Physics in Medicine & Biology

Light-weight cross-view hierarchical fusion network for joint localization and identification in Alzheimer's disease with adaptive instance-declined pruning

To cite this article before publication: Kangfu Han *et al* 2021 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/abf200

IOP Publishing Journal Title
Journal **XX** (XXXX) XXXXXX https://doi.org/XXXX/XXXX

Light-weight Cross-view Hierarchical Fusion Network for Joint Localization and Identification in Alzheimer's Disease with Adaptive Instance-declined Pruning

Kangfu Han[1, 2], Jiaxiu Luo[1, 2], Qing Xiao[1, 2], Zhenyuan Ning[1, 2,*], and 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

[*]Corresponding author

E-mail: yuzhang@smu.edu.cn (Y. Zhang) and jonnyning@foxmail.com (Z. Ning)

Abstract

Magnetic resonance imaging (MRI) has been widely used in assessing development of Alzheimer's disease (AD) by providing structural information of disease-associated regions (e.g., atrophic regions). In this paper, we propose a light-weight cross-view hierarchical fusion network (CvHF-net), consisting of local patch and global subject subnets, for joint localization and identification of the discriminative local patches and regions in the whole brain MRI, upon which feature representations are then jointly learned and fused to construct hierarchical classification models for AD diagnosis. Firstly, based on the extracted class-discriminative 3D patches, we employ the local patch subnets to utilize multiple 2D views to represent 3D patches by using an attention-aware hierarchical fusion structure in a divide-and-conquer manner. Since different local patches are with various abilities in AD identification, the global subject subnet is developed to bias the allocation of available resources towards the most informative parts among these local patches to obtain global information for AD identification. Besides, an instance declined pruning (IDP) algorithm is embedded in the CvHF-net for adaptively selecting most discriminant patches in a task-driven manner. The proposed method was evaluated on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and the experimental results show that our proposed method can achieve good performance on AD diagnosis.

Keywords: Cross-view, Hierarchical fusion, Instance-declined pruning, deep learning, Alzheimer's disease

1. Introduction

Alzheimer's disease (AD), characterized by cognitive impairment, is an irreversible neurodegenerative disorder disease caused by the accumulation of toxic protein (i.e., protein tau) [1]. Back in 2006, 26.6 million people suffered from AD around the world, and mild cognitive impairment (MCI), as the prodromal stage of AD, accounted for 56% [2]. With population aging worldwide, the number of AD patients will increase consistently and it is estimated to be over 100 million by 2050. Early diagnosis of AD is beneficial to patient care and disease management [3]. Nowadays, magnetic resonance imaging (MRI), a routine clinical imaging tool, has been widely used in assessing development of AD by providing structural information of disease-associated regions, based on which many computer-aided methods have been proposed for early diagnosis of AD [4, 5, 6].

Existing computer-aided methods for the early diagnosis of AD can be roughly categorized as conventional machine-learning based and deep-learning based approaches [7, 8]. Conventional machine learning methods [9-14] generally contain three steps: regions-of-interest (ROI) identification, feature extraction and classification model construction. For example, Zhu et al. [9] dissected the whole brain into non-overlapped 93 ROIs by wrapping it into the Jacob template [15], and then extracted morphological features (i.e., volume of gray matter) to construct a support vector machine with several relational regularizations for joint regression and classification of AD. Koikkalainen et al. [11] and Liu et al. [12] spatially normalized the whole brain into multiple atlases, and then extracted regional features from each atlas space to construct ensemble models for AD/MCI diagnosis. Wang et al. [13] and Sørensen et al. [14] extracted shape and textural features of bilateral hippocampi for AD classification, respectively. Recently, deep-learning based methods, especially convolutional neural networks (CNNs), have been successfully used for AD-related diagnosis by integrating feature extraction and model construction into a unified framework [6, 16-21]. For example, Li et al. [18] and Khvostikov et al. [20] extracted bilateral hippocampi as the most discriminative regions to train CNNs for early diagnosis of AD. Suk et al. [22] learned the shared feature representations between patches extracted from sMRI and positron emission tomography (PET) images based on deep Boltzmann machine [23] to train an ensemble SVM classifier for AD/MCI classification.

To further improve the performance for the early diagnosis of AD, some studies first identified AD-relate regions, and then used the features extracted from the patches centered at the AD-relate voxels for AD classification. Typically, in [24], patches centered at the voxel with the smallest mean p-value were extracted to train hierarchical classifiers for AD diagnosis. In [25], probability map was generated via elastic net method [26] and used to extract discriminative patches to construct mi-Graph model for AD diagnosis. In [27, 28], Liu et al. proposed a landmark-based deep multi-instance learning framework for AD/MCI classification and a weakly supervised densely connected neural network (wiseDNN) for longitudinal clinical scores regression using baseline sMRI, respectively. In [29], Lian et al. proposed a patch-based hierarchical fully convolutional network to automatically identify discriminant patches and regions by using image labels for supervised learning (on patch and region subnets, respectively), and then multi-scale feature representations were jointly learned and fused for AD diagnosis. It is to note that, to catch sufficient information, the above approaches are based on 3D patches. However, it accordingly brings huge computational cost and is time-consuming. It is also inadequate if an iterative procedure is needed in the method.

Recently, some methods extracted multiple 2D views to represent 3D ROIs and obtained promising results. In [30], multiple 2D views of pulmonary nodules were extracted to train multi-stream ConvNet, in which features of different views were combined for pulmonary nodule detection. In [31], Luo et al. proposed an integrative framework of deep learning and bag-of-feature model for preoperative prediction of sentinel lymph node metastasis by learning and fusing three 2D representative orthogonal views of 3D ROIs. In [32], a multi-view saliency-based framework was developed to detect abnormalities from MRI and classify subjects using a multiple kernel learning method. These 2D based methods also achieve promising results, by use of small number of parameters, and thus are time-saving and have less computational cost. Correspondingly, there rises a key problem: how to fuse the multi-view information to represent 3D ROIs efficiently.

