Neural Networks 150 (2022) 422–439

Contents lists available at ScienceDirect

Neural Networks

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

# Predicting progression of Alzheimer's disease using forward-to-backward bi-directional network with integrative imputation

Ngoc-Huynh Ho [a], Hyung-Jeong Yang [a,*], Jahae Kim [a,b], Duy-Phuong Dao [a], Hyuk-Ro Park [a], Sudarshan Pant [a]

[a] Department of AI Convergence, Chonnam National University, 61186, South Korea
[b] Department of Nuclear Medicine, Chonnam National University Hospital, 61469, South Korea

ARTICLE INFO

*Article history:*
Received 3 August 2021
Received in revised form 23 February 2022
Accepted 10 March 2022
Available online 17 March 2022

*Keywords:*
Alzheimer's progression
MRI biomarker forecasting
Missing value imputation
Clinical status prediction
Progressive recurrent networks

ABSTRACT

If left untreated, Alzheimer's disease (AD) is a leading cause of slowly progressive dementia. Therefore, it is critical to detect AD to prevent its progression. In this study, we propose a bidirectional progressive recurrent network with imputation (BiPro) that uses longitudinal data, including patient demographics and biomarkers of magnetic resonance imaging (MRI), to forecast clinical diagnoses and phenotypic measurements at multiple timepoints. To compensate for missing observations in the longitudinal data, we use an imputation module to inspect both temporal and multivariate relations associated with the mean and forward relations inherent in the time series data. To encode the imputed information, we define a modification of the long short-term memory (LSTM) cell by using a progressive module to compute the progression score of each biomarker between the given timepoint and the baseline through a negative exponential function. These features are used for the prediction task. The proposed system is an end-to-end deep recurrent network that can accomplish multiple tasks at the same time, including (1) imputing missing values, (2) forecasting phenotypic measurements, and (3) predicting the clinical status of a patient based on longitudinal data. We experimented on 1,335 participants from The Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE) challenge cohort. The proposed method achieved a mean area under the receiver-operating characteristic curve (mAUC) of 78% for predicting the clinical status of patients, a mean absolute error (MAE) of 3.5*ml* for forecasting MRI biomarkers, and an MAE of 6.9*ml* for missing value imputation. The results confirm that our proposed model outperforms prevalent approaches, and can be used to minimize the progression of Alzheimer's disease.



## 1. Introduction

Alzheimer's disease (AD) is an irreversibility of a neuronal disease that affects memory, cognition, and behavior (Cuingnet et al., 2011). One of the obstacles in the efficient early detection and planning of therapeutic procedures is the complex nature of AD biomarkers as well as the heterogeneity of measurements taken from various imaging modalities (Nie, Meng, Song, Chang, & Li, 2017). In the absence of significant progress in the development of modalities to treat AD, a number of researchers have examined alternative viable and cost-effective solutions to provide the care and treatment required by AD patients (Gavidia-Bovadilla, Kanaan-Izquierdo, Mataró-Serrat, Perera-Lluna, & Initiative, 2017; Huang et al., 2016; Zhou, Liu, Narayan, Yeand, & Initiative, 2013). This includes ways to facilitate AD patients with adequate and effective lifestyles as well as levels of neural training. To understand and predict how AD progressively develops in a patient is therefore essential for both early intervention and the effective provision of personalized healthcare services (Workgroup et al., 2016).

In the literature, traditional time series-based approaches and algorithms using machine learning techniques have been frequently used to model the progression of AD and classify its severity (Gavidia-Bovadilla et al., 2017; Huang et al., 2016; Liu, Zhang, Zhang, & Zhou, 2013; Sukkar, Katz, Zhang, Raunig, & Wyman, 2012; Vu, Ho, Yang, Kim, & Song, 2018; Zhou, Liu, Narayan, & Ye, 2012; Zhou et al., 2013). Despite promising progress made in studies related to disease modeling, many challenging modeling problems persist (Goldstein, Navar, Pencina,

* Corresponding author.
*E-mail addresses:* nhho@chonnam.ac.kr (N.-H. Ho), hjyang@jnu.ac.kr (H.-J. Yang), jhbt0607@daum.net (J. Kim), phuongdd.1997@gmail.com (D.-P. Dao), hyukro@jnu.ac.kr (H.-R. Park), sudarshan.pant@gmail.com (S. Pant).

https://doi.org/10.1016/j.neunet.2022.03.016

& Ioannidis, 2017). Instead of forecasting the status of diseases, the literature has focused mainly on modeling their progression. This involves predicting the status of progression of diseases at known times based on the available data, or focusing only on classifying stages of progression in case of restricted observations. Traditional studies have also ignored the problem of missing data when predicting the progression of AD (Lei, Yang, Wang, Chen, & Ni, 2017; Liu, Zhang, Adeli, , & Shen, 2019; Stonnington, Chu, Klöppel, Clifford, Ashburner, Frackowiak, & Alzheimer Disease Neuroimaging Initiative, 2010; Sukkar et al., 2012).

In applications, missing values in a time series, such as due to equipment damage or communication errors, are common owing to unforeseen accidents that can damage the performance of downstream applications. In the medical field, data may be missing due to the design of the study, delay in data collection, subject attrition, or data collection errors. Many previous studies have sought to solve the problem of missing data in statistics and machine learning. However, most of them require strong assumptions owing to the missing values. The two main approaches to handle missing data are (Schafer & Graham, 2002) "preprocessing" and "integrative". Missing data can either be solved for by imputing values through steps of preprocessing (Zhou et al., 2013), or they can be integrated directly into the given models or training strategies (Aksman et al., 2019).

In this paper, we propose a bi-directional progressive recurrent network with imputation (BiPro) for predicting AD progression. It uses biomarkers from magnetic resonance imaging (MRI) and patient demographics to tackle three tasks: (i) imputing missing values, (ii) forecasting phenotypic measurements, and (iii) predicting the clinical status of the patient. Internally, BiPro adapts recurrent neural networks (RNN) to impute missing values without any specific assumption over the data. Following work by Jung, Jun, Suk, and Initiative (2020), the proposed BiPro model can be defined as consisting of three modules: *Imputation*, *Encoding*, and *Forecasting*. The main differences between this study and Jung's work are that the *Imputation* module learns the missing values directly in a recurrent dynamical system based on the observed data and static information, which significantly boosts its final performance, and that the *Encoding* module learns latent recurrent features from imputed data based on an LSTM network, which determines the progression score of each biomarker dynamically and uses it as an input to the LSTM cell. In addition, we propose a bidirectional strategy to extract hidden features for prediction tasks and provide an accurate estimation of missing values. Lastly, the *Forecasting* module predicts and classifies the clinical status of the patient at the next timepoint into three groups: cognitively normal (CN), mild cognitive impairment (MCI), and AD dementia. It also predicts the MRI biomarkers. A multiobjective function for the proposed model is formulated and optimized in an end-to-end manner. We evaluated the performance of our proposed model on The AD Prediction Of Longitudinal Evolution (TADPOLE) challenge cohort, and observed that it outperformed competing methods on all tasks concerning imputation, forecasting, and classification. The main contributions of this study can be described as follows:

- We propose an end-to-end system to model long-term disease progression through missing value imputation, phenotypic measurement forecasting, and clinical status prediction.
- We provide a formula to combine both static and dynamic relations for missing value imputation to obtain effective estimation.
- We propose a bidirectional strategy for longitudinal data that can support a previous module for imputation and extract temporal representatives in both directions (forward and backward) for prediction tasks.
- We propose a variant of an LSTM-based cell by using the level of progression to forecast both the clinical status of patients and the MRI biomarkers. This can be used to model the trajectory of progression of the disease.
- We experimented on a public dataset and used cross-validation to verify the generalizability of the proposed system. Several aspects of disease progression analysis were exploited to better understand AD.

The remainder of this study is structured as follows: Section 2 gives a brief description of related studies on predictive modeling using RNN in healthcare, Section 3 presents details of the proposed BiPro model, and Section 4 reports on experiments that were conducted to compare the proposed method with conventional methods, and analyzes the results. Lastly, Section 5 presents the conclusion of our study.

## 2. Related work

In this section, we review the use of RNNs in healthcare applications and previous studies on AD progression. We explore the missing value problem and RNN-based solutions from past work as well. RNNs are good at capturing time series patterns (Lipton, 2015). RNN models have recently shown significant potential for use in healthcare applications. Examples include the use of gated recurrent units (GRUs) (Choi, Bahadori, Schuetz, Stewart, & Sun, 2016) as a diagnostic tool, and for predicting medication for patients' subsequent visits; predicting the early onset of cardiac failure using GRU models (Choi, Schuetz, Stewart, & Sun, 2017); LSTM models for predicting the onset of multiple conditions (Razavian, Marcus, & Sontag, 2016); and the diagnostic classification of patients in the pediatric intensive care unit (PICU) by using LSTM models (Lipton, Kale, Elkan, & Wetzel, 2017). The relevant studies have applied different RNN models for medical prediction by using longitudinal data on patient over time.

In the literature, the prediction of AD progression can be divided into three subtopics: classification (Fedorov et al., 2019; Lee, Nho, Kang, Sohn, & Kim, 2019), modeling (Bilgel, Jedynak, , & Initiative, 2019; Lorenzi et al., 2019; Venkatraghavan, Bron, Niesse, Klein, & Initiative, 2019; Wang, Qiu, & Yu, 2018), and estimation (Doody, Massman, & Dunn, 2001; Lorenzi et al., 2019; Nie et al., 2017). In the context of classification, researchers seek to identify the state of the patient in terms of either non-conversion or conversion over a long period of time. Fedorov et al. (2019) explored using Deep InfoMax (DIM) (Hjelm et al., 2018) to classify stable MCI versus progressive MCI patients by learning deep nonlinear representations of neuroimaging data as the output of a convolutional neural network. In the context of modeling, Wang et al. (2018) modeled the progression of AD in patients on subsequent visits by relying on a "many-to-one" enhanced RNN architecture to support time shift. Different numbers of visits and uneven periods of time can be covered using this approach. Furthermore, Bilgel et al. (2019) proposed a Bayesian progression score model (BPSM) to compute the trajectories of biomarkers from NL (normal) to MCI and AD. In the context of estimation, Nie et al. (2017) proposed a Multisource Multitask Learner to estimate cognitive scores, such as the mini-mental state examination (MMSE) and Alzheimer's disease assessment scale-cognitive subscale (ADAS-Cog). These scores can help estimate the severity and progression of cognitive impairment due to AD.

To solve the problem of missing data, Che, Purushotham, Cho, Sontag, and Liu (2018) proposed a deep recurrent network called the gated recurrent unit with decay (GRU-D) to capture long-term dependencies in the time series and model the missing patterns. GRU-D is an RNN-based approach to imputation that fills each missing value with the weighted combination of the last

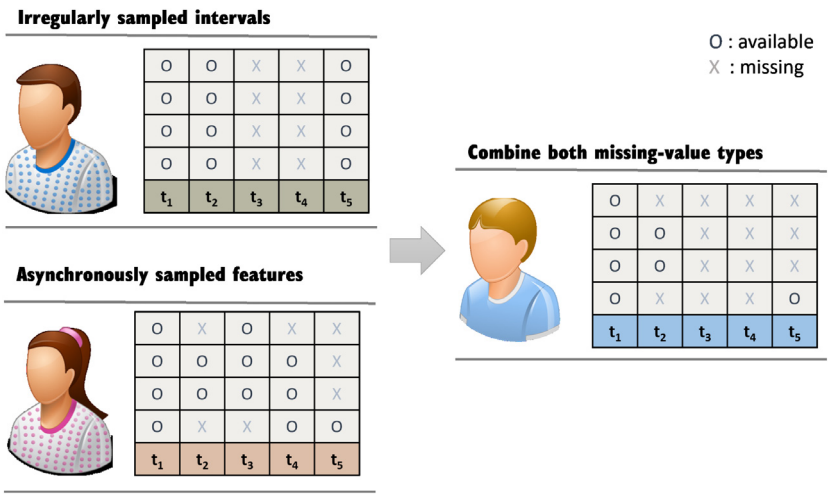



**Fig. 1.** Example of missing value types in EHR: irregularly sampled intervals (top left), asynchronously sampled features (bottom left), and a combination of two types of missing values (center right).

valid observation and the global mean (static relation), together with a recurrent component. Yoon, Zame, and van der Schaar (2019) considered a temporal relation for imputation. Contrary to a standard bidirectional RNN, the timing of the inputs to the salient layers are reversed in a forward-and-backward direction. These imputed values are dynamically updated during training until they are optimal (dynamic relation). Jung et al. (2020) used an RNN for missing value imputation in time series data, and then fed the imputed values to an LSTM cell to forecast both the diagnosis of AD and MRI biomarkers. Their model is an adaptation of the one propose in Cao et al. (2018), and inspects the temporal and multivariate relations of measurements (dynamic relations) for missing value imputation (and is called LSTM-I in our comparisons below). Nguyen et al. (2020) proposed a minimal RNN called MinRNN to predict the clinical diagnoses of patients, their cognition, and ventricular volume. Their model has fewer parameters than other RNN models, such as the LSTM, and therefore is less prone to overfitting.

In general, the above approaches do not consider combining both static and dynamic relations for the imputation task, and most models consider only the forward direction. An exception is the MRNN (multidirectional recurrent neural network) (Yoon et al., 2019). The MRNN model has a similar process to that of typical bidirectional networks, where there are two independent paths and the hidden state is encoded by previous state (in case of a forward path) or the next state (in case of a backward path). The difference is that the MRNN uses both the previous (forward path) and the next (backward path) inputs as well as the hidden state to encode the current state. We can use the backward direction, where each value in the time series can be derived from the future to the past by using another fixed, arbitrary function. In such a case, the missing values are given gradients that are delayed in both the forward and the backward directions with constraints, which increases the accuracy of estimation of missing values and improves the final prediction.

## 3. Proposed method

To predict AD progression, we introduce an end-to-end deep recurrent network (BiPro) that consists of three modules: *Imputation*, *Encoding*, and *Forecasting*. The *Imputation* module estimates missing values by using hidden representations from both the previous timepoint of the forward path and the next timepoint of the backward path. In general, the presence of missing values in time series data leads to two major problems, as shown in Fig. 1: an irregularly sampled interval pattern, where the interval between each pair of timepoints is different while all features are simultaneously collected, and an asynchronously sampled feature pattern, where not all features are periodically collected. The missing pattern of longitudinal data is composed of both these missing patterns, hence placing such data in a disadvantageous position compared with the complete data used in other statistical models.

