Landmark-based Deep Multi-Instance Learning for Brain Disease Diagnosis – *Supplementary Materials*

Mingxia Liu[a,*], Jun Zhang[a,*], Ehsan Adeli[a], Dinggang Shen[a,b,†]

[a] *Department of Radiology and BRIC, University of North Carolina at Chapel Hill, North Carolina 27599, USA*

[b] *Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea*

In what follows, we present additional experimental results. Specifically, we first report results of multi-class classification in Section 1, and then show the performance of different methods in both inter-imaging-center and intra-imaging-center learning scenarios in Section 2 and Section 3, respectively. We further visually illustrate the learned features for MRI via our proposed LDMIL method in Section 4, and analyze the influences of two parameters on LDSIL (a single-instance variant of LDMIL) in Section 5.

1. Results of Multi-class Classification

Besides two binary classification tasks (*i.e.*, AD vs. NC classification, and pMCI vs. sMCI classification) in the main text, we further study the performance of the proposed LDMIL method in multi-class classification.

[†]Corresponding author

[*]These authors contribute equally to this study

Email addresses: mxliu@med.unc.edu (Mingxia Liu), xdzhangjun@gmail.com (Jun Zhang), eadeli@gmail.com (Ehsan Adeli), dgshen@med.unc.edu (Dinggang Shen)

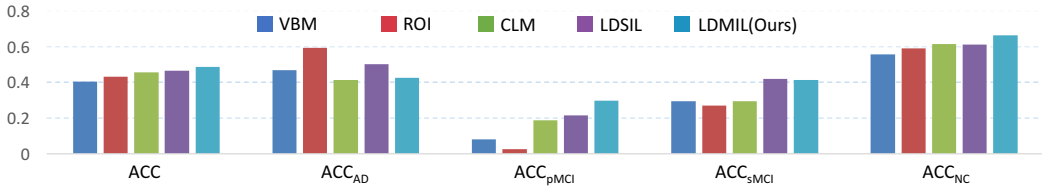

Figure S1: Results of AD vs. pMCI vs. sMCI vs. NC classification achieved by five different methods, with models trained and tested in ADNI-1 and ADNI-2, respectively. ACC: Accuracy.

Specifically, in this group of experiments, we investigate the performances of our landmark-based deep multi-instance learning (LDMIL) method and four competing methods (*i.e.*, VBM, ROI, CLM, and LDSIL) in the task of AD vs. pMCI vs. sMCI vs. NC classification. For performance evaluation, we adopt the overall classification accuracy (ACC) for four categories, as well as the accuracy for each category (*i.e.*, ACC_{AD}, ACC_{pMCI}, ACC_{sMCI}, and ACC_{NC}) as the measurement. Here, we treat subjects from the ADNI-1 dataset as the *training data*, and those from the ADNI-2 dataset as the *independent testing data*. In Fig. S1, we report the results of five different methods in the task of AD vs. pMCI vs. sMCI vs. NC classification.

From Fig. S1, we can observe that two deep learning based methods (*i.e.*, LDSIL, and LDMIL) outperform three conventional methods (*i.e.*, VBM, ROI, and CLM) that require human-engineered features for MRI, in four-class classification. It implies that integrating the feature learning into the classifier training is a good solution for improving diagnostic performance, because the learned features and models can be optimally coordinated. Also, LDMIL achieves better results than LDSIL regarding the overall accuracy (ACC), suggesting that the proposed local-to-global feature learning strategy is effective in extracting discriminative features for brain disease diagnosis.

Table S1: Results of five different methods in AD vs. NC classification on both ADNI-1 and MIRIAD datasets, where models are trained on ADNI-2.

	ADNI-1					MIRIAD				
	ROI	VBM	CLM	LDSIL	LDMIL(Ours)	ROI	VBM	CLM	LDSIL	LDMIL(Ours)
AUC	0.8652	0.8845	0.8872	0.9577	**0.9613**	0.9235	0.9414	0.9572	0.9703	**0.9731**
ACC	0.8084	0.8154	0.8201	0.9088	**0.9159**	0.8696	0.8986	0.8986	**0.9275**	**0.9275**
SEN	0.8643	0.8241	0.8241	0.9195	**0.9196**	0.8913	0.9348	0.9130	**0.9565**	0.9130
SPE	0.7598	0.8078	0.8166	0.8996	**0.9127**	0.8261	0.8261	0.8696	0.8696	**0.9565**
F-Score	0.8075	0.8059	0.8098	0.9037	**0.9104**	0.9011	0.9247	0.9231	0.9438	**0.9462**
MCC	0.6238	0.6307	0.6396	0.8177	**0.8313**	0.7100	0.7696	0.7746	0.8356	**0.8459**

