Semi-supervised Enrichment Learning to Integrate Dynamic-Static data for Improved Mortality Prediction on COVID-19 Patients

No Author Given

No Institute Given

Abstract. hi

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

Alzheimer's Disease (AD) is a chronic neurodegenerative disease that impairs patients' thinking, memory, and behavior. More than 30 million people worldwide suffer from AD, and with the increase in life expectancy this figure is projected to triple by 2050 [2]. AD typically advances from a pre-clinical level to mild cognitive impairment (MCI) and eventually to AD, along a time scale. Early identification of at-risk individuals is crucial in slowing disease progression, so many researchers have dedicated their efforts to identify the AD relevant biomarkers and develop a computer-aided system to accurately predict AD onset.

Due to its widespread availability, high spatial resolution and good contrast between different soft tissues, neuroimaging has sparked interest among researchers seeking to characterize AD progression [1, 24, 35, 28, 33, 3]. However, there are a few limitations in the existing predictive models. First, many of them routinely perform standard regression or classification at each individual time point separately, and do not take advantage of the longitudinal structure among temporal brain phenotypes. Since AD is a progressive neurodegenerative disorder, it is beneficial to explore the temporal relationships among the longitudinal measurements [30]. However, the longitudinal models [3, 29–31] only consider the order of records, not their time stamps explicitly, while the time stamps play an important role when learning the temporal relation (*e.g.* temporal locality) between records.

Second, records of neuroimaging biomarkers are often missing at some time points for some participants. This is because higher mortality rate and cognitive disability discourage older adults from staying in the studies that require multiple visits, making it difficult to consistently sample medical scans from a large participant pool. Although the missing data issue is prevalent in most medical datasets, many studies assume feature-complete data or simply remove samples with missing data [24, 16], harming the integrity of the dataset in the process.

Recently, several data imputation methods [32, 15] have been proposed to generate missing records of longitudinal AD measures, however these data imputation methods can introduce undesirable bias that degrades a model's predictive power.

Third, the biomarkers of AD can be obtained even if diagnoses for some participants are missing, and collecting the labeled data is time consuming and costly. Recently recurrent neural networks (RNNs) [17], especially Long Short Term Memory (LSTM) [21], have been successfully applied in longitudinal dataset analysis due to their flexibility to handle missing data and ability to learn the long-term dependencies. Although effective, they are supervised learning [27, 9, 25] learning methods which cannot learn from the unlabeled samples. For the LSTM application in the unsupervised learning, the previous studies [14, 23] learn the representation of dynamic data in a lower dimensional space. These unsupervised representation learning models have utilized unlabeled samples, and the proper representation of dynamic data is crucial because the dynamic data usually contains noises and redundant information from the large number of features and records [26, 14]. However, they have encoded the dynamic data into another longitudinal representation which is difficult to be integrated with the static data. In addition they often do not utilize the labeled samples, while the target labels of samples can improve the representation learning for the better predictions.

Fourth, diagnosing AD involves a number of medical tests, leading to large collections of heterogeneous multivariate results, such as voxel-based morphometry (VBM), FreeSurfer (FS), and cognitive assessments scores which are dynamic, and single-nucleotide polymorhpisms (SNPs), and demographic information which are static. As a result, AD patients are characterized by heterogeneous multi-modalities, and the integration of multi-modalities helps to identify and differentiate the subtle shifts in disease progression status and promote accurate diagnoses. However, integrating multi-modal data is challenging problem because dynamic data is represented by the matrix of varying size and the static data is represented by the fixed-length vector. The current longitudinal method [30, 14, 23] often do not fuse the complementary information from different modalities to improve diagnostic precision.

To take advantage of the full potential of heterogeneous longitudinal data, we propose a novel semi-supervised learning method to learn an enriched biomarker representation for each participant in a studied cohort. The proposed model consists of two autoencoders [12] and one predictor. The deep neural network autoencoder and long short-term memory (LSTM) autoencoder each learn the vectorial representation from static genetic data and longitudinal phenotypic data. The enriched representation of dynamic data is in a fixed-length vector format, which can be readily integrated with the enriched representation of static data as illustrated in Fig. 1. The proposed approach has the following benefits compared to the existing models: (1) LSTM autoencoder learns new representation of dynamic data in a fixed length vector format, which can be easily integrated with the static data. (2) We learn the representations of dynamic

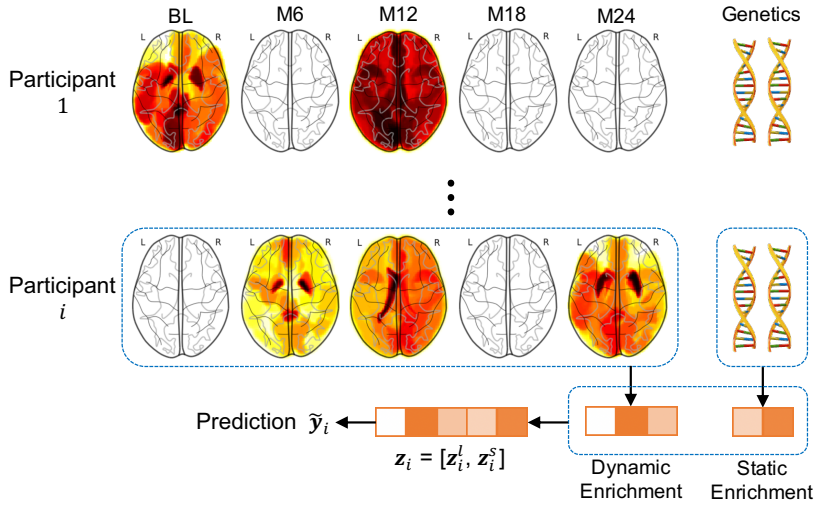

Fig. 1. Illustration of the proposed model to learn the enriched biomarker representations for each participant and predict the disease status. The blank brain plots denote the missing scans of a participant. The dynamic enrichment summarizes the longitudinal records with missing data into a fixed-length vectorial representation.

data incrementally, considering the uneven time intervals between the records. (3) The proposed semi-supervised learning model fully utilize the available labels of samples. (4) We do not rely on the imputation or discarding samples to handle the missing data.