Many studies have discarded subjects or timestamps from this problem such that a large amount of data can be lost. We define a formula to combine different relations for missing value imputation. In particular, we investigate the multivariate relations (Jung et al., 2020) among the observed values at the given timepoint, temporal relations from the recurrent network, and the empirical mean and the last valid observations up to the timepoint being considered. Moreover, to capture the underlying temporal properties in the given time series data, the *Encoding* module can encompass complete observations from the *Imputation* module. Finally, we use the *Forecasting* module to transform the encoded representations to forecast the volumetric measurements of the next timepoint and predict the future clinical status of the patient in terms of three classes: cognitively normal (CN), mild cognitive impairment (MCI), and AD dementia. Fig. 2 provides an overview of the architecture of the proposed BiPro. The novelty of the proposed BiPro can be summarized as follows:

1. We propose an integrative imputation module (Eq. (9)) to combine static and dynamic relations to estimate missing values. The fused factor is determined by using the temporal decay (Eqs. (6) and (7)) that represents the influence of the missing duration. This indicates that the longer the missing period is from the previous observation, the smaller is the contribution of the estimated value.
2. We propose a variant of an LSTM-based cell (Fig. 5), called multifeature aggregated long short-term memory with progressive score (MAPro-LSTM), to fuse multiple features, including patient information, hidden state, the estimation of missing values, mask of the valid features, and progressive scores. The progressive scores computed by Eq. (11) are modified from the studies in Bilgel, Prince, Wong, Resnick, and Jedynak (2016) and Jedynak et al. (2012) to be learned dynamically through MAPro-LSTM.
3. In the typical bidirectional LSTM (BiLSTM) and the MRNN, the forward and backward paths encode temporal features independently before concatenating at the end. This means

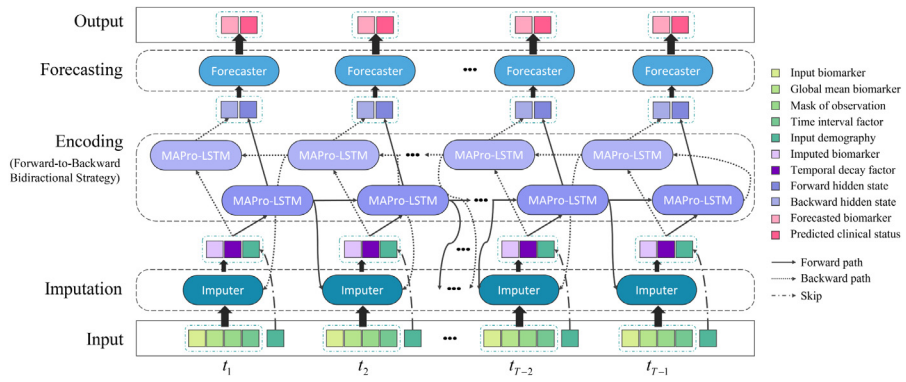



**Fig. 2.** Overall architecture of the proposed BiPro model.

$\mathbf{X} =$

| $v_1 = 0$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $v_7$ |
|---|---|---|---|---|---|---|
| 0.55 | X | 0.41 | 0.32 | X | 0.36 | X |
| 0.6 | 0.54 | X | X | 0.49 | X | X |
| 0.21 | X | 0.19 | 0.23 | X | X | 0.17 |
| 0.35 | X | X | 0.39 | X | X | X |

$\mathbf{M} =$

| $v_1 = 0$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $v_7$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 |

$\bar{\mathbf{X}} =$

| $v_1 = 0$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $v_7$ |
|---|---|---|---|---|---|---|
| 0.39 | X | 0.39 | 0.39 | X | 0.39 | X |
| 0.22 | 0.22 | X | X | 0.22 | X | X |
| 0.18 | X | 0.18 | 0.18 | X | X | 0.18 |
| 0.21 | X | X | 0.21 | X | X | X |

$\boldsymbol{\varphi} =$

| $v_1 = 0$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $v_7$ |
|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 1 | 1 | 2 | 1 |
| 0 | 1 | 1 | 2 | 3 | 1 | 1 |
| 0 | 1 | 2 | 1 | 1 | 2 | 3 |
| 0 | 1 | 2 | 3 | 1 | 2 | 3 |

**Fig. 3.** An example of the input time series $\boldsymbol{X}$, global mean $\bar{\boldsymbol{X}}$, masking observation $\boldsymbol{M}$ for $\boldsymbol{X}$, and time interval from last observation $\boldsymbol{\varphi}$.

that they use only temporal features for the imputation task and cannot update back the *Imputation* module. To solve this problem, we propose a forward-to-backward bidirectional strategy for dual association between the *Imputation* and the *Encoding* modules in both directions (forward and backward). The details are shown in Algorithm 1.

### 3.1. Notation

Following Che et al. (2018), we first denote a multivariate time series with $N$ variables of length $T$ by $\boldsymbol{X} = (x_1, x_2, \ldots, x_T)^T \in \mathbb{R}^{T\times N}$, where, for each $t \in \{1, 2, \ldots, T\}$, $x_t \in \mathbb{R}^N$ represents the $t$th measurement of all variables and $x_t^{\{n\}}$ denotes the measurement of variable $n$ of $x_t$. $\boldsymbol{X_d} = \left(x_{d,1}, x_{d,2}, \ldots, x_{d,T}\right)^T \in \mathbb{R}^{T\times N}$ represents demographic information. Let $v_t \in \mathbb{R}$ represent the timestamp when the $t$th observation is obtained, and assume that the first observation is made at timestamp 0 ($v_1 = 0$). We introduce a masking vector $\boldsymbol{M} = (m_1, m_2, \ldots, m_T)^T \in \{0, 1\}^N$ to denote variables that are missing (0) or observed (1) at time step $t$, while maintaining the time interval $\varphi_t^{\{n\}} \in \mathbb{R}$ for each variable $n$ since its last observation. Specifically, this can be expressed as

$$m_t^{\{n\}} = \begin{cases} 1, & \text{if } x_t^{\{n\}} \text{ is observed} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$\boldsymbol{\varphi}_t^{\{n\}} = \begin{cases} v_t - v_{t-1} + \varphi_{t-1}^{\{n\}}, & \text{if } t > 1, m_{t-1}^{\{n\}} = 0 \\ v_t - v_{t-1}, & \text{if } t > 1, m_{t-1}^{\{n\}} = 1 \\ 0, & \text{if } t = 1 \end{cases} \tag{2}$$

Fig. 3 shows an example of the four variables ($N = 4$) and the seven timepoints ($T = 7$) used for measurement $\boldsymbol{X}$, the global mean $\bar{\boldsymbol{X}}$, timestamps $\boldsymbol{v}$, masking $\boldsymbol{M}$, and time interval $\boldsymbol{\varphi}$. In this example, we consider how to compute $\boldsymbol{\varphi}_7^{\{4\}}$, the time interval of the fourth feature at the seventh timepoint. Because the last valid observation from $\boldsymbol{X}_7^{\{4\}}$ ($NA$) is $\boldsymbol{X}_4^{\{4\}}$ (0.39), $\boldsymbol{\varphi}_7^{\{4\}} = 7 - 4 = 3$.

### 3.2. The Imputation module

We use both static and dynamic relations for the imputation task by designing an architecture that combines temporal and multivariate relations, the global mean, and the last valid observations. The temporal and multivariate features learned from the training process represent dynamic relations while the global mean, $\bar{x}$, and the last valid observations, $x'_t$, computed from the dataset, which remain unchanged during the training process, represent static relations. The context of temporal relations, $\hat{x}_t$ in a given sequence is obtained from the hidden state values $\overrightarrow{h}_{t-1}$ of the forward path and $\overleftarrow{h}_{t+1}$ of the backward path, and is expressed as

$$\hat{x}_t = \overrightarrow{\mathbf{W}}_x \overrightarrow{h}_{t-1} + \overleftarrow{\mathbf{W}}_x \overleftarrow{h}_{t+1} + \mathbf{b}_x \tag{3}$$

where $\overrightarrow{\mathbf{W}}_x$, $\overleftarrow{\mathbf{W}}_x$, $\mathbf{b}_x$ are learnable parameters for the multivariate linear transform. The term $\overleftarrow{\mathbf{W}}_x \overleftarrow{h}_{t+1}$ is added on the backward path. While keeping the observed data, we compute the temporary vector, $\tilde{x}_t$, by replacing the missing values with those estimated from the temporal relations as follows:

$$\tilde{x}_t = m_t \odot x_t + (1 - m_t) \odot \hat{x}_t \tag{4}$$

where $\odot$ denotes the element-wise multiplication operation. We also determine the multivariate relations, $\hat{z}_t$, from Jung et al. (2020) among MRI features of the imputed temporal vector $\tilde{x}_t$ by using a linear transformation with learnable parameters $\mathbf{W}_z$ and $\mathbf{b}_z$ as follows:

$$\hat{z}_t = \mathbf{W}_z \tilde{x}_t + \mathbf{b}_x \tag{5}$$

We maintain the diagonal elements in $\mathbf{W}_z$ at zero by multiplying $\mathbf{W}_z$ with an inverse identity matrix so that we can focus on the relation between features.

To combine the imputed representations, we use the temporal decay factor $\lambda_t$ from Cao et al. (2018) to dynamically determine

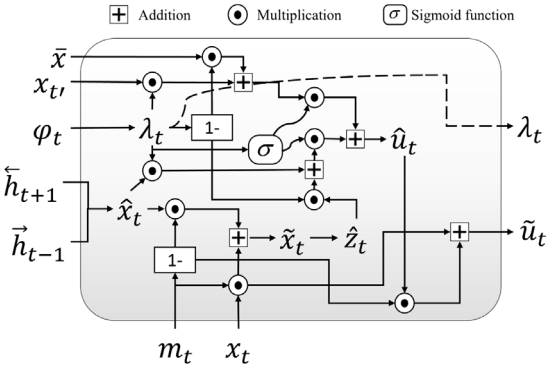



**Fig. 4.** Architecture of the *Imputation* module used for estimating missing values.

a weighting coefficient vector for each imputation. The temporal decay factor $\lambda_t \in (0, 1]^N$ is introduced as follows:

$$\lambda_t = e^{\{-\max(0, \mathbf{W}_\lambda \varphi_t + \mathbf{b}_\lambda)\}} \tag{6}$$

where $\mathbf{W}_\lambda$ and $\mathbf{b}_\lambda$ are learnable parameters. Then, we compute the weighting coefficient vector, $\chi_t$, as follows:

$$\chi_t = \sigma\left(\mathbf{W}_\chi\left[\lambda_t \oplus m_t\right] + \mathbf{b}_\chi\right) \tag{7}$$

where $\mathbf{W}_\chi$ and $\mathbf{b}_\chi$ are learnable parameters, $\oplus$ denotes the concatenation operation, and the sigmoid $\sigma$ is used as activation function. $\chi_t$ is in the range $[0, 1]$, and is used to scale the proportion between the static and dynamic relations. Lastly, we estimate the missing values using an interpolation between the static, $\bar{x}$ and $x'_t$, and the dynamic relations, $\hat{x}_t$ and $\hat{z}_t$, and their corresponding coefficients, $\chi_t$, $\lambda_{1,t}$ and $\lambda_{2,t}$, as follows:

$$\begin{aligned}\hat{u}_t =& \chi_t\left[\lambda_{1,t} \odot \hat{x}_t + \left(1 - \lambda_{1,t}\right) \odot \hat{z}_t\right] \\ &+ \left(1 - \chi_t\right)\left[\lambda_{2,t} \odot x'_t + \left(1 - \lambda_{2,t}\right) \odot \bar{x}\right]\end{aligned} \tag{8}$$

where $\lambda_{1,t}$ and $\lambda_{2,t}$ are calculated using Eq. (6). Consequently, we obtain the complete observation vector $\tilde{u}_t$ by replacing the missing values with the estimates in Eq. (8) as follows:

$$\tilde{u}_t = m_t \odot x_t + (1 - m_t) \odot \hat{u}_t \tag{9}$$

The imputation module is illustrated in Fig. 4. We utilize temporal and multivariate relations by following Jung et al. (2020). Sometimes, a feature may not be available for several timepoints, which can degrade the temporal relations. Therefore, we determine the missing values of each feature by using its correlations with another feature (multivariate relations). However, a feature occasionally has few relations that can be used to obtain both temporal and multivariate features. Therefore, we consider static relations, the last valid observation and the global mean, where the former captures characteristics of the patient and the latter represents the distribution of data. In general, we impute missing values by combining static and dynamic relations.

### 3.3. The Encoding module

We exploit the deep recurrent network for temporal encoding in the forward path of the bidirectional network, and this can be applied similarly to the backward path. We devise a computational mechanism, called multifeature aggregated long short-term memory with progressive score (MAPro-LSTM), that admits the complete value at the given timepoint $\tilde{u}_t$ in Eq. (9), the complete value at the baseline timepoint, $\tilde{u}_0$, previous hidden state $h_{t-1}$, previous cell state $c_{t-1}$, temporal decay factor $\lambda_t$, masking observation $m_t$, and patient demographics $x_{d,t}$. We first embed the hidden state $h_{t-1}$ and cell state $c_{t-1}$ with temporal decay factors $\lambda_{3,t}$ and $\lambda_{4,t}$, respectively, into Eq. (6) to capture

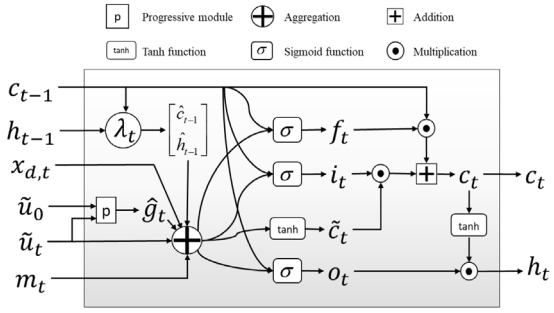



**Fig. 5.** Architecture of the *MAPro-LSTM* module for encoding temporal representation.

diverse knowledge from missing information (Che et al., 2018), $\hat{h}_{t-1}$ and $\hat{c}_{t-1}$. These can be formulated as follows:

$$\begin{aligned}\hat{h}_{t-1} &= \lambda_{3,t} \odot h_{t-1} \\ \hat{c}_{t-1} &= \lambda_{4,t} \odot c_{t-1}\end{aligned} \tag{10}$$

By taking advantage of the progressive score models in Bilgel et al. (2016) and Jedynak et al. (2012), we propose a dynamic progressive module formed by a negative exponential function as follows:

$$\hat{g}_t = \pi_t \mathbf{W}_g \sigma\left(\boldsymbol{\gamma}_g e^{-\boldsymbol{\alpha}_g\left(\tilde{u}_t - \tilde{u}_0\right)}\right) + \mathbf{b}_g \tag{11}$$

where $\mathbf{W}_g$, $\boldsymbol{\gamma}_g$, $\boldsymbol{\alpha}_g$, and $\mathbf{b}_g$ are learnable parameters, $\pi_t$ is the categorical factor used to determine whether the trajectory is decreasing or increasing, and can be expressed as a random choice between $-1$ and 1, and $\sigma$ defines a sigmoid function. The difference between Eq. (11) and the expressions in Jedynak et al. (2012) and Jung et al. (2020) is that we use features of the biomarker to compute the progression instead of age, and the parameters are learned from a deep network instead of linear models. By integrating the progressive module into the deep recurrent network, we can learn the progressive score dynamically. In addition, we insert the masking vector $m_t$ and patient demographics $x_{d,t}$ directly into the model, where $m_t$ indicates whether the input features are either observed or missing. The update functions of the proposed *MAPro-LSTM* are expressed similarly to the typical LSTM-based cell as follows:

$$\begin{aligned}f_t &= \sigma\left(\mathbf{W}_f\left[\tilde{u}_t; x_{d,t}; \hat{g}_t; \hat{h}_{t-1}; \hat{c}_{t-1}; m_{x,t}\right] + \mathbf{b}_f\right) \\ i_t &= \sigma\left(\mathbf{W}_i\left[\tilde{u}_t; x_{d,t}; \hat{g}_t; \hat{h}_{t-1}; \hat{c}_{t-1}; m_{x,t}\right] + \mathbf{b}_i\right) \\ o_t &= \sigma\left(\mathbf{W}_o\left[\tilde{u}_t; x_{d,t}; \hat{g}_t; \hat{h}_{t-1}; \hat{c}_{t-1}; m_{x,t}\right] + \mathbf{b}_o\right) \\ \tilde{c}_t &= \tanh\left(\mathbf{W}_{\tilde{c}}\left[\tilde{u}_t; x_{d,t}; \hat{g}_t; \hat{h}_{t-1}; m_{x,t}\right] + \mathbf{b}_{\tilde{c}}\right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= \tanh\left(c_t\right) \odot o_t\end{aligned} \tag{12}$$

Eq. (12) presents the forget gate $f_t$, input gate $i_t$, and output gate $o_t$ of the LSTM model, where $\mathbf{W}_f$, $\mathbf{W}_i$, $\mathbf{W}_o$, $\mathbf{b}_f$, $\mathbf{b}_i$, and $\mathbf{b}_o$ are sets of learnable parameters. The update gate $\tilde{c}_t$ is also presented with a set of learnable parameters $\{\mathbf{W}_{\tilde{c}}, \mathbf{b}_{\tilde{c}}\}$. Lastly, the cell state $c_t$ and hidden state $h_t$ are sequentially updated. The architecture of the *MAPro-LSTM* in case of the forward path is shown in Fig. 5. For the backward process, $\tilde{u}_0$ is replaced with complete features at the penultimate timepoint, and the previous states are replaced with the next ones ($h_{t+1}$ and $c_{t+1}$).

