Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

ACCEPTED MANUSCRIPT

Multi-task Multi-level Feature Adversarial Network for Joint Alzheimer's Disease Diagnosis and Atrophy Localization using sMRI

To cite this article before publication: Kangfu Han *et al* 2022 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/ac5ed5

Multi-task Multi-level Feature Adversarial Network for Joint Alzheimer's Disease Diagnosis and Atrophy Localization using sMRI

Kangfu Han[1, 2], Man He[1, 2], Feng Yang[1, 2, *], Yu Zhang[1, 2, *]

[1] School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong, China, 510515

[2] Guangdong Provincial Key Laboratory of Medical Image Processing, Guangzhou, Guangdong, China, 510515

* Authors to whom any correspondence should be addressed.

E-mail: yuzhang@smu.edu.cn and yangf@smu.edu.cn

December 2021

Abstract. Capitalizing on structural magnetic resonance imaging (sMRI), existing deep learning methods (especially convolutional neural networks, CNNs) have been widely and successfully applied to computer-aided diagnosis (CAD) of Alzheimer's disease and its prodromal stage (i.e., mild cognitive impairment, MCI). But considering the generalization capability of the obtained model trained on limited number of samples, we construct a Multi-task Multi-level Feature Adversarial Network (M^2FAN) for joint Alzheimer's disease diagnosis and atrophy localization using baseline sMRI. Specifically, the linear-aligned T1 MR images were first processed by a lightweight CNN backbone to capture the shared intermediate feature representations, which were then branched into a global subnet for preliminary dementia diagnosis and a multi instance learning (MIL) network for brain atrophy localization in multi-task learning manner. As the global discriminative information captured by the global subnet might be unstable for disease diagnosis, we further designed a module of multi-level feature adversarial learning (MFAL) that accounts for regularization to make global features robust against the adversarial perturbation synthesized by the local/instance features to improve the diagnostic performance. Our proposed method was evaluated on three public datasets (i.e., ADNI-1, ADNI-2, and AIBL), demonstrating competitive performance compared with several state-of-the-art methods in both tasks of AD diagnosis and MCI conversion prediction.

Keywords: Alzheimer's disease, Multi-level Feature Adversarial Learning, Structural MRI, Atrophy Localization.

1. Introduction

Alzheimer's Disease (AD), the most cause of dementia, is associated with the accumulation of toxic protein (i.e., beta-amyloid plaques and tau tangles) that results in the onset of memory loss and cognitive dysfunctions following the progressive brain

inflammation and atrophy Association (2019); Fox et al. (1996). As the world population ages, it is estimated that 1 out of 85 people in the global will be living with Alzheimer's by 2050 Brookmeyer et al. (2007). Recently, neuroimaging, such as magnetic resonance imaging (MRI), fluorodeoxyglucose positron emission tomography (FDG-PET), have profoundly advanced the neuroscientific research in dementia diagnosis. Especially, MRI with high-resolution offers the possibility to study pathological brain changes associated with AD in vivo based on brain morphometric pattern analysis, identifying anatomical biomarkers of Alzheimer's disease and its prodromal stage, such as mild cognitive impairment (MCI) Baron et al. (2001); Frisoni et al. (2010); Jack et al. (1999).

Based on conventional machine learning and/or emerging deep learning, varieties of methods have advanced the diagnostic performance of AD and its prodromal stage (i.e., MCI) using structural MRI (sMRI), which generally include three steps Rathore et al. (2017): 1) regions-of-interest (ROIs) identification from the whole-brain sMRI, such as brain atlas or local image patches; 2) feature extraction from ROIs; and 3) classifier construction based on extracted features. In conventional machine learning methods Klöppel et al. (2008); Liu et al. (2016); Liu et al. (2013); Sørensen et al. (2016); Wang et al. (2007); Zhang et al. (2011); Zhu et al. (2017), these three steps are performed separately, which may potentially result in suboptimal performance, due to the lack of constraints between each step. In another, most of them are confronted with the problem of large burden in data preprocessing introduced by brain tissue segmentation and non-linear image registration.

Recent studies Cui and Liu (2019); Farooq et al. (2017); Lian et al. (2020a,b); Liu et al. (2018); Liu et al. (2019, 2020); Qiu et al. (2020) have demonstrated the successful application of deep learning based approaches for AD diagnosis using sMRI, which combined the feature extraction with classifier construction in a task-oriented manner and can be further divided into four categories in coordinate with the scale of the pre-defined ROIs Wen et al. (2020), including: 1) 2D slice-level Farooq et al. (2017), 2) ROI-based Cui and Liu (2019), 3) 3D patch-level Lian et al. (2020b); Liu et al. (2018); Liu et al. (2019, 2020) and 4) 3D subject-level Lian et al. (2020a); Qiu et al. (2020). Specifically, slice-based methods extracted several 2D slices from a single 3D subject, augmenting the training samples but often neglecting the sophisticated 3D spatial information. In the patch-level or ROI-based methods, the pre-defined regions, such as bilateral hippocampi or AD-related patches, were extracted and input onto the constructed model, which achieved promising results in capturing local AD-related atrophy while sometimes lost the global information without consideration to the area out of pre-defined ROIs.

As the isolated pre-selection of potentially informative brain locations might be suboptimal in the patch-level or ROI-based methods, a few of deep learning methods Lian et al. (2020a,b); Qiu et al. (2020) have been proposed to perform discriminative atrophy localization for Alzheimer's disease diagnosis. For example, in Lian et al. (2020b), a hierarchical fully convolutional network (H-FCN) was proposed to automatically identify discriminative local patches and regions by network pruning.

In a more recent work, Lian et al. Lian et al. (2020a) developed a hybrid network to jointly learn and fuse multi-level sMRI features for classifier construction after weakly supervised localization of disease-related discriminative regions across all training samples. However, these methods paid great attention to the discriminative regions (i.e., network pruning of uninformative patches/regions Lian et al. (2020b) and generation of disease attention maps Lian et al. (2020a)) but neglected the potentiality of indiscriminative or uninformative regions and/or global information as the learn of which may potentially result in the problem of underfitting or overfitting. And recent studies with regularization, especially adversarial training Goodfellow et al. (2015); Miyato et al. (2017); Szegedy et al. (2014); Wang et al. (2021), can to some extent solve these problem. Accounting for regularization, adversarial training originally proposed in Szegedy et al. (2014) showed the effectiveness in reducing the test error by againsting adversarial perturbation and Goofellow et al Goodfellow et al. (2015) introduced an approximation of adversarial perturbation without expensive inner loop as in Szegedy et al. (2014). Further work in Miyato et al. (2017) defined the virtual adversarial direction without label information for supervised and semi-supervised learning tasks, which was successfully applied in Wang et al. (2021) for semi-supervised medical image classification.

To this end, we propose a Multi-task Multi-level Feature Adversarial Network (M²FAN) for joint Alzheimer's disease diagnosis and brain atrophy localization using baseline sMRI. Specifically, the linear-aligned T1 MR images were first processed by the lightweight CNN backbone to capture the shared intermediate feature representations, which were then branched into a global subnet for preliminary dementia diagnosis and a multi instance learning (MIL) network for brain atrophy localization in a weakly supervised manner, respectively. Based on this structure, the discriminative multi-level (i.e., local and global) features can be learned simultaneously. As the global discriminative information captured by the global subnet might be unstable for disease diagnosis due to subtle structural change in the cerebrum and inspired by adversarial training, we further designed a module of multi-level feature adversarial learning (MFAL) to make the global features robust against the adversarial perturbations synthesized by the local discriminative features to improve the generalization capability of AD diagnosis. We have evaluated our proposed methods on three public Alzheimer's disease datasets (i.e., ADNI-1, ADNI-2 and AIBL), and compared with the state-of-the-art approaches, our proposed method consistently achieves competitive performance in both tasks of AD diagnosis and MCI conversion prediction.

The remainder of this article is organized as follows. Firstly, we introduce the studied datasets (i.e., ADNI-1, ADNI-2 and AIBL), data preprocessing, as well as the proposed method in Section 2. In Section 3, we present the experimental settings, competing methods, and experimental results. We finally briefly discuss the limitations of our current work and conclude this article in Section 4 and Section 5, respectively.

Table 1. Demographic information of the subjects included three public datasets (i.e., the baseline ADNI-1, ADNI-2 and AIBL). The gender is reported as male/female. The age, education years, and mini-mental state examination (MMSE) values are reported as Mean ± Standard deviation.