In this paper, we propose a light-weight cross-view hierarchical fusion network (CvHF-net), consisting of local patch and global subject subnets, for joint localization and identification of the discriminative local patches and regions in the whole brain MRI, upon which multi-group feature representations are then jointly learned and fused to construct hierarchical classification models for AD diagnosis. As shown in Figure 1, based on the extracted class-discriminative 3D patches, we first employ the local patch subnets to utilize multiple 2D views to represent 3D patches by using an attention-aware hierarchical fusion structure in a divide-and-conquer manner. Since different local patches are with various abilities in AD diagnosis, the global subject subnet is developed to bias the allocation of available resources towards the most informative parts among these local patches via attention mechanism to obtain global information for AD/MCI classification. Besides, an instance declined pruning (IDP) algorithm is embedded in the CvHF-net for adaptively selecting most discriminant patches in a task-driven manner. The experimental results on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset show that, the proposed method can efficiently perform disease diagnosis for AD. The major contributions of this work are summarized as follows:

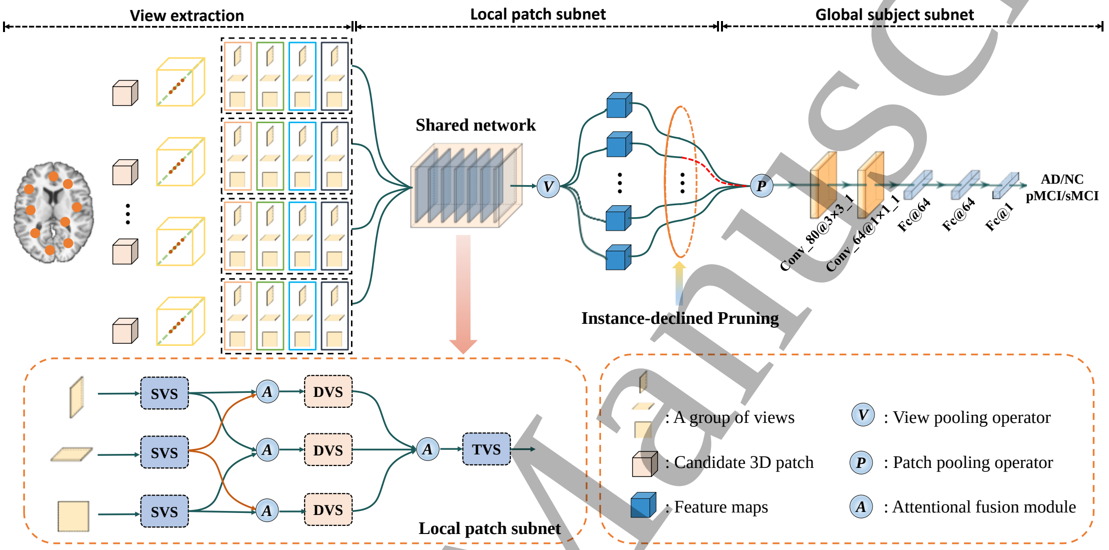


Figure 1: The architecture of CvHF-net. The multi-group 2D views are first extracted from the candidate 3D patches. Then, the local patch subnet, including single-view stream (SVS), double-view stream (DVS), and triple-view stream (TVS), utilizes multiple 2D views to represent the 3D patch using the attention-aware hierarchical fusion structure in a divide-and-conquer manner. Besides, an instance-declined pruning strategy is embedded to adaptively identify AD-associated regions at the end of the local patch subnet. Finally, the global subject subnet is developed for obtaining global information for AD identification.

Table 1. The characteristic of baseline subject

	AD	pMCI	sMCI	NC
Gender (M/F)	56/37	48/31	86/35	60/39
Age	75.43±7.34	74.97±6.68	74.85±7.48	75.68±4.73
MMSE	23.47±2.13	26.81±1.68	27.40±1.64	28.93±1.12
Education	14.80±3.00	15.80±2.70	15.77±2.8	15.80±3.10

- We propose a light-weight unified framework to perform jointly localization and identification for AD diagnosis, in which we take full advantages of local-to-global representations of disease-associated regions.
- Motivated by multi-instance learning [27], the instance-declined pruning (IDP) strategy is proposed to effectively and adaptively localizing most disease-associated regions.
- We propose to utilize multiple 2D views to represent 3D patches using the attention-aware hierarchical fusion structure in a divide-and-conquer manner. It is not only time-saving, but also provides an extra data augmentation manner for network training.

The rest of this paper is organized as follows. In Section 2, the used dataset and the proposed method are described in detail. Experimental settings and results are presented in Section 3. Finally, discussion and conclusion are provided in Sections 4.

2. Method

In this section, we first introduce the dataset and its image preprocessing pipeline. Subsequently, we describe the proposed CvHF-net, including view extraction, local patch subnet, global subject subnet, loss function, and instance-declined pruning (IDP) algorithm, respectively.

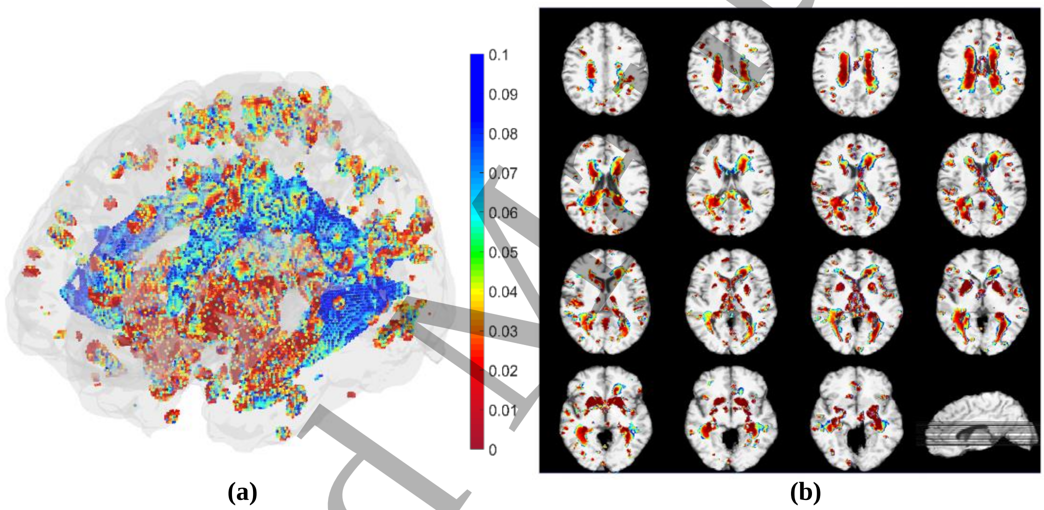


Figure 2: The visualization of the candidate regions with significant difference. a) 3D visualization of the candidate regions (the red part means more significant); b) 2D location visualization of the candidate regions. The statistically significant voxels are mostly distributed in hippocampus, corpus callosum and their surrounding regions.

2.1 *Dataset and pre-processing*

In this paper, we collected data from the public AD Neuroimaging Initiative-1 (ADNI-1, http://adni.loni.usc.edu). In total, 392 subjects with baseline sMRI were enrolled. According to some clinical criteria, such as clinical dementia rating and mini-mental state examination scores, these subjects were classified into three categories, namely, normal control (NC), mild cognitive impairment (MCI), and AD. As a part of MCI subjects would convert to AD within 36 months and the remaining are stable over time, MCI subjects can be further divided into progressive MCI (pMCI) and stable MCI (sMCI). In summary, 99 NC, 121 sMCI, 79 pMCI and 93 AD were used to train and evaluate the proposed method. The characteristics of enrolled subjects are presented in Table 1.