### 3.4. Forward-to-backward bidirectional strategy

As mentioned above, we use the hidden states from the RNN to update the imputation task. However, the typical bidirectional

RNN cannot update the inputs because both directions work independently and there is no connection between them. This necessitates the development of a new strategy to train the bidirectional network, after which a bidirectional strategy that can adapt and update inputs, as explained in Algorithm 1, is defined. Given the inputs, $\tilde{\boldsymbol{u}}$ as the outcome of the *Imputation* module, and $\mathbf{s}$ as a set of $[x_{d,t}, m_t]$, we can extract the hidden states $\mathbf{h}$. Instead of running the forward and backward directions separately, we use the idea presented in He, Kenton, Mike, and Luke (2017), where the forward path was run first to determine the representation of the sequence from the past to the future, and the first update $\Psi\left([\overrightarrow{h}_{t-1}]\right)$ was performed on the next inputs by using forward hidden states. Then, we ran the backward path to extract temporal features from the future to the past and perform the second update $\Psi\left([\overrightarrow{h}_{t-1}, \overleftarrow{h}_{t+1}]\right)$ on the inputs using both the forward and the backward hidden states. $\Psi$ is the sequence of operations from the *Imputation* module. To update the current hidden and cell states in both directions, $\left[\overrightarrow{h}_t, \overrightarrow{c}_t\right]$ and $\left[\overleftarrow{h}_t, \overleftarrow{c}_t\right]$, we use update functions in the *MAPro-LSTM* cell. Lastly, we collect all hidden states in the forward and backward directions in the same order of timepoints.

**Algorithm 1** Forward-to-backward Bidirectional Strategy

**Input:** $\tilde{\boldsymbol{u}} = \{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_T\}$, $\mathbf{s} = \{s_1, s_2, \ldots, s_T\}$
**Output:** $\mathbf{h} = \{h_1, h_2, \ldots, h_T\}$

*Run forward path*
**for** $t = 1, 2, \ldots, T$ **do**
  *Initialize hidden and cell states*
  **if** $t == 0$ **then**
    $\overrightarrow{h}_{t-1} \leftarrow 0, \overrightarrow{c}_{t-1} \leftarrow 0$
  **end if**
  *Update input*
  $\tilde{u}_t \leftarrow \Psi\left([\overrightarrow{h}_{t-1}]\right)$
  *Update hidden and cell states*
  $\overrightarrow{h}_t, \overrightarrow{c}_t \leftarrow \text{MAPro-LSTM}\left(\tilde{u}_t, \tilde{u}_0, s_t, \overrightarrow{h}_{t-1}, \overrightarrow{c}_{t-1}\right)$
**end for**
*Run backward path*
**for** $t = T, T-1, \ldots, 1$ **do**
  *Initialize hidden and cell states*
  **if** $t == T$ **then**
    $\overleftarrow{h}_{t+1} \leftarrow \overrightarrow{h}_t, \overleftarrow{c}_{t+1} \leftarrow \overrightarrow{c}_t$
  **end if**
  *Update input*
  $\tilde{u}_t \leftarrow \Psi\left([\overrightarrow{h}_{t-1}, \overleftarrow{h}_{t+1}]\right)$
  *Update hidden and cell states*
  $\overleftarrow{h}_t, \overleftarrow{c}_t \leftarrow \text{MAPro-LSTM}\left(\tilde{u}_t, \tilde{u}_{t+1}, s_t, \overleftarrow{h}_{t+1}, \overleftarrow{c}_{t+1}\right)$
**end for**
$\mathbf{h} = \{\overrightarrow{\mathbf{h}}, \overleftarrow{\mathbf{h}}\}$

### 3.5. The Forecasting module

The *Forecasting* module produces outcomes based on the hidden state representations $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ from the *ProLSTM* module for both the forward and the backward paths, respectively. The outcomes of MRI measurements $\tilde{x}_{t+1}$ and the clinical state $\tilde{y}_{t+1}$ of the next timepoint are formulated using multivariate linear and logistic models as follows:

$$\tilde{x}_{t+1} = \overrightarrow{\mathbf{W}}_{\tilde{x}}\overrightarrow{h}_t + \overleftarrow{\mathbf{W}}_{\tilde{x}}\overleftarrow{h}_t + \mathbf{b}_{\tilde{x}} \tag{13}$$

$$\tilde{y}_{t+1} = \text{softmax}\left(\overrightarrow{\mathbf{W}}_{\tilde{y}}\overrightarrow{h}_t + \overleftarrow{\mathbf{W}}_{\tilde{y}}\overleftarrow{h}_t + \mathbf{b}_{\tilde{y}}\right) \tag{14}$$

where $\overrightarrow{\mathbf{W}}_{\tilde{x}}$, $\overleftarrow{\mathbf{W}}_{\tilde{x}}$, $\overrightarrow{\mathbf{W}}_{\tilde{y}}$, $\overleftarrow{\mathbf{W}}_{\tilde{y}}$, $\mathbf{b}_{\tilde{x}}$, and $\mathbf{b}_{\tilde{y}}$ are learnable parameters.

### 3.6. Objective functions

To jointly train the three modules, i.e., *Imputation*, *ProLSTM*, and *Forecasting*, we define a blended loss function to update and optimize our objectives. We first measure the imputation loss $\mathfrak{L}_{\text{imputation}}$ by using the mean absolute error (MAE) between the observed data and the imputed values from different perspectives. By taking advantage of the proposal in Cao et al. (2018) to enhance the speed of convergence, the imputation loss can be expressed as follows:

$$\begin{aligned}\overrightarrow{\mathfrak{L}}_{impute} = \sum_{t=1}^{T}\Bigg(&\left|\overrightarrow{x}_t - \overrightarrow{\hat{h}}_t\right| \\ &+ \left|\overrightarrow{x}_t - \left(\overrightarrow{\lambda}_{\hat{x},t} \odot \overrightarrow{\hat{x}}_t + \left(1 - \overrightarrow{\lambda}_{\hat{x},t}\right) \odot \overrightarrow{\hat{z}}_t\right)\right| \\ &+ \left|\overrightarrow{x}_t - \left(\overrightarrow{\lambda}_{x,t} \odot \overrightarrow{x}_{t-1} + \left(1 - \overrightarrow{\lambda}_{x,t}\right) \odot \overrightarrow{\bar{x}}\right)\right|\Bigg) \odot \overrightarrow{m}_t\end{aligned}$$

$$\begin{aligned}\overleftarrow{\mathfrak{L}}_{impute} = \sum_{t=1}^{T}\Bigg(&\left|\overleftarrow{x}_t - \overleftarrow{\hat{h}}_t\right| \\ &+ \left|\overleftarrow{x}_t - \left(\overleftarrow{\lambda}_{\hat{x},t} \odot \overleftarrow{\hat{x}}_t \left(1 - \overleftarrow{\lambda}_{\hat{x},t}\right) \odot \overleftarrow{\hat{z}}_t\right)\right| \\ &+ \left|\overleftarrow{x}_t - \left(\overleftarrow{\lambda}_{x,t} \odot \overleftarrow{x}_{t-1} + \left(1 - \overleftarrow{\lambda}_{x,t}\right) \odot \overleftarrow{\bar{x}}\right)\right|\Bigg) \odot \overleftarrow{m}_t\end{aligned}$$

$$\mathfrak{L}_{\text{impute}} = \overrightarrow{\mathfrak{L}}_{impute} + \overleftarrow{\mathfrak{L}}_{impute} \tag{15}$$

where $\overrightarrow{\upsilon}_t$ and $\overleftarrow{\upsilon}_t$ denote any variable $\upsilon$ in the forward and backward processes, respectively. To predict the MRI biomarker, we formulate the empirical measurements $x_{t+1}$ and the predictions of the model $\tilde{x}_{t+1}$ as follows:

$$\mathfrak{L}_{\text{forecast}} = \sum_{t=1}^{T-1}\left(x_{t+1} \odot m_{t+1} - \tilde{x}_{t+1} \odot m_{t+1}\right)^2 \tag{16}$$

Eventually, we use the categorical focal loss (Lin et al., 2017) for clinical prediction. Focal Loss (FL) is an enhanced version of cross-entropy (CE) loss used to handle the problem of class imbalance by assigning more weights to examples that are difficult or easy to incorrectly classify, and to those whose weights can be easily reduced. The categorical focal loss can be expressed as follows:

$$\mathfrak{L}_{\text{predict}} = -\sum_{t=1}^{T-1} m_{y,t+1}\left[\tau y_{t+1}\left(1 - \tilde{y}_{t+1}\right)^{\varepsilon} \log\left(\tilde{y}_{t+1}\right)\right] \tag{17}$$

where $\tau$ and $\varepsilon$ are tunable focusing parameters, and $m_y$ is the label mask of clinical prognosis used to ignore non-label information. Lastly, the total loss function $\mathfrak{L}_{\text{total}}$ is expressed as follows:

$$\mathfrak{L}_{\text{total}} = \mathfrak{L}_{\text{impute}} + \mathfrak{L}_{\text{forecast}} + \mathfrak{L}_{\text{predict}} \tag{18}$$

### 3.7. Explainable AI (XAI) with Shapley values

The explainable AI (XAI) (Babic, Gerke, Evgeniou, & Cohen, 2021; Holzinger, Malle, Saranti, & Pfeifer, 2021; Ieracitano, Mammone, Hussain, & Morabito, 2021; Yang, Ye, & Xia, 2021) of models is a fundamental component of the machine learning workflow. Maintaining a machine learning model in a "black box" state is no longer an option. Fortunately, such analytical tools

as LIME (Ribeiro, Singh, & Guestrin, 2016), ExplainerDashboard,[1] Shapash,[2] and Dalex (Baniecki, Kretowicz, Piatyszek, Wisniewski, & Biecek, 2020) are quickly gaining popularity. In this study, we used the SHAP (Shapley Additive Explanations) (Lundberg & Lee, 2017) approach, which achieved the best paper award in the NeurIPS Workshop on Interpretable Machine Learning in Complex Systems in 2016 for explaining the outcomes of biomarker forecasting and clinical status prediction tasks. This method is based on game-theoretically optimum Shapley values. Shapley values are a commonly used concept in cooperative game theory that has a number of attractive qualities. The feature values of a data instance serve as members of a coalition. The Shapley value is the average marginal contribution of a feature value in all feasible coalitions.

Specifically, our proposed model includes multiple inputs (features, masks of missing values, temporal decay, etc.) and multiple outputs (phenotypic measurement forecasting and clinical status prediction); therefore, we calculate the Shapley values, $S_{in}^{\{c\}}$, of the $i$th target sample at the $n$th input feature from the $c$th outcome as follows:

$$S_{in}^{\{c\}} = \sum_{t=1}^{T} \frac{1}{P} \sum_{j=1}^{P} \nabla f_{jtn}^{\{c\}} * \left( x_{itn}^{\{c\}} - \phi_{jtn}^{\{c\}} \right) \tag{19}$$

and:

$$\nabla f = \frac{\partial f\,(\phi)}{\partial \phi} \tag{20}$$

where $x$ is target sample, $\phi$ is the reference sample, $\nabla f$ is the gradient computation of the trained model, $T$ is the number of timepoints, and $P$ is the number of reference samples. The gradient computation is carried out by using the differential equation of model prediction $f\,(\phi)$ on the reference samples. Finally, we sum the Shapley values of all outputs $K$ to get the mean absolute Shapley value for the $n$th feature over all target samples:

$$S_n = \frac{1}{Q} \sum_{i=1}^{Q} \left| \sum_{c=1}^{C} S_{in}^{\{c\}} \right| \tag{21}$$

where $Q$ is the number of target samples and $C$ is the number of outputs for a certain task.

## 4. Experimental results and analysis

### 4.1. Materials and settings

We used the TADPOLE challenge cohort[3] from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database,[4] which includes ADNI-1, ADNI-2, and ADNI-GO. The TADPOLE data contain approximately 1500 kinds of biomarkers (Jung et al., 2020), e.g., the cortical thickness and cortical volume of 1737 patients (ages 54.5 to 98.6 years), over 12,741 visits at 22 timepoints from 2005 to 2017. Since our main objective is to predict the annual progression, we collected 11 regular visits out of the 22 visits. We first sorted the raw data by deleting timepoints that had been duplicated and those with incorrect measurements, filtered irreversible patients who had constantly changed their state from AD to MCI or from MCI to CN during historical observations, and removed patients with no baseline diagnosis, or those who had already been diagnosed as having AD at the baseline timepoint. Because our main objective was to forecast both clinical status and MRI measurements, we considered only patients who had at least two visits. Similar to the work in Bilgel et al. (2019), we merged the groups of CN to MCI and MCI to CN into MCI, those of CN to dementia and MCI into dementia, and those of dementia to MCI into AD to obtain three categories: CN, MCI, and AD. We obtained data on 1335 patients in total with 11,048 timepoints, which resulted in stable data on 892 patients whose status had not been changed to AD dementia during historical observations, and 443 progressive patients, who had not been initially diagnosed as having AD dementia but whose diagnoses were changed during historical observations. The statistics are shown in Table 1.