Datasets	Category	Gender	Age	Education	MMSE
ADNI-1	AD	99/89	75.3 ± 7.5	14.7 ± 3.1	23.3 ± 2.0
	pMCI	106/69	74.6 ± 6.9	15.7 ± 2.9	26.6 ± 1.7
	sMCI	110/55	74.6 ± 7.5	15.8 ± 3.0	27.5 ± 1.7
	NC	119/110	75.9 ± 5.0	16.1 ± 2.9	29.1 ± 1.0
ADNI-2	AD	86/62	74.5 ± 8.1	15.7 ± 2.7	23.1 ± 2.1
	pMCI	56/38	72.5 ± 7.1	16.2 ± 2.5	27.2 ± 1.8
	sMCI	153/127	71.1 ± 7.4	16.4 ± 2.6	28.3 ± 1.6
	NC	89/96	73.3 ± 6.2	16.5 ± 2.6	29.0 ± 1.3
AIBL	AD	27/45	73.4 ± 8.0	-	20.2 ± 5.6
	NC	156/203	72.3 ± 6.4	-	28.7 ± 1.2

2. Materials and Methods

2.1. Datasets and Preprocessing

Three public datasets with 1895 baseline sMRI scans were studied in this work, including: 1) AD Neuroimaging Initiative-1 (ADNI-1); 2) ADNI-2 and 3) Australian Imaging, and Biomarker and Lifestyle Flagship Study of Aging (AIBL). The demographic information of the subjects included in this work are summarized in Table 1. These subjects were divided into three categories (i.e., normal control: NC, mild cognitive impairment: MCI, and Alzheimer's disease: AD) in terms of the standard clinical criteria, including mini-mental state examination (MMSE) scores and/or clinical dementia rating. Subjects in the MCI group were classified as progressive MCI (pMCI) if the subjects converted to AD during a 3 years follow-up visit, otherwise they were classified as stable MCI (sMCI). To summarize, the baseline ADNI-1 dataset contains 229 NC, 165 sMCI, 175 pMCI, and 188 AD subjects and the ADNI2 contains 185 NC, 280 sMCI, 94 pMCI, and 148 AD subjects. To another, the baseline AIBL dataset contains 1.5T/3.0T T1-weighted sMRI scans with ADNI-compliant series acquired from totally 431 subjects, where 72 subjects were diagnosed as AD and the remaining 359 subjects are NCs.

For each structural MR image corresponding to a specific subject, we first perform intensity inhomogeneity correction using N3 algorithm Sled et al. (1998), skull stripping and cerebellum removal Dale et al. (1999) via Freesurfer ‡. Next, we linearly align each image to a common Colin27 template Holmes et al. (1996) using FLIRT method

‡ http://www.freesurfer.net/

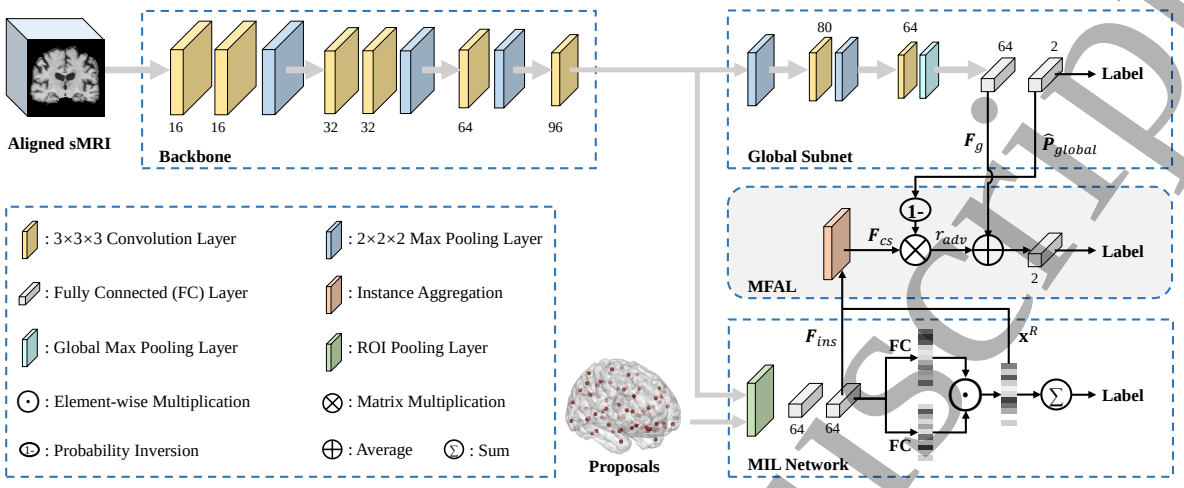

Figure 1. The architecture of our proposed Multi-task Multi-level Feature Adversarial Network (M²FAN) for jointly Alzheimer's disease diagnosis and atrophy localization, which mainly includes four parts: the backbone, global subnet, multi instance learning (MIL) network (local subnet for atrophy localization) and the module of multi-level feature adversarial learning (MFAL). The $\mathbf{F}_g$, $\mathbf{F}_{cs}$, $\mathbf{P}_{global}$,r_{adv}, $\mathbf{F}_{ins}$ and $\mathbf{x}^R$ indicates the global features, class-specific features, global prediction, adversarial perturbation, instance features and corresponding scores of all proposals, respectively.

Jenkinson et al. (2002); Jenkinson and Smith (2001) in the FSL package § to remove global linear difference and also to resample all images for having an identical spatial resolution (i.e., $1 \times 1 \times 1$ mm^3). Finally, all linear-aligned sMRIs were cropped to have the identical size of $158 \times 195 \times 153$ to reduce computational burden without any loss of cerebrum structures.

2.2. Methods

As introduced in Cipolla et al. (2018) that learning multiple tasks improves the model's representation and individual task performance, we propose Multi-task Multi-level Feature Adversarial Network (M²FAN) for joint Alzheimer's disease diagnosis and brain atrophy localization using baseline sMRI, which mainly includes four parts: the backbone, the global subnet, the multi instance learning (MIL) network as well as the module of multi-level feature adversarial learning (MFAL), as shown in Fig. 1. Specifically, the linear-aligned T1 MR images with the size of $158 \times 195 \times 153$ were first processed by the backbone to capture the shared intermediate feature representations, which were then branched into a global subnet for preliminary dementia diagnosis and a multi instance learning (MIL) network for brain atrophy localization in a weakly supervised manner, respectively. Based on this structure, the discriminative multi-level (i.e., local and global) features can be learned simultaneously. To further improve the generalization capability of the proposed methods, the module of multi-level feature adversarial learning (MFAL) that accounts for regularization was designed to make

§ https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT

global features robust against the adversarial perturbations for brain disease diagnosis.

2.2.1. ***Backbone*** The backbone applies a lightweight CNNs architecture to capture the shared feature representations from the whole brain sMRI. As shown in Fig. 1, in our current implementation, the backbone consisted of six convolutional layers and each convolution layer was followed by batch normalization (BN) and rectified linear unit (ReLu) activations. All convolutional layers have identical kernel size of $3 \times 3 \times 3$ and the number of channels is 16, 16, 32, 32, 64, 96, respectively. In addition, three $2 \times 2 \times 2$ max-pooling layers were adopted to down-sample the feature maps following the second, fourth, fifth convolutional layer, respectively. Therefore, given an input whole brain MR image with the size of $W \times H \times D$, the backbone yielded a set of 96 feature maps with the size of $(W/8) \times (H/8) \times (D/8)$. It is worth noting that the backbone was designed to be lightweight to limit the number of learning parameters, especially considering a relatively small number of the training samples.

2.2.2. ***Global Subnet*** The global subnet attempts to capture the discriminative information from the whole brain sMR images for the diagnostic task with the input of the intermediate feature maps extracted from the backbone, which consists of three pooling layer, two convolutional layer and two fully connected (FC) layers as shown in Fig. 1. The former two pooling layers were max pooling layer with pool size of $2 \times 2 \times 2$ and the last pooling layer was global max pooling layer. The convolutional layers followed by BN and ReLu activation were placed between any two pooling layers with kernel size of $3 \times 3 \times 3$ and the number of channels is 80 and 64, respectively. After that, a FC layer with units of 64 and ReLu activation was utilized to extract global discriminate features $\mathbf{F}_g$, which were then fed into another FC layer with softmax activation to produce global prediction probability $\widehat{P}_{global}$.

2.2.3. ***MIL Network*** In the datasets mentioned in Sec. 2.1, we only have subject-level labels that indicates whether patients suffered from AD or whether patients with MCI deteriorated into AD within 36 months after baseline visit. To identify individual-level disease-related regions and meanwhile extract discriminative local information, we construct the MIL network Bilen and Vedaldi (2016) to obtain the instance classifier in a weak supervision manner.

