**
wR (larger = better)	Dropping MRI	0.7670 ± 0.0158 (7.6×10^{-1})	0.7091 ± 0.0156 (2.1×10^{-1})	0.7821 ± 0.0131 (9.5×10^{-1})	0.7250 ± 0.0153 (6.1×10^{-1})
	Dropping PET	0.7668 ± 0.0126 (7.3×10^{-1})	0.7057 ± 0.0231 (1.8×10^{-1})	0.7818 ± 0.0139 (1×10^{0})	0.7273 ± 0.0151 (1×10^{0})
	Dropping Demographics	0.7667 ± 0.0147 (8.1×10^{-1})	0.7081 ± 0.0151 (1.7×10^{-1})	**0.7841 ± 0.0128** (6.3×10^{-1})	0.7256 ± 0.0112 (4.7×10^{-1})
	Complete Modalities (Our)	**0.7680 ± 0.0150**	**0.7178 ± 0.0183**	0.7818 ± 0.0163	**0.7273 ± 0.009**

Table 10
MAE results of predicting the progression for MRI and PET.

Method	Modality (mean ± std (*p*-value))	
	MRI	PET
GRU-ModelF	0.0671 ± 0.1031 (5.4×10^{-1})	0.0806 ± 0.0759 (4.3×10^{-1})
LSTM-T	0.0815 ± 0.0056 (3×10^{-8})*	0.0930 ± 0.0051 (1.3×10^{-8})*
MinimalRNN	0.2152 ± 0.0168 (1.8×10^{-10})*	0.3143 ± 0.0208 (3.2×10^{-11})*
Our	**0.0466 ± 0.0008**	**0.0606 ± 0.0007**

module may even disturb the model learning rather than enhance the model performance.

4.5.4. Importance of different modalities

In this part, we explore the importance of each modality by successively removing one of the modalities and modeling AD progression with the other two modalities. We analyze the importance of each modality by comparing the corresponding decrease in the model performance when a certain modality is discarded. Considering that some individuals do not have any MRI or PET data at historical visits, we employ the selection rule in Section 4.5.1, and compare the performance of different strategies based on the subset of 384 selected individuals.

As shown in Table 9, the most obvious change occurs in dropping demographics, which exhibits even better performance than the complete modalities in some cases. The possible reason is that the demographics are not sample-specific data for predicting AD progression, and it may be somewhat redundant to the proposed model. In contrast, the performance decreases in both dropping MRI and PET, indicating the importance of imaging modalities.

4.5.5. Predicting modality progression

It is noteworthy that our model can predict the progression of both clinical scores and modalities. Specifically, when predicting the score trajectories, the estimated variable $\tilde{\mathbf{s}}_t$ contains not only the clinical scores $\tilde{\mathbf{y}}_t$, but also the comprehensive representation $\tilde{\mathbf{h}}_t$ at the target time point. As shown in Fig. 3, with the estimated representation $\tilde{\mathbf{h}}_t$, the multi-modal data at the target time points can be reconstructed by the degradation networks. In this experiment, we apply the proposed framework to predict the progression of MRI and PET.

We use the first half of the historical visits of each individual to predict his/her MRI and PET at the second half of the historical visits. Since GRU-ModelF, LSTM-T, and MinimalRNN are directly fed with the concatenated multi-modal data, the corresponding values of MRI and PET at target time points can be directly estimated by the "Model Filling" strategy. Therefore, we compare with the above methods and use MAE as the measure to evaluate the prediction results. Table 10 shows the MAE results in our experiments. From the table, one may observe that our method achieved the best performance in predicting the modality progression.

5. Discussion

In practice, disease progression modeling suffers from both issues of label missing and modality missing. The convex Fused Sparse Group Lasso (cFSGL) model (Zhou et al., 2013) is the first work to deal with the label missing issue in the training process for disease progression modeling, which ingeniously formulates disease progression prediction as a multi-task learning problem. However, it does not consider the possible modality missing issue in the longitudinal learning. To obtain a complete solution, Zhang et al. propose a novel two-stage Multi-Resemblance Multi-Target Low-Rank Coding (MMLC) framework (Zhang et al., 2021) to simultaneously handle both label missing and modality missing issues. Specifically, in the first stage, the MMLC innovatively presents an online multi-resemblant low-rank sparse coding method to immune to incomplete longitudinal modality data and maintain a low computational cost. In the second stage, the MMLC employs a simple yet effective multi-target learning method to address the label missing issue, which can actually be recognized as an ablated version of the cFSGL model.

