2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669428

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

Hoon Seo
Department of Computer Science
Colorado School of Mines
Golden, USA
seohoon@mymail.mines.edu

Hua Wang
Department of Computer Science
Colorado School of Mines
Golden, USA
huawangcs@gmail.com

for the Alzheimer's Disease Neuroimaging Initiative

***Abstract*—Alzheimer's Disease (AD) is a progressive memory disorder that causes irreversible cognitive declines, therefore early diagnosis is imperative to prevent the progression of AD. To this end, many biomarker analysis models have been presented for early AD detection. However, these models may not realize the full data potential due to their failure to integrate longitudinal (dynamic) phenotypic data with (static) genetic data. Sometimes, they may not fully utilize both labeled and unlabeled samples either. To overcome these limitations, we propose a semi-supervised enrichment learning method to learn a fixed-length vectorial representation for each participant, by which the static data record can be integrated with the dynamic data records. We have applied our new method on the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort and achieved 75% accuracy on multiclass AD progression prediction by one year in advance.**

***Index Terms*—Alzheimer's Disease, Representation Enrichment, Longitudinal Learning, Semi-Supervised Learning**

I. 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 its widespread availability that takes advantage of high spatial resolution and good contrast between different soft tissues [2]–[6].

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 [7], including a deep neural network autoencoder and an LSTM autoencoder, each of which learns the vectorial representation from static genetic data and dynamic phenotypic data respectively. 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. We have conducted the experiments to evaluate the proposed model on the real world dataset in the ablation study and comparison with the other prediction models, in which promising results have validate the effectiveness of our new method.

Corresponding author: Hua Wang (huawangcs@gmail.com). This work was supported in part by the National Science Foundation (NSF) under the grants of IIS 1652943, IIS 1849359.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DoD ADNI (Department of Defense award number W81XWH-12-2-0012). A full list of funding sources for ADNI is provided in the document "Alzheimers Disease Neuroimaging Initiative (ADNI) Data Sharing and Publication Policy" available through adni.loni.usc.edu/.

II. Our Method

In the following of this paper, we denote a vector as a bold lower case letter, and a matrix as a bold upper case letter. For a matrix $\mathbf{X}$, we use $[\mathbf{X}]^r$, $[\mathbf{X}]_c$, $[\mathbf{X}]_c^r$ to denote the r-th row, c-th column, an element at the 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 $\mathcal{X}_i = \{\mathbf{x}_i^b, \mathbf{x}_i^s, \mathbf{X}_i, \mathbf{M}_i, \mathbf{t}_i\}$ as follows:

- $\mathbf{x}_i^b \in \Re^{D_b}$ is a vector of basic demographic information (age and gender, $D_b = 2$), and $\mathbf{x}_i^s \in \Re^{D_s}$ is a vector of SNPs ($D_s = 1223$);
- $\mathbf{X}_i = [\mathbf{x}_i^1; \mathbf{x}_i^2; \cdots; \mathbf{x}_i^{n_i}] \in \Re^{n_i \times D_l}$ are the longitudinal records collected across the n_i time points. We note that n_i vary across the participants;
- $\mathbf{M}_i = [\mathbf{m}_i^1; \mathbf{m}_i^2; \cdots; \mathbf{m}_i^{n_i}] \in \{1, 0\}^{n_i \times D_l}$ are the binary masks of longitudinal records $\mathbf{X}_i$, where 1 and 0 indicates the observed and unobserved entry;
- $\mathbf{t}_i = [t_i^1; t_i^2; \cdots; t_i^{n_i}] \in \Re^{n_i}$ are the time stamps of n_i records.

