EARLY DIAGNOSIS OF ALZHEIMER'S DISEASE WITH DEEP LEARNING

Siqi Liu[1], Sidong Liu[1], Student Member, IEEE, Weidong Cai[1], Member, IEEE, Sonia Pujol[2], Ron Kikinis[2], Dagan Feng[1], Fellow, IEEE

[1]BMIT Research Group, School of IT, University of Sydney, Australia
[2]Surgical Planning Lab, Brigham & Women's Hospital, Harvard Medical School, Boston, USA

ABSTRACT

The accurate diagnosis of Alzheimer's disease (AD) plays a significant role in patient care, especially at the early stage, because the consciousness of the severity and the progression risks allows the patients to take prevention measures before irreversible brain damages are shaped. Although many studies have applied machine learning methods for computer-aided-diagnosis (CAD) of AD recently, a bottleneck of the diagnosis performance was shown in most of the existing researches, mainly due to the congenital limitations of the chosen learning models. In this study, we design a deep learning architecture, which contains stacked auto-encoders and a softmax output layer, to overcome the bottleneck and aid the diagnosis of AD and its prodromal stage, Mild Cognitive Impairment (MCI). Compared to the previous workflows, our method is capable of analyzing multiple classes in one setting, and requires less labeled training samples and minimal domain prior knowledge. A significant performance gain on classification of all diagnosis groups was achieved in our experiments.

Index Terms—Alzheimer's disease, neuroimaging, classification

1. INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia, which is a progressive brain disorder mostly occurring in the late life [1]. Comparing with the patient's previous functions, a decline in memory and other cognitive functions is noted as a primary dementia syndrome. In 2006, the worldwide prevalence of AD was 26.6 million, and this number is expected to double in every 20 years. By 2046, 1.2% of the global population will be affected by AD [2]. The early diagnosis of AD is primarily associated to the detection of Mild Cognitive Impairment (MCI), a prodromal stage of AD. Though the memory complaints and deficits of MCI do not notably affect the patients' daily activities, it has been reported that MCI has a high risk of progression to AD or other forms of dementia [3]. The accurate early diagnosis AD, especially identifying the risk of progression of MCI to AD, affords the AD patients awareness of the severity and allows them to take prevention measures, e.g., lifestyle changing and medications [4].

Many machine learning methods have been proposed to aid the diagnosis of AD based on high dimensional features extracted from various neuroimaging biomarkers, subclass MRI and PET. These machine learning methods not only need to identify the AD subjects from the normal control (NC) subjects automatically, but also predict the risk of MCI subjects evolving to AD, thus MCI instances can be labeled as MCI non-converters (ncMCI) or MCI converters (cMCI), depending on the risk of progression. Therefore, the early diagnosis of AD can be naturally modeled to be a multi-class classification problem.

Some previous studies simplified the problem into a binary classification task [5-7]. The workflow in [5] combined features from multiple biomedical modalities using a multi-kernel SVM classifier. However it is difficult for SVM to classify subjects with more than two classes in one setting. Some methods embedded the prior knowledge in the model designing [8-10]. For example, an optimized graph cut algorithm was proposed in [8] with parameters adjusted corresponding to the distribution of particular classes in the training dataset. The dependence of prior knowledge may be also sensitive to the changes of the dataset and hard to configure.

Targeting at the constraints in previous studies, we believe the existing workflows can be efficiently optimized. In this paper, we proposed a novel early diagnosis method for AD based on a deep learning architecture, consisting of stacked sparse auto-encoders and a softmax regression layer. The proposed method has a multi-class nature and could reduce the reliance on prior knowledge about the data. Furthermore, our method is semi-supervised that can be extended to use unlabeled training samples, which are easier and cheaper to obtain.

2. METHODS

2.1. Representation Learning with Auto-encoders

The performance of machine learning models depends heavily on the representation quality of the original training samples. The representative learning models challenge the other algorithms by the capabilities of disentangling the complex and structural dependencies in the high dimensional feature spaces. The feature representation can benefit from the depth of the learning structure, which learns more profound representations from the m

anifolds extracted by the previous hidden layers [11].