Similar to the MMLC model, our proposed framework considers the hybrid data-missing issue on both input and output sides. In our method, an indicator matrix is used to handle the label missing issue. Besides, our framework further imputes missing target scores during the training process. Specifically, the missing values in $\mathbf{y}_t$ can be imputed with the corresponding entries in $\tilde{\mathbf{s}}_t$ via the imputation module. To handle the modality missing issue, both cases of "partial-modality-missing" and "visit-missing" are fully considered in our framework. The proposed multi-modality fusion module can handle the "partial-modality-missing" issue and exploit the complementary information from the remaining available modalities. Regarding the "visit-missing" issue, since the complementary information is quite limited at these time points, we propose to utilize the "Model Filling" strategy to predict the comprehensive representations at these time points.

Although our method exhibits superior performance compared with other methods, it is subject to two limitations. First, we only explored the modality importance in Sections 4.5.1 and 4.5.4, whereas the importance of different features (brain regions) may be of more interest to clinicians, which has not yet been considered in our current model. Second, our method uses multi-modal hand-crafted features, which may not be well coordinated with prediction models and therefore degrades the prediction performance. Recently, deep learning methods have been proposed to automatically learn feature representations from medical images in an end-to-end manner. One of the representative works is wiseDNN (Liu et al., 2020), which is a weakly supervised densely connected neural network for task-oriented feature extraction and joint progression prediction. However, wiseDNN uses only complete baseline MRI scans, ignoring the temporal changes of imaging features without considering the data missing issue. To address these problems, the longitudinal-diagnostic generative adversarial network (LDGAN) (Ning et al., 2020) is proposed for joint longitudinal image synthesis and clinical score prediction based on incomplete MRIs and incomplete clinical scores. A major concern of existing end-to-end weakly supervised (i.e., visit missing or score missing or both) methods is that they only focus on single-modal data. However, for multi-modal data, the inherent

inter-modality correlation and the large data size will inevitably bring new challenges to both data utilization and computational cost. The comprehensive representation learning strategy in our method provides a feasible solution for utilizing multi-modal data in DPM models. In our future work, we will develop effective DPM methods with incomplete longitudinal multi-modal data using an end-to-end manner.

6. Conclusion

In this paper, we proposed a deep latent representation collaborated sequence learning framework for AD progression modeling based on incomplete variable-length longitudinal multi-modal data. The multi-modality fusion module is first introduced to learn comprehensive representations at each time point based on multi-modal data (even individuals with incomplete modalities). Then RNN based sequence learning module is used to flexibly process variable-length input and predict longitudinal trajectories of cognitive scores. We utilized the "Model Filling" strategy to handle the visit missing issue and trained the fusion module and sequence learning module collaboratively in a unified framework, so as to facilitate both representation learning and parameter learning. The experimental results on the ADNI dataset verified the superiority of the proposed model.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared the link to the data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61872190. Data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. The investigators within ADNI did not participate in the analysis or writing of this study. The complete listing of ADNI investigators can be found online.[8]

References

Association, A., 2019. 2019 Alzheimer's disease facts and figures. Alzheimer's Demen. 15 (3), 321–387.

Bergstra, J., Yamins, D., Cox, D.D., 2013. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proc. ICML, Vol. 28. pp. 115–123.

Brookmeyer, R., Abdalla, N., 2018. Estimation of lifetime risks of Alzheimer's disease dementia using biomarkers for preclinical disease. Alzheimer's Dement. 14, 981–988.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. EMNLP. pp. 1724–1734.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10 (7), 1895–1923.

Duchesne, S., Caroli, A., Geroldi, C., Collins, D.L., Frisoni, G.B., 2009. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. NeuroImage 47 (4), 1363–1370.

El-Sappagh, S., Abuhmed, T., Islam, S., Kwak, K., 2020. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. Neurocomputing 412, 197–215.

Fleet, B., Deller, J., Goodman, E., 2016. Initial results in alzheimer's disease progression modeling using imputed health state profiles. In: Proc. CSCI. pp. 7–12.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12 (3), 189–198.

Gauthier, S., Rosa-Neto, P., Morais, J., Webster, C., 2021. World alzheimer report 2021. Alzheimer's Dis. Int.