2 Datasets and Problem Formulation

2.1 Datasets

We obtain the data used in this experiment from the ADNI database [20]. We download 1.5 T MRI scans, single nucleotide polymorphism (SNP), and demographic information (age and gender) of 821 ADNI-1 participants. For the SNP data, the quality control steps discussed in [22] were followed, and 1223 SNPs for each participant are provided. We perform voxel-based morphometry (VBM) and FreeSurfer (FS) on the MRI data following [20] and extract mean modulated gray matter measure for 90 target regions of interest. We also downloaded the longitudinal scores of the participants in five independent cognitive assessments including Alzheimer's Disease Assessment Scale (ADAS), Mini-Mental State Examination (MMSE), Fluency test (FLU), Rey's Auditory Verbal Learning Test (RAVLT) and Trail making test (TRAILS). In this analysis, the time points for both imaging records and cognitive assessments include baseline (BL), month 6 (M6), month 12 (M12), month 18 (M18), and month 24 (M24). We use the

diagnosis at month 36 (M36) in Alzheimer's disease (AD), mild cognitive impairment (MCI), and healthy control(HC) as predictive target in our studies. The participants with no missing SNPs, demographic information, and AD diagnosis at M36 were included, providing a set of 379 subjects (104 AD, 115 MCI, 160 HC). Our dataset includes 231 male and 148 female participants, and the average age is 75.35.

In the following pages, we denote a vector as a bold lower case letter, and matrix as a bold upper case letter. For the arbitrary matrix $\mathbf{X}$, $[\mathbf{X}]^r$, $[\mathbf{X}]_c$, $[\mathbf{X}]_c^r$ denotes the r-th row, c-th column, an element of r-th row and c-th column respectively. We use i and j to index the participant and record respectively. We describe the records of i-th participant as $\mathcal{X}_i = \{\mathbf{x}_i^b, \mathbf{x}_i^s, \mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i\}$. $\mathbf{x}_i^b \in \Re^{D_b}$ is a vector of basic demographic information (age and gender, $D_b = 2$), $\mathbf{x}_i^s \in \Re^{D_s}$ is a vector of SNPs ($D_s = 1223$), and $\mathbf{X}_i = [\mathbf{x}_i^1; \mathbf{x}_i^2; \cdots; \mathbf{x}_i^{n_i}] \in \Re^{n_i \times D_l}$ are the longitudinal records collected across the n_i time points, and $\mathbf{M}_i = [\mathbf{m}_i^1; \mathbf{m}_i^2; \cdots; \mathbf{m}_i^{n_i}] \in \{1, 0\}^{n_i \times D_l}$ are the binary masks of longitudinal records $\mathbf{X}_i$, where 1 and 0 indicates the observed and unobserved entry respectively. To fully utilize the multi-modal information, each longitudinal record at j-th time point $\mathbf{x}_i^j$ $(1 \leq j \leq n_i)$ is the concatenation of multi-modal neuroimagings and cognitive assessments, such that $\mathbf{x}_i^j = [\mathbf{x}_{i,VBM}^j, \mathbf{x}_{i,FS}^j, \mathbf{x}_{i,ADAS}^j, \mathbf{x}_{i,MMSE}^j, \mathbf{x}_{i,FLU}^j, \mathbf{x}_{i,RAVLT}^j, \mathbf{x}_{i,TRAILS}^j] \in \Re^{D_l}$, and the missing records are filled with the constant -1. $\mathbf{t}_i = [t_i^1; t_i^2; \cdots; t_i^{n_i}] \in \Re^{n_i}$ are the time stamps of n_i records. The target label $\mathbf{y}_i \in \Re^{D_y}$ of i-th participant is given if that participant is in training set, such that $i \in \Omega_{train}$. The target label is one-hot encoded, such that $[1, 0, 0]$, $[0, 1, 0]$, and $[0, 0, 1]$ represent AD, MCI, and HC respectively. The input and output of the proposed model is described in Fig. 2.

2.2 Static Data Enrichment

We leverage the autoencoder [12] to learn the abstract representation of genotypic biomarkers. The autoencoder consists of two deep neural networks: an encoder $\phi_{SNP} : \Re^{D_s} \mapsto \Re^{d_s}$ that encodes a vector of SNPs into the abstract representation such that $\phi_{SNP}(\mathbf{x}_i^s;\ \theta_E^s) = \mathbf{z}_i^s$, and an decoder $\psi_{SNP} : \Re^{d_s} \mapsto \Re^{D_s}$ that decodes the encoded representation into the reconstructed vector of SNPs such that $\psi_{SNP}(\mathbf{z}_i^s;\ \theta_D^s) = \tilde{\mathbf{x}}_i^s$. θ_E^s and θ_D^s is the set of trainable variables for encoder and decoder respectively. The deep neural network is defined as the K consecutive fully connected layers where the output of k-th layer is:

$$\mathbf{h}_k = \sigma(\mathbf{h}_{k-1}\mathbf{W}_k + \mathbf{b}_k), \tag{1}$$

where σ is an activation function, and weight matrix $\mathbf{W}_k$ and bias vector $\mathbf{b}_k$ $(1 \leq k \leq K)$ are the traiable variables in θ_E^s or θ_D^s. The input $\mathbf{h}_0$ of the network is then forwarded to the last layer $\mathbf{h}_K = \tilde{\mathbf{x}}_i^s$, which is the output of the network. We train the weights matrices and bias vectors in θ_E^s or θ_D^s to minimize the following reconstruction loss:

$$\mathcal{L}_{static}(\mathbf{x}_i^s, \tilde{\mathbf{x}}_i^s;\ \theta_E^s, \theta_D^s) = \|\mathbf{x}_i^s - \tilde{\mathbf{x}}_i^s\|_F^2, \tag{2}$$

