Learning Deeply Enriched Representations of Longitudinal Imaging-Genetic Data to Predict Alzheimer's Disease Progression

Abstract. Alzheimer's Disease (AD) is a progressive memory disorder that causes irreversible cognitive decline. Early diagnosis is imperative to prevent the progression of AD and many biomarker analysis models have been presented to detect the disease in its early stages. However, these models often lack reliability due to their failure to integrate longitudinal (dynamic) phenotypic data with (static) genetic data, and/or to fully utilize both labeled and unlabeled samples. To overcome these difficulties, we propose a semi-supervised enrichment learning method to learn the fixed-length vectorial representation of dynamic data for each participant. Armed with the enriched representation in fixed-length vector format, the static data can be readily integrated with the dynamic data. We have applied our new method on the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort and achieved 75% accuracy on multiclass AD progression prediction one year in advance. In addition to the improved prediction performance when compared against several widely used machine learning algorithms, the proposed model identifies the most relevant disease biomarkers.

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 [1]. 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 and many researchers have dedicated their efforts to identify AD relevant biomarkers and develop a computer-aided system to accurately predict AD onset. Neuroimaging has sparked interest among researchers seeking to characterize AD progression due to it's widespread availability that takes advantage of high spatial resolution and good contrast between different soft tissues [18, 22, 2]. However, there are a few limitations to the existing neuroimaging analysis models. First, since AD is a progressive neurodegenerative disorder, many longitudinal models [2, 23–25] have been proposed to explore the temporal relationships among neuroimaging records. However, they only consider the order of records, not their time intervals between the records.

Second, diagnosing AD involves a number of medical tests, leading to large collections of heterogeneous, multivariate results of dynamic and static data. As

a result, AD patients are characterized by heterogeneous multi-modalities which does help to identify and differentiate the subtle shifts in disease progression and promote accurate diagnoses. However, records of dynamic data are often missing at certain time points for some participants. This is because higher mortality rate and cognitive disability discourage older adults from staying in studies that require multiple visits, making it difficult to consistently sample medical scans from a large participant pool. As a result, integrating multi-modal data is challenging because the dynamic data is represented by a matrix of varying size and the static data is represented by the fixed-length vector. Current longitudinal methods [24, 10, 17] often do not combine the complementary information from different modalities to improve diagnostic precision.

Third, the biomarkers of AD can be obtained even if the diagnoses for some participants is missing, but collecting the labeled data is time consuming and costly. Recent recurrent neural networks (RNNs) [11], especially Long Short Term Memory (LSTM) [15], have been successfully applied in longitudinal dataset analysis due to their flexibility to handle missing data and ability to learn long-term dependencies. Although effective, they are supervised learning [21, 5, 19] models which cannot learn from unlabeled samples. In the application of LSTM to unsupervised learning, the previous studies [10, 17] learn the representation of dynamic data in a lower dimensional space. The proper representation of dynamic data is crucial because dynamic data usually contains noises and redundant information from the large number of features and records [20, 10]. While these models do utilize unlabeled samples, they encode the dynamic data into another longitudinal representation which is difficult to integrate with the static data. In addition, they often do not utilize the labeled samples, which is a missed opportunity to improve the representation learning for better predictions.

To take full advantage 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 [8]; The deep neural network autoencoder and long short-term memory (LSTM) autoencoder each learn the vectorial representation from static genetic data and dynamic 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.

2 The Dataset and Problem Formalization

2.1 Datasets

We obtained the data used in this experiment from the ADNI database [14]. We downloaded 1.5 T MRI scans, single nucleotide polymorphisms (SNPs), and demographic information (age and gender) of 821 ADNI-1 participants. For the SNPs data, the quality control steps discussed in [16] were followed, and 1223 SNPs for each participant are extracted. We performed voxel-based morphometry (VBM) and FreeSurfer (FS) on the MRI data following [14] and extracted mean modulated gray matter measures for 90 target regions of interest. We