**Table 1**
The statistics of raw and clean TADPOLE data.

| Characteristic | Raw data | Clean data |
|---|---|---|
| No. of patients | 1737 | 1335 |
| No. of timepoints | 12,741 | 11,048 |
| Maximum visits/patient | 22 | 11 |
| Average visits/patient | 5.76 | 5.64 |
| Male/Female | 957/780 | 740/595 |
| Range of age | 54.5∼98.6 | 55.0∼91.4 |

By following the work in Jung et al. (2020), Oxtoby et al. (2018) and Ghazi et al. (2019), we considered the volumes of the ventricles, hippocampus, fusiform gyrus, middle temporal gyrus, entorhinal cortex, and the entire brain, which are extracted from T1-weighted MRI, as biomarker features (Bio.). We also included patient demographics (age (Age), gender (Gen.), education (Edu.), race (Race), ethnicity (Eth.), and marital status (M.S.)) as patient demographic features. An example of the data structure is shown in Fig. 6. To compensate for intersubject variability in brain sizes, as in Ghazi et al. (2019), we normalized the volumetric MRI biomarkers with an individual intracranial volume (ICV). We then standardize the MRI features based on the mean and standard deviation (standard score) so that they were within a standard normal distribution. For the demographic data, we encoded gender, race, ethnicity, and marital status by one-hot encoding, encoded age by z-score normalization, and education by dividing it with the maximum value. The missing rates of the six volumes are shown in Fig. 7.

Of the 11 collected timepoints, we used the first 10 visits, including the baseline, to forecast for several years. For each timepoint, the forecast was conducted based on, if unacknowledged or missing, all observed historical values and those predicted or imputed. We reported the average results of *five*-fold cross-validation, which involved randomly separating the dataset into five partitions: One part was used as the test set and the rest for training, with a ratio of 8 : 2. To evaluate performance on the imputation task, we randomly removed 20% of the real observation values in the samples used for training validation, and testing, and then used them as the ground truth for the missing values. We used all observations for the clinical status prediction and MRI biomarker forecasting tasks.

For quantitative evaluation, we used the same metrics as in Jung et al. (2020): the MAE and mean relative error (MRE) for the imputation task, the MAE for the MRI biomarker forecasting task, and multiclass area under the receiver-operating characteristic curve (mAUC), accuracy (ACC), precision (PRE), and recall (REC) scores for the clinical status prediction task. We compared our model with those reported in Che et al. (2018), Jung et al. (2020), Nguyen et al. (2020) and Yoon et al. (2019). We conducted experiments by using static imputation methods, such as zero-padding

[1] https://explainerdashboard.readthedocs.io/.

[2] https://shapash.readthedocs.io/.

[3] Available at https://tadpole.grand-challenge.org/Data/.

[4] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership led by Principal Investigator Michael W. Weiner, MD. Its primary goal has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, the interested reader can see www.adni-info.org.

| ID | Visit | Age | Gender | Edu. | Race | Ethnic | Marr. Sta. | Ventricles | Hippocampus | Fusiform | Mid. Temp. | Entorhinal | Whole-brain | ICV | Diagnosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 78.2 | Male | 18 | White | Not Hisp./Latio | Married | 48.933 | 4.087 | 16.454 | 16.009 | 2.784 | 952.780 | 1,519.690 | MCI |
| 1 | 1 | 79.1 | Male | 18 | White | Not Hisp./Latio | Married | 52.933 | 3.858 | 17.696 | 15.444 | 2.773 | 920.277 | 1,520.850 | AD |
| 1 | 3 | 81.1 | Male | 18 | White | Not Hisp./Latio | Married | 58.368 | 3.490 | 17.079 | 15.745 | 2.350 | 910.787 | 1,509.600 | AD |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| N | 0 | 74.5 | Female | 16 | Asian | Not Hisp./Latio | Married | 24.880 | 7.945 | 18.497 | 19.969 | 5.430 | 967.608 | 1,527.570 | AD |
| N | 2 | 76.4 | Female | 16 | Asian | Not Hisp./Latio | Married | 30.357 | 7.653 | 17.629 | 17.327 | 4.259 | 919.675 | 1,528.880 | AD |

Edu. : *education* – Marr. Sta. : *marrage status* – Mid. Temp : *middle temporal* – ICV : *intracranial volume* – Hisp./Latio : *Hispanic and Latino*
Volume unit in *milliliter* - Visit 0 means baseline observation

**Fig. 6.** Example of TADPOLE data structure.

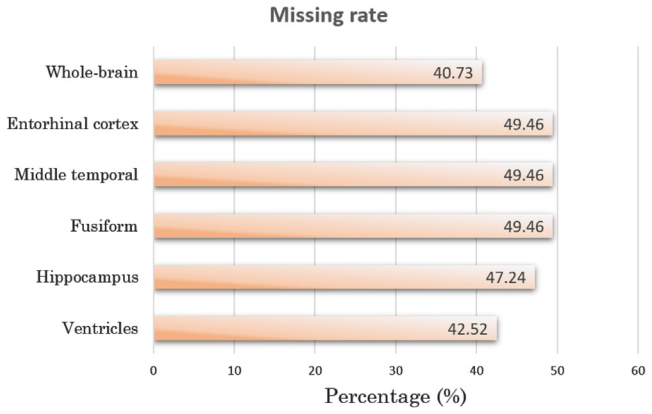



**Fig. 7.** Missing rate of volumetric MRI biomarkers.

**Table 2**
Hyper-parameter settings.

| Characteristic | Parameter |
|---|---|
| RNN hidden units | 64 |
| Optimizer | ADAM |
| No. of epochs | 100 |
| Batchsize | 32 |
| Learning rate | $3 \times 10^{-3}$ |
| Focal loss $(\tau; \varepsilon)$ (Lin et al., 2017) | (0.25; 2) |

(Simple – *S*), mean value (Mean – *M*), and the most recent observation (Forward – *F*) followed by the GRU, GRU-S, GRU-M, and GRU-F. The parameter settings for all models are presented in Table 2. We also report the number of parameters and the time required for inference in the training process for each model in Fig. 8. The calculation of the time needed for inference was based on an average of 100 epochs with a batchsize of 32. Fig. 8 shows that due to the use of bidirectional networks, our proposed BiPro incurred a higher computational cost (in terms of both the number of parameters and inference time) than the other methods. This is a limitation of the proposed method. However, since the model has fewer than one million parameters converged quickly, in fewer than 1000 epochs, the training time was less than 200 seconds. It can thus be used in practice.

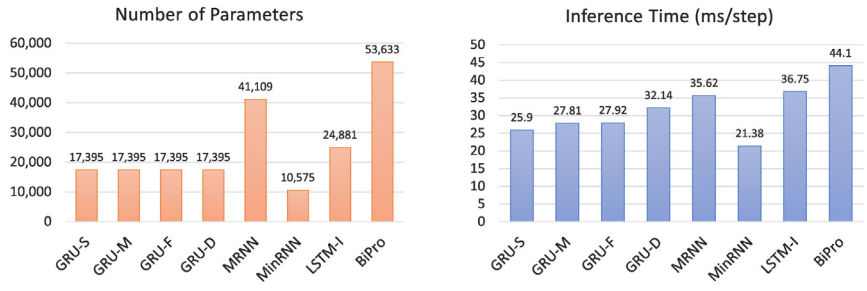
### 4.2. Performance of recent methods based on machine learning and deep learning on AD detection

Various methods have been used to diagnose Alzheimer's disease in the literature, and can be categorized into two groups: machine learning-based methods and deep learning-based methods. In this research, we highlight the most recent and efficient strategies that have been recommended in the literature from 2018 to 2020. Recently published studies on Alzheimer's disease diagnosis using machine learning are shown in Table 3 and those that use deep learning-based are illustrated in Table 4. The machine learning-based approaches use the support vector machine (SVM) with radial basis function (RBF) (Tabarestani, Aghili, Shojaie, Freytes, & Adjouadi, 2018), principal component analysis (PCA) (Cui et al., 2018), generalized linear model (GLM) (Shahbaz, Ali, Guergachi, Niazi, & Umer, 2019), and personalized Gaussian process experts (pGPE) (Utsumi et al., 2019) to classify AD status or estimate cognitive measurements. Of deep learning-based approaches, the 3D convolution neural network (3DCNN) (Xia et al., 2020) and the high-level layer concatenation autoencoder (HiLCAE) (Vu et al., 2018) have been used to learn spatial features, and the deep neural network (DNN) (Park, Ha, & Park, 2020), 3DVGG16 (Vu et al., 2018), sparse autoencoder (SAE) (Martinez-Murcia, Ortiz, Gorriz, Ramirez, & Castillo-Barnes, 2019), deep ensemble learning (DELearning) (An, Ding, Yang, Au, & Ang, 2020), 3D convolutional long short-term memory (3DCLSTM) (Xia et al., 2020), and bidirectional long short-term memory (BiLSTM) (Abuhmed, El-Sappagh, & Alonso, 2021) have been used to detect the AD state or localize AD-related regions. In general, all methods used features of patients obtained in recent visits to diagnose them instead of predicting progression.

### 4.3. Performance in terms of missing value imputation

The experimental performance of the methods on imputing the missing values in the data is shown in Table 5. Our proposed method achieved the lowest MAE and MRE scores, thus outperforming all other methods. Its MAE was 6.861 ml and MRE was 9.816%, improvements of 0.74 ml and 1.124%, respectively. We did not compare our model with those developed in Nguyen et al. (2020) since their model did not support the imputation task. We also performed a t-test analysis to compare our proposed BiPro model with the other models. For all comparisons, $p < 0.05$ as calculated by ANOVA, verifying that our proposed model was significantly different from the other models.

Our proposed method considered systematically missing patterns based on temporal relations, multivariate relations, the empirical mean and final observation, and time delay through a series of operations. This systematic imputation is beneficial, and significantly enhanced performance. In addition, we show the results of imputation in terms of individual volume error in Fig. 9. Fig. 9(a) presents the mean absolute error in the estimated whole-brain volume in the right-vertical axis and the other volumes




(a) Number of parameters (b) Inference time (ms/step)

**Fig. 8.** Number of parameters and inference times of all models.

**Table 3**
Machine learning-based approaches for Alzheimer's disease detection.

| Author | Year | Target | Method | Modality | Dataset | Performance |
|---|---|---|---|---|---|---|
| Tabarestani et al. (2018) | 2018 | MMSE estimation | SVM-RBF | Clinical data | ADNI | RMSE of 3.06 |
| Cui et al. (2018) | 2018 | NC vs AD | PCA | fMRI | ADNI | Accuracy of 91.3% |
| Bucholc et al. (2019) | 2019 | Multiclass AD classification | SVM | Cognitive assessments | ADNI | Accuracy of 83% |
| Shahbaz et al. (2019) | 2019 | Multiclass AD classification | GLM | Clinical data | ADNI | Accuracy of 88.24% |
| Utsumi et al. (2019) | 2019 | Cognitive changes | pGPE | Clinical data | ADNI | MAE of 2.65 |
| Afzal et al. (2019) | 2019 | Multiclass AD classification | SVM | MRI | ADNI | Accuracy of 92.4% |

**Table 4**
Deep learning-based approaches for Alzheimer's disease detection.

| Author | Year | Target | Method | Modality | Dataset | Performance |
|---|---|---|---|---|---|---|
| Vu et al. (2018) | 2018 | AD classification | HiLCAE-3DVGG16 | MRI PET | ADNI | Accuracy of 98.8% |
| Martinez-Murcia et al. (2019) | 2019 | AD classification | SAE | MRI | ADNI | Accuracy of 86.47% |
| Huang et al. (2019) | 2019 | AD classification | 3DCNN | MRI PET | ADNI | Accuracy of 90.1% |
| An et al. (2020) | 2020 | AD classification | DELearning | Clinical data | NACC | Accuracy of 85% |
| Basher et al. (2019) | 2019 | Hippocampus localization | Ensemble Hough-CNN | MRI | ADNI, GARD | RMSE of 2.24 mm |
| Maqsood et al. (2019) | 2019 | Multiclass AD classification | CNN | MRI | OASIS | Accuracy of 92.85% |
| Xia et al. (2020) | 2020 | AD classification | 3DCNN-3DCLSTM | MRI | ADNI | Accuracy of 94.19% |
| Park et al. (2020) | 2020 | AD classification | DNN | Genes DNA | RMH 44k 1.1, IHM 450 | Accuracy of 82.3% |
| Abuhmed et al. (2021) | 2021 | AD progression detection | BiLSTM | Clinical data | ADNI | Accuracy of 86.08% |

**Table 5**
Performance on the imputation task in terms of MAE and MRE (MEAN±STD).

| Model | MAE (ml) | MRE (%) |
|---|---|---|
| GRU-S | 11.322 ± 0.480 | 15.289 ± 0.857 |
| GRU-M | 14.554 ± 0.950 | 15.475 ± 0.736 |
| GRU-F | 9.378 ± 1.774 | 12.208 ± 0.592 |
| GRU-D (Che et al., 2018) | 13.784 ± 0.374 | 15.160 ± 0.729 |
| MRNN (Yoon et al., 2019) | 11.78 ± 0.832 | 15.383 ± 0.78 |
| LSTM-I (Jung et al., 2020) | 7.601 ± 0.361 | 10.94 ± 0.343 |
| BiPro (Proposed) | **6.861 ± 0.396** | **9.816 ± 0.511** |

The best performance is indicated in boldface.

in the left-vertical axis. In addition, Fig. 9(b) presents the mean relative error in the volume of the ventricle in the right-vertical axis and the other volumes in the left-vertical axis. These figures show that our proposed model obtained the lowest error rates of all models.

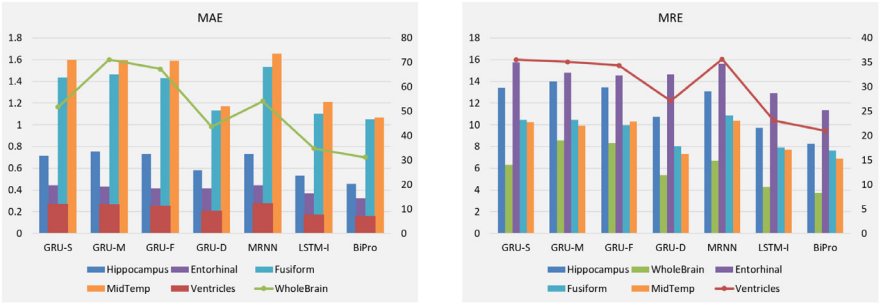
### 4.4. Performance in terms of MRI biomarker forecasting

The forecasting errors in terms of the MAE over the MRI biomarkers are presented in Table 6. It expresses combinations of patient-related information and the biomarkers. By capturing the temporal and multivariate relations inherent in the data as well as the global mean and the last observed values, our proposed model outperformed the state-of-the-art method proposed by Jung et al. (2020), with a margin of 3.11 ml when using only biomarkers and 2.6 ml when combining all patient-related information with the biomarkers. Table 6 shows that our model achieved the best performance with a combination of age and biomarkers. In general, patient information did not significantly support the task of MRI biomarker forecasting compared with when only biomarkers were used. All pairwise comparisons were statistically significant ($p < 5 \times 10^{-4}$ as calculated by ANOVA).