Gers, F.A., Schmidhuber, J., Cummins, F.A., 2000. Learning to forget: continual prediction with LSTM. Neural Comput. 12 (10), 2451–2471.

Green, C., Shearer, J., Ritchie, C., Zajicek, J., 2011. Model-based economic evaluation in Alzheimer's disease: a review of the methods available to model Alzheimer's disease progression. Value Health 14, 621–630.

Ito, K., Ahadieh, S., Corrigan, B., French, J., Fullerton, T., Tensfeldt, T., null, n., 2010. Disease progression meta-analysis model in Alzheimer's disease. Alzheimer's Dement. 6, 39–53.

Ito, K., Corrigan, B., Zhao, Q., French, J., Miller, R., Soares, H., Katz, E., Nicholas, T., Billing, B., Anziano, R., et al., 2011. Disease progression model for cognitive deterioration from Alzheimer's Disease Neuroimaging Initiative database. Alzheimer's Dement. 7 (2), 151–160.

Jack Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.

Jung, W., Jun, E., Suk, H.I., 2021. Deep recurrent model for individualized prediction of Alzheimer's disease progression. NeuroImage 237, 118143.

Kabani, N.J., MacDonald, D.J., Holmes, C.J., Evans, A.C., 1998. 3D anatomical atlas of the human brain. NeuroImage 7 (4, Part 2), S717.

Kanekiyo, T., Bu, G., 2016. Apolipoprotein E and amyloid-β-independent mechanisms in alzheimer's disease. Genes, Environ. Alzheimer's Dis. 171–196.

Kim, K.W., Woo, S.Y., Kim, S., Jang, H., Kim, Y., Cho, S.H., Kim, S.E., Kim, S.J., Shin, B.S., Kim, H.J., et al., 2020. Disease progression modeling of Alzheimer's disease according to education level. Sci. Rep. 10 (1), 1–9.

Liu, Y., Fan, L., Zhang, C., Zhou, T., Xiao, Z., Geng, L., Shen, D., 2021. Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. Med. Image Anal. 69, 101953.

Liu, M., Zhang, J., Lian, C., Shen, D., 2020. Weakly supervised deep learning for brain disease prognosis using MRI and incomplete clinical scores. IEEE Trans. Cybern. 50 (7), 3381–3392.

Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Golland, P., Klein, S., Alexander, D.C., 2019. TADPOLE challenge: accurate alzheimer's disease prediction through crowdsourced forecasting of future data. In: Proc. Predict. Intell. Med. pp. 1–10.

McDonnell, J., Redekop, W., Roer, N., Goes, E., Ruitenberg, A., Busschbach, J., Breteler, M., Rutten, F., 2012. The cost of treatment of alzheimer's disease in the netherlands. PharmacoEconomics 19, 379–390.

Mehdipour-Ghazi, M., Nielsen, M., Pai, A., Cardoso, M.J., Modat, M., Ourselin, S., Sørensen, L., 2019. Training recurrent neural networks robust to incomplete data: Application to Alzheimer-s disease progression modeling. Med. Image Anal. 53, 39–46.

Nguyen, M., He, T., An, L., Alexander, D.C., Feng, J., Yeo, B.T.T., 2020. Predicting Alzheimer's disease progression using deep recurrent neural networks. NeuroImage 222, 117203.

Nguyen, M., Sun, N., Alexander, D.C., Feng, J., Yeo, B.T.T., 2018. Modeling Alzheimer's disease progression using deep recurrent neural networks. In: Proc. PRNI. pp. 1–4.

Nie, L., Zhang, L., Meng, L., Song, X., Chang, X., Li, X., 2017. Modeling disease progression via multisource multitask learners: a case study with alzheimer's disease. IEEE Trans. Neural Networks Learn. Syst. 28 (7), 1508–1519.

Ning, Z., Zhang, Y., Pan, Y., Zhong, T., Liu, M., Shen, D., 2020. LDGAN: longitudinal-diagnostic generative adversarial network for disease progression prediction with missing structural MRI. In: Machine Learning in Medical Imaging. pp. 170–179.

Petrella, J., Coleman, R.E., Doraiswamy, P., 2003. Neuroimaging and early diagnosis of Alzheimer disease: a look to the future. Radiology 226 (2), 315–336.

Rahimi, J., Kovacs, G., 2014. Prevalence of mixed pathologies in the aging brain. Alzheimer's Res. Ther. 6 (9).