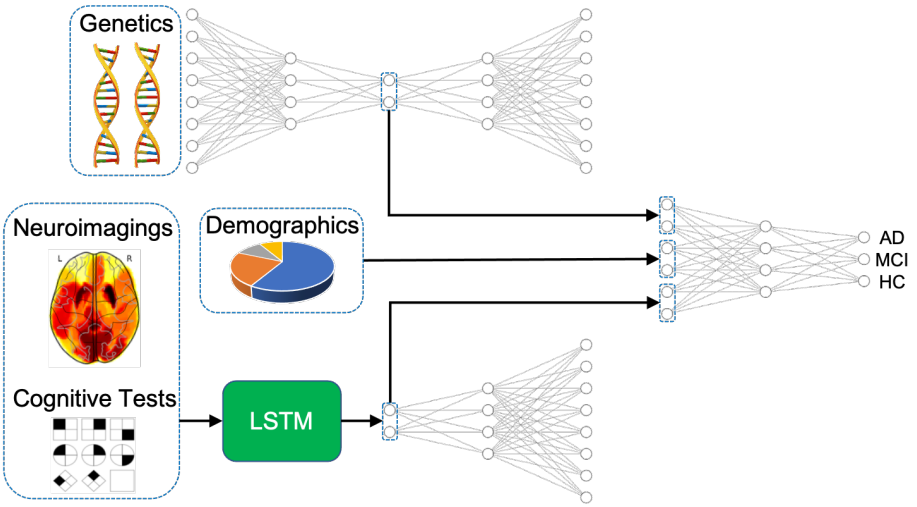

Fig. 2. Armed with the enriched representations of dynamic and static data, we can fully utilize the information in the dataset. The demographic information is provided without enrichment, as it's dimensionality is relatively small.

where squared Frobenious norm $\|\cdot\|_F^2$ is defined as the summation of all the entries squared. By minimizing the reconstruction error, the encoded representation $\mathbf{z}_i^s$ preserves the as much information as possible while removing the redundant noises in SNPs $\mathbf{x}_i^s$.

2.3 Dynamic Data Enrichment

We choose LSTM encoder $\phi_{dynamic} : \Re^{n_i \times (2D_l+1)} \mapsto \Re^{d_l}$ to summarize the longitudinal records $\mathbf{X}_i$ in the fixed length vector $\mathbf{z}_i^l$. The time stamp of each record is crucial in learning the temporal relation between records (e.g. temporal locality), and the missingness pattern of the entries may help LSTM encoder to interpret the input. Thus we provide the concatenation of longitudinal records, masks, and time stamps, $[\mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i] = [\hat{\mathbf{x}}_i^1; \hat{\mathbf{x}}_i^2; \cdots; \hat{\mathbf{x}}_i^{n_i}] = \hat{\mathbf{X}}_i \in \Re^{n_i \times (2D_l+1)}$, as an input of LSTM encoder such that $\phi_{dynamic}(\mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i;\ \theta_E^l) = \mathbf{z}_i^l$.

The concatenated longitudinal record at the j-th time step $\hat{\mathbf{x}}_i^j$ $(1 \le j \le n_i)$ is processed by the following LSTM architecture [34]:

$$\mathbf{k}_i^j = \sigma(\hat{\mathbf{x}}_i^j \mathbf{W}_{xk} + \mathbf{h}_i^{j-1} \mathbf{W}_{hk} + \mathbf{c}_i^{j-1} \mathbf{W}_{ck} + \mathbf{b}_k), \tag{3}$$

$$\mathbf{f}_i^j = \sigma(\hat{\mathbf{x}}_i^j \mathbf{W}_{xf} + \mathbf{h}_i^{j-1} \mathbf{W}_{hf} + \mathbf{c}_i^{j-1} \mathbf{W}_{cf} + \mathbf{b}_f), \tag{4}$$

$$\mathbf{c}_i^j = \mathbf{f}_i^j \odot \mathbf{c}_i^{j-1} + \mathbf{k}_i^j \odot \tanh(\hat{\mathbf{x}}_i^j \mathbf{W}_{xc} + \mathbf{h}_i^{j-1} \mathbf{W}_{hc} + \mathbf{b}_c), \tag{5}$$

$$\mathbf{o}_i^j = \sigma(\hat{\mathbf{x}}_i^j \mathbf{W}_{xo} + \mathbf{h}_i^{j-1} \mathbf{W}_{ho} + \mathbf{c}_i^j \mathbf{W}_{co} + \mathbf{b}_o), \tag{6}$$