In addition, we tested the performance in terms of forecasting MRI biomarkers for different MRI volumes with each combination, as shown in Fig. 10. In this figure, the right-vertical axis measures the MAE of the ventricles and the whole-brain volumes, while and the left-vertical axis measures the MAEs of the hippocampus, entorhinal, fusiform, and middle temporal volumes. In all cases, the proposed model obtained the lowest error on every predicted volume.

### 4.5. Performance in terms of clinical status prediction

For the clinical status prediction task, we also report the ACC, PRE, and REC scores along with the mAUC, as shown in Table 7. We present the comparison between the proposed method and the other methods by using a combination of all information and biomarkers. Our proposed BiPro delivered the best performance on all metrics. In particular, it outperformed all other methods to achieve the highest ACC, PRE, REC, and mAUC scores, with margins of 3.25%, 2.93%, 3.23%, and 2.58%, respectively, compared with the LSTM-I (Jung et al., 2020) method. The pairwise comparisons were statistically significant ($p < 0.002$ as calculated by ANOVA).




(a) Mean Absolute Error (b) Mean Relative Error

**Fig. 9.** Performance in terms of missing value imputation for individual volumes. MAE was calculated in milliliters and MRE in percentages.

**Table 6**
Performance on forecasting MRI biomarkers for the next year visit in terms of the MAE (MEAN± STD) in milliliter (ml).

| Feature | GRU-S | GRU-M | GRU-F | GRU-D (Che et al., 2018) | MRNN (Yoon et al., 2019) | MinRNN (Nguyen et al., 2020) | LSTM-I (Jung et al., 2020) | BiPro (Proposed) |
|---|---|---|---|---|---|---|---|---|
| Bio. | 7.51 ± 1.3 | 7.26 ± 1.05 | 7.4 ± 1.53 | 7.55 ± 1.57 | 9.33 ± 2.04 | 6.77 ± 1.06 | 6.64 ± 0.97 | **3.53 ± 0.47** |
| Age + Bio. | 8.18 ± 1.3 | 7.98 ± 1.18 | 7.84 ± 1.44 | 8.22 ± 1.48 | 8.38 ± 1.94 | 6.76 ± 0.87 | 6.63 ± 1.24 | **3.49 ± 0.25** |
| Gen. + Bio. | 8.1 ± 1.62 | 8.04 ± 1.23 | 8.02 ± 1.42 | 8.18 ± 1.53 | 8.91 ± 1.32 | 6.98 ± 1.09 | 6.87 ± 1.43 | **3.69 ± 0.64** |
| Edu. + Bio. | 8.07 ± 1.31 | 7.65 ± 1.15 | 7.84 ± 1.51 | 8.05 ± 1.6 | 8.62 ± 1.37 | 6.76 ± 1.21 | 6.5 ± 1.11 | **3.62 ± 0.4** |
| Race + Bio. | 7.85 ± 1.71 | 7.62 ± 1.21 | 7.75 ± 1.41 | 7.6 ± 1.33 | 8.77 ± 1.59 | 6.94 ± 1.1 | 6.5 ± 1.24 | **3.59 ± 0.31** |
| Eth. + Bio. | 7.59 ± 1.11 | 7.9 ± 1.05 | 7.83 ± 1.26 | 8.19 ± 1.76 | 8.22 ± 0.97 | 7.07 ± 0.86 | 6.48 ± 1.21 | **3.66 ± 0.5** |
| M.S. + Bio. | 7.71 ± 1.16 | 7.72 ± 1.06 | 7.86 ± 1.2 | 7.68 ± 1.12 | 8.7 ± 1.59 | 6.69 ± 1.02 | 6.82 ± 1.01 | **3.77 ± 0.68** |
| All + Bio. | 8.37 ± 1.37 | 8.06 ± 1.22 | 8.47 ± 1.72 | 8.31 ± 1.6 | 9.07 ± 1.52 | 7.59 ± 1.06 | 6.78 ± 1.23 | **4.18 ± 0.23** |

The smallest MAE is highlighted in boldface.

**Table 7**
Performance of multiple classification (CN vs. MCI vs. AD) for the next one-year patient visits in terms of ACC, PRE, REC, and mAUC (MEAN± STD).

| Model | ACC | PRE | REC | mAUC |
|---|---|---|---|---|
| GRU-S | 52.48 ± 3.37 | 53.81 ± 2.09 | 52.35 ± 3.64 | 70.61 ± 2.41 |
| GRU-M | 53.52 ± 2.58 | 54.02 ± 2.7 | 53.51 ± 2.14 | 74.03 ± 2.3 |
| GRU-F | 54.49 ± 2.58 | 55.28 ± 2.67 | 53.63 ± 3.27 | 74.77 ± 2.13 |
| GRU-D | 53.22 ± 2.54 | 53 + .94 ± 3.23 | 52.94 ± 2.92 | 74.24 ± 2.23 |
| MRNN | 54.06 ± 3.13 | 56.44 ± 3.98 | 53.58 ± 3.19 | 74.77 ± 2.74 |
| MinRNN | 54.37 ± 1.84 | 54.96 ± 1.69 | 53.56 ± 2.06 | 75.14 ± 2.07 |
| LSTM-I | 55.22 ± 3.12 | 56.32 ± 2.94 | 54.55 ± 2.96 | 75.66 ± 1.92 |
| BiPro (Proposed) | **58.47 ± 2.94** | **59.25 ± 3.7** | **57.78 ± 3.44** | **78.24 ± 1.11** |

The best performance on each metric is highlighted in boldface.

For further comparison, we show the confusion matrices over all methods as well as the proposed BiPro when all patient-related information and biomarkers were used, as shown in Fig. 11. The results confirm that the proposed model can deliver balanced performance on the three classes: CN, MCI, and AD. The other models performed well for only one class, e.g., GRU-M, GRU-F, GRU-D, and MRNN performed well on the CN class, and MinRNN and LSTM-I on the MCI class. We show the performance of the models when only biomarkers were used, and when a combination of the biomarkers and one demographic feature was used, in Fig. 12 in terms of accuracy, precision, recall, and mAUC. In general, the proposed BiPro model outperformed all other models, with the patient age having the greatest influence on performance.

Furthermore, to understand how partial loss functions are learned, we visualized the clinical data, biomarkers, and imputation losses over 100 epochs by using the MRNN, LSTM-I, and the proposed BiPro models, as shown in Fig. 13. Because the GRU-D and MinRNN did not update the imputation loss, we show only the histories of the losses of clinical and biomarker data for these models. The figure presents the case of using biomarkers with all patient demographics. It shows that the proposed BiPro learned all three tasks better than the other models, and yielded fewer outliers and a lower overall cost over time.

### 4.6. Additional verification

#### 4.6.1. Results without the Imputation module

To efficiently validate the proposed *BiPro*, we excluded the *Imputation* module from the system. To this end, all missing values were imputed by using the global mean values for all models. The results for MRI biomarker forecasting are presented in Tables 8 and 9. They show that the proposed BiPro model slightly outperformed the other, conventional, models without the Imputation module, with the lowest MAE for MRI biomarker forecasting (Table 8) in most cases except those involving combinations of (race + biomarker) and (all patient information + biomarker), and the highest mAUC in all combinations for clinical status prediction (Table 9). The Imputation module thus significantly supports MRI biomarker forecasting, and supports the prediction of the clinical status of the patient to a lesser extent.

#### 4.6.2. Results on data for stable and progressive patients

In this subsection, we examine the performance of the methods on groups with data on stable and progressive patients. Stable

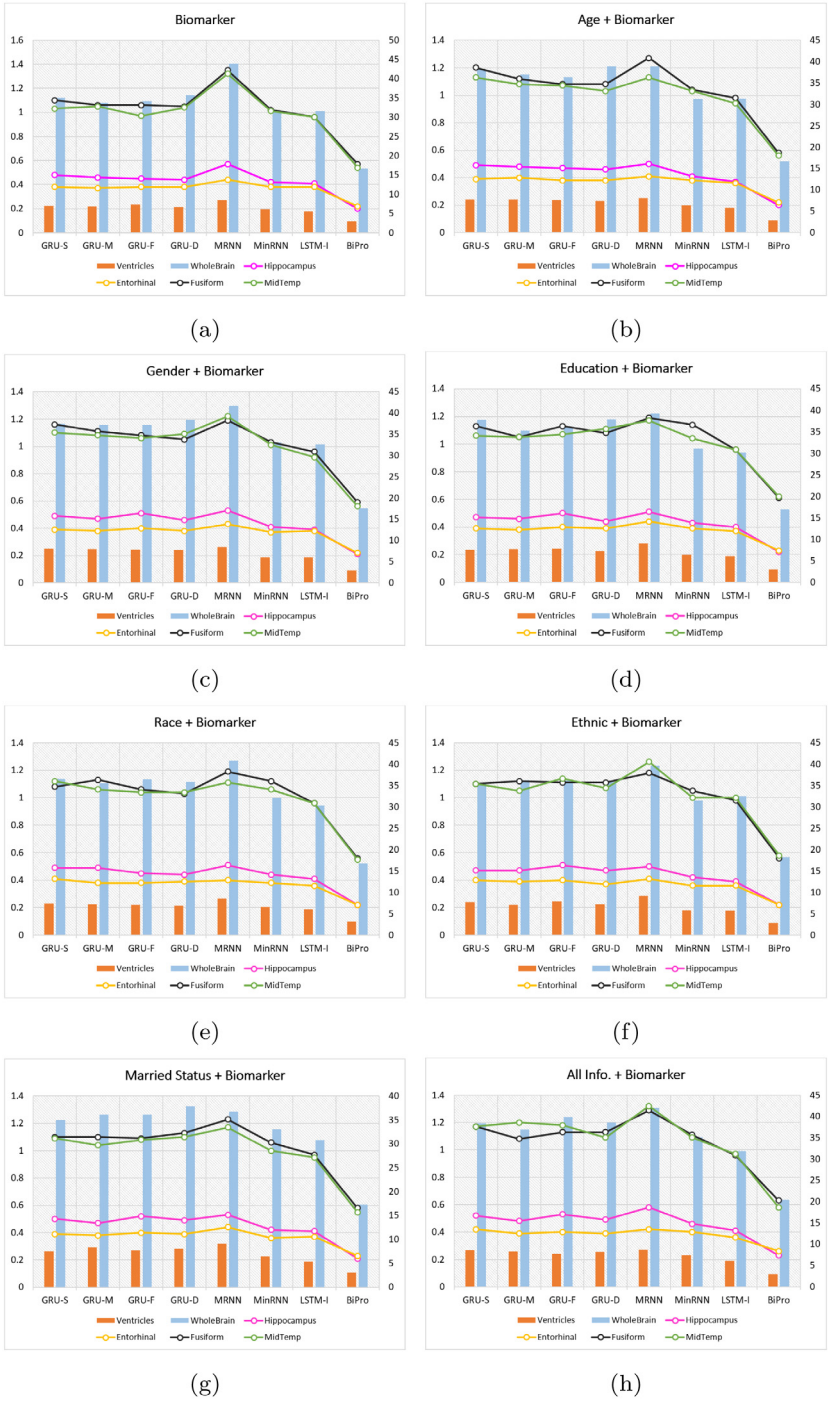



**Fig. 10.** Performance in terms of MRI biomarker forecasting using each combination. The right-vertical axis measures the MAE of the ventricles and whole-brain volumes while the left-vertical axis measures the MAE of the hippocampus, entorhinal, fusiform, and middle temporal volumes.

**Table 8**
Performance of the methods on MRI biomarker forecasting without missing value imputation. Values are in the form (MEAN± STD).

| Feature | GRU-D (Che et al., 2018) | MRNN (Yoon et al., 2019) | MinRNN (Nguyen et al., 2020) | LSTM-I (Jung et al., 2020) | BiPro (Proposed) |
|---|---|---|---|---|---|
| Bio. | 7.12 ± 0.92 | 8.73 ± 1.77 | 6.92 ± 1.28 | 7.91 ± 1.24 | **6.23 ± 0.8** |
| Age + Bio. | 8.01 ± 1.46 | 9.06 ± 1.73 | 6.87 ± 1.41 | 8.52 ± 1.46 | **6.81 ± 1.33** |
| Gen. + Bio. | 7.58 ± 1.16 | 9.06 ± 1.09 | 6.96 ± 1.29 | 8.16 ± 1.2 | **6.89 ± 1.23** |
| Edu. + Bio. | 7.38 ± 1.11 | 8.94 ± 1.76 | 6.59 ± 0.87 | 8.06 ± 1.18 | **6.32 ± 1.12** |
| Race + Bio. | 7.31 ± 0.84 | 9.47 ± 2.3 | **7.02 ± 1.07** | 8.16 ± 1.31 | 7.2 ± 1.46 |
| Eth. + Bio. | 7.66 ± 0.99 | 9.05 ± 1.88 | 7.01 ± 1.18 | 8.15 ± 1.21 | **6.72 ± 0.83** |
| M.S. + Bio. | 7.55 ± 1.46 | 8.8 ± 1.92 | 7.0 ± 1.22 | 7.9 ± 0.96 | **6.34 ± 1.08** |
| All + Bio | **8.25 ± 1.35** | 8.84 ± 1.89 | 7.67 ± 1.32 | 8.66 ± 1.3 | 8.35 ± 0.8 |

The best performance on each metric is highlighted in boldface.