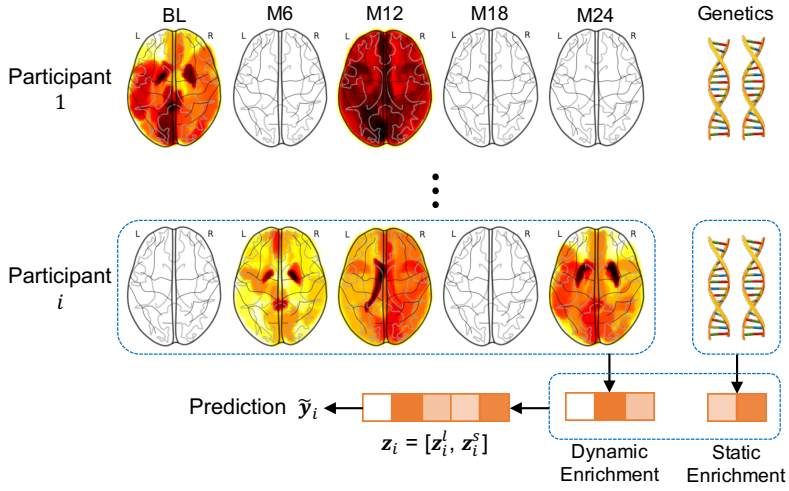

Fig. 1. Illustration of the proposed model to integrate dynamic and static data via enrichment. The blank brain plots denote the missing scans of a participant.

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). The studied cohort 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 a 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 the 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. $\mathbf{M}_i = [\mathbf{m}_i^1; \mathbf{m}_i^2; \cdots; \mathbf{m}_i^{n_i}] \in \{1, 0\}^{n_i \times D_l}$ are the binary masks of longitudinal records $\mathbf{X}_i$, where 1 and 0 indicates the observed and unobserved entry. Each longitudinal record at the 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 the i-th participant is given for training if that participant is in training set, such that $i \in \Omega_{train}$. The input and output of the proposed model is described in Fig. 2.

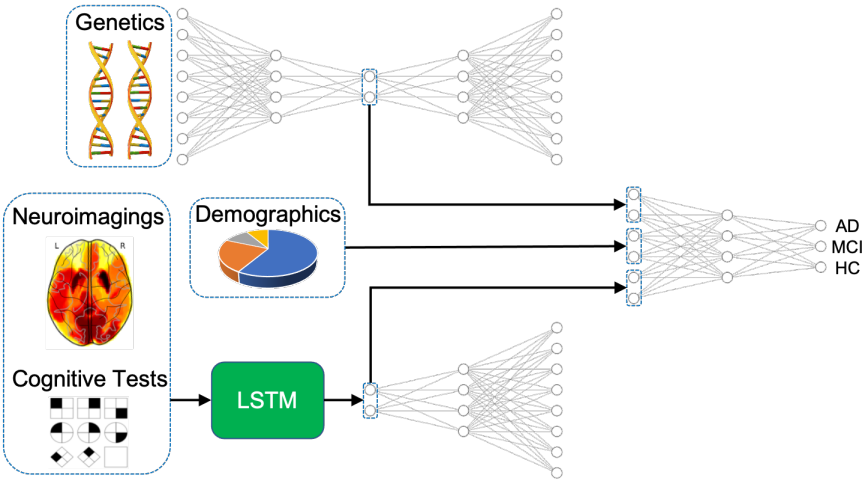

Fig. 2. Armed with the enriched representations of dynamic and static data, we can fully utilize the information in the dataset.

2.2 Enrichment and Objective Formulation

Static Data Enrichment We leverage the autoencoder [8] to learn the enriched 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 enriched 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 enriched 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 weights matrices and bias vectors for the encoder and decoder respectively. The encoded representation $\mathbf{z}_i^s$ preserves as much information as possible while removing the redundant noises in SNPs $\mathbf{x}_i^s$ by updating θ_E^s or θ_D^s to minimize the following reconstruction loss:

$$\mathcal{L}_{static}(\mathbf{x}_i^s, \tilde{\mathbf{x}}_i^s;\ \theta_E^s, \theta_D^s) = \|\mathbf{x}_i^s - \tilde{\mathbf{x}}_i^s\|_F^2\,, \tag{1}$$

where squared Frobenious norm $\|\cdot\|_F^2$ is summation of all the entries squared.