Our learning structure can be demonstrated by two primary components: stacked sparse auto-encoders and a softmax regression layer. The auto-encoders obtain deep representations of the original input. The softmax regression layer classifies instances by selecting the highest predicted probabilities of each label.

The sparse auto-encoder is an encoding structure, which consists of a neural network with multiple hidden layers as shown in Fig. 1. Neurons of the input layer represent the original input vector. Each hidden layer can be seen as a higher level representation of the previous layer, though it is usually nontrivial to define the exact meaning of each layer. The output layer is a sparse representation of the input layer with the same dimensionality as the input [12-14].

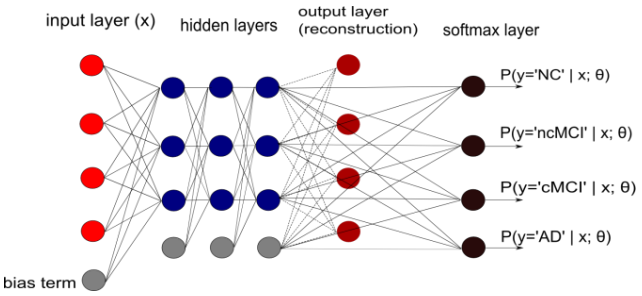

Figure 1. Illustration of the deep learning structure with a multilayered neural network. The input layer feed-forwards pre-computed features from MRI data and PET data to the hidden layers. Each hidden layer obtains nonlinear transformations from the previous layer and is optimized to reconstruct the original instance. The softmax layer takes the activations of the last hidden layer as inputs and gives the probability of each AD stage.

The activation signals are iteratively propagated forward through the network using Eq. (1), until the the output layer is reached. The neuron activation a of each layer can be computed by

$$\begin{cases} a_{(l)}^{(i)} = x^{(i)}, l = 1 \\ a_{(l)}^{(i)} = \sigma(W^l a + b), l > 1 \end{cases} \quad (1)$$

$$h(W, b, x) = a^{(N)}, \quad (2)$$

where x is the unlabeled data $\{x^{(i)}\}_{i=1}^{m}$; W is the weight matrix controls the activation effect between neurons on neighbor layers; b is the bias term; σ is the activation function, which can be set to sigmoid function or hyperbolic tangent function to introduce non-linearity for the network to model complex relationships [15]; and $h(W, b, x)$ is the representation of the input data, as well as the activations at the output layer.

By altering the number of the neurons at each hidden layer, we are able to perform the feature dimensionality reduction or over-completion. It is also possible to combine features of different views or modalities by concatenating features into one input vector. The sparse auto-encoder is proven to work well in data fusion, catching the synergy between different modalities.

To train this unsupervised model, the representation loss is used as the objective function for optimization as in Eq. (3):

$$L(W, b, x, z) = \min_{W,b} E(W, b, x, z) + \gamma \|W\|_2^2 + \beta K(W, b, x), \quad (3)$$

where $E(W, b, x, z) = \|h(W, b, x) - z\|_2^2$ is the representation loss with squared error; The second term is the weight decay that lead to small weights; the third term is the sparsity penalty regularized by β with a target activation of ρ close to 0, which enforces a sparse representation by penalizing the objective function using Kullback-Leibler divergence across n training samples [16]

$$K(W, b, x) = \sum_{j}^{n} ID_{KL}\left(\rho \| \frac{1}{m} \sum_{i=1}^{m} h_j\left(x^{(i)}; W, b\right)\right). \quad (4)$$

The gradients of the objective function can be computed exactly by applying back-propagation algorithm; therefore, the cost function in Eq. (3) can be optimized by gradient descent based algorithms. Considering the limitation on the quantity of biomedical data, we applied L-BFGS in this study for better performance [13].

We train the hidden layers of sparse auto-encoder one at a time and stack them to form a complete neural network by removing the temporary output layer. It was proven to benefit the first and last few hidden layers more than propagating the entire network [17].