$$\mathbf{h}_i^j = \mathbf{o}_i^j \odot \tanh(\mathbf{c}_i^j), \tag{7}$$

where σ and tanh is the logistic sigmoid and hyperbolic tangent activation function respectively, and $\mathbf{k}_i^j$, $\mathbf{o}_i^j$, $\mathbf{f}_i^j$ are input, output, forget gate of j-th time step respectively. $\{\mathbf{W}_{xk}, \mathbf{W}_{hk}, \mathbf{W}_{ck}, \mathbf{W}_{xf}, \mathbf{W}_{hf}, \mathbf{W}_{cf}, \mathbf{W}_{xc}, \mathbf{W}_{hc}, \mathbf{W}_{xo}, \mathbf{W}_{ho}, \mathbf{W}_{co}\} \subset \theta_E^l$ are trainable weight matrices and $\{\mathbf{b}_k, \mathbf{b}_f, \mathbf{b}_c, \mathbf{b}_o\} \subset \theta_E^l$ are trainable bias vectors. $\mathbf{c}_i^j$ and $\mathbf{h}_i^j$ denote the cell state and hidden representation. The hidden representation $\mathbf{h}_i^{n_i}$ at the last time step n_i represents summarization of the longitudinal records $\hat{\mathbf{X}}_i$, such that $\mathbf{h}_i^{n_i} = \mathbf{z}_i^l \in \Re^{d_l}$. Since the hidden representation at j-th time point aims to summarize the records from first time step to j-th time step, LSTM cell needs to refer to the cell state $\mathbf{c}_i^j$ and reflect past records to $\mathbf{h}_i^j$. Since the cell state $\mathbf{c}_i^j$ is guided by the input gate $\mathbf{k}_i^j$ and forget gate $\mathbf{f}_i^j$, which control how much information came from previous step should be preserved, the cell state $\mathbf{c}_i^j$ enables the hidden representation $\mathbf{h}_i^j$ to learn long term dependencies. For example, LSTM encoder can capture the cognitive decline from the temporal trends in the scores of cognitive assessments.

We propose a decoder for dynamic data enrichment with a deep neural network instead of another LSTM. A previous study [23] that attempted to enrich longitudinal records with a recurrent neural network, did so by using RNNs for both the encoder and decoder, where the output (reconstructed record) of the decoder at each time step depends on the output at the previous time step. However, since no additional information is provided to the decoder other than a time stamp and a learned representation that is no longer longitudinal, there should not be dependency between the outputs of the decoder. Since the enriched representation $\mathbf{z}_i^l$ summarizes *whole* longitudinal records, the decoder for dynamic data enrichment $\psi_{dynamic} : \Re^{d_l+1} \mapsto \Re^{D_l}$ should be able to reconstruct the j-th record $\mathbf{x}_i^j$ given time stamp t_i^j without any additional information, such that $\psi_{dynamic}(\mathbf{z}_i^l, t_i^j;\ \theta_D^l) = \tilde{\mathbf{x}}_i^j \approx \mathbf{x}_i^j$, where θ_D^l is a set of weight matrices and bias vectors of the decoder. This architecture, to the best of our knowledge, has not yet been proposed. We update θ_E^l and θ_D^l to minimize the error between the reconstructed records and original records for the observed entries indicated by the mask $\mathbf{M}_i$:

$$\mathcal{L}_{dynamic}(\mathbf{X}_i, \tilde{\mathbf{X}}_i, \mathbf{M}_i;\ \theta_E^l, \theta_D^l) = \frac{\left\| (\tilde{\mathbf{X}}_i - \mathbf{X}_i) \odot \mathbf{M}_i \right\|_F^2}{\sum_{q=1}^{D_l} \sum_{j=1}^{n_i} [\mathbf{M}_i]_q^j}, \tag{8}$$

where $\tilde{\mathbf{X}}_i \in \Re^{n_i \times D_l}$ is the stack of reconstructed n_i records for i-th participant.

2.4 Prediction and Loss Function

From the enriched representations $\mathbf{z}_i^l$ and $\mathbf{z}_i^s$ of dynamic and static data, another fully connected layers $\psi_{pred} : \Re^{(d_s+d_l+D_b)} \mapsto \Re^{D_y}$ predicts the target label $\mathbf{y}_i$, such that $\psi_{pred}(\mathbf{z}_i^l, \mathbf{z}_i^s, \mathbf{x}_i^b;\ \theta_D^p) = \tilde{\mathbf{y}}_i$. We design the loss function to induce the enriched representation to convey the useful information to reconstruct the

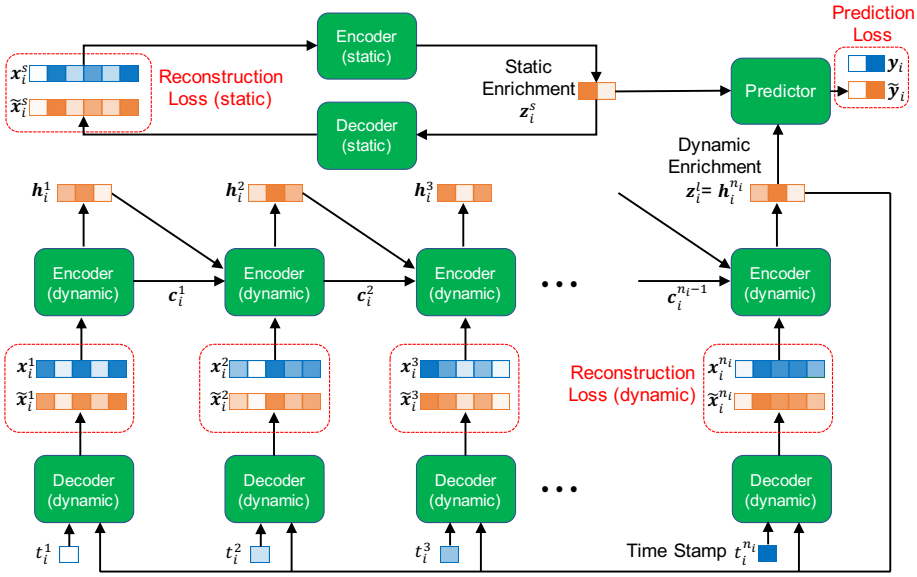

Fig. 3. An schematic illustration about loss function. Our semi-supervised learning autoencoder minimizes the reconstruction loss for the labeled or unlabeled samples, and prediction loss only for the labeled samples.