Dynamic Data Enrichment LSTM encoder $\phi_{dynamic} : \Re^{n_i \times (2D_l+1)} \mapsto \Re^{d_l}$ summarizes 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 (e.g. temporal locality) between records, and the pattern of missing entries may help the 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 the 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 $\hat{\mathbf{x}}_i^j$ $(1 \leq j \leq n_i)$ is processed by the following LSTM architecture [26]:

$$\mathbf{k}_i^j = \sigma(\hat{\mathbf{x}}_i^j \mathbf{W}_{xk} + \mathbf{h}_i^{j-1} \mathbf{W}_{hk} + \mathbf{c}_i^{j-1} \mathbf{W}_{ck} + \mathbf{b}_k), \tag{2}$$

$$\mathbf{f}_i^j = \sigma(\hat{\mathbf{x}}_i^j \mathbf{W}_{xf} + \mathbf{h}_i^{j-1} \mathbf{W}_{hf} + \mathbf{c}_i^{j-1} \mathbf{W}_{cf} + \mathbf{b}_f), \tag{3}$$

$$\mathbf{c}_i^j = \mathbf{f}_i^j \odot \mathbf{c}_i^{j-1} + \mathbf{k}_i^j \odot \tanh(\hat{\mathbf{x}}_i^j \mathbf{W}_{xc} + \mathbf{h}_i^{j-1} \mathbf{W}_{hc} + \mathbf{b}_c), \tag{4}$$

$$\mathbf{o}_i^j = \sigma(\hat{\mathbf{x}}_i^j \mathbf{W}_{xo} + \mathbf{h}_i^{j-1} \mathbf{W}_{ho} + \mathbf{c}_i^j \mathbf{W}_{co} + \mathbf{b}_o), \tag{5}$$

$$\mathbf{h}_i^j = \mathbf{o}_i^j \odot \tanh(\mathbf{c}_i^j), \tag{6}$$

where $\mathbf{k}_i^j$, $\mathbf{o}_i^j$, $\mathbf{f}_i^j$ are the input, output, and forget gate of the 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 a summary of *whole* 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 the j-th time point aims to summarize the records from the first time step to the j-th time step, it refers to the cell state $\mathbf{c}_i^j$ containing information about the previous records. 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 from the previous step should be preserved, therefore cell state $\mathbf{c}_i^j$ enables the hidden representation $\mathbf{h}_i^j$ to learn long term dependencies.

We propose a decoder for dynamic data enrichment with a fully connected layers instead of another LSTM. A previous study [17] 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. Therefore the decoder $\psi_{dynamic} : \Re^{d_l+1} \mapsto \Re^{D_l}$ 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 $\psi_{dynamic}(\mathbf{z}_i^l, t_i^j;\ \theta_D^l) = \tilde{\mathbf{x}}_i^j \approx \mathbf{x}_i^j$. 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 and original records for the observed entries defined as:

$$\mathcal{L}_{dynamic}(\mathbf{X}_i, \tilde{\mathbf{X}}_i, \mathbf{M}_i;\ \theta_E^l, \theta_D^l) = \frac{\left\| (\tilde{\mathbf{X}}_i - \mathbf{X}_i) \odot \mathbf{M}_i \right\|_F^2}{\sum_{q=1}^{D_l} \sum_{j=1}^{n_i} [\mathbf{M}_i]_q^j}, \tag{7}$$

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

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 layer $\psi_{pred} : \Re^{(d_s+d_l+D_b)} \mapsto \Re^{D_y}$ 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 inducing the enriched representation to convey the useful information to reconstruct the original records and predict the target label:

$$\begin{aligned}\theta_E^s, \theta_D^s, \theta_E^l, \theta_D^l, \theta^p = \underset{\theta_E^s, \theta_D^s, \theta_E^l, \theta_D^l, \theta^p}{\arg\min} \big(&\gamma_1 \mathcal{L}_{predict}(\mathbf{y}_i, \tilde{\mathbf{y}}_i;\ \theta^p) \\ &+ \gamma_2 \mathcal{L}_{static}(\mathbf{x}_i^s, \tilde{\mathbf{x}}_i^s;\ \theta_E^s, \theta_D^s) + \gamma_3 \mathcal{L}_{dynamic}(\mathbf{X}_i, \tilde{\mathbf{X}}_i, \mathbf{M}_i;\ \theta_E^l, \theta_D^l)\big),\end{aligned} \tag{8}$$

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:

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

where $\mathbf{1}$ is a vector of 1's and log is an element-wise logarithm function. The illustration of loss function is provided in supplementary.