2. Results using Models Trained on ADNI-2

Besides the experiments in the main text (with ADNI-1 as the training set, and ADNI-2 and MIRIAD as independent testing sets), we have now added a new group of experiments to study the generalization capability of our proposed LDMIL method. Specifically, we treat subjects in ADNI-2 as the training data, and those in ADNI-1 and MIRIAD as the testing data. Since there are unbalanced MCI subjects (*i.e.*, 38 pMCI, and 239 sMCI) in ADNI-2, it is difficult to train a good model for pMCI vs. sMCI classification using ADNI-2 as training data. Hence, in this groups of experiments, we only perform the task of AD vs. NC classification, with results given in Table S1. From Table S1, we can see that the proposed LDMIL method yields better results in most cases, compared with four competing methods. Considering that MR images in ADNI-2 were scanned using 3T scanners and those from both ADNI-1 and MIRIAD were acquired by using 1.5T scanners, these results further demonstrate that the proposed LDMIL method generalizes well across different imaging centers.

Table S2: Results of five different methods in AD vs. NC classification on both ADNI-1 and ADNI-2 datasets, where models are trained on ADNI-1 and ADNI-2 via cross validation, respectively.

	ADNI-1					ADNI-2				
	ROI	VBM	CLM	LDSIL	LDMIL(Ours)	ROI	VBM	CLM	LDSIL	LDMIL(Ours)
AUC	0.8804	0.8921	0.9107	0.9591	**0.9622**	0.8632	0.8850	0.8862	0.9431	**0.9537**
ACC	0.8248	0.8364	0.8411	0.9112	**0.9182**	0.8083	0.8139	0.8194	0.8944	**0.9028**
SEN	0.8191	0.8040	0.8492	**0.9196**	**0.9196**	0.8113	0.8491	0.7987	**0.8931**	**0.8931**
SPE	0.8297	0.8646	0.8341	0.9039	**0.9170**	0.8060	0.7861	0.8358	0.8955	**0.9105**
F-Score	0.8130	0.8205	0.8325	0.9059	**0.9127**	0.7890	0.8012	0.7962	0.8820	**0.8903**
MCC	0.6482	0.6709	0.6820	0.8222	**0.8359**	0.6144	0.6308	0.6342	0.7867	**0.8030**

3. Results via Cross Validation

In this group of experiments, we perform 5-fold cross validation (CV) (Liu et al., 2016) on the ADNI-1 and ADNI-2 datasets, respectively. Here, we do not perform 5-fold CV on the MIRIAD dataset, since there are very limited (*i.e.*, 69) subjects in this dataset. The experimental results are reported in Table S2. It can be seen from Table S2 that the proposed LDMIL and LDSIL consistently outperform the conventional methods (*i.e.*, ROI, VBM, and CLM) that adopt human-engineered feature representations for MR images. Hence, from Tables $2-3$ in the main text and Tables S1-S2, we can observe that the proposed LDMIL method has good generalization capability in both inter-imaging-center and intra-imaging-center learning scenarios.

4. Learned Feature Representations for MRI

In the proposed multi-instance convolutional neural network (MICNN), each layer combines the extracted low layer feature maps to learn higher level feature at the next layer in a hierarchical manner. Such architecture helps to describe more abstract anatomical variations of the brain in MRI.

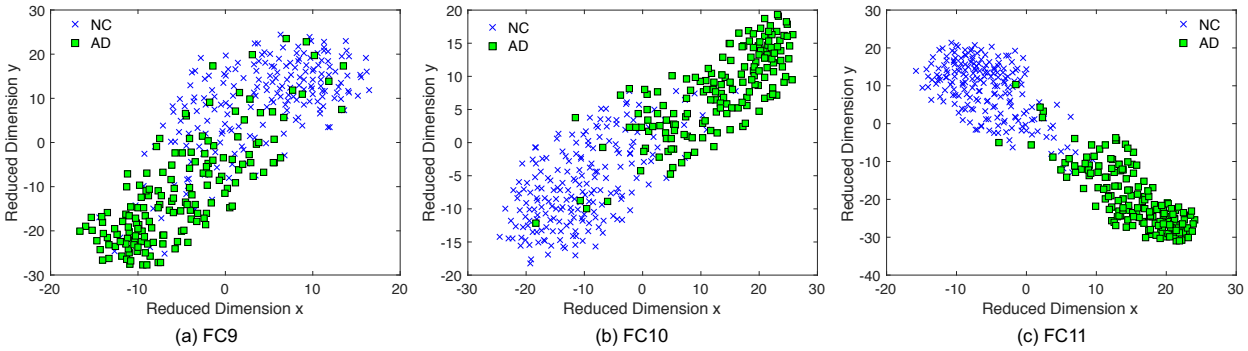

Figure S2: Manifold visualization of AD and NC subjects in ADNI-2 via t-SNE projection (Maaten and Hinton, 2008) in the learned layers of (a) FC9, (b) FC10, and (c) FC11 in the proposed MICNN (shown in Fig. 3 of the main text).