original records and predict the target label:

$$\begin{aligned}\theta_E^s, \theta_D^s, \theta_E^l, \theta_D^l, \theta^p = \underset{\theta_E^s, \theta_D^s, \theta_E^l, \theta_D^l, \theta^p}{\arg\min} \big(&\gamma_1 \mathcal{L}_{predict}(\mathbf{y}_i, \tilde{\mathbf{y}}_i;\ \theta^p) \\ &+ \gamma_2 \mathcal{L}_{static}(\mathbf{x}_i^s, \tilde{\mathbf{x}}_i^s;\ \theta_E^s, \theta_D^s) + \gamma_3 \mathcal{L}_{dynamic}(\mathbf{X}_i, \tilde{\mathbf{X}}_i, \mathbf{M}_i;\ \theta_E^l, \theta_D^l)\big),\end{aligned} \tag{9}$$

where γ_1, γ_2, and γ_3 are hyperparameters to adjust the impact of each loss. We choose the cross entropy for the prediction loss defined as follows:

$$\mathcal{L}_{predict}(\mathbf{y}_i, \tilde{\mathbf{y}}_i;\ \theta_D^p) = \begin{cases} 0 & i \notin \Omega_{train}, \\ -\|\mathbf{y}_i \odot \log(\tilde{\mathbf{y}}_i) + (\mathbf{1} - \mathbf{y}_i) \odot \log(\mathbf{1} - \tilde{\mathbf{y}}_i)\|_1 & i \in \Omega_{train}, \end{cases}$$

where $\mathbf{1}$ is a vector of 1's and log is an element-wise logarithm function.

3 Experiments

Our experiments consist of two parts; (1) we evaluate the prediction performance of the proposed model, and (2) we identify the biomarkers which are most important for AD progression prediction.

3.1 Competing Models and Hyperparameters

We use the following hyperparameters found by the grid search. For our model, semi-supervised autoencoder (SAE), the static encoder ϕ_{SNP} and decoder ψ_{SNP}

has 2 fully connected layers (FC) each with the tanh activation function at the first to third layer and logistic sigmoid at the fourth layer. The dynamic decoder $\psi_{dynamic}$ has 3 FCs with activation function of leaky rectified linear unit (alpha = 0.1) at the first layer and tanh at the second and third layer. The dynamic encoder $\phi_{dynamic}$ is the LSTM with 64 units and tanh activation function. We set $\gamma_1 = 1e+2$, $\gamma_2 = 1e+1$, $\gamma_3 = 1$ in Eq. (9). To minimize the loss function in Eq. (9), we adapt Adam optimizer [11] at a fixed learning rate of 0.0003 and the other parameters kept at their default values. We do not use any regularization or dropout technique, as they degrade the performance.

For an ablation study to observe the effectiveness of our semi-supervised enrichment learning, we introduce baseline LSTM (BLSTM) model as a competing model by removing decoders ψ_{SNP} and $\psi_{dynamic}$ from our model SAE. In addition to the longitudinal model, we introduce the random forest [8] (RF) with 34 max depth and deep neural network (DNN) with 5 FCs as competing models. Since these baseline models are not longitudinal models, we provide the concatenation of most recent record $[\mathbf{x}_i^b, \mathbf{x}_i^s, \mathbf{x}_i^{n_i}]$ to them. Both the training and test set are provided to train SAE in a semi-supervised manner, while other competing models are trained only with the training set. Although the order of participants is randomly shuffled to avoid the bias, we use the same training and test data across all the competing methods for the fair comparison.

3.2 Classification Performance

We conduct the multiclass AD progression prediction task for 1 year ahead and evaluate the performance of the predictive models. We evaluate the predictive models with the following metrics:

$$\begin{aligned} &\text{Accuracy} = \frac{\sum_{c\in C}(TP_c + TN_c)}{\sum_{c\in C}(TP_c + TN_c + FP_c + FN_c)}, \\ &\text{Precision of class } c = \frac{TP_c}{TP_c + FP_c}, \ \text{Recall of class } c = \frac{TP_c}{TP_c + FN_c}, \end{aligned} \tag{10}$$

where C is a set of classes {AD, MCI, HC}. In table 2, the average and standard deviation of performance results across k groups are reported following k-fold cross validation scheme. We split the dataset with $k = 2,\ 3,\ 5$ subgroups, such that 50%, 66%, 80% of participants belong to the training set respectively.

In table 2, nan (not a number) indicates the denominator $TP_c + FP_c$ in Eq. (10) is zero, meaning the model never predicted class c. Considering the prediction of random forest is biased towards AD (high AD Recall), our model SAE generally outperforms other competing models. Interestingly, the performance gap between SAE and BLSTM increases as the size of training set decreases. We suppose that this is because our semi-supervised learning approach can learn from the unlabeled samples in test set, while BLSTM cannot. This finding shows the robustness of our model against number of unlabeled samples, as well as its promise in early diagnosis of AD.