2.2. Softmax Regression

For AD classification, a softmax output layer is added on the top of the trained auto-encoder stack containing only previous hidden layers [14, 18]. The softmax layer uses a different activation function, which might have nonlinearity, different from the one applied in previous layers. The softmax activation function is

$$h_i^l = \frac{e^{W_i^l h^{l-1} + b_i^l}}{\sum_j e^{W_j^l h^{l-1} + b_j^l}}, \quad (5)$$

where W_i^l is the i-th row of W^l and b_i^l is the i-th bias term of last layer. We can use h_i^l as an estimator of $P(Y = i|x)$, where Y is the associated label of input data vector x. In our case, four output neurons at the softmax layer can be interpreted as the probabilities of diagnosing an example as NC, ncMCI, cMCI or AD.

Similar to the procedure of training the Deep Belief Net (DBN) [19], we can further fine-tune all the parameters in the network with respect to the overall classification loss by unfolding all the auto-encoders and applying the back-propagation algorithm on the entire network [14, 20].

2.3. ROI Sensitivity Evaluation

Hidden neurons at the first layer of our network are trained to catch different patterns of the input data. The features acquired from the first hidden layer can be examined to identify ROIs that are sensitive to the AD progression.

Based on Eq. (1), we can derive the input pattern x_j^*, which maximally activate the hidden neuron a_i

$$x_{ij}^* = \frac{W_{ij}^{(1)}}{\|W\|_2}. \quad (6)$$

In our case, we are able to map x_{ij}^* to the ROI where this feature was extracted. By splitting the pattern x into m feature views, we compute the variance $D_j^{(m)}$ of all $x_j^{(m)}$ of the same ROI, measuring how the ROI activate different hidden neurons. When $D_j^{(m)}$ is low, we consider the features extracted from region j are more stable for AD diagnosis than the high-variance regions. The overall feature stability S_j of the j-th ROI can be computed as

$$S_j = \sum_m \frac{\sum_j D_j^{(m)}}{D_j^{(m)}}. \quad (7)$$

S can be convolved with a Gaussian filter to exaggerate the distinctions between each ROI.

3. EXPERIMENT

3.1. Data Acquisition and Feature Extraction

Our experiment used the neuroimaging data obtained from Alzheimer's disease Neuroimaging Initiative (ADNI) database[1][21]. We recruited the MRI images of 311 subjects from the ADNI baseline cohort, including 65 AD subjects, 67 cMCI subjects, 102 ncMCI subjects and 77 normal control subjects. All the MRI images are nonlinearly registered to the ICBM_152 template [8, 22, 23] and further segmented into 83 functional regions. We extracted the grey matter volumes from MRI [24] and CMRGlc patterns from PET [25]. The features were further selected with Elastic Net [26] before each classification task. To support sigmoidal decoder, all the features are normalized to zero-mean and between 0 and 1.

3.2. Evaluation

We implemented the deep learning framework described in this paper on Matlab 2013a. Random search in a log-domain was applied to choose the hyper-parameters that could be sensitive to the results when training sets are small [27].

The widely used single-kernel SVM (SK-SVM) and multi-kernel SVM (MK-SVM) [5] were chosen to compare with our proposed method. All SVM based experiments were conducted using LIBSVM library, with the radial basis function (RBF) kernel implemented [28]. We applied the 'one against all' approach to allow SVM to perform the four-class classification problem [29]. Grid search was used to adjust the parameters of SVM. All the experiments were carried out with the same features.

The structure was evaluated using 10-fold cross-validation on the softmax layer. To maximally avoid the 'lucky trails', we randomly sampled the training and testing instances from each class to ensure they have similar distributions as the original dataset. For all methods in each fold of cross validation, about 90% subjects were used for training (including the pre-training of the deep neural nets) and the rest subjects were used for testing.

[1] The ADNI database is available at http://www.loni.ucla.edu/ADNI.

4. RESULTS

4.1. Visualization of Sensitive ROIs

Figure 2 is a mapping from feature stability to ROIs on a masked 3D MRI image (83 ROIs). The distinctions between various ROIs were clearly visualized. The darker regions tend to be more sensitive to the progression of AD and MCI than the lighter ROIs, since features extracted from these RIOs tend to benefit all the hidden neurons equally. We convolved the image with a Gaussian filter for sharper distinctions. The dark regions are not denoted to be totally trivial, but carrying less predictive information.