Given an input baseline linear-aligned sMRI with only subject-level label, we regarded the ROI centroid of AAL atlas with fixed patch size as instance proposals. Different from Lian et al. (2020a) that developed an independent H-FCN to extract local feature of patches/regions, the instance features corresponding to each proposal were first extracted from the output of the lightweight CNN backbone by ROI pooling layer, upon which the intermediate feature representations were reused, Thus each instance proposals have identical size of $H' \times W' \times D'$ (e.g., $5 \times 5 \times 5$) in the embedding space (also identical size in original space) and the ROI pooling layer applies max pooling strategy to convert the features inside any valid proposals into one-dimensional

vectors $\in \mathbb{R}^{|R|\times 64}$, $|R|$ denotes the number of instance proposals and is 90 in this paper. Subsequently, two fully connected (FC) layers with identical unit of 64 were adopted to obtain instance features $\mathbf{F}_{ins} \in \mathbb{R}^{|R|\times 64}$, which were then branched into two FC streams (classification stream f_{cls} and detection stream f_{det})with activation through different dimensions to produce two matrices $\mathbf{x}^{cls}$, $\mathbf{x}^{det} \in \mathbb{R}^{|R|\times C}$,respectively, C denotes the number of objects (atrophy) ($C = 1$ in this paper as AD/pMCI-related atrophy only occurs in AD/pMCI subjects ideally):

$$\mathbf{x}^{cls} = \sigma(f_{cls}(\mathbf{F}_{ins})), \mathbf{x}^{det} = \sigma(f_{det}(\mathbf{F}_{ins})) \tag{1}$$

where $\mathbf{x}^{cls}$ denotes the classification of the instances, $\mathbf{x}^{det}$ denotes the localization of the instances. The $\sigma(\cdot)$ is the activation function, which was Sigmoid function for classification stream and Softmax function for detection stream in this paper, respectively. After that, the scores of all instance proposals $\mathbf{x}^R$ are generated by element-wise product:

$$\mathbf{x}^R = \mathbf{x}^{cls} \odot \mathbf{x}^{det} \tag{2}$$

Finally, the c-th class prediction score at the subject-level can be obtained by summing up the scores over all proposals: $\widehat{P}_{mil} = \sum_{r=1}^{|R|} \mathbf{x}^r$.

2.2.4. ***Multi-level Feature Adversarial Learning (MFAL)*** Since the global discriminative information captured by the global subnet might be unstable for disease diagnosis due to subtle structural abnormalities in the cerebrum, in the literature, such as in Lian et al. (2020a), multi-level features (i.e., local and global) are fused to improve the diagnostic performance by fully utilizing complementary information after transforming the local features of multiple patches into local-to-global representations by use of max-pooling, mean-pooling and concatenation. While simple transformations may suppress the discriminability of the local-to-global representations as it is perturbed by label-unrelated or indiscriminative features of multiple patches detected by the MIL network, instead of pruning them, we designed a module of multi-level feature adversarial learning (MFAL) to make full use of uninformative patches and enable multi-level features fusion as adversarial attack, so as to render the global information robust against the adversarial perturbations synthesized by local features in a gradient-free adversarial training manner, leading to diagnostic performance improvement.

To construct the module of MFAL, the critical step is to generate the adversarial perturbations by utilizing the local features of multiple instances. Specifically, the unpruned local features of each instance and the corresponding confidence score of different severities (each bag contains positive and negative instances) enables the learning of discriminative abnormality/normality-related features that we called class-specific features $\mathbf{F}_{cs}$ by the instance aggregation layer. With input of instance features $\mathbf{F}_{ins}$ of all proposals and the corresponding scores $\mathbf{x}^R$, the instance aggregation layer averages k instance features corresponding to top-ranked k instance scores as abnormality-related (AD/pMCI-related) and averages k instance features corresponding to bottom-ranked k instance scores as normality-related (NC/sMCI-related) features due

to C=1 in this article to generate the class-specific features $\mathbf{F}_{cs}$, which were formulated as:

$$\mathbf{F}_1 = \frac{1}{k}\mathbf{F}_{ins} \otimes \mathbb{I}(Top(\mathbf{x}^R, k)) \tag{3}$$

$$\mathbf{F}_0 = \frac{1}{k}\mathbf{F}_{ins} \otimes \mathbb{I}(Bottom(\mathbf{x}^R, k)) \tag{4}$$

$$\mathbf{F}_{cs} = [\mathbf{F}_0, \mathbf{F}_1] \tag{5}$$

Where $\mathbb{I}$ is the indicator and k is the hyper-parameter to decide how many instances are utilized to represent the corresponding class. To this end, we can obtain each kind of the feature representations corresponding to the specific class and query the global prediction $\widehat{P}_{global}$ learned from the global subnet in a black-box manner. Even though the local-to-global features can be learned by the operation of matrix multiplication between $\mathbf{F}_{cs}$ and $\widehat{P}_{global}$ for multi-level feature learning, to make the global features robust for AD diagnosis, we firstly apply the simple probability inversion operation on the global prediction before the operation of matrix multiplication, making the local-to-global representations to be an adversary (adversarial perturbations). The inversed probability $\overline{\widehat{P}_{global}}$ and adversarial perturbations r_{adv} were formulated as:

$$\overline{\widehat{P}_{global}} = \frac{1}{Z}(1 - \widehat{P}_{global}) \tag{6}$$

$$r_{adv} = \mathbf{F}_{cs} \otimes \overline{\widehat{P}_{global}} \tag{7}$$

Where Z is partition function and defined as $\sum_{c=1}^{C+1} \overline{\widehat{P}_{global,c}}$. In this way, the adversarial perturbations r_{adv} were prone to the normality-related (NC/sMCI-related) feature representations when the inputs were predicted as the abnormality-related (AD/pMCI-related) by the global subnet and vice versa.

Therefore, in the black-box adversarial training settings and with the input of the global feature $\mathbf{F}_g$ smoothed by the r_{adv}, we construct a surrogate classifier by utilizing a FC layer with softmax activation as in this settings we can't access to the learning parameters of the global classifier (the last FC layer in global subnet). It is worth noting that the proposed module of MFAL is of time-saving and low computation since the adversarial perturbations were computed without backpropagation and it is no need to repeatedly compute the intermediate feature representations as in Miyato et al. (2017); Rasmus et al. (2015) the r_{adv} was applied on the latent features.

2.2.5. Implementation Details We designed a hybrid cross entropy loss to effectively train the proposed M²FAN in one stage, which was formulated as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{mil} + \beta\mathcal{L}_{global} + \gamma\mathcal{L}_{MFAL} \tag{8}$$

where the $\mathcal{L}_{mil}$, $\mathcal{L}_{global}$, $\mathcal{L}_{MFAL}$ is the cross entropy loss for the global subnet, MIL network and MFAL, respectively. And the parameters α, β and γ was empirically set to 2, 1 and 1, respectively.

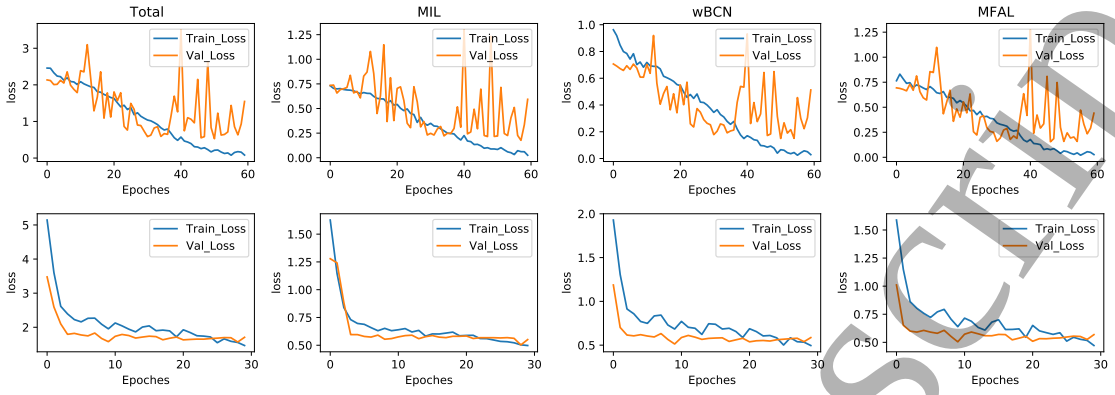

Figure 2. Training history of the proposed of M²FAN trained on ADNI-1 dataset in the task of AD diagnosis (the first row) and MCI conversion prediction (the second row), respectively

To validate the generalization capability of the proposed model, two different tasks were studied in this article, including AD diagnosis (i.e., AD *vs.* NC classification) and MCI conversion prediction (i.e., pMCI *vs.* sMCI classification). We randomly selected 10% of one complete dataset (e.g., ADNI-1) as validation set and the rest of which was used to train the proposed M²FAN, which was then evaluated by the other independent dataset (e.g., ADNI-2). In addition, we trained the M²FAN with parameters randomly initialized by 'glorot normal' Glorot and Bengio (2010) for 60 epochs by setting batch size as 2 and applying Adam optimizer with learning rate of 0.0001 for AD diagnosis using Python based on Keras API of Tensorflow and a single GPU (i.e., NVIDIA GeForce RTX 2080Ti 11GB). For MCI conversion prediction, as it's more challenging since subtle brain change of patients with MCI caused by dementia, we transferred the network parameters learned from AD diagnosis as training initialization of the network for pMCI *vs.* sMCI classification, which was then trained for 30 epochs by applying Adam optimizer with learning rate of 0.00001.

3. Experiments

In this section, experiments were implemented to validate the effectiveness and robustness of the proposed M²FAN in two different tasks (AD diagnosis and MCI conversion prediction), in which we trained the classification model on ADNI-1 and evaluate them on the other two independent datasets, including ADNI-2 and AIBL. The classification performances were evaluated by four metrics, including classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under receiver operating characteristic curve (AUC).

Table 2. Results of AD *vs.* NC classification on ADNI-2 and AIBL, respectively, obtained by the models trained on ADNI-1.