All MR images have been reviewed and corrected by ADNI researchers for spatial distortion caused by B1 field inhomogeneity and gradient nonlinearity. Our image pre-processing contains following procedures: (1) anterior commissure-posterior commissure (ACPC) correction via MIPAV software (http://mipav.cit.nih.gov/clickwrap.php), (2) intensity inhomogeneity correction using N3 algorithm [33], (3) skull stripping, cerebellum removal with aBEAT (https://www.nitrc.org/projects/abeat), and manual confirmation for cleanly skull and dura removal, (4) the linear and non-linear registration [34, 35] were used to align each MR image to the Colin27 template [36], and (5) resampling all MR images to the fixed size of 256 × 256 × 256 and spatial resolution of 1 × 1 × 1 mm^3.

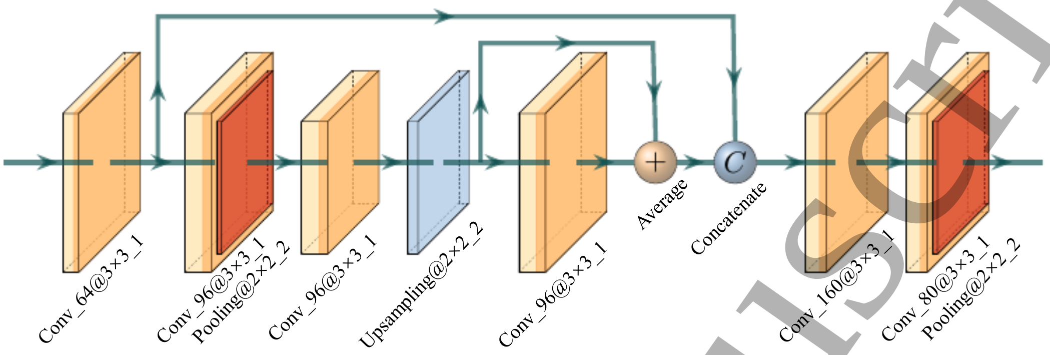


Figure 3: The detailed architecture of SVS/TVS in the local patch subnet. The number of channels (e.g., 64), kernel size (e.g., 3 × 3), and stride (e.g., 1) in each convolutional layer is denoted as "Conv_64@3 × 3_1".

2.2 Architecture

2.2.1 *View Extraction*

To better utilize local regional information, we first extract class-discriminative 3D patches in the whole brain MRI image. Following [22, 24], we perform group-wise analysis to exploit the voxels with statistical significance in each patch. Specifically, by performing group comparison (e.g., AD and NC, pMCI and sMCI), we obtain p-value for all voxels and find these statistically significant voxels (i.e., p-value < 0.05). As p-value distribution map shown in Figure 2, we can observe that these statistically significant voxels are mostly distributed in hippocampus, corpus callosum and their surrounding regions, which has been verified to be related to the development of AD [37]. Subsequently, given a fixed patch with a size of $w \times w \times w$ (the patch size is experimentally set to 32 × 32 × 32 in our study, see section 3.5 for details), we use it to scan all statistically significant voxels and select class-discriminative patches in a greedy manner with the following rules: (1) The candidate patch should be overlapped less than 25% with any of the selected patches, and (2) Among the candidate patches that satisfy the rule above, we select patches which cover the more number of statistically significant voxels. Considering the difference in pathological changes between AD and MCI subjects, the aforementioned steps are implemented on two tasks (AD vs. NC and sMCI vs. pMCI) independently. In total, 36 and 30 3D patches are chosen for AD and MCI diagnosis, respectively. Unlike other view extraction methods [30-32], we extract multi-group 2D views for each 3D patch. As shown in Figure 1, we first find all main-diagonal points of the 3D patch, based on which the axial, coronal, and sagittal planes across each main-diagonal point are extracted as a group of views. The view extraction approach here has some advantages: (1) It reduces the computational cost, (2) It can retain sufficient information of the 3D patch, and (3) A patch can generate multi-group views, which provides an extra data argumentation way for network training.

2.2.2 *Local patch subnet*

To better utilize multiple 2D views to represent the 3D patch, we develop a local patch subnet to learn discriminative view representations using the attention-aware hierarchical fusion structure in a divide-and-conquer manner. As shown in Figure 1, the local patch subnet consists of three-level attentional streams, namely, single-view stream (SVS), double-view stream (DVS), and triple-view stream (TVS).

Single-view stream (SVS). As Figure 1 and 3 shown, the SVS is used to independently extract the specific features for each view, which basically contains several convolutional layers, two skip connections, and a tiny U-net-like block. Specifically, the SVS first uses a convolutional layer with a kernel size of 3 × 3 to extract features for each view. Then, a tiny U-net-like block is placed to capture global view representations, increase the receptive field, and reduce the number of parameters, which consists of a 3 × 3 convolutional layer, a 2 × 2 pooling layer, a 3 × 3 convolutional layer, and a 2 × 2 up-sampling layer in sequence. Subsequently, a 3 × 3 convolutional layer is used to further learn high-level features and a skip connection is applied for fusing the input and output of this layer. To retain low-level and high-level feature representations, the feature maps output by the first convolutional layer are concatenated with the output of the skip connection, followed by two consecutive convolutional layers (with kernel size of 3 × 3 and 1 × 1, respectively) and a 2 × 2 pooling layer. The filter number of 6 convolutional layers is set to 64, 96, 96, 96, 160, and 80, respectively.

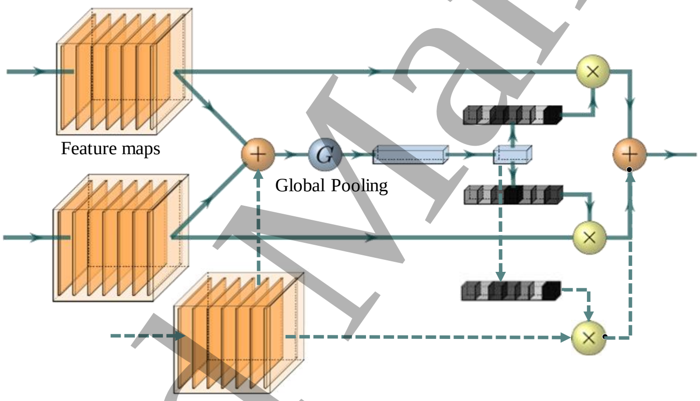


Figure 4: The architecture of attention fusion module. The solid lines denote the process of this module in DVS for any two views, and the dotted lines show the extended version in TVS for three views.

Double-view stream (DVS). As the features of each view are extracted independently by SVS, it may loss the inherent information between any two of views. To this end, we construct a module of DVS to fuse the information from any two views. In particular, to take advantage of the interdependence between the channel maps from two views, we can emphasize the interdependent feature maps from the views and improve the feature representation of specific semantics. Herein, similar to the attention strategy used in SENet [38], we build an attention fusion module to excavate the interdependencies between views, as shown in Figure 4. Specifically, the SVS output feature maps from any two of views are first averaged. And then, the channel attention mechanism is applied on these averaged feature maps. Subsequently, we perform a matrix multiplication between channel-attention-weights and the feature maps from each view, and obtain the channel-attention-weighted feature maps for each view. The attention-weighted feature maps of any two of views are then fused into output features by pixel-wise average.