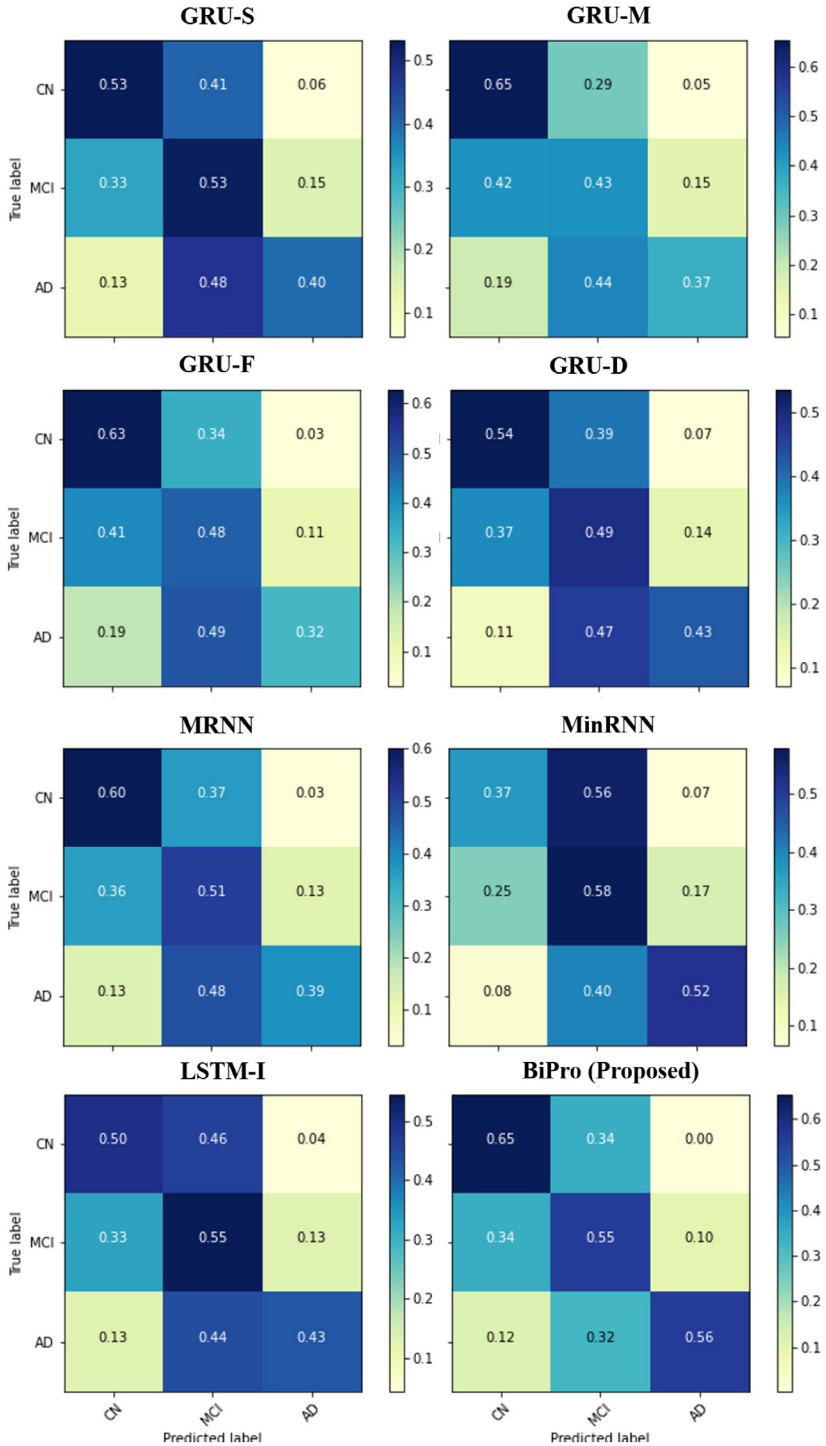



**Fig. 11.** Confusion matrices of the eight methods obtained by using all patient information and biomarkers.

patients included those with stable CN who had remained in the CN state throughout the observed time, and those with stable MCI who had begun with an MCI state that had not transformed into AD. Progressive patients included those who had begun with the CN state and had later transitioned to MCI or AD (progressive CN), or who had begun with the MCI state and later transitioned to AD (progressive MCI). Fig. 14(a) shows the box plot of the methods on both stable and progressive patients in terms of MRI biomarker forecasting, where the true median of the proposed BiPro model differs from those of the other models with 95% confidence. The box-plot predictions of the clinical status of both groups of patients are shown in Fig. 14(b), where the difference between BiPro's median and those of the other models is greater than 2%. This supports our conclusion that the true median of the proposed BiPro model differed from those of the other models with 95% confidence. In a comparison between data on stable and progressive patients, the performance of the methods on the stable group was better than on the progressive group on both tasks, as more daunting challenges are involved in capturing progressive patterns of AD.

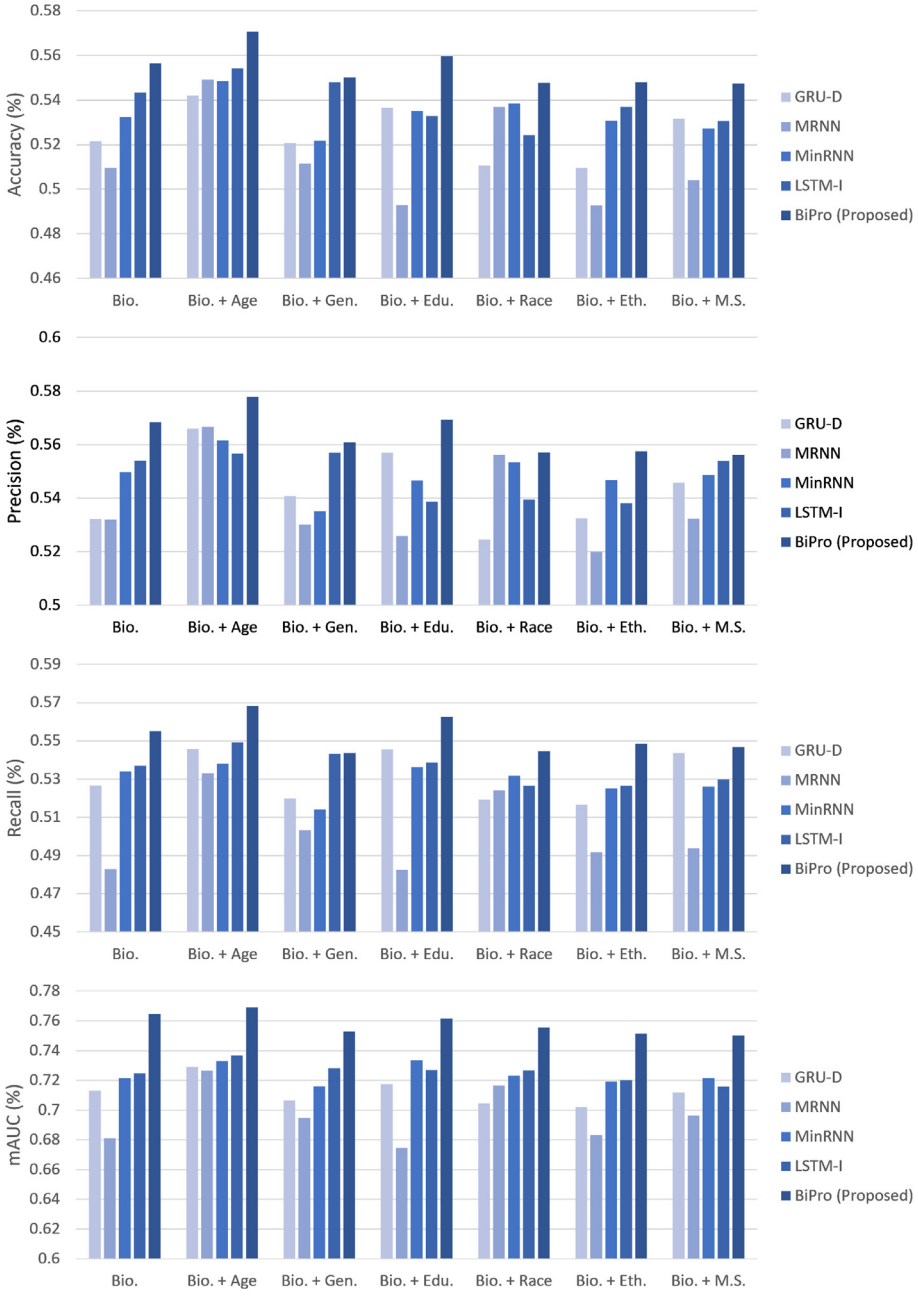



**Fig. 12.** Comparison of different combinations of patient information and biomarkers.

**Table 9**
Performance of all methods on clinical status prediction without missing value imputation. Values are in the form (MEAN± STD).

| Feature | GRU-D (Che et al., 2018) | MRNN (Yoon et al., 2019) | MinRNN (Nguyen et al., 2020) | LSTM-I (Jung et al., 2020) | BiPro (Proposed) |
|---|---|---|---|---|---|
| Bio. | 68.91 ± 2.79 | 66.67 ± 5.28 | 68.29 ± 1.72 | 70.47 ± 2.14 | **74.91 ± 2.95** |
| Age + Bio. | 71.04 ± 1.74 | 70.96 ± 5.94 | 71.13 ± 2.11 | 70.78 ± 1.24 | **75.74 ± 1.36** |
| Gen. + Bio. | 70.37 ± 1.46 | 66.03 ± 2.31 | 70.76 ± 2.09 | 71.13 ± 1.11 | **73.29 ± 1.63** |
| Edu. + Bio. | 68.39 ± 1.43 | 65.47 ± 2.34 | 70.53 ± 1.66 | 70.6 ± 1.59 | **74.66 ± 2.09** |
| Race + Bio. | 69.03 ± 1.99 | 70.89 ± 2.58 | 70.96 ± 1.98 | 69.8 ± 1.37 | **74.38 ± 1.74** |
| Eth. + Bio. | 68.04 ± 2.32 | 67.79 ± 2.19 | 68.08 ± 2.22 | 70.85 ± 2.25 | **73.28 ± 2.73** |
| M.S. + Bio. | 70.73 ± 2.71 | 68.68 ± 2.3 | 70.61 ± 2.11 | 69.46 ± 0.93 | **72.26 ± 1.24** |
| All + Bio. | 72.18 ± 2.24 | 73.91 ± 1.95 | 74.19 ± 1.93 | 74.67 ± 1.1 | **77.02 ± 0.68** |

The best performance in each metric is highlighted in boldface.

*4.6.3. Results with different timepoint prediction*

In this section, we examine the predictions of the methods at different timepoints. Given a timepoint $t$, we used data from the baseline up to the $t$th timepoint to make 10 year predictions of the clinical status of patients and their MRI biomarkers. Fig. 15(a) illustrates the results of MRI biomarker forecasting and 15(b)

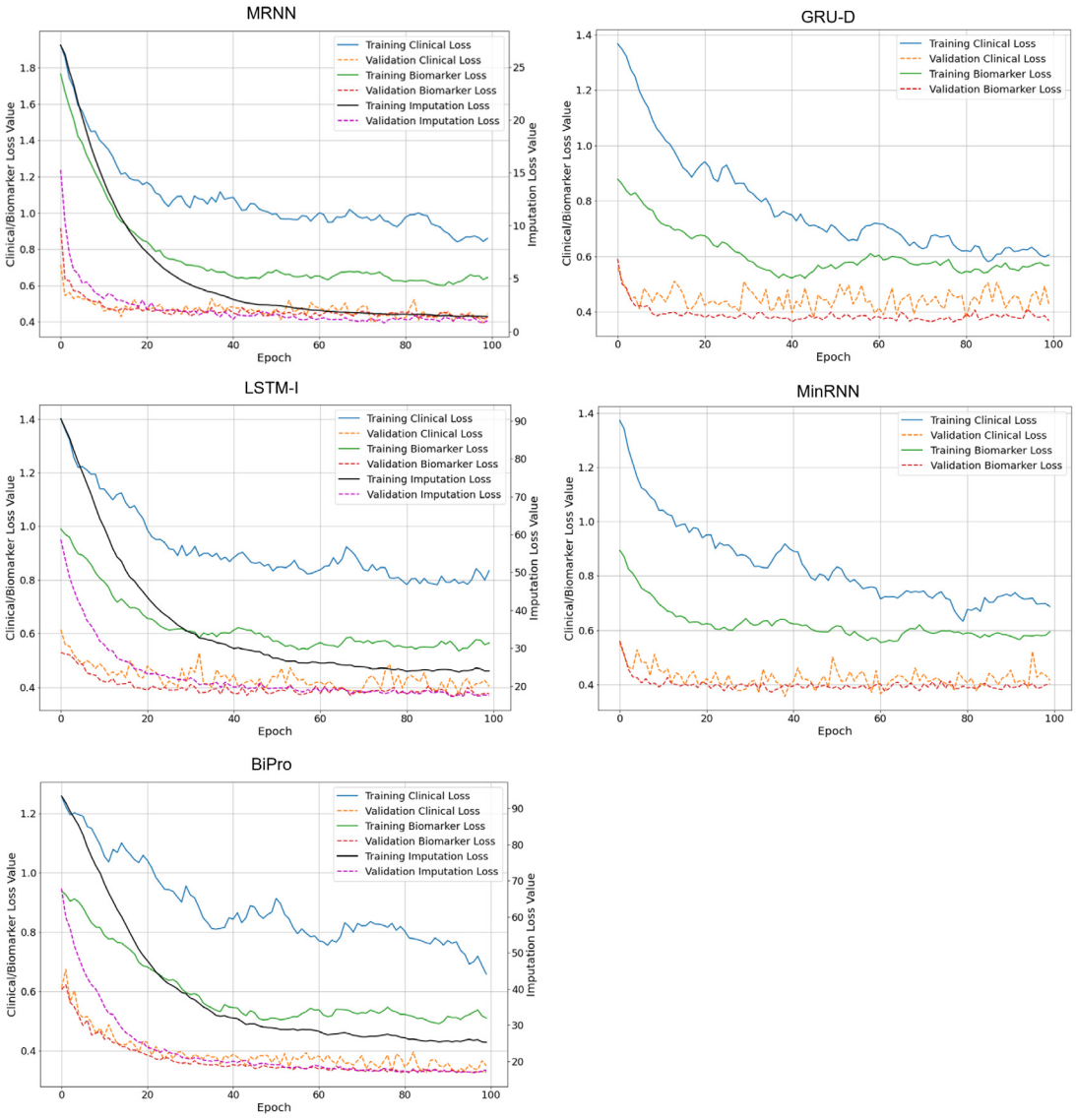
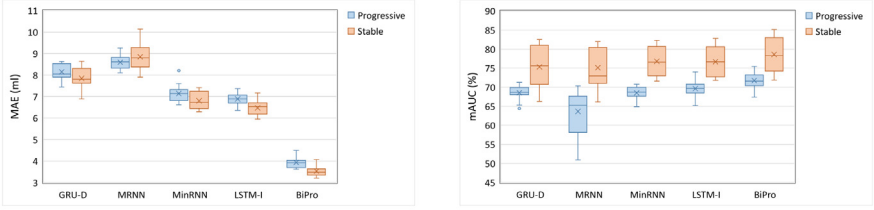



**Fig. 13.** Visualizing the partial losses on three learning tasks.




(a) MRI Biomarker Forecast (b) Clinical Status Prediction

**Fig. 14.** Box-plot of predictions of AD progression in case of data on stable and progressive patients.

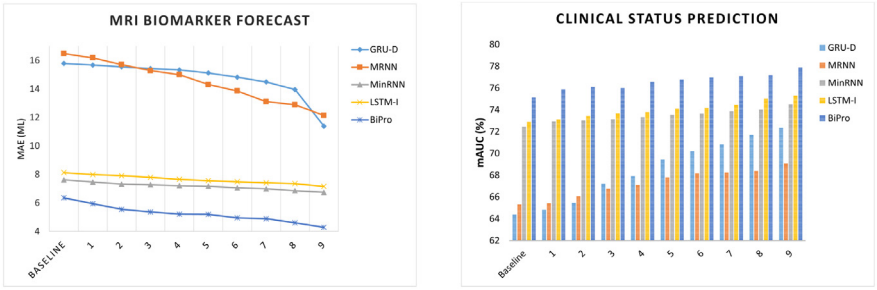
shows the performance of the methods in terms of predicting clinical status using data accumulated up to time $t$. The performance of the models improved proportionally with the number of timepoints used for prediction.