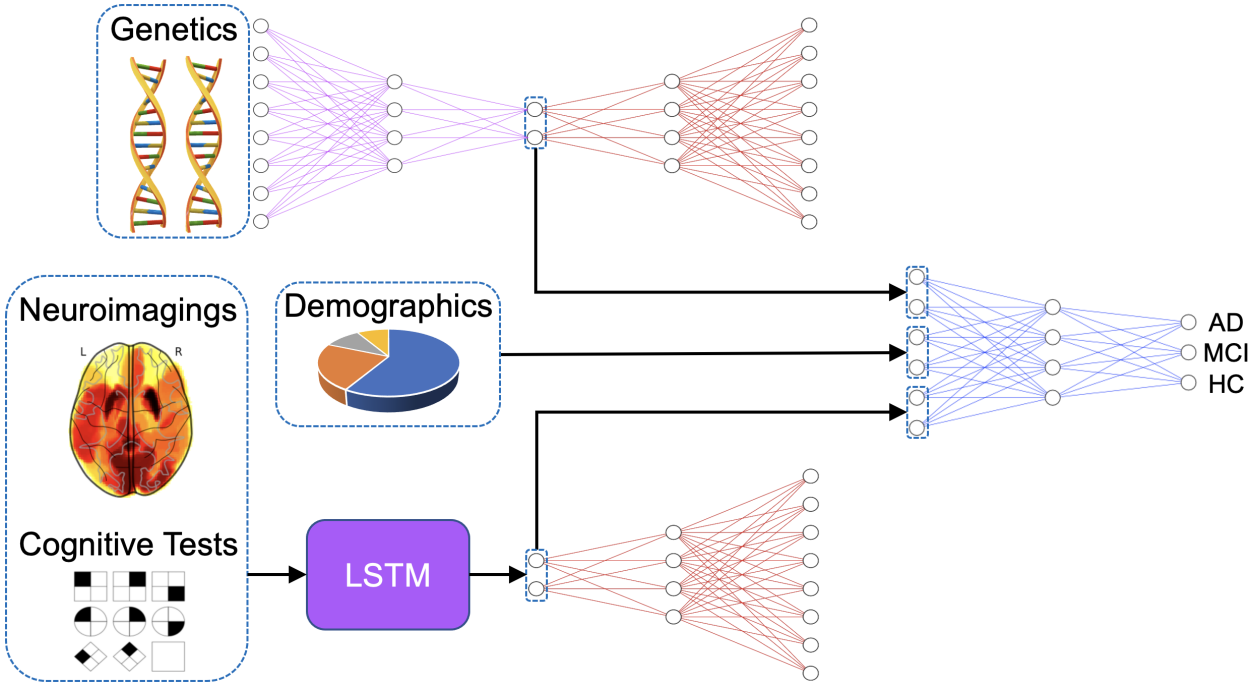

Fig. 1. In the proposed semi-supervised learning model, Encoder, Decoder, and Predictor are trained jointly for the labeled samples, while only the Encoder and Decoder are trained for the unlabeled samples.

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 $\mathbf{x}_i^j = [\mathbf{x}_{i,VBM}^j, \mathbf{x}_{i,FS}^j, \dots]$, and the missing entries are filled with the constant 0. 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 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 overview of the proposed model is described in Fig. 1.

We use the autoencoder [7] to learn the intrinsic representations of the 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 $\phi_{SNP}(\mathbf{x}_i^s;\ \theta_E^s) = \mathbf{z}_i^s$, and an decoder $\psi_{SNP} : \Re^{d_s} \mapsto \Re^{D_s}$ that decodes the enriched representation into reconstructed vector of SNPs $\psi_{SNP}(\mathbf{z}_i^s;\ \theta_D^s) = \tilde{\mathbf{x}}_i^s$, where θ_E^s and θ_D^s is set of trainable weights matrices and bias vectors for encoder and decoder respectively. The deep neural network is defined as the K consecutive fully connected layers where the output of the k-th layer is:

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

where σ is the activation function. 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. 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:

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

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

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 especially when the time intervals between the records are uneven,. The pattern of missing entries aids 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 [8]:

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

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

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

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

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

where $\mathbf{k}_i^j$, $\mathbf{o}_i^j$, $\mathbf{f}_i^j$ 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 of the previous records. In Eq. (5), 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 the long term temporal trends. The output gate $\mathbf{o}_i^j$ in Eq. (6) and Eq. (7) refers to the cell state $\mathbf{c}_i^j$ and generates the hidden representation $\mathbf{h}_i^j$ conveying the useful information for reconstruction of previous records and prediction. Here the time stamps play an important role, and if the drastic change in the participant's records has been observed in a short period of time, it indicates the significant changes in the disease status and the output gate will reflect it in the learned representation $\mathbf{h}_i^j$. For example, the LSTM encoder can capture the cognitive decline from the temporal trends in the scores of cognitive assessments. The LSTM encoder may have more than one LSTM layer stacked on each other. In stacked LSTMs, the hidden states across time points $[\mathbf{h}_i^1, \mathbf{h}_i^2, \cdots, \mathbf{h}_i^{n_i}]$ of previous LSTM layer is passed to the next LSTM layer.