Triple-view stream (TVS). The DVS only fuses the feature representations from any two views, based on which the TVS is constructed to further learn the shared representations of three views. Similar to

DVS (as the dotted part of Figure 4 illustrated), TVS also uses an attentional-aware view fusion module with three input channels that matched with the output of DVS, to generate the corresponding weighted vectors for the feature maps at each input channel. Finally, the attention-weighted feature maps of all three views (i.e. axial, coronal, and sagittal plane) are then fused into output features by pixel-wise average. It is worth to note that this output features only represent one-group views extracted at one main-diagonal point. We then develop a view pooling approach at the end of TVS to integrate the features of all group 2D views extracted from all main-diagonal points for representing each 3D patch. Specifically, let C and V denote the channel of fused feature maps (with the size of $W \times H$) for each group views and the number of group views of a 3D patch, respectively. For V group of views, the view pooling layer first extracts the feature maps at c-th (c ranging from 1 to C) channel to form a $V \times W \times H$ feature tensor and averages features along the first-dimensional direction to obtain the feature map with the size of $W \times H$. The aforementioned process is repeated for each channel and finally forms a $C \times W \times H$ feature tensor to represent each 3D patch.

2.2.3 *Global subject-net*

Finally, all patch level feature representations (size $L \times C \times W \times H$, L is the number of extracted class-discriminative patches, which is 36 for AD classification and 30 for MCI conversion prediction, respectively) are averaged by using a patch pooling operation [39] along patch direction. Specifically, given L candidate patch representations with the size of $C \times W \times H$ yielded by local patch subnet, that is, the size of the input of the global subject-net is $L \times C \times W \times H$, the patch pooling strategy is implemented on the corresponding features of each patch, and averages the patch representations to obtain the fused features with the size of $C \times W \times H$. Subsequently, two convolutional layers with a kernel size of 3 × 3 and 1 × 1, respectively, and three fully-connected (FC) layers with the number of neural units of 64, 64, 1, respectively, are used to make the final diagnosis for each subject. It is worth mentioning that the dropout operator with dropout rate of 0.3 is armed with the three fully connected layers for avoiding overfitting.

2.3 *Loss function*

We design a hybrid loss function to train the proposed CvHF-net efficiently. Specifically, let $\left\{(\mathcal{X}_n, \mathrm{y}_n)\right\}_{n=1}^{N}$ be the training set, where $\mathcal{X}_i = \{\mathbf{X}_{il}^{1}, \ldots, \mathbf{X}_{il}^{M}\}_{l=1}^{L}$ denotes the ROI of i-th subject and each ROI has L 3D patches that consists of M 2D views groups (denoted as $\mathbf{X}_{il}^{m} = \{\mathbf{x}_{il}^{m,a}, \mathbf{x}_{il}^{m,c}, \mathbf{x}_{il}^{m,s}\}$, M is the number of group for each 3D patch), and $\mathrm{y}_i = \{\mathrm{y}_l\}_{l=1}^{L}$ denotes the corresponding class label. The loss function is designed as follows:

$$\mathcal{L}\left(\mathbf{W}^{p}, \mathbf{W}^{s}\right) = L_s(\mathbf{W}^{s} \middle| \mathbf{W}^{p}, \mathcal{X}_n, \mathrm{y}_n) + L_p(\mathbf{W}^{p}, \mathbf{X}_l, \mathrm{y}_l)$$

where $\mathbf{W}^{p}$ and $\mathbf{W}^{s}$ denote the learnable parameters for the patch- and subject-level subnet, L_s is the binary cross entropy, and the L_p is defined as:

$$\begin{aligned} L_p(\mathbf{W}^{p}, \mathbf{X}_i, \mathrm{y}_i) = {} & L_{TVS}(\mathbf{W}^{TVS} | \mathbf{W}^{DVS} | \mathbf{W}^{SVS}, \{\mathbf{x}_l^{a}, \mathbf{x}_l^{c}, \mathbf{x}_l^{s}\}, \mathrm{y}_l) \\ & + L_{DVS}(\mathbf{W}^{DVS} | \mathbf{W}^{SVS}, \{\mathbf{x}_l^{a}, \mathbf{x}_l^{c}\}, \{\mathbf{x}_l^{a}, \mathbf{x}_l^{s}\}, \{\mathbf{x}^{s}, \mathbf{x}_l^{c}\}) \\ & + L_{SVS}(\mathbf{W}^{SVS}, \mathbf{x}_l^{a}, \mathbf{x}_l^{c}, \mathbf{x}_l^{s}) \end{aligned}$$

where L_{TVS}, L_{DVS}, and L_{SVS} denote the binary cross entropy of TVS, DVS and SVS, respectively.

2.4 *Instance declined pruning (IDP) algorithm for adaptive localization*

Based on the resulting diagnostic/classification scores on the training set for each local patch and global subject sub-networks, we further refine the initial CvHF-net by using instance declined pruning (IDP) algorithm to remove uninformative patches. The advantages of IDP algorithm is: 1) providing a reversible pruning way during the iterative process; 2) selecting discriminative patches in an adaptive manner. In IDP, we look at all patches of one subject as a bag, and each patch is an instance in the bag. The number of instances used to represent the bag is gradually tapered during iterative process, and all instances can be backtracked for each pruning. In addition, IDP uses instance-wise cross-entropy loss instead of accuracy (commonly used in MIL) as a pruning criteria, in which the instances with a small loss are considered to be discriminative patches. Specifically, for the j-th iteration, the instances with top j × k loss (k = 4 for AD vs. NC and 3 for sMCI vs. pMCI) are pruned. Then, the remaining instances are used to continue training the network. Subsequently, all instances (including pruned and remaining instances) with top (j +1) × k loss are pruned at the (j +1)-th iteration. The process stops until the score at a certain iteration consistently outperforms those at subsequent 3 iterations or the number of iteration reaches to a given number (the given number is set to 9 in our experiments).

3. **Experiments and Results**

In this section, we first introduce the experiment settings, including competing methods and evaluation strategy. Subsequently, the experimental results are presented, including predictive performance of all comparison methods, effectiveness of each part by ablation experiments, and the influence of parameters.

3.1 *Experiment settings*

To verify the efficiency of our proposed framework, we first conduct comparison experiments with several state-of-the-art methods, including 1) ROI-based method [40], 2) VBM-based method [37], 3) multi-view convolutional network (denoted as Mv-net) [30], 4) landmark-based deep multi-instance convolutional network (denoted as LDMI-net) [27].