3 Experiments

Our experiments consist of two parts: (1) we compare the prediction performance of the proposed model to the widely used prediction models, and (2) we identify the biomarkers which are most predictive for AD progression.

3.1 Classification Performance

Competing models For an ablation study to observe the effectiveness of our semi-supervised autoencoder (SAE), we introduce a baseline LSTM (BLSTM) model by removing decoders ψ_{SNP} and $\psi_{dynamic}$ from our model SAE. In addition to the longitudinal model, we use the random forest [4] (RF) with 34 max depth and deep neural network (DNN) with 5 fully connected layers. Since these competing models are not longitudinal, we provide the concatenation of the most recent record $[\mathbf{x}_i^b, \mathbf{x}_i^s, \mathbf{x}_i^{n_i}]$ to them. Both the training and test sets are provided to train SAE in a semi-supervised manner, while the 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 a fair comparison. The hyperparameters of our model are provided in the supplementary.

Result and Evaluation We conduct the multiclass AD progression prediction task for 1 year ahead. In table 1, the average and standard deviation of performance results across k groups are reported following a k-fold cross validation scheme. We split the dataset into $k = 2,\ 3,\ 5$ subgroups, such that 50%, 66%, 80% of participants belong to the training set respectively. In table 1, nan (not a

number) in precision on specific class means the model never predicted that class in one of k subgroups. Our model SAE generally outperforms other competing models especially when the proportion of training set is small. We suppose that this is because our semi-supervised learning approach can learn from the unlabeled samples in test set, while the other models cannot. This finding shows our model's promise in the early diagnosis of AD.

Table 1. The prediction performance from k-fold cross validation. Receiver operating characteristic curves of this result is provided in supplementary.

Training Set	Metric	**SAE** (Ours)	**BLSTM**	**RF**	**DNN**
80%	Accuracy	**74.95±4.72**	71.57±5.29	71.79±6.33	49.45±2.91
	AD Precision	**71.28±11.89**	71.19±16.79	70.55±9.22	55.13±6.30
	MCI Precision	59.84±10.11	**60.07±11.74**	59.64±8.15	33.84±4.97
	HC Precision	**89.69±7.33**	86.60±8.15	87.28±9.88	47.02±7.9
	AD Recall	69.30±7.30	65.92±21.31	**71.18±12.64**	74.49±7.45
	MCI Recall	**66.00±11.83**	60.00±8.57	61.47±11.93	22.89±2.93
	HC Recall	85.36±8.98	**89.37±6.53**	86.58±9.50	36.33±8.67
66%	Accuracy	**73.34±1.71**	69.92±1.94	61.83±4.02	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	53.91±8.85	31.03±1.37
	HC Precision	**88.62±3.47**	85.94±3.09	71.74±3.14	46.64±6.77
	AD Recall	**69.14±6.16**	63.90±7.23	59.48±8.8	70.09±6.63
	MCI Recall	**56.97±8.32**	53.85±10.16	39.09±11.54	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

3.2 AD Relevant Biomarkers Identification

It is vital to identify AD relevant biomarkers for early detection and the treatment of people at high risk of developing AD. Despite the promising performance of deep neural networks, their predictions are notoriously difficult to interpret. To identify which biomarkers largely affect the predictions, we add the perturbation to the input data and observe the changes in prediction. The details of this identification method is described in supplementary. In Fig 3 and Fig 4, we plot the importance distribution over the brain regions and AlzGene groups of SNPs. The AlzGene groups of SNPs have been constructed by the multiple genome-wide association studies listed on the website (http://www.alzgene.org/). From our

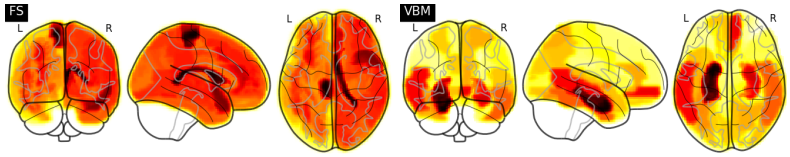

Fig. 3. Importance distribution over the brain regions. The darker color indicates the greater importance. The top five most important regions identified are **FS:** Right Lateral Ventricle, Left Para-Hippocampal, Left Amygdala, Left Cerebral White Matter, Left Hippocampus, and **VBM:** Left Para-Hippocampal, Left Amygdala, Left Hippocampus, Right Hippocampus, Right Amygdala.