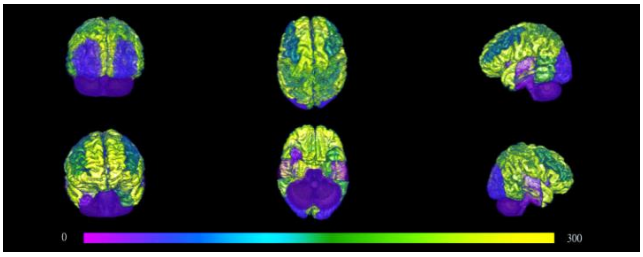

Figure 2. The variance map denotes that brighter ROIs tend to be more affected by the progression of AD. The image was generated using 3D Slicer (V4.3) [30].

4.2. Classification Performance

The performance comparisons between the proposed method and SVM-based methods are shown in Table 1 and Table 2.

Table 1: The mean values of the binary classification performance (%) with pre-computed features from MRI and PET images

Methods	AD vs. NC			MCI vs. NC		
	ACC	SEN	SPE	ACC	SEN	SPE
SK-SVM	84.40	84.64	84.31	76.81	56.14	86.24
MK-SVM	86.42	84.98	87.83	77.25	55.48	87.10
Proposed	87.76	88.57	87.22	76.92	74.29	78.13

Table 2: The mean values of the 4-class classification performance (%) with pre-computed features from MRI and PET images

Methods	NC	cMCI	cMCI	AD	ACC	SEN	SPE
SK-SVM	40.00	40.00	41.67	49.25	43.82	61.96	71.71
MK-SVM	45.00	38.57	52.63	48.36	45.28	70.54	74.03
Proposed	55.43	43.02	31.37	51.96	47.42	65.71	83.75

All figures displayed in both Table 1 and Table 2 were mean values obtained from experiments with the optimized settings. As shown in Table 1, the deep learning method produced a better overall accuracy (87.76%) in binary

classification of AD. When classifying NC and MCI, the proposed method demonstrated almost even accuracy as SVM, because the training set for this task has unbalanced proportion of each class (77 NC subjects and 169 MCI subjects), that is harder for training parametric model. In addition to the classification accuracy, higher sensitivity values (88.57% and 74.29%) were observed. The greater sensitivity is also known to be beneficial to diagnosis, because the cost of misclassifying between different groups usually varies, *e.g.*, diagnosing AD or MCI patients to NC may cause more severe effects than the reverse [5]. In Table 2, the former 4 columns represent the average classification precisions achieved on each class and the latter 3 columns represent the overall performance. Better precisions were observed on three classes (55.43% on NC, 32.02% on ncMCI and 51.96% on AD), except cMCI which contains much fewer subjects than its sibling class ncMCI. A performance gain has been obtained on the overall accuracy and the overall specificity (47.42% and 83.75%) comparing to SVMs.

5. CONCLUSION

In this study, we proposed a novel method for the early diagnosis of AD and MCI based on deep learning. Compared to the conventional binary classification methods, such as SVM, our method conducts AD diagnosis as a multi-class classification task, with minimal prior knowledge dependency in the model optimization. The proposed method also performs dimensionality reduction and data fusion at the same time to reserve the synergy between data modalities. A performance gain was achieved in both binary and four-class classifications. We have also proven that multi-layered parametric learning model can be applied on biomedical datasets with smaller size to extract high-level biomarkers. This study may have a great potential to lead to a new perspective for computer-aided diagnosis in other biomedical fields.

6. ACKNOWLEDGEMENTS

This work was supported in part by the ARC, AADRF, NA-MIC (NIH U54EB005149), and NAC (NIH P41EB015902).

7. REFERENCES

[1] G. McKhann, D. Drachman, *et al.*, "Clinical diagnosis of Alzheimer's disease Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease," *Neurology,* vol. 34, pp. 939-939, 1984.

[2] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of Alzheimer's disease," *Alzheimer's & dementia,* vol. 3, pp. 186-191, 2007.