1) ROI-based method [40]. After skull and cerebellum removing, each sMRI was first segmented into three tissue types, i.e., gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), by using aBEAT package [41]. Then, following previous studies [40], the anatomical automatic labeling (AAL) atlas [42], with 90 pre-defined ROIs in the cerebrum, was aligned to each subject. Finally, the features of GM volumes in the 90 ROIs were extracted to train linear SVM classifiers.

2) VBM-based method [37]. In line with [37], all sMRI data were spatially normalized to the Colin27 template to extract local GM density in a voxel-wise manner. After that, a statistical group comparison based on t-test was performed to extract voxel-level feature representations for SVM-based classification.

3) Mv-net [30]. The Mv-net performed prediction by using several views extracted from the 3D patches. Specifically, nine views were first extracted from each candidate 3D patch as the input of Mv-net. Following [30], the architecture of Mv-net includes nine channels to extract features for each view, and each channel contains three convolutional layers, three max-pooling layers as well as one fully-connected layer. The deep features generated by each channel were concatenated for patch-level prediction using two fully connected layers. Finally, a naïve ensemble strategy (i.e., averaging the predicted probability of patches from the same subject) was used for subject-level classification. In our experiments, the batch size and epoch size were set to 64 and 80, respectively.

4) LDMI-net [27]. The LDMI-net performed classification for AD diagnosis by learning the local-

Table 2: The comparison results with other methods for AD classification and MCI conversion prediction tasks in terms of four metrics, including accuracy (ACC), sensitivity (SEN), specificity (SPE), as well as the area under the curve of receiver operating characteristic (AUC), which are reported as mean ± Standard deviation (Std).

Task	Metrics	Methods				
		ROI-based	VBM-based	MV-net	LDMI-net	Ours
AD vs. NC	ACC	0.760±0.059	0.787±0.088	0.847±0.029	0.911±0.030	**0.937±0.014**
	SEN	0.707±0.120	0.753±0.068	0.878±0.099	**0.900±0.046**	0.889±0.039
	SPE	0.808±0.024	0.819±114	0.820±0.076	0.920±0.057	**0.980±0.027**
	AUC	0.840±0.057	0.881±0.093	0.899±0.040	0.948±0.035	**0.951±0.034**
pMCI vs. sMCI	ACC	0.625±0.047	0.605±0.057	0.740±0.042	0.770±0.041	**0.800±0.035**
	SEN	0.253±0.040	0.318±0.129	0.613±0.162	0.613±0.143	**0.650±0.130**
	SPE	0.868±0.090	0.794±0.118	0.825±0.080	0.875±0.083	**0.900±0.048**
	AUC	0.584±0.067	0.606±0.053	0.696±0.068	**0.761±0.040**	0.745±0.059
Time(h)		-	-	0.22	**1.62**	1.07
Para		-	-	307K	**33M**	2.5M

to-global structural information in an end-to-end way. Specifically, the LDMI-net consists of several channels (equal to the number of 3D patches). Each channel contains six convolutional layers with the kernel size of 3 × 3 × 3, three 2 × 2 × 2 "max" pooling layers (each one is placed after every two convolutional layers), and two FC layers for generating local representations. By concatenating these local level representations as global-level representations, three fully-connected layers are further used to perform final prediction. In our experiments, the batch size and epoch size were set to 4 and 40, respectively.

We use Python3.7 to implement all experiments and evaluations and Tensorflow to build all deep learning networks. All computationally intensive calculations are offloaded to a 12 GB NVIDIA Pascal Titan X GPU. All comparison experiments are conducted on the same data partition of 5-fold hold-out strategy. The average predictive performance of all methods is assessed by four metrics, including accuracy (ACC), sensitivity (SEN), specificity (SPE), as well as the area under the curve of receiver operating characteristic (AUC), which are defined as $\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$, $\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ and $\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN+FP}}$, where TP, TN, FP, and FN denote the true positive, true negative, false positive, and false negative values, respectively. The AUC is calculated based on all possible pairs of SEN and SPE obtained by changing the thresholds performed on the classification scores yielded by the trained networks.

3.2 *Comparison results*

The comparison results are listed in Table 2. From Table 2, we can observe that: 1) For both diagnosis tasks, the CvHF-net outperforms other three methods (i.e., the ROI, VBM, and Mv-net methods) with relatively large margin, demonstrating that our proposed hierarchical fusion network can learn more discriminative features which are beneficial for AD and MCI classification tasks. 2) Compared with the state-of-the-art LDMI-net, our proposed CvHF-net method also has competitive performance in the tasks of AD and MCI classification. Specifically, our method yields better results on ACC and SPE. The LDMI-net outperforms our CvHF-net for SEN and AUC, especially in pMCI vs. sMCI classification task. It is perhaps due to the reason that we construct shared CNNs for all extracted patches, this approach is help for obtaining the light-weight network, but may loss a few specific information for classification. 3) Compared with LDMI-net, the CvHF-net shows the advantage of light-weight. The number of parameters of CvHF-net is far fewer than LDMI-net (2.5M/33M), and it takes less time to train CvHF-net than LDMI-net (1.07/1.62 hours). It suggests that our proposed model can obtain state-of-the-art results with less computational cost.

Table 3. The results of ablation study for AD classification and MCI conversion prediction tasks in terms of four metrics, including accuracy (ACC), sensitivity (SEN), specificity (SPE), as well as the area under the curve of receiver operating characteristic (AUC), which are reported as mean ± Standard deviation (Std).

SVS		√	√	√	√	√
DVS		×	×	√	×	√
TSV		×	×	×	√	√
subject -net		×	√	√	√	√
AD *vs* NC	ACC	0.858±0.040	0.874±0.051	0.884±0.030	0.884±0.040	**0.895±0.037**
	SEN	0.811±0.101	0.844±0.072	0.867±0.091	0.867±0.063	**0.867±0.101**
	SPE	0.900±0.079	0.900±0.079	0.900±0.084	0.900±0.071	**0.920±0.057**
	AUC	0.893±0.046	0.917±0.042	**0.917±0.032**	0.915±0.037	0.916±0.036
pMCI *vs* sMCI	ACC	0.740±0.034	0.750±0.031	0.755±0.045	0.765±0.045	**0.765±0.038**
	SEN	**0.613±0.195**	0.588±0.114	0.575±0.149	0.600±0.034	0.588±0.175
	SPE	0.825±0.095	0.858±0.048	0.875±0.051	0.875±0.066	**0.883±0.054**
	AUC	0.702±0.072	0.717±0.051	0.719±0.051	0.715±0.051	**0.721±0.075**

Table 4. The performance (ACC) of CvHF-net during iteration process ('Iter k' denotes the k-th iteration, and the top and bottom is for AD classification and MCI conversion prediction task, respectively).