Table 1. The prediction results of SAE and the other competing models from k-fold cross validation. The best prediction is denoted as bold font.

k	**Metric**	**SAE** (Ours)	**BLSTM**	**RF**	**DNN**
5	Accuracy	**74.95±4.72**	72.57±5.29	53.55±4.33	49.45±2.91
	AD Precision	**71.28±11.89**	71.19±16.79	51.55±5.22	55.13±6.30
	MCI Precision	59.84±10.11	**60.07±11.74**	nan±nan	33.84±4.97
	HC Precision	**89.69±7.33**	86.60±8.15	66.28±2.88	47.02±7.9
	AD Recall	69.30±7.30	65.92±21.31	**99.18±1.64**	74.49±7.45
	MCI Recall	**66.00±11.83**	60.00±8.57	0.47±0.93	22.89±2.93
	HC Recall	85.36±8.98	**89.37±6.53**	34.58±5.50	36.33±8.67
3	Accuracy	**73.34±1.71**	69.92±1.94	53.83±1.02	47.81±1.65
	AD Precision	**63.47±7.84**	62.10±5.53	51.63±1.29	55.26±1.61
	MCI Precision	**63.97±8.74**	55.39±7.85	nan±nan	31.03±1.37
	HC Precision	**88.62±3.47**	85.94±3.09	65.74±3.14	46.64±6.77
	AD Recall	69.14±6.16	63.90±7.23	**98.48±0.8**	70.09±6.63
	MCI Recall	**56.97±8.32**	53.85±10.16	0.00±0.00	27.62±5.09
	HC Recall	**88.55±3.44**	85.48±3.04	36.56±2.84	32.44±6.52
2	Accuracy	**72.29±2.44**	47.29±23.60	53.69±2.04	47.68±0.13
	AD Precision	**60.79±2.25**	50.38±26.69	51.56±2.68	52.87±2.26
	MCI Precision	**60.93±6.38**	nan±nan	nan±nan	33.12±6.56
	HC Precision	**86.99±2.20**	nan±nan	64.99±2.28	44.38±0.06
	AD Recall	72.90±8.45	81.36±18.64	**98.14±0.08**	75.56±5.73
	MCI Recall	**45.21±11.24**	30.19±30.18	0.53±0.53	20.51±2.42
	HC Recall	**89.85±4.13**	42.21±42.20	36.10±0.17	30.70±7.16

3.3 AD Relevant Biomarkers Identification

It is vital to identify AD relevant biomarkers for early detection and treatment of AD in people at high risk of developing AD. Despite of the promising performance of deep neural networks, the predictions of deep neural networks are notoriously difficult to be interpreted. To identify which biomarkers (features) largely affect to the predictions, we add the perturbation to the input data and observe the changes in prediction.

For the longitudinal records of q-th biomarker ($1 \leq q \leq D_l$) and i-th participant, we sample the column vector of perturbations $\mathbf{p}_{i,q} \in \Re^{n_i}$ from the normal distribution $\mathcal{N}(0, \sigma_q^2)$ with zero mean and the same standard deviation as the observed distribution of q-th biomarker across all n participants, and add the perturbation to the records as follows:

$$N = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{q=1}^{D_l}[\mathbf{M}_i]_q^j,\ \mu_q = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{q=1}^{D_l}[\mathbf{X}_i \odot \mathbf{M}_i]_q^j,\ \sigma_q^2 = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{q=1}^{D_l}$$
$$[\mathbf{M}_i]_q^j([\mathbf{X}_i]_q^j - \mu_q)^2,\ \mathbf{X}_i' = [[\mathbf{X}_i]_1, [\mathbf{X}_i]_2, \cdots, [\mathbf{X}_i]_q + \mathbf{p}_{i,q}, \cdots, [\mathbf{X}_i]_{D_l}]. \tag{11}$$

Table 2. The prediction performance of SAE and the competitive models from k-fold cross validation. The best prediction is denoted as bold font.

Training Set	Metric	**SAE** (Ours)	**BLSTM**	**RF**	**DNN**
80%	Accuracy	**74.95±4.72**	72.57±5.29	53.55±4.33	49.45±2.91
	AD Precision	**71.28±11.89**	71.19±16.79	51.55±5.22	55.13±6.30
	MCI Precision	59.84±10.11	**60.07±11.74**	nan±nan	33.84±4.97
	HC Precision	**89.69±7.33**	86.60±8.15	66.28±2.88	47.02±7.9
	AD Recall	69.30±7.30	65.92±21.31	**99.18±1.64**	74.49±7.45
	MCI Recall	**66.00±11.83**	60.00±8.57	0.47±0.93	22.89±2.93
	HC Recall	85.36±8.98	**89.37±6.53**	34.58±5.50	36.33±8.67
66%	Accuracy	**73.34±1.71**	69.92±1.94	53.83±1.02	47.81±1.65
	AD Precision	**63.47±7.84**	62.10±5.53	51.63±1.29	55.26±1.61
	MCI Precision	**63.97±8.74**	55.39±7.85	nan±nan	31.03±1.37
	HC Precision	**88.62±3.47**	85.94±3.09	65.74±3.14	46.64±6.77
	AD Recall	69.14±6.16	63.90±7.23	**98.48±0.8**	70.09±6.63
	MCI Recall	**56.97±8.32**	53.85±10.16	0.00±0.00	27.62±5.09
	HC Recall	**88.55±3.44**	85.48±3.04	36.56±2.84	32.44±6.52
50%	Accuracy	**72.29±2.44**	47.29±23.60	53.69±2.04	47.68±0.13
	AD Precision	**60.79±2.25**	50.38±26.69	51.56±2.68	52.87±2.26
	MCI Precision	**60.93±6.38**	nan±nan	nan±nan	33.12±6.56
	HC Precision	**86.99±2.20**	nan±nan	64.99±2.28	44.38±0.06
	AD Recall	72.90±8.45	81.36±18.64	**98.14±0.08**	75.56±5.73
	MCI Recall	**45.21±11.24**	30.19±30.18	0.53±0.53	20.51±2.42
	HC Recall	**89.85±4.13**	42.21±42.20	36.10±0.17	30.70±7.16

Then the prediction changes by the perturbation is:

$$\Delta\tilde{\mathbf{y}}_i = \|\psi_{pred}\big(\phi_{dynamic}(\mathbf{X}'_i, \mathbf{M}_i, \mathbf{t}_i;\ \theta^l_\phi),\ \psi_{SNP}(\phi_{SNP}(\mathbf{x}^s_i;\ \theta^s_\phi);\ \theta^s_\psi),\ \mathbf{x}^b_i;\ \theta^p_\psi\big) - \psi_{pred}\big(\phi_{dynamic}(\mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i;\ \theta^l_\phi),\ \psi_{SNP}(\phi_{SNP}(\mathbf{x}^s_i;\ \theta^s_\phi);\ \theta^s_\psi),\ \mathbf{x}^b_i;\ \theta^p_\psi\big)\|_1. \tag{12}$$

The importance of q-th biomarker is the average of prediction changes across all the participants: $\frac{1}{n}\sum_{i=1}^{n}\Delta\tilde{\mathbf{y}}_i$. Similarly, we can calculate the input genetic data whose q-th SNP is perturbed and it's prediction changes.