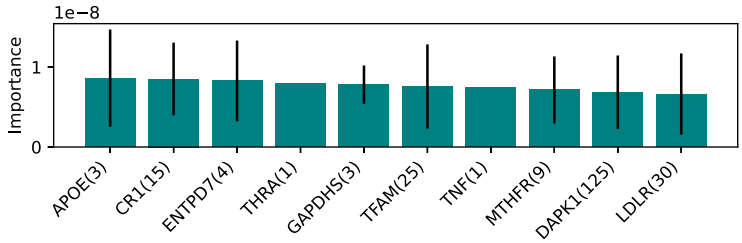

Fig. 4. Importance of each AlzGene group. The standard deviation and number of SNPs in each group is denoted by the line and the number next to the group name

model SAE, ventricular [3], hippocampus [12], and amygdala [13] are identified as important brain regions, and apolipoprotein E [6] and complement receptor 1 [9] are identified as important AlzGene groups. The identified biomarkers have been shown in the literature to be related to AD, thus they provide substantial evidence that our approach can identify the biomarkers associated with AD.

4 Conclusion

We present a semi-supervised enrichment learning method that integrates the longitudinal multi-modal dataset and is clinically applicable for use in real-time, automatic AD diagnosis. The novel LSTM autoencoder compresses longitudinal records with missing data into a fixed-length vectorial representation. Armed with this enriched representation, one can fully utilize the genotypic and phenotypic data. Our experiments show that our model outperforms competing predictive models. When combined with the perturbation based feature identification method, our model also discovers the neuroimaging and genetic biomarkers associated with AD, adding further value to our approach.

References

1. 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)
2. 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)
3. 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)
4. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
5. 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)
6. Kim, J., Basak, J.M., Holtzman, D.M.: The role of apolipoprotein e in alzheimer's disease. Neuron **63**(3), 287–303 (2009)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Kramer, M.A.: Nonlinear principal component analysis using autoassociative neural networks. AIChE journal **37**(2), 233–243 (1991)
9. 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)
10. 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)
11. Medsker, L.R., Jain, L.: Recurrent neural networks. Design and Applications **5** (2001)
12. Mu, Y., Gage, F.H.: Adult hippocampal neurogenesis and its role in alzheimer's disease. Molecular neurodegeneration **6**(1), 85 (2011)
13. 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)
14. 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)
15. Schmidhuber, J., Hochreiter, S.: Long short-term memory. Neural Comput **9**(8), 1735–1780 (1997)
16. 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)
17. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International conference on machine learning. pp. 843–852 (2015)

18. 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)
19. 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)
20. Tuncel, K.S., Baydogan, M.G.: Autoregressive forests for multivariate time series modeling. Pattern recognition **73**, 202–215 (2018)
21. 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)
22. 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)
23. 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)
24. 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)
25. 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)
26. 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)

Supplementary Materials: Learning Deeply Enriched Representations of Longitudinal Imaging-Genetic Data to Predict Alzheimer's Disease Progression

S1 Perturbation Based Biomarker Identification

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:

$$N = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{q=1}^{D_l}[\mathbf{M}_i]_q^j,\ \mu_q = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{q=1}^{D_l}[\mathbf{X}_i \odot \mathbf{M}_i]_q^j,\ \sigma_q^2 = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{q=1}^{D_l}$$
$$[\mathbf{M}_i]_q^j([\mathbf{X}_i]_q^j - \mu_q)^2,\ \mathbf{X}_i' = [[\mathbf{X}_i]_1, [\mathbf{X}_i]_2, \cdots, [\mathbf{X}_i]_q + \mathbf{p}_{i,q}, \cdots, [\mathbf{X}_i]_{D_l}]. \tag{S1}$$

Then the prediction changes by the perturbation are:

$$\Delta\tilde{\mathbf{y}}_i = \|\psi_{pred}\big(\phi_{dynamic}(\mathbf{X}_i', \mathbf{M}_i, \mathbf{t}_i;\ \theta_\phi^l),\ \psi_{SNP}(\phi_{SNP}(\mathbf{x}_i^s;\ \theta_\phi^s);\ \theta_\psi^s),\ \mathbf{x}_i^b;\ \theta_\psi^p\big)$$
$$- \psi_{pred}\big(\phi_{dynamic}(\mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i;\ \theta_\phi^l),\ \psi_{SNP}(\phi_{SNP}(\mathbf{x}_i^s;\ \theta_\phi^s);\ \theta_\psi^s),\ \mathbf{x}_i^b;\ \theta_\psi^p\big)\|_1. \tag{S2}$$

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.