	Iter0	Iter1	Iter2	Iter3	Iter4	Iter5	Iter6	Iter7	Iter8
Fold1	0.842	0.842	0.842	0.868	0.868	0.842	0.868	0.895	**0.921**
Fold2	0.921	0.895	0.895	0.921	0.921	0.921	0.921	**0.947**	0.895
Fold3	0.921	0.868	0.921	0.868	0.895	0.895	0.921	**0.947**	0.921
Fold4	0.921	0.921	0.921	0.947	0.947	**0.947**	0.895	0.921	0.842
Fold5	0.868	0.868	0.868	**0.921**	0.868	0.895	0.868	-	-
Fold1	0.750	**0.775**	0.675	0.700	0.725	-	-	-	-
Fold2	0.825	0.825	0.800	**0.850**	0.800	0.800	0.800	-	-
Fold3	0.750	0.750	0.725	0.700	0.750	0.750	0.675	**0.775**	0.750
Fold4	0.725	0.725	**0.775**	0.725	0.750	0.750	-	-	-
Fold5	0.775	0.800	0.775	0.750	0.800	**0.825**	0.775	0.800	0.775

3.3 *Ablation Study*

We conducted a series of ablation experiments to investigate the effectiveness of each component in the proposed method.

3.3.1 *Effectiveness of hierarchical fusion strategy*

In this experiment, we consider SVS as the backbone of the network (i.e., baseline model). The ablation study compares five models and all of them are trained without the IDP strategy. Notably, a na ïve ensemble strategy is used to perform subject-level prediction when global subject subnet is not available. As Table 3 shown, we can observe that: 1) The model with subject subnet (the 2-rd row) obtains better results than the baseline model almost on all metrics, which owes to local-to-global fusion strategy. 2) Integrating either DVS or TVS into the model with subject subnet can further improve the performance (such as ACC and SEN for AD classification task, and ACC and SPE for MCI conversion prediction task). A potential explanation is that DVS and TVS can explore inner characteristics exits in any two views and three views, respectively. 3) Utilizing local patch and global subject subnets, the proposed architecture achieves the best results almost on all metrics, which benefits from local patch

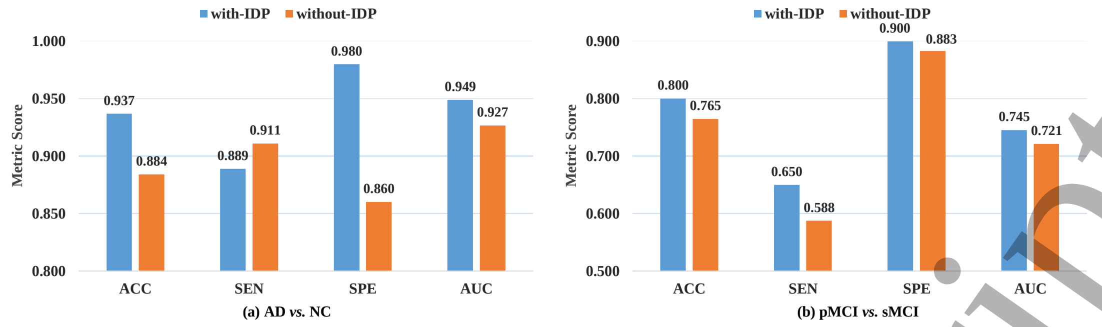


Figure 5: The comparison results of CvHF-net with and without IDP in terms of four metrics, including accuracy (ACC), sensitivity (SEN), specificity (SPE), as well as the area under the curve of receiver operating characteristic (AUC), for AD classification (a) and MCI conversion prediction (b), respectively.

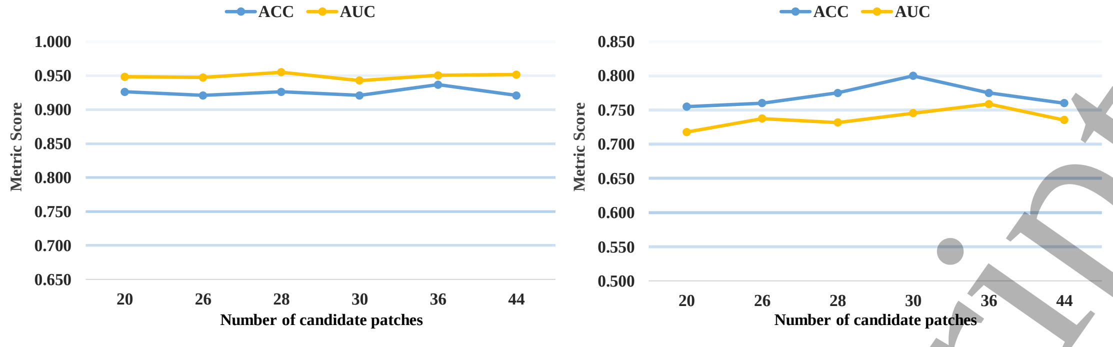
subnet for efficiently representing 3D patches in a novel divide-and-conquer manner and global patch subnet for fusing local-to-global representations.

3.3.2 *Effectiveness of IDP algorithm*

As introduced in Section 2.4, a key component of our proposed method is the IDP strategy to iteratively prune uninformative patch-level subnetworks, and ultimately boosting the diagnostic performance. To validate the effectiveness of IDP algorithm, we first have an insight into the performance of network during iteration processing. As Table 4 shown, we can observe that the ACC of CvHF-net fluctuates during the iteration process before falling into the optimal solution (the ACC outperforms those at subsequent 3 iterations or the number of iteration reaches to 9). We tail after the distribution of the remaining instances after each pruning, and find that some instances are reappear for current optimal network, even though they were pruned at previous certain pruning process, which indicates that reversible pruning is crucial for discriminative region identification. On the other hand, the selected regions (regardless of number and location) are different for AD classification and MCI conversion prediction task, which is practically reasonable due to distribution discrepancy of pathology regions for these two tasks. Also, it demonstrates that localizing these specific and discriminative regions in a task-driven and adaptive manner is significantly meaningful for AD/MCI diagnosis. We further compare the performance of CvHF-net with- and without-IDP. In Figure 5, we can see that CvHF-net with IDP outperforms the CvHF-net without IDP with significant improvement for AD classification (improved 5.3%, 12.0% and 2.2% in ACC, SPE and AUC, respectively) for MCI conversion prediction (improved 3.5%, 6.2%, 1.7% and 2.4% in ACC, SEN, SPE and AUC, respectively). Overall, IDP can efficiently and adaptively locate disease-related regions in a task-driven manner and boost the performance of CvHF-net based on the prior knowledge that derived from the former iteration.