Methods	ADNI-2				AIBL			
	ACC	**SEN**	**SPEC**	**AUC**	**ACC**	**SEN**	**SPEC**	**AUC**
ROI-based	0.826	0.804	0.843	0.873	0.843	0.833	0.844	0.872
VBM-based	0.829	0.811	0.843	0.892	0.831	0.778	0.841	0.886
MIL	0.859	0.797	0.908	0.906	0.852	0.861	0.850	0.912
wBCN	0.874	0.818	0.919	0.926	0.865	0.847	0.869	0.921
DMIL	0.889	**0.905**	0.876	0.952	0.875	0.792	0.891	0.923
VAT	0.889	0.892	0.886	0.950	0.901	0.872	0.906	0.929
M²FAN@MIL	0.883	0.878	0.886	0.942	0.845	0.875	0.838	0.926
M²FAN@wBCN	**0.916**	0.892	0.935	**0.966**	0.907	**0.903**	0.908	0.949
M²FAN@MFAL	0.913	0.872	**0.946**	0.965	**0.923**	0.889	**0.930**	**0.950**

3.1. AD diagnosis and MCI Conversion Prediction

In this group of experiments, we compare our M²FAN with two conventional machine learning methods, including 1) region-of-interest-based (ROI-based) method Zhang et al. (2011) and 2) voxel-based morphometry (VBM-based) method Baron et al. (2001), and four deep learning methods, including 3) multi instance learning (MIL) network Bilen and Vedaldi (2016), 4) whole brain convolutional network (wBCN), 5) deep multi instance learning (DMIL) Liu et al. (2018) and 6) virtual adversarial training (VAT) Miyato et al. (2017), in the task of AD diagnosis (AD *vs.* NC) and MCI conversion prediction (pMCI *vs.* sMCI) using the models trained on ADNI-1, in which 10% subjects were stratified sampled for validation. The quantified classification performances in terms of four different metrics (i.e., ACC, SEN, SPE, and AUC) evaluated on ADNI-2 and AIBL are summarized in Table 2 and 3, respectively.

1) **ROI-based**. Following previous studies Zhang et al. (2011), the whole brain sMRI data were partitioned into multiple regions to extract region-level features for SVM-based classification. More specifically, using aBEAT software ‖, each sMRI was first segmented into three tissue types, i.e., gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), and then AAL atlas Tzourio-Mazoyer et al. (2002) with 90 pre-defined ROIs in the cerebrum was aligned to each subject to extract regional features for linear SVM classification.

2) **VBM-based**. In VBM-based method Baron et al. (2001), voxel-level handcrafted features were quantified from the whole brain sMRI to construct SVM classifier. Firstly, each sMRI data was wrapped into Colin27 template to extract local GM density as features and then a statistical group comparison based on t-test was

‖ htttp://www.nitrc.org/projects/abeat

Table 3. Results of pMCI *vs.* sMCI classification on ADNI-2, respectively, obtained by the models trained on ADNI-1.

Methods	ACC	SEN	SPEC	AUC
ROI-based	0.674	0.553	0.714	0.651
VBM-based	0.644	0.670	0.636	0.685
MIL	0.684	0.713	0.675	0.725
wBCN	0.746	0.702	0.761	0.761
DMIL	0.757	0.649	0.793	0.777
VAT	0.762	0.660	0.796	0.773
M²FAN@MIL	0.786	0.702	0.814	0.796
M²FAN@wBCN	0.797	0.734	0.818	0.810
M²FAN@MFAL	**0.802**	**0.745**	**0.821**	**0.814**

performed to reduce the dimensionality of which. Finally, linear SVM classifiers were constructed for AD diagnosis.

3) **MIL**. As an object detection method, MIL network Bilen and Vedaldi (2016) also has the capability of classification which extracted local information by sharing learning parameters for each proposal. Specifically, the spatially normalized MR images were first branched into several convolutional layers to extract discriminate feature maps, which were then fed into ROI pooling layer and two FC layers to obtain corresponding features with respect to the given proposals. Finally, two FC layers with activation through different dimension were adopted to generate the final prediction probability via a weighted sum pooling strategy.

4) **wBCN**. The wBCN method constructed a CNN-based method to extract global features for brain disease diagnosis with input of whole brain sMRI. Without any difference, the sMRI was firstly spatially normalized onto Colin27 template and then cropped to have identical size of 158 × 195 × 153, which was branched into a series of convolutional layers and FC layers for brain disease diagnosis. The MIL and wBCN was train on 60 epochs using Adam optimizer with learning rate of 0.0001.

5) **DMIL**. In DMIL method Liu et al. (2018), Liu et al. developed a CNN-based multi instance learning model to extract local-to-global representations for brain disease diagnosis. We firstly extracted 30 patches with size of 45 × 45 × 45 and each of them was fed into an independent CNN to yield a set of local features, which were then concatenated and fused by FC layers for AD diagnosis and MCI conversion prediction. the DMIL was train using Adam optimizer with learning rate of 0.001.

6) **VAT**. In VAT Miyato et al. (2017), we applied the wBCN as the classification model and calculated the virtual adversarial loss in line with Miyato et al. (2017), which was train on 80 epochs using Adam optimizer in the task of AD classification and MCI conversion prediction.

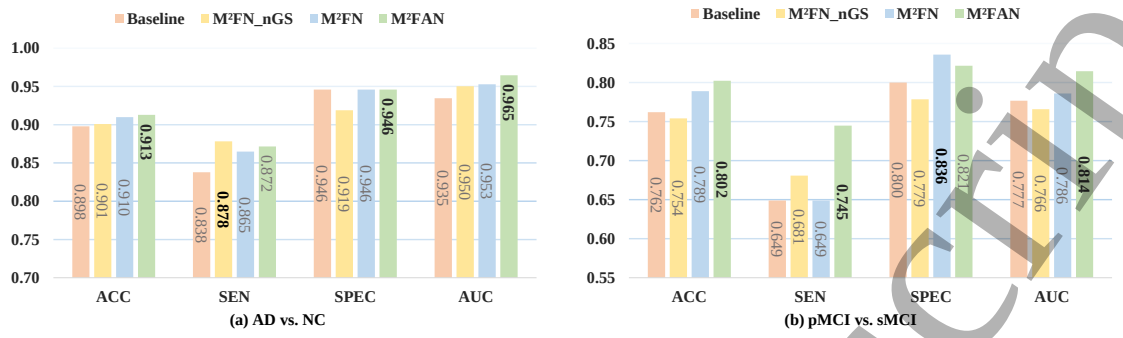

Figure 3. Results of (a) AD diagnosis (AD *vs.* NC classification) and (b) MCI conversion prediction (pMCI *vs.* sMCI classification) on the ADNI-2 dataset, obtained by the nMFAL, nGS, MFL and M^{2}FAN trained on ADNI-1 dataset, respectively

From Table 2 and 3, several observations can be summarized as followed, in which the classification performances of the global subnet, MIL network as well as the module of MFAL was denoted as M^{2}FAN@MIL, M^{2}FAN@wBCN and M^{2}FAN@MFAL, respectively. *First*, compared with conventional machine learning methods (i.e., ROI-based and VBM-based), the deep-learning approaches (i.e., MIL, wBCN, DMIL and our M^{2}FAN) largely improved the diagnostic performance on both ADNI-2 and AIBL datasets, which demonstrates the significance in integrating feature extraction and classifier construction for both AD diagnosis MCI conversion prediction. *Second*, compared with wBCN which outperformed MIL on both ADNI-2 and AIBL datasets, DMIL that adopted local-to-global representation learning approach for brain disease diagnosis yields better classification performance, which implies the great potentiality of local information regarding subtle brain changes. *Third*, compared with wBCN and MIL, M^{2}FAN@MIL and M^{2}FAN@wBCN generally yields better performance on both ADNI-2 and AIBL datasets, especially in AUC. For example, the AUC value of M^{2}FAN@MIL improved from 0.906 to 0.942 on ADNI-2 dataset while improved from 0.912 to 0.926 on AIBL dataset and the AUC value of M^{2}FAN@wBCN improved from 0.926 to 0.966 on ADNI-2 dataset while improved from 0.921 to 0.949 on AIBL dataset. And in another, compared with M^{2}FAN@wBCN, the M^{2}FAN@MFAL also yields competitive performance in both tasks of AD diagnosis and MCI conversion prediction on both two datasets, especially on the diagnostic performance of M^{2}FAN@MFAL also consistently surpassed in all four metrics (i.e., ACC, SEN, SPEC and AUC) on MCI conversion prediction, for example, the AUC values were improved from 0.810 to 0.814, demonstrating that multi-task learning can improve the individual's performance and to another multi-level (i.e., local and global) feature adversarial learning can improve the generalization and robustness capability of the classification model.