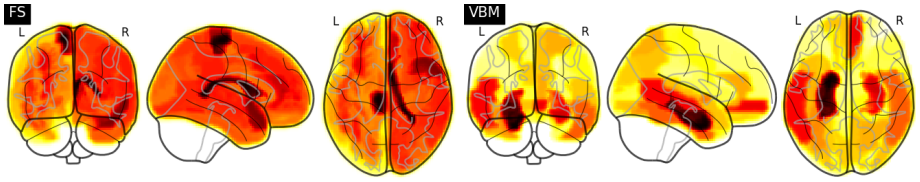

Fig. 4. Importance distribution over the brain regions. The darker color indicates the larger importance. The top five most important regions identified are **FS:** Right Lateral Ventricle, Left Para-Hippocampal, Left Amygdala, Left Cerebral White Matter, Left Hippocampus, and **VBM:** Left Para-Hippocampal, Left Amygdala, Left Hippocampus, Right Hippocampus, Right Amygdala.

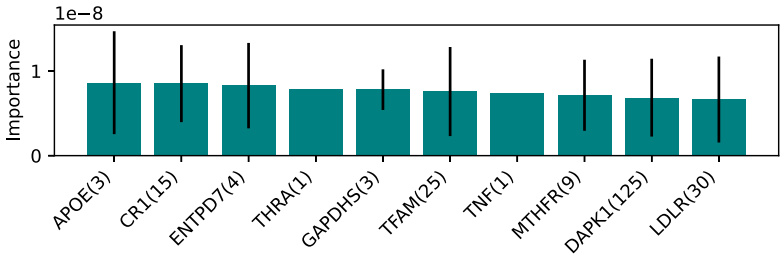

Fig. 5. Importance of each AlzGene group. The standard deviation and number of SNPs of each group is denoted by the line length at the head of each bar and the number next to the group name respectively.

Identifying AD relevant imaging biomarkers In Fig 4, we plot the importance distribution over the brain regions calculated by Eq. (12). The identified regions have been shown in the literature to be related to AD. For example, the previous studies [4] found that ventricular volume and its rate of change is related with vulnerability to cognitive decline and dementia. They observe that the

larger ventricles in healthy participants increase the probability to the progression of dementia-related disease in the future. The hippocampus is vulnerable to be damaged from AD [18] and has been shown to affect long-term memory and spatial navigation in patients with AD. Finally, the amygdala region, also identified by our approach, is also severely affected by AD [19] and is associated with emotional response and decision-making.

Identifying AD relevant genetic biomarkers We plot the importance distribution over the AlzGene groups of SNPs in Fig 5. The AlzGene groups of SNPs have been constructed by the multiple genome-wide association studies listed on the website (http://www.alzgene.org/). Apolipoprotein E (APOE) group is identified by our approach, and APOE genes involve in amyloid beta peptide (Aβ) aggregation and clearance [10]. The accumulation of Aβ is commonly observed in the progress of AD [5] and amyloid hypothesis suggests reasonable mechanism how the accumulation of Aβ can result neuronal malfunction [7]. In addition, $\epsilon 4$ allele of APOE gene increase risk factor for AD and decrease the age of AD onset [6]. For complement receptor 1 (CR1) group, genome-wide analysis [13] reported CR1 association with late-onset AD. The subset of biomarkers identified in the FreeSurfer, VBM, and SNP modalities, provides the substantial evidence that our approach can identify the biomarkers associated with AD.

4 Conclusion

We present the semi-supervised enrichment learning method to integrate the longitudinal multi-modal dataset, which fits to the clinical applicability and can be used in a real-time automatic AD diagnosis. The novel LSTM autoencoder is introduced to compress longitudinal records with missing data into a fixed-length vectorial representation. Armed with the enriched representation, we can fully utilize the genotypic and phenotypic data. Our experiments show that our model beats the performance of the competing predictive models in AD prediction. In addition, combined with the perturbation based feature identification method, our model discovers the neuroimaging and genetic biomarkers associated with AD, further adding value to our approach.

References

1. Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., et al.: The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimer's & dementia **7**(3), 270–279 (2011)
2. Barnes, D.E., Yaffe, K.: The projected effect of risk factor reduction on alzheimer's disease prevalence. The Lancet Neurology **10**(9), 819–828 (2011)

