MRI-based Alzheimer's disease prediction via distilling the knowledge in multi-modal data

Hao Guan [a], Chaoyue Wang [a,*], Dacheng Tao [a,b]

[a] *School of Computer Science, The University of Sydney, Australia*
[b] *JD Explore Academy, China*

ARTICLE INFO

ABSTRACT

Mild cognitive impairment (MCI) conversion prediction, *i.e.,* identifying MCI patients of high risks converting to Alzheimer's disease (AD), is essential for preventing or slowing the progression of AD. Although previous studies have shown that the fusion of multi-modal data can effectively improve the prediction accuracy, their applications are largely restricted by the limited availability or high cost of multi-modal data. Building an effective prediction model using only magnetic resonance imaging (MRI) remains a challenging research topic. In this work, we propose a multi-modal multi-instance distillation scheme, which aims to distill the knowledge learned from multi-modal data to an MRI-based network for MCI conversion prediction. In contrast to existing distillation algorithms, the proposed multi-instance probabilities demonstrate a superior capability of representing the complicated atrophy distributions, and can guide the MRI-based network to better explore the input MRI. To our best knowledge, this is the first study that attempts to improve an MRI-based prediction model by leveraging extra supervision distilled from multi-modal information. Experiments demonstrate the advantage of our framework, suggesting its potentials in the data-limited clinical settings.

1. Introduction

Mild cognitive impairment (MCI) is a transitional stage between cognitively normal and dementia (Petersen, 2004). In each year, around 10%-15% of MCI patients convert to AD (Mitchell and Shiri-Feshki, 2008), while the other MCI patients tend to remain stable, develop other forms of dementia, or even revert to normal cognition (Pandya et al., 2016). The therapeutic interventions are thought to be feasible for MCI before it progressing to the irreversible AD (Carretti et al., 2013; Sherman et al., 2017; Yao et al., 2020). Therefore, it is of great importance to identify the MCI patients who have high risk of converting to AD, allowing early treatments to prevent or delay the progression of the disease.

The AD process is associated with plaques and neurofibrillary tangles in the brain (Ballard et al., 2011). AD-related neuropathology can be identified via a series of factors several years before AD clinical manifestation (Ewers et al., 2011; Hampel et al., 2018; Reiman et al., 2010; Teipel et al., 2018). For example, MRI detects grey matter volume changes caused by gross neuronal loss and atrophy (Ewers et al., 2011). And the E4 variant of apolipoprotein E (APOe4) is the main susceptibility gene for Alzheimer's disease (Corder et al., 1993; Genin et al., 2011; Roses, 1996). Multi-modal data, including imaging and non-imaging clinical data, provide rich and complementary information about the disease. Previous studies have identified some genetic basis of phenotypic neuroimaging markers (Shen et al., 2010), and brain imaging has revealed correlations between brain structure and cognition in AD and MCI (Braskie and Thompson, 2013). Based on these findings, extensive computer-aided diagnosis (CAD) studies have demonstrated that employing multi-modal data can largely improve the performance of AD diagnosis and MCI conversion prediction (Liu et al., 2017; Moradi et al., 2015; Qiu et al., 2020; Spasov et al., 2019; Zhou et al., 2019; Zhu et al., 2019). However, it is difficult to meet the setting of multi-modal data in practice, because the medical data of some certain modalities are of high cost or limited availability. The requirement of multi-modal data restricts the applicability of the above multi-modal based CAD methods (Liu et al., 2017; Moradi et al., 2015; Qiu et al., 2020; Spasov et al., 2019; Zhou et al., 2019; Zhu et al., 2019).