3.2. Effectiveness of Multi-level Feature Adversarial Learning

As introduced in Sec. 2.2.4, MFAL collaborates the global and local representation learning and synthesizes adversary using local knowledge to attack global information

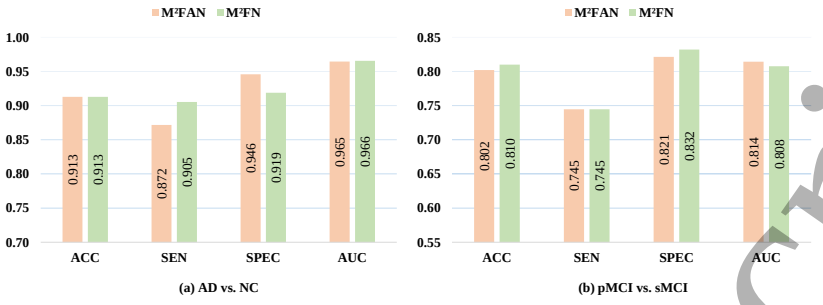

Figure 4. The results of proposed M²FAN in the attack evasion stage by transfering the learned parameters from M²FAN to M²FN without finetuning on AD diagnosis (a) and MCI conversion prediction (b), respectively.

so that improve the robustness and margin of multi-level information. To evaluate the effectiveness of the MFAL, several ablated models were designed for comparison, including 1) the **Baseline** model which removes the module of MFAL from the proposed model M²FAN, 2) **M²FN_nGS** in which removes the global supervision on global subnet and the weight of class-specific features was learned from unsupervised learning and 3) **M²FN** in which the invert operation of global prediction probability in the proposed M²FAN is removed so that generated local-to-global representations highly related to global information for multi-level feature learning. Specifically, we trained the Baseline, M²FN_nGS, M²FN and the proposed M²FAN on ADNI-1 for the task of AD diagnosis as well as MCI conversion prediction and the classification performances were evaluated on ADNI-2 in terms of ACC, SEN, SPEC and AUC as summarized in Fig. 3 and Fig. 4.

From Fig. 3 and Fig. 4, we can at least have the following observations. *First*, compared with the Baseline model in Fig. 3, the M²FN_nGS model achieves better performance in the task of the AD diagnosis while unfits in the task of MCI conversion prediction, for example, the AUC value of the M²FN_nGS improved from 0.935 to 0.950 in AD *vs.* NC classification while reduced from 0.777 to 0.766 in pMCI *vs.*sMCI classification. But the M²FN model outperforms these two methods in both task of AD diagnosis and MCI conversion prediction, which indicates that the global supervised learning is of great importance in brain disease diagnosis and can also enhance the generalization ability of deep learning approach. *Second*, in Fig. 3, compared with M²FN our proposed M²FAN method with MFAL yields better performance on both AD diagnosis and MCI conversion prediction. For example, in the task of AD diagnosis, the ACC and AUC of M²FAN improved from 0.910 to 0.913 and from 0.953 to 0.965, respectively. In the task of MCI conversion prediction, the ACC and AUC of M²FAN improved from 0.789 to 0.802 and from 0.786 to 0.814, respectively. It implies that the stimulated attack generated from local class-specific features $\mathbf{F}_{cs}$ can improve the discriminative ability of learned features. *Third*, to further evaluate the effectiveness of the proposed M²FAN using multi-level feature adversarial learning, in Fig. 4, we can observe that the proposed method achieves competitive performance in the attack

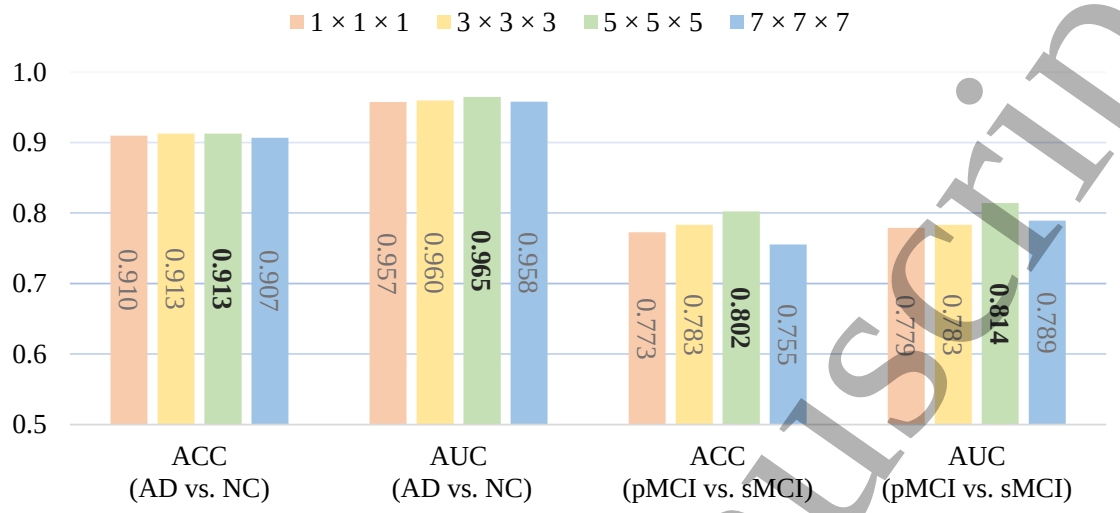

Figure 5. The results of AD diagnosis and MCI conversion prediction in terms of ACC and AUC obtained by our proposed method with different patch size of the proposals in ROI Pooling layer (i.e., $1 \times 1 \times 1$, $3 \times 3 \times 3$, $5 \times 5 \times 5$, $7 \times 7 \times 7$).

evasion stage by transfering the learned parameters from M²FAN to M²FN without finetuning on both AD diagnosis and MCI conversion prediction, for example, the ACC value in the task of MCI conversion prediction improved from 0.802 to 0.810, suggesting the anti-interference ability (not sensitive to the local features) and rationality of the proposed method.

3.3. Influence of Patch Size in ROI Pooling

In the above-mentioned experiments, we adopt a fixed patch size (i.e. , $5 \times 5 \times 5$) located at the ROI centroid of AAL atlas for our proposed M²FAN method. We now investigate the influence of the patch size on the performance of M²FAN by varying the patch size in the embedding space and testing all the values in the set $\{1 \times 1 \times 1, 3 \times 3 \times 3, 5 \times 5 \times 5, 7 \times 7 \times 7\}$ in terms of ACC and AUC on both AD diagnosis and MCI conversion prediction and the quantified results of the obtained model trained on ADNI1 and tested on ADNI2 were summarized in Fig. 5.

From Fig. 5, we can see that the best results are obtained by M²FAN using the patch size of $5 \times 5 \times 5$ in the embedding space on both AD diagnosis and MCI conversion prediction. Specifically, the proposed M²FAN is not very sensitive to the patch size setting in ROI pooling layer on AD *vs.* NC classification, for example, the AUC value of which improved from 0.957 to 0.965 when changing the patch size from $1 \times 1 \times 1$ to $7 \times 7 \times 7$, potentially due to the deliberate consideration of global information. While the value of the AUC improved from 0.779 to 0.814 in the task of MCI conversion prediction when varying the patch size from $1 \times 1 \times 1$ to $5 \times 5 \times 5$ and the value of ACC greatly decreased when the patch size was set to $7 \times 7 \times 7$, indicating that relatively small or large local patches are not capable of capturing enough structural

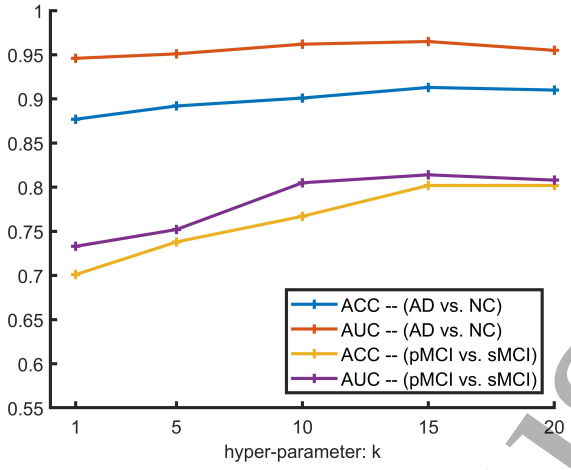

Figure 6. The results of AD diagnosis and MCI conversion prediction in terms of ACC and AUC obtained by our proposed method with different hyperparameter k in instance aggregation (i.e., k = 1, 5, ..., 20).

information from the brain for pMCI *vs.* sMCI classification, especially when the patch size was set to $7 \times 7 \times 7$ as the proposals are with too large overlapping, thus it is reasonable to choose a moderate patch size (i.e., $5 \times 5 \times 5$) in the embedding space for the proposed M^{2}FAN. It worth noting that considerable large patch size does not bring too much computational burden in the proposed M^{2}FAN, as the patch-level features were extracted in the embedding space for the downstream task.

3.4. Influence of Hyperparameter in Instance Aggregation

As introduced in Sec. 2.2.4, the module of instance aggregation was designed to generate the local class-specific features $\mathbf{F}_{cs}$ with the input of local instance features and corresponding scores, in which the hyperparameter k was utilized to determine how many instances represent the corresponding class. In this group of experiments, we investigate the influence of hyperparameter k in instance aggregation on the classification performances achieved by our M^{2}FAN method and orderly selected k from {1; 5; 10; 15; 20} in instance aggregation. The corresponding results quantified by ACC and AUC are summarized in Fig. 6 on both AD diagnosis and MCI conversion prediction.