*4.6.4. Explainable artificial intelligence analysis*

In order to be feasible, AI not only needs to exhibit good decision-making performance, but also needs to explain these decisions and convince us that they are correct. We analyzed the features that were the most important for successful predictions by the models. We use the SHAP (SHapley Additive exPlanations) toolbox.[5] to visualize feature importance as shown in Fig. 16.

Based on the outcomes on biomarker forecasting (see Fig. 16(a)), the ventricles were the most important biomarker, with an average SHAP value of 0.85. Biomarker-related features clearly made a stronger contribution to biomarker forecasting than features related to patient information. The outcome

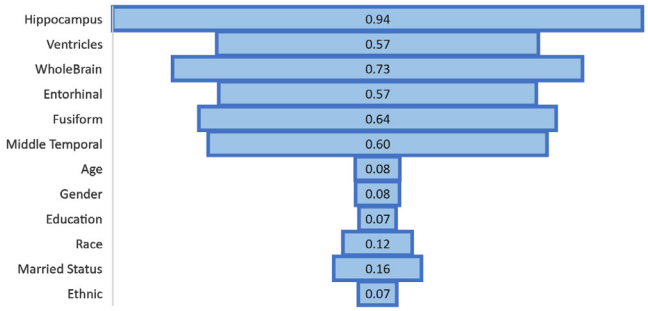
5 https://github.com/slundberg/shap



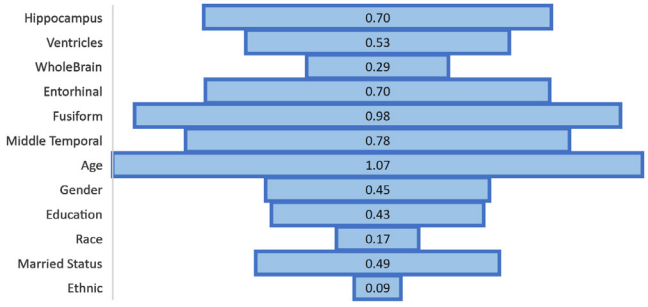

(a) MRI biomarker forecasting (b) Clinical status prediction

**Fig. 15.** Performance obtained using data accumulated up to time *t*.

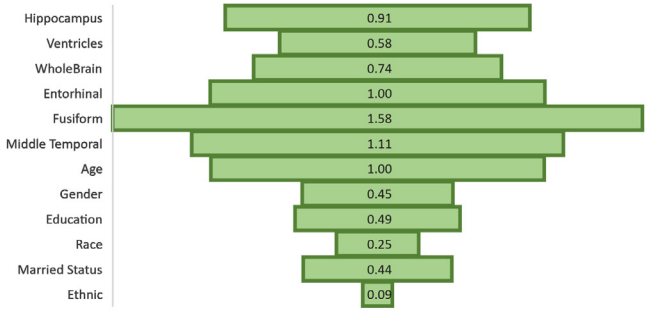
of clinical status prediction (see Fig. 16(b)) shows that the hippocampus was the most important feature, with an average SHAP value of 1.66, followed by the middle temporal gyrus (1.4), entorhinal cortex (1.21), patient age (1.2), and fusiform gyrus (0.9). The other features, such as gender, ventricles, whole-brain volume, education, race, and ethnicity were less important, and SHAP values of less than 0.4. A combination of the two outcomes (see Fig. 16(c)) shows that the five most important features for model prediction were those related to the hippocampus, fusiform gyrus, entorhinal cortex, middle temporal gyrus, and patient age.




(a) Biomarker forecasting outcome




(b) Clinical status prediction outcome




(c) Both outcomes

**Fig. 16.** Visualization of feature importance using Shapley values based on the outcomes.

## 5. Conclusions and future work

In this study, we proposed BiPro, a model for disease progression based on the RNN architecture with progressive LSTM cells, for tracking the progression of AD. We tested it on the TADPOLE challenge cohort at several future timepoints in comparison with other methods in the area. BiPro performs three tasks at the same time: missing value imputation, MRI biomarker forecasting, and clinical status prognosis across multiple timepoints. It has the capacity to anticipate both patient data and status for future visits when these three operations are combined, and can model the 10 year course of Alzheimer's disease. We used TADPOLE to analyze the proposed model and verify its feasibility. The following are the primary contributions of this study:

- We proposed an end-to-end bidirectional RNN-based model to execute three concurrent tasks to predict the progression of Alzheimer's disease using long-term observations. We also proposed a forward-to-backward bidirectional strategy for encoding sequential characteristics.
- To the best of our knowledge, the BiPro model is the first that can perform the integrative imputation of missing values in longitudinal data, and considers both static and dynamic relations. As a result, it can provide reliable estimations of missing data that can be used in subsequent prediction. Consider the following scenario: Given a patient subjected to a first test, we may use these measurements in the BiPro model to forecast their data as well as their status with regard to Alzheimer's disease at the next visit. We can then use the prediction as input for the next timepoint, and so on. In this way, we can predict the course of illness of the patient even if only data from only one time point are available.
- By utilizing the expression for the progression score in Jedynak et al. (2012), we adapted the LSTM cell with a progressive module to dynamically learn the evolution of Alzheimer's disease.

Overall, the results show that the proposed BiPro model can predict the statuses of AD patients. Moreover, we inspected various scenarios of disease progression. However, our model still has certain limitations. The feature dimensions of clinical data are limited (five for demographics and six for biomarkers), which makes it difficult to fully use the deep learning model. Meanwhile, imaging data, such as MRI or PET, provide a large amount of information that is essential for diagnosing Alzheimer's

**Table A.10**
Table of notations.

| Symbol | Description | Symbol | Description |
|---|---|---|---|
| $T$ | Maximum number of timepoints of time series data | $t$ | $t$th Timepoint |
| $N$ | Number of biomarker features | $n$ | Biomarker $n$ |
| $\boldsymbol{X}$ | Input biomarker features | $x_t^{\{n\}}$ | Biomarker $n$ at $t$th timepoint |
| $\boldsymbol{X_d}$ | Demographic features | $x_{d,t}$ | Demographic features at $t$th timepoint |
| $\nu_t$ | Timestamp when features are obtained at $t$th timepoint | $\varphi_t^{\{n\}}$ | Missing time interval of biomarker $n$ at $t$th timepoint |
| $\boldsymbol{M}$ | Mask of observation (0: missing; 1: observed) | $m_t^{\{n\}}$ | Observation mask of biomarker $n$ at $t$th timepoint |
| $\overrightarrow{h}_{t-1}$ | Hidden state of recurrent network from the previous $(t-1)$th timepoint in the forward direction | $\overleftarrow{h}_{t+1}$ | Hidden state of recurrent network from the next $(t+1)$th timepoint in the backward direction |
| $\overrightarrow{\upsilon}_t$ | Forward values of a variable $\upsilon$ at $t$th timepoint | $\overleftarrow{\upsilon}_t$ | Backward values of a variable $\upsilon$ at $t$th timepoint |
| *Imputation* | | | |
| $\bar{x}_t$ | Global mean biomarkers at $t$th timepoint | $x'_t$ | Last valid biomarkers from $t$th timepoint |
| $\hat{x}_t$ | Temporal relation estimation at $t$th timepoint | $\overrightarrow{\mathbf{W}}_x$, $\overleftarrow{\mathbf{W}}_x$, $\mathbf{b}_x$ | Learnable parameters of $\hat{x}_t$ |
| $\tilde{x}_t$ | Imputed values calculated by temporal relations at $t$th timepoint | | |
| $\hat{z}_t$ | Imputed values calculated by multivariate relations at $t$th timepoint | $\mathbf{W}_z$, $\mathbf{b}_z$ | Learnable parameters of $\hat{z}_t$ |
| $\lambda_t$ | Temporal decay factor | $\mathbf{W}_\lambda$, $\mathbf{b}_\lambda$ | Learnable parameters of $\lambda_t$ |
| $\chi_t$ | Weighting coefficient factor of temporal decay at $t$th timepoint | $\mathbf{W}_\chi$, $\mathbf{b}_\chi$ | Learnable parameters of $\chi_t$ |
| $\hat{u}_t$ | Imputed values calculated by last observed, global mean, temporal and multivariate features | $\tilde{u}_t$ | Complete biomarker after imputation at $t$th timepoint |
| *Encoding* | | | |
| $\hat{h}_t$ | Embedded hidden state by temporal decay factor at $t$th timepoint | $\pi_t$ | Categorical trajectory factor at $t$th timepoint |
| $\hat{g}_t$ | Dynamic progressive score at $t$th timepoint | $\mathbf{W}_g$, $\boldsymbol{\gamma}_g$, $\boldsymbol{\alpha}_g$, $\mathbf{b}_g$ | Learnable parameters of $\hat{g}_t$ |
| $f_t$ | Forget gate of the MAPro-LSTM cell $t$ | $\mathbf{W}_f$, $\mathbf{b}_f$ | Learnable parameters of $f_t$ |
| $i_t$ | Input gate of the MAPro-LSTM cell $t$ | $\mathbf{W}_i$, $\mathbf{b}_i$ | Learnable parameters of $i_t$ |
| $o_t$ | Output gate of the MAPro-LSTM cell $t$ | $\mathbf{W}_o$, $\mathbf{b}_o$ | Learnable parameters of $o_t$ |
| $\tilde{c}_t$ | Update gate of the MAPro-LSTM cell $t$ | $\mathbf{W}_{\tilde{c}}$, $\mathbf{b}_{\tilde{c}}$ | Learnable parameters of $\tilde{c}_t$ |
| $h_t$ | Hidden state of the MAPro-LSTM cell $t$ | $c_t$ | Cell state of the MAPro-LSTM cell $t$ |
| $\Psi$ | Sequence of imputation operation | $\mathbf{s}$ | Set of recurrent inputs |
| *Forecasting* | | | |
| $\overrightarrow{h}_t$ | Hidden state of forward direction at $t$th timepoint | $\overleftarrow{h}_t$ | Hidden state of backward direction at $t$th timepoint |
| $\tilde{x}_{t+1}$ | Clinical status prediction at the next $(t+1)$th timepoint | $\overrightarrow{\mathbf{W}}_{\tilde{x}}$, $\overleftarrow{\mathbf{W}}_{\tilde{x}}$, $\mathbf{b}_{\tilde{x}}$ | Learnable parameters of $\tilde{x}_{t+1}$ |
| $\tilde{y}_{t+1}$ | MRI biomarker forecasting at the next $(t+1)$th timepoint | $\overrightarrow{\mathbf{W}}_{\tilde{y}}$, $\overleftarrow{\mathbf{W}}_{\tilde{y}}$, $\mathbf{b}_{\tilde{y}}$ | Learnable parameters of $\tilde{y}_{t+1}$ |
| *Objective Function* | | | |
| $\overrightarrow{\mathfrak{L}}_{impute}$ | Imputation loss of the forward direction | $\overleftarrow{\mathfrak{L}}_{impute}$ | Imputation loss of the backward direction |
| $\mathfrak{L}_{\text{impute}}$ | Summation of imputation loss from both directions | $\mathfrak{L}_{\text{forecast}}$ | Biomarker forecasting loss |
| $\mathfrak{L}_{\text{predict}}$ | Clinical status prediction loss | $\mathfrak{L}_{\text{impute}}$ | Total loss |
| $m_{y,t}$ | Label mask of clinical prognosis of observed timepoints | $\tau$, $\varepsilon$ | Tunable focusing parameters |
| *Shapley Values* | | | |
| $C$ | Number of output dimension | $\nabla f$ | Gradient computation |
| $Q$ | Number of target samples | $P$ | Number of reference samples |
| $x_{itn}^{\{c\}}$ | Input feature $n$ of target sample $i$th for output $c$ at $t$th timepoint | $\phi_{jtn}^{\{c\}}$ | Input feature $n$ of reference sample $j$th for output $c$ at $t$th timepoint |
| $S_{in}^{\{c\}}$ | Shapley value of input feature $n$ of target sample $i$th for output $c$ | | |

disease, and are not routinely gathered for long-term diseases (10 year monitoring). Another disadvantage of the proposed approach is that it employs equity loss functions for multitask learning, which can easily cause the model to focus on a specific task with a higher cost, and to neglect to optimize the training process.

In future work, we plan to collect imaging data for the proposed model by using a multimodal strategy. We were unable to acquire complete clinical data, and might have missed images in many periods. We plan to generate images at missing timepoints using either an autoencoder model (Andrew et al., 2021; Dong, Gan, Mao, Yang, & Shen, 2018; Mani, Aggarwal, Ghosh, & Jacob, 2020; Saravanan & Sujitha, 2020) or a generative adversarial network (Dang, Khurana, & Tiwari, 2020; Gu et al., 2020; Singh & Raza, 2021). Because using entire reconstructed volume images may burden the memory, we intend to infer features in latent space to minimize computational costs. Another possibility is to extract characteristics from AD-associated regions (Huang et al., 2011; Lancour et al., 2020; Zhang et al., 2015) and use them to impute data at the missing timepoints. Furthermore, we used the equity loss function for multitask learning. However, it is preferable to conduct additional ablation studies to analyze the effect of loss functions on a separate basis as well as to select appropriate weight values for each loss function (Dosovitskiy & Djolonga, 2019; Groenendijk, Karaoglu, Gevers, & Mensink, 2021).

## 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.

## Acknowledgments

This work was supported by a National Research Foundation of Korea (NRF) grant (MSIT)(NRF-2020R1A2B5B01002085) and a Bio & Medical Technology Development Program of the National Research Foundation (NRF) grant funded by the Korean government (MSIT)(NRF-2019M3E5D1A02067961).

## Appendix. Notation

All notations used in this paper are shown in Table A.10.

## References

Abuhmed, T., El-Sappagh, S., & Alonso, J. M. (2021). Robust hybrid deep learning models for Alzheimer's progression detection. *Knowledge-Based Systems*, *213*, Article 106688. http://dx.doi.org/10.1016/j.knosys.2020.106688.

Afzal, S., Javed, M., Maqsood, M., Aadil, F., Rho, S., & Mehmood, I. (2019). A segmentation-less efficient Alzheimer detection approach using hybrid image features. In *Handbook of multimedia information security: Techniques and applications* (pp. 421–429). Springer, http://dx.doi.org/10.1007/978-3-030-15887-3_20.

Aksman, L. M., Scelsi, M. A., Marquand, A. F., Alexander, D. C., Ourselin, S., Altmann, A., et al. (2019). Modeling longitudinal imaging biomarkers with parametric Bayesian multi-task learning. *Human Brain Mapping*, *40*(13), http://dx.doi.org/10.1101/593459.

An, N., Ding, H., Yang, J., Au, R., & Ang, T. F. (2020). Deep ensemble learning for Alzheimer's disease classification. *Journal of Biomedical Informatics*, *105*, Article 103411. http://dx.doi.org/10.1016/j.jbi.2020.103411.