We propose a decoder for dynamic data enrichment with the fully connected layers instead of another LSTM which is traditionally used. A previous study [9] 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 dependencies between the output records 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 the records previously reconstructed, such that $\psi_{dynamic}(\mathbf{z}_i^l, t_i^j;\ \theta_D^l) = \tilde{\mathbf{x}}_i^j \approx \mathbf{x}_i^j$. To recover the original record at the specific time point, the decoder needs to be provided with that time point. Thus the input of the decoder is the concatenation of the enriched representation $\mathbf{z}_i^l$ and the time stamp t_i^j, which is $[\mathbf{z}_i^l, t_i^j] \in \Re^{d_l+1}$. By forwarding the input to the decoder's fully connected layers, we can generate the reconstructed record $\tilde{\mathbf{x}}_i^j$, and we have the stack of reconstructed records of the i-th participant: $\tilde{\mathbf{X}}_i = [\tilde{\mathbf{x}}_i^1; \tilde{\mathbf{x}}_i^2; \cdots; \tilde{\mathbf{x}}_i^{n_i}]$. We update θ_E^l and θ_D^l to minimize the difference between the reconstructed and original records for the observed entries:

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

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

From the enriched representations $\mathbf{z}_i^l$ and $\mathbf{z}_i^s$ of dynamic and static data, additional 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 inducing the enriched representation to convey the useful information to reconstruct the original records and predict the target label:

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

where γ_1, γ_2, and γ_3 are hyperparameters to adjust the impact of each loss. We emphasize that all the components and losses optimized simultaneously through the labeled and unlabeled samples both as described in Fig. 1. We choose the cross entropy for prediction loss $\mathcal{L}_{predict}(\mathbf{y}_i, \tilde{\mathbf{y}}_i;\ \theta_D^p)$ defined as:

$$\begin{cases} 0 & i \notin \Omega_{train},\\ -\left\|\mathbf{y}_i \odot \log(\tilde{\mathbf{y}}_i) + (\mathbf{1} - \mathbf{y}_i) \odot \log(\mathbf{1} - \tilde{\mathbf{y}}_i)\right\|_1 & i \in \Omega_{train}, \end{cases}$$

where $\mathbf{1}$ is a vector of 1's and log is an element-wise logarithm function. The high capacity unsupervised autoencoder may suffer from the tendency to learn trivial identity mapping and memorize the input [9] which is not useful for predicting the target label. In our semi-supervised learning model, the addition of the prediction loss can prevent this memorization problem.

III. Experiments

We obtained the data used in this experiment from the ADNI database [10], by which we compare the prediction performance of the proposed model to the widely used prediction models. We use AD progression in AD, MCI and HC as predictive targets in our studies.

A. Competing models

To evaluate the effectiveness of proposed semi-supervised autoencoder (SAE), we compare our new method against the following competing models:

- Baseline LSTM (BLSTM) by removing decoders ψ_{SNP} and $\psi_{dynamic}$ from SAE for comparison between supervised vs. semi-supervised approaches,
- SAE-woS (SAE without static data) by removing static data enrichment ϕ_{SNP} and ψ_{SNP} to observe the impact of inclusion of static data,
- Non-longitudinal models, such as Random Forest [11] (RF) with 34 max depth, Ridge Classifier (RC) with regularization parameter of 1000, and deep neural network (DNN) with 5 fully connected layers (the number of units are 150, 125, 100, 50, and 25) to assess the benefits of longitudinal approaches.

For the non-longitudinal models, we provide the concatenation of the most recent record $[\mathbf{x}_i^b, \mathbf{x}_i^s, \mathbf{x}_i^{n_i}]$. SAE and SAE-woS are trained with the training and test sets both in a semi-supervised manner, while the other competing models are trained only with training set. Although the order of participants is randomly shuffled to avoid bias, we use the same training and test data across all the competing methods for a fair comparison.

B. Hyperparameters of proposed model

For our model SAE, the static encoder ϕ_{SNP} and decoder ψ_{SNP} have 2 fully connected layers (FC) with tanh activation function. 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 single LSTM with 64 units and tanh activation function. We set $\gamma_1 = 100$, $\gamma_2 = 10$, $\gamma_3 = 1$ in Eq. (9). To minimize the loss function in Eq. (9), we adapt the Adam optimization policy [12] at a learning rate of 0.0003. We do not use any regularization or dropout techniques as they degrade the performance, but we prevent overfitting by early stopping the training when the prediction loss does not decrease during the last three epochs. The accuracy on the validation set in 5-fold cross validation scheme is used as a criterion for selecting hyperparameters of our model and competing models.