To analyze the discriminative capability of the representations extracted by MICNN, we visualize the learned features at three fully connected layers in MICNN. Specifically, we project those learned features in each fully connected layer down to 2 dimensions using the t-SNE dimension reduction algorithm (Maaten and Hinton, 2008), with results shown in Fig. S2. As can be seen from Fig. S2 (a-c), the sequential fully connected layers (*i.e.*, FC9, FC10, and FC11) gradually enhance the separability between AD and NC subjects along the hierarchy. Also, features at the top-most FC layer (*i.e.*, FC11) are the most discriminative.

5. Parameter Analysis for LDSIL

Different from our proposed landmark based multi-instance learning (LDMIL) method, the landmark based single-instance learning (LDSIL) can learn only patch-level representations for brain MR images. Now we investigate the influence of the number of landmarks and the patch size on the performance of LDSIL. Specifically, we vary the number of landmarks in $\{1, 10, 20, \cdots, 60\}$

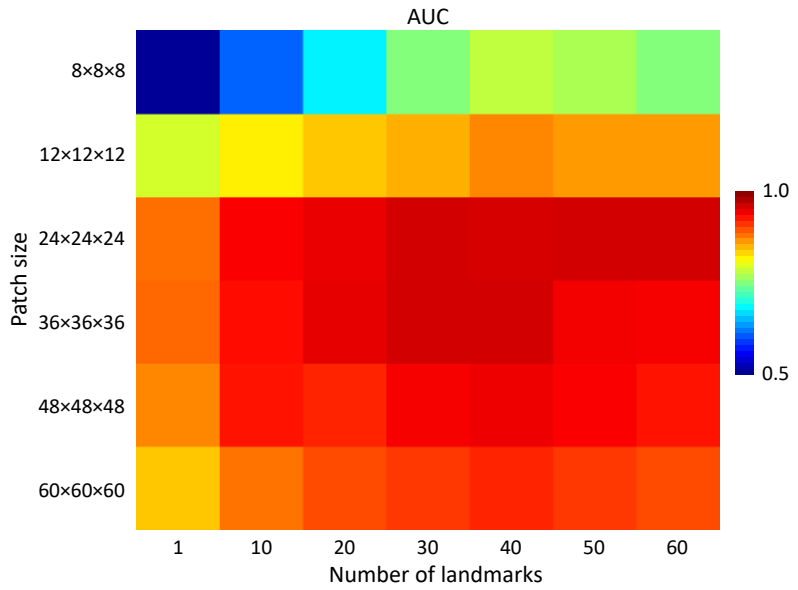

Figure S3: Influence of the number of landmarks and the size of image patches on the performance of LDSIL in tasks of AD vs. NC classification in terms of AUC values. Here, models are trained on ADNI-1 and tested on ADNI-2, respectively.

and the size of image patches in $\{8 \times 8 \times 8, 12 \times 12 \times 12, 24 \times 24 \times 24, 36 \times 36 \times 36, 48 \times 48 \times 48, 60 \times 60 \times 60\}$, and report the AUC values achieved by LDSIL method in AD vs. NC classification on ADNI-2 in Fig. S3, with models trained on ADNI-1.

From Fig. S3, we can observe that LDSIL achieves good results when the number of landmarks is within the range $[30, 60]$ and the size of image patches is within the range $[24 \times 24 \times 24, 36 \times 36 \times 36]$. It suggests that LDSIL requires a relatively larger number of landmarks and larger image patches to yield better results in identifying AD from normal control subjects. In contrast, using fewer landmarks (*e.g.*, 1 and 10) and smaller patches (*e.g.*, $8 \times 8 \times 8$ and $12 \times 12 \times 12$), LDSIL cannot yield satisfying results. The possible reason could be that LDSIL only extracts features from local image patches, without

explicitly modeling the global structural information of the whole brain. In such a case, LDSIL requires using more landmarks and larger image patches to cover the global structural information of brain MR images. In contrast, our proposed LDMIL method can explicitly model both the local and global structural information via the proposed MICNN architecture shown in Fig. 3 in the main text.

References

Liu, M., Zhang, D., Shen, D., 2016. Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. IEEE Transactions on Medical Imaging 35, 1463–1474.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.