From Fig. 6, we can observe that both the values of ACC and AUC are clearly increased when changing k from 1 to 15 and slightly decreased at $k = 20$. For example, on AD diagnosis, the ACC and AUC of the proposed method improved from 0.877 to 0.913 and from 0.946 to 0.965, respectively, when changing k from 1 to 15, while on MCI conversion prediction, the ACC and AUC improved from 0.701 to 0.802 and from 0.733 to 0.814, respectively. This implies that relatively large k which indicated more instances were selected to represent a specific category is appropriate, potentially because that 1) too few instances cannot comprehensively characterize information of specific class, and 2) too small k may make the model incline to a few particular instances, leading

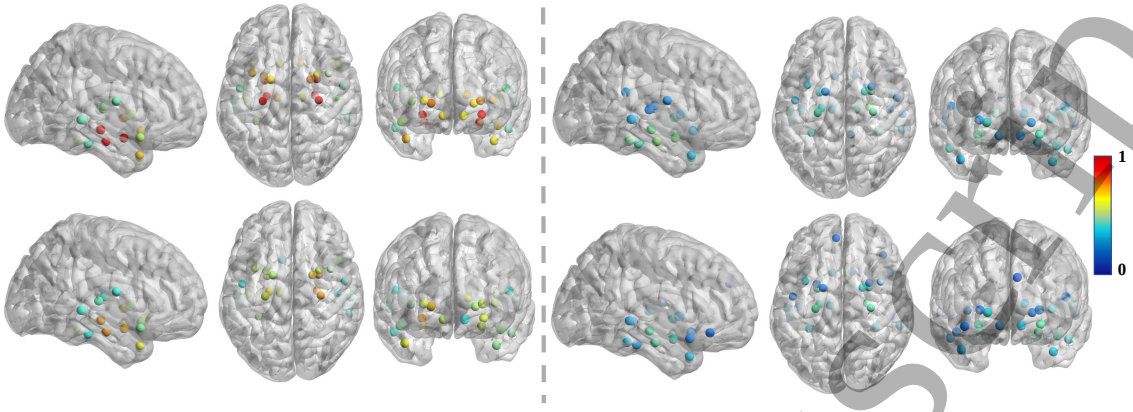

Figure 7. 30 instances with top-ranked statistical difference (obtained by true positive minus false positive) between groups (i.e., AD *vs.* NC, pMCI *vs.* sMCI) on the training set (ADNI-1) and testing set (ADNI-2) generated by instance classifier in the tasks of AD *vs.* NC classification (two rows on the left) and pMCI *vs.* sMCI classification (two rows on the right), respectively.

to local optima. In another, the values of ACC and AUC reduced slightly at $k = 20$ on both AD diagnosis and MCI conversion prediction, which implies that too large k in instance aggregation is inappropriate because that excessive instances may magnify the risk of learning specific-class features from instance of different category. Hence, it is reasonable to choose $k = 15$ in our implementations.

3.5. Automatic Brain Atrophy Localization

As introduced in Sec. 2.2.3, our proposed method can automatically identify discriminative brain regions from the whole-brain sMRIs by MIL network. In Fig. 7, we present the 30 instances with top-ranked statistical variance (obtained by true positive minus false positive) between groups (i.e., AD *vs.* NC, pMCI *vs.* sMCI) for the training set (ADNI-1) and testing set (ADNI-2) generated by instance classifier in the tasks of AD *vs.* NC classification and pMCI *vs.* sMCI classification, respectively. Additionally, in Fig. 8 and Fig. 9, we present individual-level brain atrophy localization with respect to top-ranked 30 instances in Fig. 7 through some examples on both AD diagnosis and MCI conversion prediction, which is shown in 2D projection from three different views.

Specifically, the two rows on the left of Fig. 7 presents the locations with respect to top-ranked 30 instances on the training set and testing set identified by the model trained for AD diagnosis, respectively, and the two rows on the right for MCI conversion prediction. From Fig. 7, we can have the following observations. *First*, our proposed method consistently emphasized the locations at temporal lobe and sub-cortical area, such as hippocampus, amygdala, fusiform gyrus, which has been verified in the previous study Baron et al. (2001); Lian et al. (2020b); Zhang et al. (2011); Zhang et al. (2016), accounting for the feasibility of the proposed method in automatic brain atrophy localization. *Second*, the instances identified by the proposed method trained for MCI

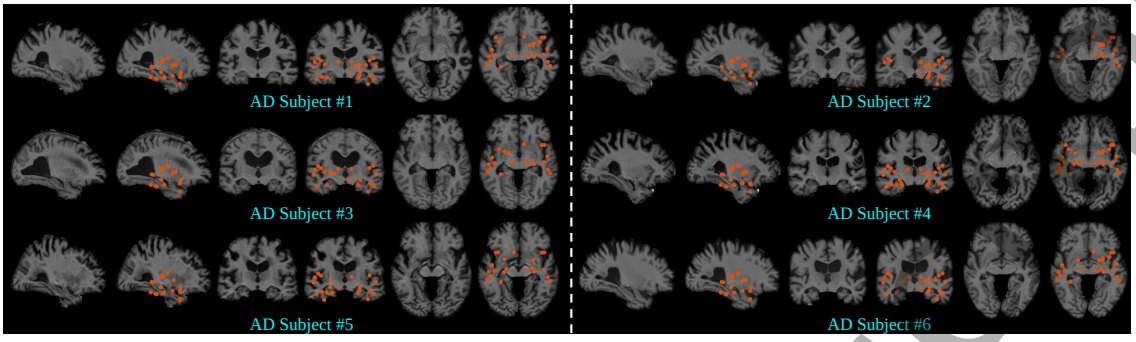

Figure 8. Illustration of individual-level brain atrophy localization obtained by instance classifier through six AD subjects from testing set (ADNI-2) with respect to 30 instances with top-ranked statistic difference on training set (ADNI-1) in the task of AD diagnosis.

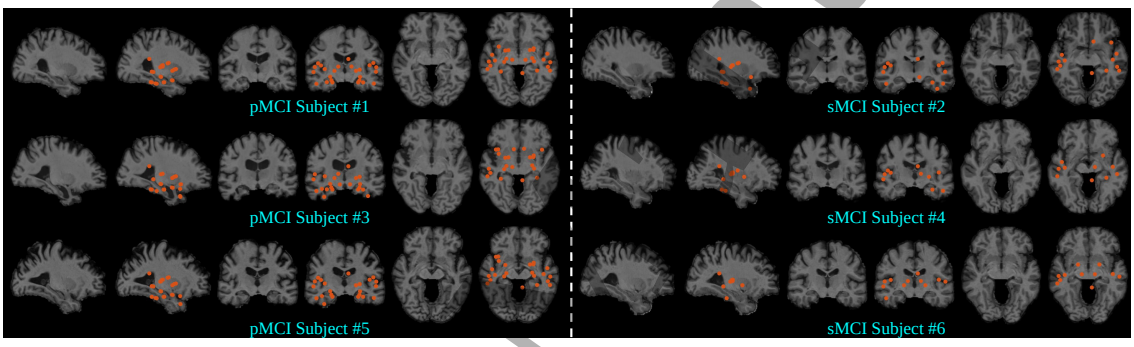

Figure 9. Illustration of individual-level brain atrophy localization obtained by instance classifier through three pMCI subjects and 3 sMCI subjects from testing set (ADNI-2) with respect to 30 instances with top-ranked statistic difference on training set (ADNI-1) in the task of MCI conversion prediction.

conversion prediction were consistent with those identified by the model trained for AD diagnosis, but on the other hand, compared with the instances detected in the task of AD diagnosis, the instances detected in the task of pMCI *vs.* sMCI classification were with relatively small statistical difference between pMCI and sMCI, which implies the robustness and effectiveness of the proposed method in localizing discriminative brain atrophy area using sMRI.

Moreover, to further validate the generalization ability of the proposed method, in Fig. 8 and Fig.9, we present the results of individual-level brain atrophy localization through some examples from testing set (ADNI-2) with respect to 30 instances obtained by statistics from training set (ADNI-1), specifically, 6 subjects with AD were shown in Fig.8, while 3 subjects with pMCI and 3 subjects with sMCI were shown in Fig. 9. From Fig. 8 and Fig. 9, we can observe that relatively more instances were predicted as positive by the instance classifier for subjects with AD and pMCI, while, on MCI conversion prediction, a few of instances were also predicted as positive, indicating that although brain atrophy localization is more challenging in the task of pMCI *vs.* sMCI classification since atrophy caused by dementia also occurred in patients with sMCI,

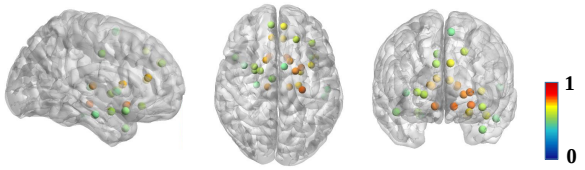

Figure 10. 30 instances with top-ranked statistical difference (obtained by true positive minus false positive) between groups (i.e., FTD *vs.* NC) on the training set generated by instance classifier.

the proposed method is capable of identifying brain atrophy of different severity. In another, instances with positive prediction located at sub-cortical area are consistently related to ventricular enlargement, which is in line with the previous studies Frisoni et al. (2010); Nestor et al. (2008), suggesting the rationality of the proposed method in automatically localizing brain atrophy.