3. Brand, L., Wang, H., Huang, H., Risacher, S., Saykin, A., Shen, L., et al.: Joint high-order multi-task feature learning to predict the progression of alzheimer's disease. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 555–562. Springer (2018)
4. Carmichael, O.T., Kuller, L.H., Lopez, O.L., Thompson, P.M., Dutton, R.A., Lu, A., Lee, S.E., Lee, J.Y., Aizenstein, H.J., Meltzer, C.C., et al.: Ventricular volume and dementia progression in the cardiovascular health study. Neurobiology of aging **28**(3), 389–397 (2007)
5. Chen, G.f., Xu, T.h., Yan, Y., Zhou, Y.r., Jiang, Y., Melcher, K., Xu, H.E.: Amyloid beta: structure, biology and structure-based therapeutic development. Acta Pharmacologica Sinica **38**(9), 1205–1235 (2017)
6. Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G., Roses, A., Haines, J., Pericak-Vance, M.A.: Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. Science **261**(5123), 921–923 (1993)
7. Hardy, J., Selkoe, D.J.: The amyloid hypothesis of alzheimer's disease: progress and problems on the road to therapeutics. science **297**(5580), 353–356 (2002)
8. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
9. Hong, X., Lin, R., Yang, C., Zeng, N., Cai, C., Gou, J., Yang, J.: Predicting alzheimer's disease using lstm. IEEE Access **7**, 80893–80901 (2019)
10. Kim, J., Basak, J.M., Holtzman, D.M.: The role of apolipoprotein e in alzheimer's disease. Neuron **63**(3), 287–303 (2009)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Kramer, M.A.: Nonlinear principal component analysis using autoassociative neural networks. AIChE journal **37**(2), 233–243 (1991)
13. Lambert, J.C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Combarros, O., Zelenika, D., Bullido, M.J., Tavernier, B., et al.: Genome-wide association study identifies variants at clu and cr1 associated with alzheimer's disease. Nature genetics **41**(10), 1094–1099 (2009)
14. Längkvist, M., Karlsson, L., Loutfi, A.: A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters **42**, 11–24 (2014)
15. Li, Y., Wang, L., Zhou, J., Ye, J.: Multi-task learning based survival analysis for multi-source block-wise missing data. Neurocomputing **364**, 95–107 (2019)
16. Liu, M., Zhang, J., Adeli, E., Shen, D.: 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 (2018)
17. Medsker, L.R., Jain, L.: Recurrent neural networks. Design and Applications **5** (2001)
18. Mu, Y., Gage, F.H.: Adult hippocampal neurogenesis and its role in alzheimer's disease. Molecular neurodegeneration **6**(1), 85 (2011)
19. Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C., Initiative, A.D.N., et al.: Amygdala atrophy is prominent in early alzheimer's disease and relates to symptom severity. Psychiatry Research: Neuroimaging **194**(1), 7–13 (2011)
20. Risacher, S.L., Shen, L., West, J.D., Kim, S., McDonald, B.C., Beckett, L.A., Harvey, D.J., Jack Jr, C.R., Weiner, M.W., Saykin, A.J., et al.: Longitudinal mri atrophy biomarkers: relationship to conversion in the adni cohort. Neurobiology of aging **31**(8), 1401–1418 (2010)

21. Schmidhuber, J., Hochreiter, S.: Long short-term memory. Neural Comput **9**(8), 1735–1780 (1997)
22. Shen, L., Kim, S., Risacher, S.L., Nho, K., Swaminathan, S., West, J.D., Foroud, T., Pankratz, N., Moore, J.H., Sloan, C.D., et al.: Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in mci and ad: A study of the adni cohort. Neuroimage **53**(3), 1051–1063 (2010)
23. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International conference on machine learning. pp. 843–852 (2015)
24. Stonnington, C.M., Chu, C., Klöppel, S., Jack Jr, C.R., Ashburner, J., Frackowiak, R.S., Initiative, A.D.N., et al.: Predicting clinical scores from magnetic resonance scans in alzheimer's disease. Neuroimage **51**(4), 1405–1413 (2010)
25. Tabarestani, S., Aghili, M., Shojaie, M., Freytes, C., Cabrerizo, M., Barreto, A., Rishe, N., Curiel, R.E., Loewenstein, D., Duara, R., et al.: Longitudinal prediction modeling of alzheimer disease using recurrent neural networks. In: 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). pp. 1–4. IEEE (2019)
26. Tuncel, K.S., Baydogan, M.G.: Autoregressive forests for multivariate time series modeling. Pattern recognition **73**, 202–215 (2018)
27. Vieira, S., Pinaya, W.H., Mechelli, A.: Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neuroscience & Biobehavioral Reviews **74**, 58–75 (2017)
28. Wang, H., Nie, F., Huang, H., Yan, J., Kim, S., Nho, K., Risacher, S.L., Saykin, A.J., Shen, L., Initiative, A.D.N.: From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps. Bioinformatics **28**(18), i619–i625 (2012)
29. Wang, H., Nie, F., Huang, H., Yan, J., Kim, S., Risacher, S., Saykin, A., Shen, L.: High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In: Advances in neural information processing systems. pp. 1277–1285 (2012)
30. Wang, X., Shen, D., Huang, H.: Prediction of memory impairment with mri data: a longitudinal study of alzheimer's disease. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 273–281. Springer (2016)
31. Wang, X., Yan, J., Yao, X., Kim, S., Nho, K., Risacher, S.L., Saykin, A.J., Shen, L., Huang, H., et al.: Longitudinal genotype-phenotype association study via temporal structure auto-learning predictive model. In: International Conference on Research in Computational Molecular Biology. pp. 287–302. Springer (2017)
32. Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P.M., Ye, J., Initiative, A.D.N., et al.: Bi-level multi-source learning for heterogeneous block-wise missing data. NeuroImage **102**, 192–206 (2014)
33. Yan, J., Li, T., Wang, H., Huang, H., Wan, J., Nho, K., Kim, S., Risacher, S.L., Saykin, A.J., Shen, L., et al.: Cortical surface biomarkers for predicting cognitive outcomes using group l2, 1 norm. Neurobiology of aging **36**, S185–S193 (2015)
34. Yu, Y., Si, X., Hu, C., Zhang, J.: A review of recurrent neural networks: Lstm cells and network architectures. Neural computation **31**(7), 1235–1270 (2019)
35. Zhang, D., Shen, D., Initiative, A.D.N., et al.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. NeuroImage **59**(2), 895–907 (2012)