Andrew, J., Mhatesh, T., Sebastin, R. D., Sagayam, K. M., Eunice, J., Pomplun, M., et al. (2021). Super-resolution reconstruction of brain magnetic resonance images via lightweight autoencoder. *Informatics in Medicine Unlocked*, *26*, Article 100713. http://dx.doi.org/10.1016/j.imu.2021.100713.

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science*, *373*(6552), 284–286. http://dx.doi.org/10.1126/science.abg1834.

Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., & Biecek, P. (2020). Dalex: Responsible machine learning with interactive explainability and fairness in Python. arXiv preprint arXiv:2012.14406.

Basher, A., Choi, K. Y., Lee, J. J., Lee, B., Kim, B. C., Lee, K. H., et al. (2019). Hippocampus localization using a two-stage ensemble hough convolutional neural network. *IEEE Access*, *7*, 73436–73447. http://dx.doi.org/10.1109/ACCESS.2019.2920005.

Bilgel, M., Jedynak, B. M., & Initiative, A. D. N. (2019). Predicting time to dementia using a quantitative template of disease progression. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *11*(1), 205–215.

Bilgel, M., Prince, J. L., Wong, D. F., Resnick, S. M., & Jedynak, B. M. (2016). A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging. *NeuroImage*, *134*, 658–670.

Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G., et al. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert Systems with Applications*, *130*, 157–171. http://dx.doi.org/10.1016/j.eswa.2019.04.022.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., & Li, Y. (2018). Brits: Bidirectional recurrent imputation for time . In *Advances in neural information processing systems* (pp. 6775–6785).

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, *8*(1), 1–12.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. arXiv preprint arXiv:1511.05942.

Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, *24*(2), 361–370. http://dx.doi.org/10.1093/jamia/ocw112.

Cui, X., Xiang, J., Guo, H., Yin, G., Zhang, H., Lan, F., et al. (2018). Classification of Alzheimer's disease, mild cognitive impairment, and normal controls with subnetwork selection and graph kernel principal component analysis based on minimum spanning tree brain functional network. *Frontiers in Computational Neuroscience*, *12*, 31. http://dx.doi.org/10.3389/fncom.2018.00031.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M. -O., et al. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, *56*(2), 766–781. http://dx.doi.org/10.1016/j.neuroimage.2010.06.013.

Dang, N., Khurana, M., & Tiwari, S. (2020). MirGAN: Medical image reconstruction using generative adversarial networks. In *2020 5th International conference on computing, communication and security* (pp. 1–5). IEEE, http://dx.doi.org/10.1109/ICCCS49678.2020.9277127.

Dong, L. -F., Gan, Y. -Z., Mao, X. -L., Yang, Y. -B., & Shen, C. (2018). Learning deep representations using convolutional auto-encoders with symmetric skip connections. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 3006–3010). IEEE, http://dx.doi.org/10.1109/ICASSP.2018.8462085.

Doody, R. S., Massman, P., & Dunn, J. K. (2001). A method for estimating progression rates in Alzheimer disease. *Archives of Neurology*, *58*, 449–454. http://dx.doi.org/10.1001/archneur.58.3.449.

Dosovitskiy, A., & Djolonga, J. (2019). You only train once: Loss-conditional training of deep networks. In *International conference on learning representations*.

Fedorov, A., Hjelm, R. D., Abrol, A., Fu, Z., Du, Y., Plis, S., et al. (2019). Prediction of progression to Alzheimer's disease with deep InfoMax. In *Proceedings of the IEEE international conference on biomedical & health informatics* (pp. 1–5). IEEE.

Gavidia-Bovadilla, G., Kanaan-Izquierdo, S., Mataró-Serrat, M., Perera-Lluna, A., & Initiative, A. D. N. (2017). Early prediction of Alzheimer's disease using null longitudinal model-based classifiers. *PLoS One*, *12*(1), Article e0168011. http://dx.doi.org/10.1371/journal.pone.0168011.

Ghazi, M. M., Nielsen, M., Pai, A., Cardoso, M. J., Modat, M., Ourselin, S., et al. (2019). Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling. *Medical Image Analysis*, *53*, 39–46. http://dx.doi.org/10.1016/j.media.2019.01.004.

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, *24*(1), 198–208. http://dx.doi.org/10.1093/jamia/ocw042.

Groenendijk, R., Karaoglu, S., Gevers, T., & Mensink, T. (2021). Multi-loss weighting with coefficient of variations. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1469–1478).

Gu, Y., Zeng, Z., Chen, H., Wei, J., Zhang, Y., Chen, B., et al. (2020). MedSRGAN: Medical images super-resolution using generative adversarial networks.. *Multimedia Tools & Applications*, *79*, http://dx.doi.org/10.1007/s11042-020-08980-w.

He, L., Kenton, L., Mike, L., & Luke, Z. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th annual meeting of the association for computational linguistics: Vol. 1*, (pp. 473–483).

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., et al. (2018). Learning deep representations by mutual information estimation and maximization. arXiv Preprint arXiv:1808.06670.

Holzinger, A., Malle, B., Saranti, A., & Pfeifer, B. (2021). Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. *Information Fusion*, *71*, 28–37. http://dx.doi.org/10.1016/j.inffus.2021.01.008.

Huang, L., Jin, Y., Gao, Y., Thung, K. -H., Shen, D., & Initiative, A. D. N. (2016). Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of Aging*, *46*, 180–191. http://dx.doi.org/10.1016/j.neurobiolaging.2016.07.005.

Huang, S., Li, J., Ye, J., Wu, T., Chen, K., Fleisher, A., et al. (2011). Identifying Alzheimer's disease-related brain regions from multi-modality neuroimaging data using sparse composite linear discrimination analysis. *Advances in Neural Information Processing Systems*, *24*.

Huang, Y., Xu, J., Zhou, Y., Tong, T., Zhuang, X., Alzheimer's Disease Neuroimaging Initiative (ADNI), et al. (2019). Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network. *Frontiers in Neuroscience*, *13*, 509. http://dx.doi.org/10.3389/fnins.2019.00509.

Ieracitano, C., Mammone, N., Hussain, A., & Morabito, F. C. (2021). A novel explainable machine learning approach for EEG-based brain-computer interface systems. *Neural Computing and Applications*, 1–14. http://dx.doi.org/10.1007/s00521-020-05624-w.

Jedynak, B. M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B. T., et al. (2012). A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort. *Neuroimage*, *63*(3), 1478–1486.

Jung, W., Jun, E., Suk, H. -I., & Initiative, A. D. N. (2020). Deep recurrent disease progression model for conversion-time prediction of Alzheimer's disease. arXiv preprint arXiv:2005.02643.

Lancour, D., Dupuis, J., Mayeux, R., Haines, J. L., Pericak-Vance, M. A., Schellenberg, G. C., et al. (2020). Analysis of brain region-specific co-expression networks reveals clustering of established and novel genes associated with Alzheimer disease. *Alzheimer's Research & Therapy*, *12*(1), 1–11. http://dx.doi.org/10.1186/s13195-020-00674-7.

Lee, G., Nho, K., Kang, B., Sohn, K. -A., & Kim, D. (2019). Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific Reports*, *9*(1), 1–12. http://dx.doi.org/10.1038/s41598-018-37769-z.

Lei, B., Yang, P., Wang, T., Chen, S., & Ni, D. (2017). Relational-regularized discriminative sparse learning for Alzheimer's disease diagnosis. *IEEE Transactions on Cybernetics*, *47*(4), 1102–1113. http://dx.doi.org/10.1109/TCYB.2016.2644718.

Lin, T. -Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Lipton, Z. C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019.

Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2017). Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677.

Liu, M., Zhang, J., Adeli, E., & Shen, D. (2019). Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. *IEEE Transactions on Biomedical Engineering*, *66*(5), 1195–1206. http://dx.doi.org/10.1109/TBME.2018.2869989.

Liu, W., Zhang, B., Zhang, Z., & Zhou, X. -H. (2013). Joint modeling of transitional patterns of Alzheimer's disease. *PLoS One*, *8*(9), Article e75487. http://dx.doi.org/10.1371/journal.pone.0075487.

Lorenzi, M., Filippone, M., Frisoni, G. B., Alexander, D. C., Ourselin, S., & Initiative, A. D. N. (2019). Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease. *NeuroImage*, *190*, 56–68. http://dx.doi.org/10.1016/j.neuroimage.2017.08.059.

Lundberg, S. M., & Lee, S. -I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777).

Mani, M. P., Aggarwal, H. K., Ghosh, S., & Jacob, M. (2020). Model-based deep learning for reconstruction of joint kq under-sampled high resolution diffusion MRI. In *2020 IEEE 17th international symposium on biomedical imaging* (pp. 913–916). IEEE, http://dx.doi.org/10.1109/isbi45749.2020.9098593.

Maqsood, M., Nazir, F., Khan, U., Aadil, F., Jamal, H., Mehmood, I., et al. (2019). Transfer learning assisted classification and detection of Alzheimer's disease stages using 3D MRI scans. *Sensors*, *19*(11), 2645. http://dx.doi.org/10.3390/s19112645.

Martinez-Murcia, F. J., Ortiz, A., Gorriz, J. -M., Ramirez, J., & Castillo-Barnes, D. (2019). Studying the manifold structure of Alzheimer's disease: A deep learning approach using convolutional autoencoders. *IEEE Journal of Biomedical and Health Informatics*, *24*(1), 17–26. http://dx.doi.org/10.1109/JBHI.2019.2914970.

Nguyen, M., He, T., An, L., Alexander, D. C., Feng, J., & Alzheimer's Disease Neuroimaging Initiative (2020). Predicting Alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, *222*(15), http://dx.doi.org/10.1016/j.neuroimage.2020.117203.

Nie, L., Meng, L. Z. L., Song, X., Chang, X., & Li, X. (2017). Modeling disease progression via multisource multitask learners: A case study with Alzheimer's disease. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(7), 1508–1519. http://dx.doi.org/10.1109/TNNLS.2016.2520964.

Oxtoby, N. P., Young, A. L., Cash, D. M., Benzinger, T. L., Fagan, A. M., Morris, J. C., et al. (2018). Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain*, *141*(5), 1529–1544. http://dx.doi.org/10.1093/brain/awy050.

Park, C., Ha, J., & Park, S. (2020). Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. *Expert Systems with Applications*, *140*, Article 112873. http://dx.doi.org/10.1016/j.eswa.2019.112873.

Razavian, N., Marcus, J., & Sontag, D. (2016). Multi-task prediction of disease onsets from longitudinal laboratory tests. In *Machine learning for healthcare conference* (pp. 73–100).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). http://dx.doi.org/10.1145/2939672.2939778.

Saravanan, S., & Sujitha, J. (2020). Deep medical image reconstruction with autoencoders using deep Boltzmann machine training. *EAI Endorsed Transactions on Pervasive Health and Technology*, *6*(24), Article e2. http://dx.doi.org/10.4108/eai.24-9-2020.166360.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Shahbaz, M., Ali, S., Guergachi, A., Niazi, A., & Umer, A. (2019). Classification of Alzheimer's disease using machine learning techniques.. In *Data* (pp. 296–303). http://dx.doi.org/10.5220/0007949902960303.

Singh, N. K., & Raza, K. (2021). Medical image generation using generative adversarial networks: A review. *Health Informatics: A Computational Perspective in Healthcare*, 77–96. http://dx.doi.org/10.1007/978-981-15-9735-0_5.

Stonnington, C. M., Chu, C., Klöppel, S., Clifford, R. J., Jr., Ashburner, J., Frackowiak, R. S., et al. (2010). Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *NeuroImage*, *51*(4), 1405–1413. http://dx.doi.org/10.1016/j.neuroimage.2010.03.051.

Sukkar, R., Katz, E., Zhang, Y., Raunig, D., & Wyman, B. T. (2012). Disease progression modeling using hidden Markov models. In *2012 annual international conference of the IEEE engineering in medicine and biology society* (pp. 2845–2848). IEEE, http://dx.doi.org/10.1109/EMBC.2012.6346556.

Tabarestani, S., Aghili, M., Shojaie, M., Freytes, C., & Adjouadi, M. (2018). Profile-specific regression model for progression prediction of Alzheimer's disease using longitudinal data. In *2018 17th IEEE international conference on machine learning and applications* (pp. 1353–1357). IEEE, http://dx.doi.org/10.1109/ICMLA.2018.00220.

Utsumi, Y., Guerrero, R., Peterson, K., Rueckert, D., Picard, R. W., et al. (2019). Meta-weighted gaussian process experts for personalized forecasting of AD cognitive changes. In *Machine learning for healthcare conference* (pp. 181–196). PMLR.

Venkatraghavan, V., Bron, E. E., Niesse, W. J., Klein, S., & Initiative, A. D. N. (2019). Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *NeuroImage*, *186*, 518–532. http://dx.doi.org/10.1016/j.neuroimage.2018.11.024.

Vu, T. -D., Ho, N. -H., Yang, H. -J., Kim, J., & Song, H. -C. (2018). Non-white matter tissue extraction and deep convolutional neural network for Alzheimer's disease detection. *Soft Computing*, *22*, 6825–6833. http://dx.doi.org/10.1007/s00500-018-3421-5.

Wang, T., Qiu, R. G., & Yu, M. (2018). Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks. *Scientific Reports*, *8*(1), 1–12. http://dx.doi.org/10.1038/s41598-018-27337-w.

Workgroup, S. M., Borson, S., Boustani, M. A., Buckwalter, K. C., Burgio, L. D., Chodosh, J., et al. (2016). Report on milestones for care and support under the US national plan to address Alzheimer's disease. *Alzheimer's & Dementia*, *12*(2), 334–369. http://dx.doi.org/10.1016/j.jalz.2016.01.005.

Xia, Z., Yue, G., Xu, Y., Feng, C., Yang, M., Wang, T., et al. (2020). A novel end-to-end hybrid network for Alzheimer's disease detection using 3D CNN and 3D CLSTM. In *2020 IEEE 17th international symposium on biomedical imaging* (pp. 1–4). IEEE, http://dx.doi.org/10.1109/ISBI45749.2020.9098621.

Yang, G., Ye, Q., & Xia, J. (2021). Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. arXiv preprint arXiv:2102.01998.

Yoon, J., Zame, W. R., & van der Schaar, M. (2019). Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, *66*(5).

Zhang, Y., Dong, Z., Phillips, P., Wang, S., Ji, G., Yang, J., et al. (2015). Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. *Frontiers in Computational Neuroscience*, *9*, 66. http://dx.doi.org/10.3389/fncom.2015.00066.

Zhou, J., Liu, J., Narayan, V. A., & Ye, J. (2012). Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1095–1103). ACM, http://dx.doi.org/10.1145/2339530.2339702.

Zhou, J., Liu, J., Narayan, V. A., Yeand, J., & Initiative, A. D. N. (2013). Modeling disease progression via multi-task learning. *NeuroImage*, *78*, 233–248. http://dx.doi.org/10.1016/j.neuroimage.2013.03.073.