S2 Hyperparameters of proposed model

For our model, semi-supervised autoencoder (SAE), the static encoder ϕ_{SNP} and decoder ψ_{SNP} have 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 a leaky rectified linear unit (alpha = 0.1) activation function 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. (8). To minimize the loss function in Eq. (8), we adapt the Adam optimizer [7] at a fixed learning rate of 0.0003 and the other parameters are kept at their default values. We do not use any regularization or dropout techniques, as they degrade the performance.

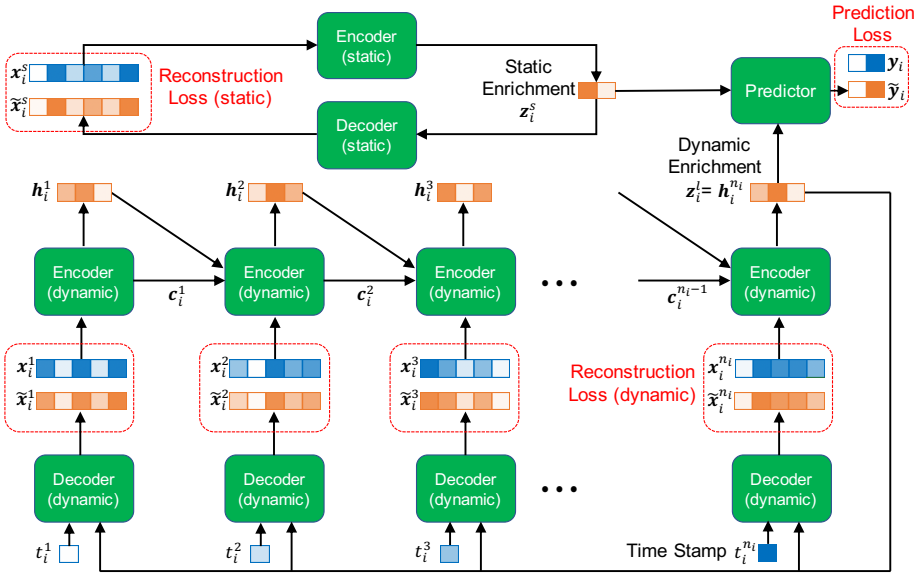

Fig. S1. An schematic illustration about loss function. Our semi-supervised learning autoencoder minimizes the reconstruction loss for the labeled or unlabeled samples, and prediction loss only for the labeled samples.

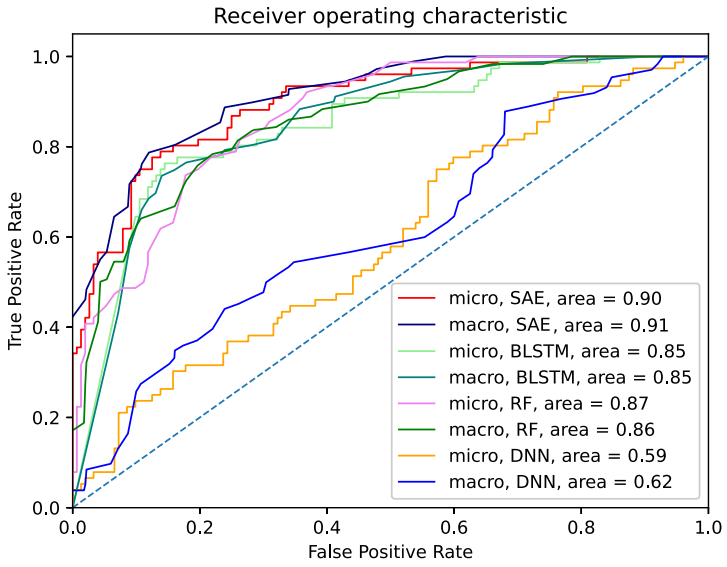

Fig. S2. Micro and macro receiver operating characteristic curves (ROC) averaged across the classes and their area under the curve (AUC). The proportion of training set is 80% and the AUC shows SAE outperforms the other competing models.