Rosen, W., Mohs, R., Davis, K., 1984. A new rating scale for Alzheimer's disease. Am. J. Psychiatry 141 (11), 1356–1364.

Sabuncu, M., Bernal-Rusiel, J.L., Reuter, M., Greve, D., Fischl, B., 2014. Event time analysis of longitudinal neuroimage data. NeuroImage 97, 9–18.

Samtani, M.N., Farnum, M., Lobanov, V., Yang, E., Raghavan, N., DiBernardo, A., Narayan, V., the Alzheimer's Disease Neuroimaging Initiative, 2012. An improved model for disease progression in patients from the alzheimer's disease neuroimaging initiative. J. Clin. Pharmacol. 52 (5), 629–644.

Spasov, S.E., Passamonti, L., Duggento, A., Liò, P., Toschi, N., 2019. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. NeuroImage 189, 276–287.

Stekhoven, D.J., Bühlmann, P., 2012. MissForest-non-parametric missing value imputation for mixed-type data. Bioinf. 28 (1), 112–118.

[8] https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Stonnington, C.M., Chu, C., Klöppel, S., Jr., C.R.J., Ashburner, J., Frackowiak, R.S., 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. NeuroImage 51 (4), 1405–1413.

Sukkar, R., Katz, E., Zhang, Y., Raunig, D., Wyman, B., 2012. Disease progression modeling using hidden Markov models. In: Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. pp. 2845–2848.

Tabarestani, S., Aghili, M., Eslami, M., Cabrerizo, M., Barreto, A., Rishe, N., Curiel, R.E., Loewenstein, D., Duara, R., Adjouadi, M., 2020. A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. NeuroImage 206, 116317.

Thung, K., Yap, P., Adeli, E., Lee, S., Shen, D., 2018. Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion. Med. Image Anal. 45, 68–82.

Vemuri, P., Wiste, H., Weigand, S., Shaw, L., Trojanowski, J., Weiner, M., Knopman, D., Petersen, R., Jack, C., Alzheimer's Disease Neuroimaging Initiative, 2009. MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. Neurology 73 (4), 294–301.

Wang, X., Sontag, D., Wang, F., 2014. Unsupervised learning of disease progression models. In: Proc. SIGKDD. pp. 85–94.

Wang, M., Zhang, D., Shen, D., Liu, M., 2019. Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data. Med. Image Anal. 53, 111–122.

White, I.R., Royston, P., Wood, A.M., 2011. Multiple imputation using chained equations: Issues and guidance for practice. Stat. Med. 30 (4), 377–399.

Williams, J.W., Plassman, B.L., Burke, J., Benjamin, S., 2010. Preventing Alzheimer's disease and cognitive decline. Evid. Rep. Technol. Assess. (193), 1–727.

Xie, Q., Wang, S., Zhu, J., Zhang, X., 2016. Modeling and predicting AD progression by regression analysis of sequential clinical data. Neurocomputing 195, 50–55.

Yang, M., Elazab, A., Yang, P., Xia, Z., Wang, T., Lei, B., 2019. Joint and long short-term memory regression of clinical scores for alzheimer's disease using longitudinal data. In: Proc. EMBC. pp. 281–284.

Zhang, C., Han, Z., Cui, Y., Fu, H., Zhou, J.T., Hu, Q., 2019. CPM-nets: cross partial multi-view networks. In: Proc. NIPS. pp. 557–567.

Zhang, J., Wu, J., Li, Q., Caselli, R.J., Thompson, P.M., Ye, J., Wang, Y., 2021. Multi-resemblance multi-target low-rank coding for prediction of cognitive decline with longitudinal brain images. IEEE Trans. Med. Imaging 40 (8), 2030–2041.

Zhou, J., Liu, J., Narayan, V.A., Ye, J., 2013. Modeling disease progression via multi-task learning. NeuroImage 78, 233–248.

Zhou, T., Liu, M., Thung, K.H., Shen, D., 2019. Latent representation learning for alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. IEEE Trans. Med. Imaging 38 (10), 2411–2422.

Zhou, G., Wu, J., Zhang, C., Zhou, Z., 2016. Minimal gated unit for recurrent neural networks. Int. J. Autom. Comput. 13 (3), 226–234.

Zhu, X., Suk, H.I., Wang, L., Lee, S.W., Shen, D., 2017. A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. Med. Image Anal. 38, 205–214.