[3] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, *et al.*, "Mild cognitive impairment," *The Lancet,* vol. 367, pp. 1262-1270, 2006.

[4] E. D. Roberson and L. Mucke, "100 years and counting: prospects for defeating Alzheimer's disease," *Science,* vol. 314, pp. 781-784, 2006.

[5] D. Zhang, Y. Wang, L. Zhou, *H. Yuan, and D. Shen*, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *Neuroimage,* vol. 55, pp. 856-867, 2011.

[6] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer*, et al.*, "Automatic classification of MR scans in Alzheimer's disease," *Brain,* vol. 131, pp. 681-689, 2008.

[7] N. Singh, A. Y. Wang, *et al.*, "Genetic, structural and functional imaging biomarkers for early detection of conversion from MCI to AD," in *MICCAI 2012*,, *Part I, LNCS* 7510, pp. 132-140.

[8] S. Liu, W. Cai, L. Wen, and D. Feng, "Neuroimaging Biomarker based Prediction of Alzheimer's Disease Severity with Optimized Graph Construction," *ISBI 2013*, 2013.

[9] S. Liu, L. Zhang, W. Cai, Y. Song, Z. Zhang, L. Wen, D. Feng, "A supervised multi-view spectral embedding method for neuroimaging classification," in *ICIP 2013*, 2013, pp. 602-605.

[10] S. Liu, W. Cai, Y. Song, *et al.*, "Localized Sparse Code Gradient in Alzheimer's Disease Staging", in *EMBC 2013*, 2013 pp.5398-5401

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," 2013.

[12] C. Poultney, S. Chopra, and Y. L. Cun, "Efficient learning of sparse representations with an energy-based model," in *Advances in neural information processing systems*, 2006, pp. 1137-1144.

[13] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, A. Ng, and Q. V. Le, "On optimization methods for deep learning," *ICML-11*, 2011, pp. 265-272.

[14] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning,* vol. 2, pp. 1-127, 2009.

[15] G. E. Hinton, "Connectionist learning procedures," *Artificial intelligence,* vol. 40, pp. 185-234, 1989.

[16] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum,* vol. 9, 2010.

[17] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems,* vol. 19, p. 153, 2007.

[18] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*, 1990, pp. 227-236.

[19] Y.-l. Boureau and Y. L. Cun, "Sparse feature learning for deep belief networks," in *Advances in neural information processing systems*, 2007, pp. 1185-1192.

[20] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science,* vol. 313, pp. 504-507, 2006.

[21] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging,* vol. 27, pp. 685-691, 2008.

[22] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles*, et al.*, "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)," *Phil. Trans. of Royal Soci. London. B: Biol. Sci.*, vol. 356, pp. 1293-1322, 2001.

[23] J. A. Schnabel, D. Rueckert, M. Quist, J. M. Blackall, A. D. Castellano-Smith, T. Hartkens*, et al.*, "A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations," in *MICCAI 2001, LNCS 2208*, pp. 573-581.

[24] P. G. Batchelor, A. D. Castellano Smith, D. L. G. Hill, D. J. Hawkes, T. C. S. Cox, and A. Dean, "Measures of folding applied to the development of the human fetal brain," *Medical Imaging, IEEE Transactions on,* vol. 21, pp. 953-965, 2002.

[25] W. Cai, S. Liu, L. Wen, S. Eberl, M. J. Fulham, D. Feng, "3D neurological image retrieval with localized pathology-centric CMRGlc patterns," in *ICIP 2010*, 2010, pp.3201,3204.

[26] Zou, H. and Hastie, T., "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, vol. 67, pp. 301–320, 2005.

[27] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Machine Learning Research,* vol. 13, pp. 281-305, 2012.

[28] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, p. 27, 2011.

[29] J. Weston and C. Watkins. "Multi-class support vector machines," *Technical Report CSD-TR-98-04*, Department of Computer Science, Royal Holloway, University of London, 1998.

[30] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, *et al.*, "3D Slicer as an image computing platform for the Quantitative Imaging Network," *Magn. Reson. Imaging,* vol. 30, no. 9, pp. 1323 – 1341, 2012.