3.4 *Influence of the number of candidate patches*

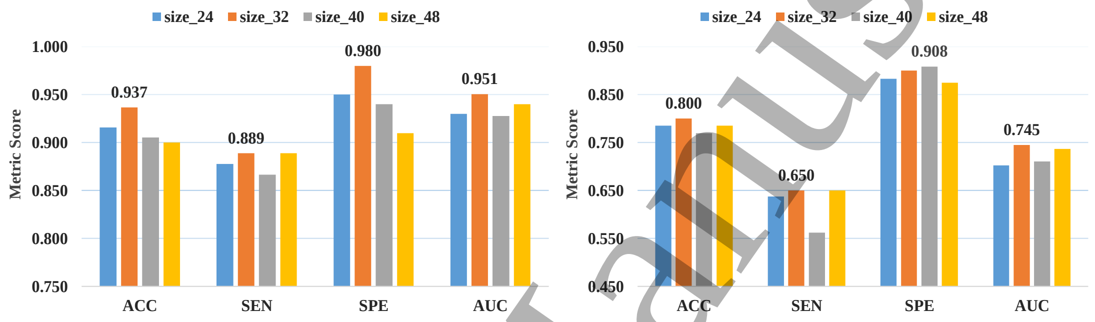
In the previous experiments, we use the candidate patches with the fixed number (i.e., 36 and 30 for AD classification and MCI conversion prediction task, respectively) to train CvHF-net. To investigate the influence of the number of the candidate patches, we compare a set of parameter settings (i.e. 20, 26, 28, 30, 36, 44) based on the same partition of dataset. The experimental results are shown in Figure 6. The results indicate that CvHF-net achieves a relatively stable performance for AD classification task (ACC ranges from 0.921 to 0.937; AUC ranges from 0.943 to 0.957). But the performance is fluctuating (ACC ranges from 0.755 to 0.800; AUC ranges from 0.728 to 0.759) for MCI conversion prediction task.



(a) AD *vs.* NC

(b) pMCI *vs.* sMCI

Figure 6: Influence of the number of candidate patches to the proposed CvHF-net in terms of accuracy (ACC) and the area under the curve of receiver operating characteristic (AUC) for AD classification (a) and MCI conversion prediction (b), respectively.



(a) AD *vs.* NC

(b) pMCI *vs.* sMCI

Figure 7: Influence of the size of image patches on the performance of the proposed CvHF-net in terms of four metrics, including accuracy (ACC), sensitivity (SEN), specificity (SPE), as well as the area under the curve of receiver operating characteristic (AUC) for AD vs NC classification (a) and pMCI vs sMCI classification (b), respectively.

A possible reason is that the structural difference on MR images for pMCI *vs.* sMCI task is much subtler than those for AD *vs.* NC task, which indicates that the MCI conversion prediction task is more difficult than AD classification task [27].

3.5 Influence of the size of patches

In the previous experiments, the size of patches is fixed as 32 ×32 ×32. We set a group of parameters to study the influence of the patch size, varying the size of 24 ×24 ×24, 32 ×32 ×32, 40 ×40 ×40, and 48 ×48 ×48. The results are showed in Figure 7, from which we can observe that the CvHF-net with the patch size of 32 × 32 × 32 achieves the optimal performance for both of AD classification and MCI conversion prediction tasks. By contrast, the CvHF-net using patches with relatively larger sizes (i.e., 40 × 40 × 40, and 48 × 48 × 48) obtains slightly inferior performance, as more redundant information is included in the large patch and affects the subtle brain changes identification [37]. On the other hand, the performance of CvHF-net using small patch (i.e., 24 × 24 × 24) is also decreased. It may due to less information contained in a small patch.

4. Discussion

4.1 Compare with previous work

Different from conventional brain morphometric analysis methods [9-14, 43] using manually-engineered imaging features, the proposed CVHF-Net can automatically learn high-nonlinear features, which can be seamlessly integrated for classifier construction. Also, different from the existing patch-level methods [22, 25, 27-29] which adopted 3D patches as input, the proposed CVHF-Net utilizes multiple 2D views from 3D patches to capture the local-to-global representation. Specifically, the local patch subnet first utilizes multiple 2D views to represent 3D patches using the attention-aware hierarchical fusion structure in a divide-and-conquer manner. Since different local regions are with various abilities in AD identification, the global subject-net is developed to bias the allocation of available resources towards the most informative parts among these local regions to obtain global information for AD identification. In addition, since not all candidate patches extracted from an MR image are significantly affected by Alzheimer's disease so that hampers the diagnostic performance, the IDP algorithm is introduced to train the proposed CvHF-net for adaptively localizing discriminant regions in a task-driven manner and removing the uninformative patches, resulting in reducing the computation consumption.

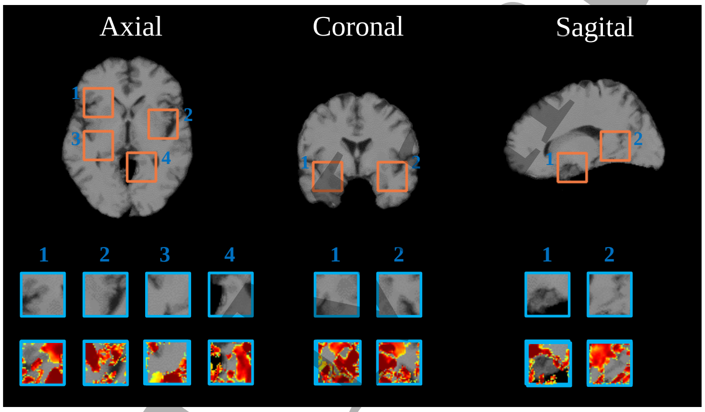


Figure 8. Discriminative disease-related regions identified by our proposed method in the task of AD diagnosis. The first to third rows correspond to the identified location yielded by the proposed CVHF-Net, the image patches and the corresponding p-value map, respectively.

4.2 Discriminative Disease-Associated Regions

The proposed CVHF-Net have the potential capacity in identifying features with diagnostic power by adopting the IDP algorithm in the training stage, upon which the uninformative patches with respect to the top instance-wise cross-entropy loss were pruned iteratively and the remaining patches were fed into local patch subnet and global subject-net for AD/MCI diagnosis. In Figure 8, we visually present the identified discriminative AD-related region by the proposed methods in the task of AD diagnosis, upon which the first to third rows correspond to the identified location yielded by the proposed CVHF-Net, the image patches and the corresponding p-value map, respectively. From the first row of Figure 8, we can observe that the proposed method revealed the discriminative regions with diagnostic power in temporal lobes and insula, which is in line with the previous work [29, 37], suggesting the rationality of the IDP algorithm. In another, from the second and third row of Figure 8, we can observe that, regardless of the information deficiency in extracting multiple 2D views for AD/MCI diagnosis, the proposed CVHF-Net still can explore the discriminative disease-associated regions effectively.