4. Discussion

4.1. Evaluation on Frontotemporal Dementia with sMRI

Apart from AD, the proposed method is also applied for another sMRI-based brain disease (i.e., frontotemporal dementia (FTD)) diagnosis. FTD is the second most common cause of presenile early onset dementia Vieira et al. (2013), which can be classified into three types based on the distinct patterns of signs and symptoms, including behavioral variant FTD (bvFTD), semantic variant primary progressive aphasia (svPPA) as well as non-fluent/agrammatic variant primary progressive aphasia (nfvPPA). The proposed M^2FAN was trained on the public datasets, that is, frontotemporal lobar degeneration neuroimaging initiative (FTLDNI)¶, for FTD diagnosis. Briefly, the FTLDNI dataset consists of the T1-weighted sMR images acquired from 121 FTD and 123 NC subjects, in which we randomly selected 10% subjects (12 FTDs and 12 NCs) as the test set and used the remaining subjects as the training set.

In Fig. 10, we present the 30 instances with top-ranked statistical variance (obtained by true positive minus false positive) between groups (i.e., FTD *vs.* NC) for the training set generated by instance classifier. In Fig. 11, we present individual-level brain atrophy localization with respect to top-ranked 30 instances in Fig. 10 through three subjects with different types of FTD (i.e., bvFTD, svPPA and nfvPPA) on FTD diagnosis, which is shown in 2D projection from three orthogonal views. From Fig. 10 and Fig. 11, we can have the following observations. *First*, our proposed method consistently highlighted the locations of brain temporal lobe, frontal lobe and subcortical area, which has been validated to be related to FTD by previous studies Beyer et al. (2021); Manera et al. (2019), further demonstrating the feasibility of MIL network in brain atrophy localization. *Second*, in Fig. 11, different instances were

¶ http://4rtniftldni.ini.usc.edu/

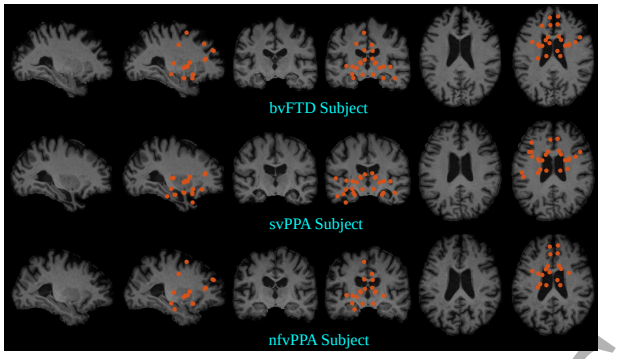

Figure 11. Illustration of individual-level brain atrophy localization obtained by instance classifier through three FTD subjects from testing set with respect to 30 instances with top-ranked statistic difference on training set in the task of FTD diagnosis.

Table 4. Results of AD *vs.* NC classification on ADNI-2 and AIBL, respectively, obtained by the models trained on ADNI-1.

Methods	AD *vs.* NC				pMCI *vs.* sMCI			
	ACC	SEN	SPEC	AUC	ACC	SEN	SPEC	AUC
Salvatore et al. (2015)	0.76	-	-	-	0.66	-	-	-
Tong et al. (2014)	0.90	0.86	0.93	-	0.72	0.69	0.74	-
Zhang et al. (2016)	0.831	0.805	0.851	0.828	-	-	-	-
Cao et al. (2017)	0.886	0.857	0.904	0.898	0.704	0.677	0.718	0.705
Liu et al. (2018)	0.911	0.881	0.935	0.959	0.769	0.421	0.824	0.776
Lian et al. (2020b)	0.903	0.824	**0.965**	0.951	0.809	0.526	0.854	0.781
Qiu et al. (2020)	0.834	0.767	0.889	-	-	-	-	-
Lian et al. (2020a)	**0.919**	0.887	0.945	0.965	**0.827**	0.579	**0.866**	0.793
Guan et al. (2021)	0.899	0.877	0.917	0.940	0.780	0.534	**0.866**	0.788
Proposed@wBCN	0.916	**0.892**	0.935	**0.966**	0.797	0.734	0.818	0.810
Proposed@MFAL	0.913	0.872	0.946	0.965	0.802	**0.745**	0.821	**0.814**

predicted as positive among FTD subjects of different types, for example, the positive instances of nfvPPA subject were mainly located at the temporal area, suggesting the potential interpretability of different symptoms by the proposed method.

4.2. Compare with previous work

For a comparison between our method and related studies on the performance of AD diagnosis and MCI conversion prediction using baseline sMRI, in Table 4, we briefly summarize four conventional learning based methods Cao et al. (2017); Salvatore et al.

(2015); Tong et al. (2014); Zhang et al. (2016) and five state-of-the-art deep learning based methods Guan et al. (2021); Lian et al. (2020a,b); Liu et al. (2018); Qiu et al. (2020).

It worth noting that the direct comparison between these methods is impossible due to the different utilization of the ADNI dataset (i.e., in subject enrollment, the definition of pMCI/sMCI and train/test set partition). However, by roughly comparing our proposed M^{2}FAN (the last two row of Table 4) with these state-of-the-art methods, we can observe that: *First*, the proposed methods surpassed the compared conventional methods (i.e., voxel-based Salvatore et al. (2015), ROI-based Cao et al. (2017), patch-based Tong et al. (2014); Zhang et al. (2016) methods) by evaluating on a much larger cohort of 1464 subjects from both ADNI1 and ADNI2, which impartially indicates the superiority of convolutional neural networks in extracting high-level features for AD classification. *Second*, though we only selected the ROI centroid of AAL atlas as the patch proposals without the identification of AD-related landmarks Lian et al. (2020b); Liu et al. (2018) or iterative network pruning Lian et al. (2020b) by assigning the subject-level label to the image patches/regions for local feature learning, our methods achieved competitive performance on both AD diagnosis and MCI conversion prediction and even was superior to some subject-level deep learning based methods (i.e., Guan et al. (2021); Qiu et al. (2020)). *Third*, compared with the most similar work Lian et al. (2020a) which developed an independent H-FCN for local representation learning after identifying the discriminative AD-related landmarks by another backbone FCN, our lightweight one-stage methods achieve comparable performance by extracting local features from embedding space of the backbone CNN for atrophy localization and multi-level feature learning. This implies that, our proposed method, to some extents, is more capable to identifying subtle structural changes in sMRI for AD diagnosis due to the use of multi-task learning (i.e., AD diagnosis and atrophy localization) and multi-level (i.e., local and global) feature adversarial learning.

4.3. Limitations and Future Work

Although our proposed method achieves promising results in automatically brain atrophy localization and dementia diagnosis, several limitations should be carefully considered in the future to further improve its performance and practical value.

First, to identify brain atrophy resulting from AD, our proposed method adopted a branch of MIL network which performs simultaneously instance selection and classification, but MIL network is sensitive to initialization as the optimization of which is non-convex. Hence, it should be a promising direction to further improve the diagnostic performance and generalization capability of the proposed method by alleviating the non-convexity problem. For example, in Wan et al. (2019), a continuation optimization method was introduced into MIL by partitioning instances into spatially related and class related subsets with a series of smoothed loss functions defined within them. *Second*, in our implementations, the instance proposals with identical fixed size

were located at the centroid of ROIs with respect to AAL atlas and automatically identified by the basic MIL network, but the structural changes caused by dementia vary across different instances and not all of the pre-defined locations were significantly correlated with brain atrophy so that aggravated instance imbalance. Therefore, it is reasonable to design a sub-network to proactively localize anatomical changes at multiple scales in a data-driven manner. *Third*, previous studies Jie et al. (2018); Tong et al. (2014) have shown that topological information is conducive to brain disease classification. For example, brain connectivity network analysis was applied in Jie et al. (2018) using functional MRI for disease diagnosis. In our proposed method, the instance of different severity can be identified by the instance classifier, in which the high-level discriminative features of them were extracted, therefore it is potential to integrate topological analysis based on which in our future work.

5. Conclusion

In this work, we proposed a Multi-task Multi-level Feature Adversarial Network (M^2FAN) to simultaneously localize brain atrophy and perform disease diagnosis using sMRI, in which a module of multi-level feature adversarial learning was designed for global features to confront the attack synthetized by local/instance features so as to render it robust to perform the Alzheimer's disease diagnosis. The effectiveness of our proposed method has been evaluated on three public datasets (i.e., ADNI-1, ADNI-2 and AIBL) consisting of 1895 subjects. The experimental results have demonstrated the competitive performance of our method compared with several state-of-the-art methods in both the tasks of AD diagnosis and MCI conversion prediction.

6. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61971213 and Grant 61671230, in part by the Basic and Applied Basic Research Foundation of Guangdong Province under Grant 2019A1515010417, and in part by the Guangdong Provincial Key Laboratory of Medical Image Processing under Grant No.2020B1212060039. The authors have no relevant conflicts of interest to disclose.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) database (https://aibl.csiro.au/adni/index.html). As such, the investigators within the ADNI and AIBL contributed to the design and implementation and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List .pdf and the AIBL researchers are listed at www.aibl.csiro.au.

References

Association, A., 2019. 2019 alzheimer's disease facts and figures. Alzheimer's & Dementia 15, 321 – 387.

Baron, J., Chételat, G., Desgranges, B., Perchey, G., Landeau, B., de la Sayette, V., Eustache, F., 2001. In vivo mapping of gray matter loss with voxel-based morphometry in mild alzheimer's disease. NeuroImage 14, 298 – 309.

Beyer, L., Meyer-Wilmes, J., Schönecker, S., Schnabel, J., Sauerbeck, J., Scheifele, M., Prix, C., Unterrainer, M., Catak, C., Pogarell, O., Palleis, C., Perneczky, R., Danek, A., Buerger, K., Bartenstein, P., Levin, J., Rominger, A., Ewers, M., Brendel, M., 2021. Cognitive reserve hypothesis in frontotemporal dementia: A fdg-pet study. NeuroImage: Clinical 29, 102535.

Bilen, H., Vedaldi, A., 2016. Weakly supervised deep detection networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M., 2007. Forecasting the global burden of alzheimer's disease. Alzheimer's & Dementia 3, 186 – 191.

Cao, P., Liu, X., Yang, J., Zhao, D., Huang, M., Zhang, J., Zaiane, O., 2017. Nonlinearity-aware based dimensionality reduction and over-sampling for ad/mci classification from mri measures. Computers in Biology and Medicine 91, 21–37.

Cipolla, R., Gal, Y., Kendall, A., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7482–7491.

Cui, R., Liu, M., 2019. Hippocampus analysis by combination of 3-d densenet and shapes for alzheimer's disease diagnosis. IEEE Journal of Biomedical and Health Informatics 23, 2099–2107.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. NeuroImage 9, 179 – 194.

Farooq, A., Anwar, S., Awais, M., Rehman, S., 2017. A deep cnn based multi-class classification of alzheimer's disease using mri, in: 2017 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–6.

Fox, N.C., Warrington, E.K., Freeborough, P.A., Hartikainen, P., Kennedy, A.M., Stevens, J.M., Rossor, M.N., 1996. Presymptomatic hippocampal atrophy in Alzheimer's disease: A longitudinal MRI study. Brain 119, 2001–2007.

Frisoni, G.B., Fox, N.C., Jr, C.R.J., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural mri in alzheimer's disease. Nature reviews. Neurology 6, 67–77. doi:10.1038/nrneurol.2009.215.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256.

Goodfellow, I., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples, pp. 1–10.

Guan, H., Liu, Y., Yang, E., Yap, P.T., Shen, D., Liu, M., 2021. Multi-site mri harmonization via attention-guided deep domain adaptation for brain disorder identification. Medical Image Analysis 71, 102076.

Holmes, C.J., Hoge, R., Collins, L., Evans, A.C., 1996. Enhancement of t1 mr images using registration for signal averaging. NeuroImage 3, S28.

Jack, C., Petersen, R., Xu, Y., O'Brien, P., Smith, G., Ivnik, R., Boeve, B., Waring, S., Tangalos, E., Kokmen, E., 1999. Prediction of ad with mri-based hippocampal volume in mild cognitive impairment. Neurology 52, 1397—1403. doi:10.1212/wnl.52.7.1397.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825 – 841.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Medical Image Analysis 5, 143 – 156.

Jie, B., Liu, M., Zhang, D., Shen, D., 2018. Sub-network kernels for measuring similarity of brain connectivity networks in disease diagnosis. IEEE Transactions on Image Processing 27, 2340–2353.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, Clifford R., J., Ashburner, J., Frackowiak, R.S.J., 2008. Automatic classification of mr scans in alzheimer's disease. Brain 131, 681–689.

Lian, C., Liu, M., Pan, Y., Shen, D., 2020a. Attention-guided hybrid network for dementia diagnosis with structural mr images. IEEE Transactions on Cybernetics doi:10.1109/TCYB.2020.3005859.

Lian, C., Liu, M., Zhang, J., Shen, D., 2020b. Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 880–893.

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.

Liu, M., Zhang, J., Adeli, E., Shen, D., 2018. Landmark-based deep multi-instance learning for brain disease diagnosis. Medical Image Analysis 43, 157 – 168.

Liu, M., Zhang, J., Adeli, E., Shen, D., 2019. Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis. IEEE Transactions on Biomedical Engineering 66, 1195–1206.

Liu, M., Zhang, J., Lian, C., Shen, D., 2020. Weakly supervised deep learning for brain disease prognosis using mri and incomplete clinical scores. IEEE Transactions on Cybernetics 50, 3381–3392.

Liu, X., Tosun, D., Weiner, M.W., Schuff, N., 2013. Locally linear embedding (lle) for mri based alzheimer's disease classification. NeuroImage 83, 148 – 157.

Manera, A.L., Dadar, M., Collins, D.L., Ducharme, S., 2019. Deformation based morphometry study of longitudinal mri changes in behavioral variant frontotemporal dementia. NeuroImage: Clinical 24, 102079.

Miyato, T., Dai, A., Goodfellow, I., 2017. Adversarial training methods for semi-supervised text classification.

Nestor, S.M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J.L., Fogarty, J., Bartha, R., the Alzheimer's Disease Neuroimaging Initiative, 2008. Ventricular enlargement as a possible measure of alzheimer's disease progression validated using the alzheimer's disease neuroimaging initiative database. Brain 131, 2443–2454.

Qiu, S., Joshi, P., Miller, M., Xue, C., Zhou, X., Karjadi, C., Chang, G., Joshi, A., Dwyer, B., Zhu, S., Kaku, M., Zhou, Y., Alderazi, Y., Swaminathan, A., Kedar, S., Saint-Hilaire, M.H., Auerbach, S., Yuan, J., Sartor, E., Kolachalama, V., 2020. Development and validation of an interpretable deep learning framework for alzheimer's disease classification. Brain : a journal of neurology 143, 1920–1933.

Rasmus, A., Valpola, H., Honkala, M., Berglund, M., Raiko, T., 2015. Semi-supervised learning with ladder networks, in: Proceedings of the 28th International Conference on Neural Information Processing Systems.

Rathore, S., Habes, M., Iftikhar, M.A., Shacklett, A., Davatzikos, C., 2017. A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages. NeuroImage 155, 530 – 548.

Salvatore, C., Cerasa, A., Battista, P., Gilardi, M.C., Quattrone, A., Castiglioni, I., 2015. Magnetic resonance imaging biomarkers for the early diagnosis of alzheimer's disease: a machine learning approach. Frontiers in Neuroscience 9, 307.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in mri data. IEEE transactions on medical imaging 17, 87–97.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R., 2014. Intriguing properties of neural networks. ICLR .

Sørensen, L., Igel, C., Liv Hansen, N., Osler, M., Lauritzen, M., Rostrup, E., Nielsen, M., for the Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle Flagship Study of Ageing , 2016. Early detection of alzheimer's disease using mri hippocampal texture. Human Brain Mapping 37, 1148–1161.

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J.V., Rueckert, D., 2014. Multiple instance learning for classification of dementia in brain mri. Medical Image Analysis 18, 808 – 818.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in

spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. NeuroImage 15, 273 – 289.

Vieira, R., Caixeta, L., Machado, S., Cardoso, A., Nardi, A., Arias-Carrión, O., Carta, M., 2013. Epidemiology of early-onset dementia: A review of the literature. Clinical practice and epidemiology in mental health 9, 88–95.

Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q., 2019. C-mil: Continuation multiple instance learning for weakly supervised object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Wang, L., Beg, F., Ratnanather, T., Ceritoglu, C., Younes, L., Morris, J.C., Csernansky, J.G., Miller, M.I., 2007. Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the alzheimer type. IEEE Transactions on Medical Imaging 26, 462–470.

Wang, X., Chen, H., Xiang, H., Lin, H., Lin, X., Heng, P.A., 2021. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. Medical Image Analysis 70, 102010.

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., Colliot, O., 2020. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. Medical Image Analysis 63, 101694.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of alzheimer's disease and mild cognitive impairment. NeuroImage 55, 856 – 867.

Zhang, J., Gao, Y., Gao, Y., Munsell, B.C., Shen, D., 2016. Detecting anatomical landmarks for fast alzheimer's disease diagnosis. IEEE Transactions on Medical Imaging 35, 2524–2533.

Zhu, X., Suk, H.I., Wang, L., Lee, S.W., Shen, D., 2017. A novel relational regularization feature selection method for joint regression and classification in ad diagnosis. Medical Image Analysis 38, 205 – 214.