C. Result and Evaluation

From the results reported in Fig. 2, the proposed model SAE generally outperforms the other competing models across the different proportions of training set. To be specific, the semi-supervised approaches SAE and SAE-woS show the better predictions compared to the supervised approaches especially when proportion of training set is small. We suspect this is due to our semi-supervised learning approach that allows our model to learn from unlabeled samples while still fully utilizing the benefits of labeled samples. In addition, the predictions of the semi-supervised model are more stable (with a smaller standard deviation) when compared to the supervised learning models possibly due to their supervised nature that rely on labels of the partial dataset.

When SAE is compared to SAE-woS, the inclusion of static (genetic) data improves the predictions. This comparison shows that the integrated representation of static and dynamic data improves predictions on the disease status of participants. Overall, the longitudinal models SAE, SAE-woS, and BLSTM perform better than non-longitudinal models RF, DNN, and RC. These results validate the usefulness of the proposed longitudinal semi-supervised learning approach integrating static and dynamic data, and show our model's promise in the early diagnosis of AD.

IV. 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. We have conducted experiments on the ADNI dataset and the results show that our model outperforms competing predictive models and semi-supervised longitudinal enrichment learning improves the prediction.

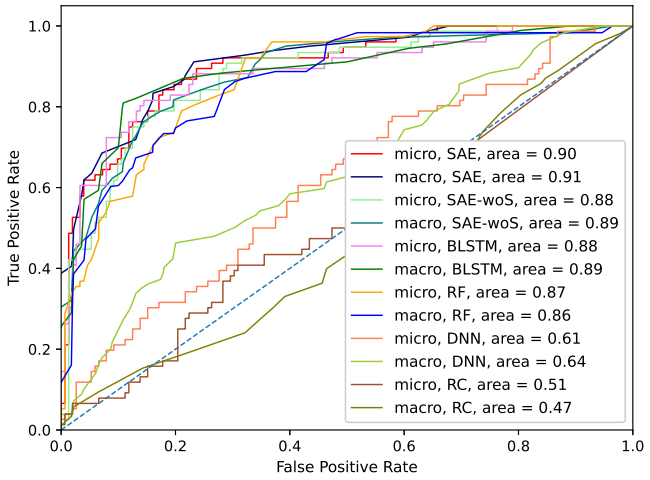

Fig. 2. Micro and macro receiver operating characteristic curves (ROC) averaged across the classes and their area under the curve (AUC). The proportion of training set is 80% and the AUC shows SAE outperforms the other competing models.

References

[1] D. E. Barnes and K. Yaffe, "The projected effect of risk factor reduction on alzheimer's disease prevalence," *The Lancet Neurology*, vol. 10, no. 9, pp. 819–828, 2011.

[2] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack Jr, J. Ashburner, R. S. Frackowiak, A. D. N. Initiative *et al.*, "Predicting clinical scores from magnetic resonance scans in alzheimer's disease," *Neuroimage*, vol. 51, no. 4, pp. 1405–1413, 2010.

[3] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, and A. D. N. Initiative, "From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps," *Bioinformatics*, vol. 28, no. 18, pp. i619–i625, 2012.

[4] L. Brand, H. Wang, H. Huang, S. Risacher, A. Saykin, L. Shen *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*. Springer, 2018, pp. 555–562.

[5] L. Brand, K. Nichols, H. Wang, L. Shen, and H. Huang, "Joint multi-modal longitudinal regression and classification for alzheimer's disease prediction," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1845–1855, 2019.

[6] L. Lu, S. Elbeleidy, L. Baker, H. Wang, L. Shen, and H. Huang, "Improved prediction of cognitive outcomes via globally aligned imaging biomarker enrichments over progressions," *IEEE Transactions on Biomedical Engineering*, 2021.

[7] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[8] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[9] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*, 2015, pp. 843–852.

[10] S. L. Risacher, L. Shen, J. D. West, S. Kim, B. C. McDonald, L. A. Beckett, D. J. Harvey, C. R. Jack Jr, M. W. Weiner, A. J. Saykin *et al.*, "Longitudinal mri atrophy biomarkers: relationship to conversion in the adni cohort," *Neurobiology of aging*, vol. 31, no. 8, pp. 1401–1418, 2010.

[11] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.