4.3 Limitations

Although our proposed CvHF-Net achieves promising results in both AD classification and MCI conversion prediction, there are several technical issues to be addressed in the future. First, preliminary landmark detection is conducted according to the p-value map computed by group comparison, which is a standalone task and achieves consistent localization. However, pathological and anatomical atrophy of the brain varies greatly among patients, which may lead to sub-optimal learning performance. Future works will try to integrate the process of landmark detection and the training of classification models into a unified framework to avoid uncertainty caused by the discrepancy of brain atrophy lesion. Second, in the present IDP strategy, the correlation among landmarks, which relates to the topological information of brain structure may be neglected. As such, topological learning can be embedded in IDP in the future. Third, we could extend our proposed model for the prediction of brain disease progression, a more challenging task compared with disease diagnosis. Furthermore, our work only considers the problem of AD diagnosis via the proposed CvHF-net based on the baseline MRI data. It is interesting to develop a deep-learning framework for predicting the longitudinal progression of AD.

5. **Conclusion**

In this paper, a light-weight cross-view hierarchical fusion network (CvHF-net), consisting of local patch and global subject subnets, is proposed to perform adaptive localization and identification with instance-declined pruning (IDP) for AD diagnosis and MCI conversion prediction. Experimental results on the ADNI dataset demonstrate the effectiveness of our proposed model on joint discriminative localization and disease diagnosis. In the future, we will extend our proposed model for brain disease progression prediction using longitudinal data.

Reference

[1] 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.

[2] 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.

[3] Alzheimer's Association, 2019. 2019 alzheimer’s disease facts and figures. Alzheimer’s & Dementia 15, 321–387.

[4] 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.

[5] Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., 2012. The alzheimer’s disease neuroimaging initiative: a review of papers published since its inception. Alzheimers & Dementia the Journal of the Alzheimers Association 8.

[6] Liu, S., Cai, W., Liu, S., Zhang, F., Fulham, M., Feng, D., Pujol, S., Kikinis, R., 2015. Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders. Brain Informatics 2, 167–180.

[7] Tanveer, M., Richhariya, B., Khan, R.U., Rashid, A.H., Lin, C.T., 2020. Machine learning techniques for the diagnosis of alzheimer’s disease: A review. Acm Transactions on Multimedia Computing Communications & Applications 16, 35.

[8] Khan, A., Usman, M., 2016. Early diagnosis of alzheimer's disease using machine learning techniques: A review paper, in: International Joint Conference on Knowledge Discovery.

[9] 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.

[10] Xiang, S., Yuan, L., Fan,W.,Wang, Y., Thompson, P.M., Ye, J., 2014. Bi-level multi-source learning for heterogeneous block-wise missing data. NeuroImage 102, 192–206.

[11] Koikkalainen, J., Lötjönen, J., Thurfjell, L., Rueckert, D., Waldemar, G., Soininen, H., 2011. Multi-template tensor-based morphometry: Application to analysis of alzheimer's disease. NeuroImage 56, 1134– 1144.

[12] Liu, M., Zhang, D., Shen, D., 2016. Relationship induced multitemplate learning for diagnosis of alzheimer's disease and mild cognitive impairment. IEEE Transactions on Medical Imaging 35, 1463–1474.

[13] 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.

[14] 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.

[15] Kabani, N.J., Macdonald, D.J., Holmes, C.J., Evans, A.C., 1998. 3d anatomical atlas of the human brain. NeuroImage 7.

[16] LeCun Y, Bengio Y, H.G., 2015. Deep learning. Nature 521, 436–444.

[17] Selvikvåg Lundervold, A., Lundervold, A., 2018. An overview of deep learning in medical imaging focusing on mri. Zeitschrift Für Medizinische Physik.

[18] Li, H., Habes, M., Fan, Y., 2017. Deep ordinal ranking for multi-category diagnosis of alzheimer's disease using hippocampal mri data. ArXiv.

[19] Yang, Y., Yan, L.F., Zhang, X., Han, Y., Nan, H.Y., Hu, Y.C., Hu, B., Yan, S.L., Zhang, J., Cheng, D.L., Ge, X.W., Cui, G.B., Zhao, D., Wang,W., 2018. Glioma grading on conventional mr images: A deep learning study with transfer learning. Frontiers in Neuroscience 12, 804.

[20] Khvostikov, A., Aderghal, K., Benois-Pineau, J., Krylov, A.S., Catheline, G., 2018. 3d cnn-based classification using smri and MD-DTI images for alzheimer disease studies. ArXiv abs/1801.05968.

[21] Wang, H., Shen, Y., Wang, S., Xiao, T., Deng, L., Wang, X., Zhao, X., 2019. Ensemble of 3d densely connected convolutional network for diagnosis of mild cognitive impairment and alzheimer's disease. Neurocomputing 333, 145 – 156.

[22] Suk, H.I., Lee, S.W., Shen, D., 2014. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. NeuroImage 101, 569 – 582.

[23] Salakhutdinov, R., Larochelle, H., 2010. Efficient learning of deep boltzmann machines. Journal of Machine Learning Research - Proceedings Track 9, 693–700.

[24] Liu, M., Zhang, D., Shen, D., 2014. Hierarchical fusion of features and classifier decisions for alzheimer's disease diagnosis. Human Brain Mapping 35, 1305–1319.

[25] 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.

[26] Koritakova, E., Vounou, M., Wolz, R., Gray, K., Rueckert, D., Montana, G., 2012. Biomarker discovery for sparse classification of brain images in alzheimer's disease. Annals of the BMVA 2012, 1–11.

[27] 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.

[28] 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.

[29] Lian, C., Liu, M., Zhang, J., Shen, D., 2020. 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.

[30] [Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S.J., Wille, M.M.W., Naqibullah, M., Sánchez, C.I., van Ginneken, B., 2016. Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks. IEEE Transactions on Medical Imaging 35, 1160–1169.

[31] Luo, J., Ning, Z., Zhang, S., Feng, Q., Zhang, Y., 2018. Bag of deep features for preoperative prediction of sentinel lymph node metastasis in breast cancer. Physics in Medicine & Biology 63, 245014.

[32] Ben-Ahmed, O., Lecellier, F., Paccalin, M., Fernandez-Maloigne, C., 2017. Multi-view visual saliency-based mri classification for alzheimer's disease diagnosis, in: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6.

[33] 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.

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

[35] 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.

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

[37] 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.

[38] Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.

[39] Ning, Z., Luo, J., Li, Y., Han, S., Feng, Q., Xu, Y., Chen, W., Chen, T., Zhang, Y., 2019. Pattern classification for gastrointestinal stromal tumors by integration of radiomics and deep convolutional features. IEEE Journal of Biomedical and Health Informatics 23, 1181–1191.

[40] 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.

[41] Wang, L., Shi, F., Li, G., Shen, D., 2012. 4d segmentation of longitudinal brain mr images with consistent cortical thickness measurement, pp. 63–75.

[42] 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.

[43] J Shi, X Zheng, Y Li, Q Zhang, S Yin, 2018. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE journal of biomedical and health informatics 22 (1), 173-183.