MRI is non-invasive without using ionizing radiation. The comparatively low cost makes MRI be widely applied in the clinical practice. In recent years, MRI-based CAD methods have achieved promising performance in AD diagnosis (Bron et al., 2015; Lian et al., 2018; Liu et al., 2016; 2018). However, since the disease-related patterns in MCI brains are more subtle than those in AD brains (Driscoll et al., 2009), MCI conversion prediction appears to be a more challenging task (Lian et al., 2018; Liu et al., 2016; 2018; Pan et al., 2020). Compared with multi-modal based CAD methods (Lin et al., 2018; Moradi et al., 2015;
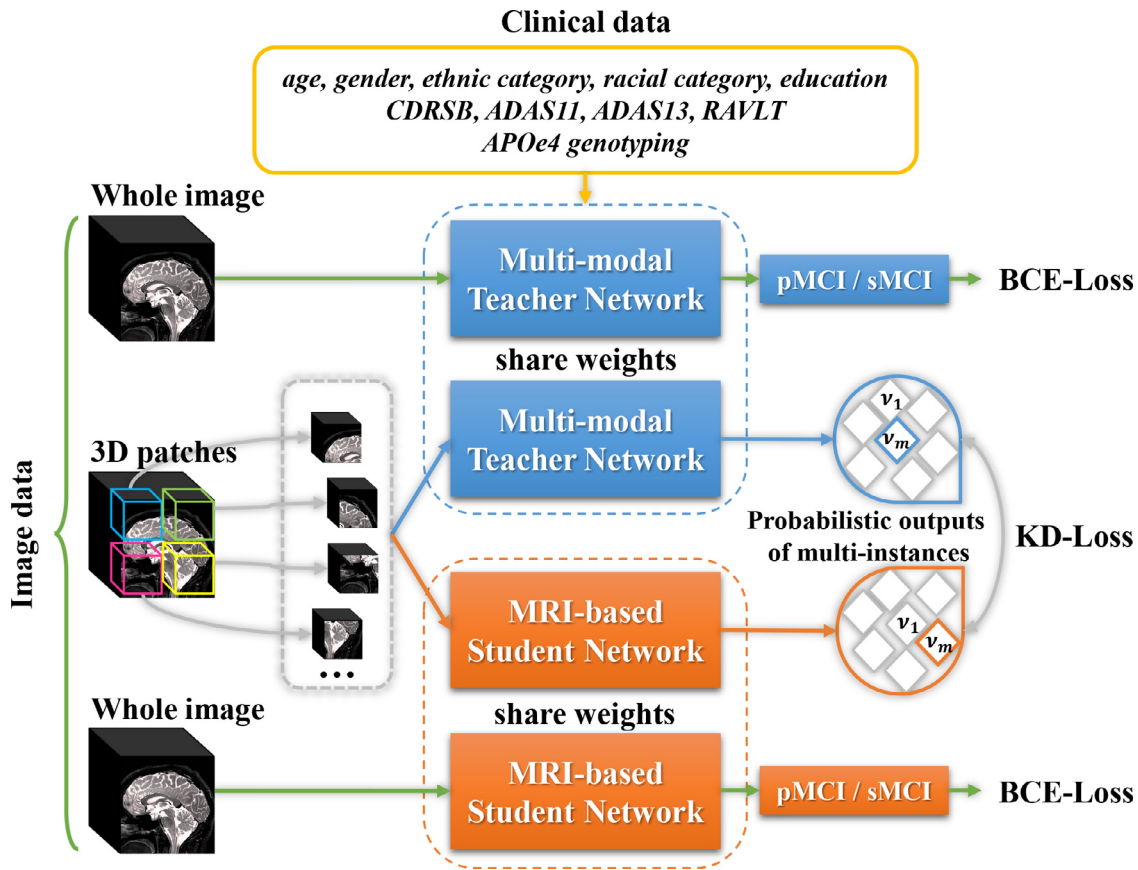
---

* Corresponding author.

E-mail addresses: chaoyue.wang@sydney.edu.au (C. Wang), dacheng.tao@sydney.edu.au (D. Tao).

Fig. 1. The schematic of the proposed framework. The teacher network takes MRI and clinical data as input. The student network is only fed with MRI images. Knowledge is encoded and transferred in the form of probabilistic outputs of multi-instances. Abbreviations: APOe4, Apolipoprotein E; CDRSB, Clinical Dementia Rating Sum of Boxes; ADAS, Alzheimer's disease Assessment Scale; RAVLT, Ray Auditory Verbal Learning Test; BCE-Loss: binary cross-entropy loss; KD-Loss: knowledge distillation loss.

Pan et al., 2020; Spasov et al., 2019), existing MRI-based methods usually perform worse in MCI conversion prediction (Lian et al., 2018; Liu et al., 2016; 2018). This suggests that when merely using MRI data, extracting disease-related features from MCI brain images is difficult. To achieve better prediction performance, a CAD algorithm has to extract fine-grained features that contain more information about the disease patterns and less disease-irrelevant information.

Knowledge distillation is typically used to transfer knowledge from a larger teacher network to a small student network (Gou et al., 2020; Hinton et al., 2015; Ruffy and Chahal, 2019). Similarly, the knowledge learned by a teacher from privileged data can also be transferred to a student without accessing that data (Lopez-Paz et al., 2016; Vapnik and Izmailov, 2015). For a classification task, the relative probabilities of incorrect classes provide the knowledge that guides the student network to generalize (Hinton et al., 2015). By minimizing the Kullback Leibler (KL) divergence between the outputs of a teacher and a student, the student receives extra supervision beyond the conventional training loss. However, for a binary classification problem like MCI conversion prediction, the output probabilities of a teacher network are not sufficiently informative for a student to mimic (Shen et al., 2016), since the relationship between classes is not leveraged (Tang et al., 2020). Meanwhile, the gap between data of different modalities raises the challenge of how to transfer knowledge from a multi-modal teacher network to an MRI-based student network.

In this paper, instead of directly using the multi-modal data for MCI conversion prediction, we propose a novel multi-modal multi-instance distillation (M3ID) method for achieving better MRI-based prediction. The schematic diagram of our framework is shown in Fig. 1. We perform knowledge transfer between a multi-modal teacher network (using

multi-modal data as input) and a single-modal student network (using only MRI as input). Specifically, multi-modal data include MRI and non-imaging clinical data. Different from existing distillation methods, we devise a probabilities matching scheme regarding multi-instances: we use the KL-Divergence to reduce the distance between the teacher's and the student's probabilistic estimates of multi-instances. Specifically, we partition each MRI scan into multiple 3D image patches, which make individual instances. The probability estimates of multi-instances are obtained by feeding the networks with image patches, and used as the extra supervision for the student network. Thus, the student is guided to mimic the teacher's modeling of subtle disease patterns and achieve better prediction performance. Overall, we forge a connection between *knowledge distillation* and *multi-instance learning*. Based on this connection, we are able to leverage knowledge learned from multi-modal data to improve the MRI-based model's performance. We conducted experiments on three datasets for MCI conversion prediction. Experimental results demonstrate our framework's superior prediction performance.

Our contributions are summarized as follows:

1. This is the first study that aims to improve the MRI-based prediction of MCI conversion by leveraging extra supervision distilled from multi-modal information;
2. Our work facilitates the clinical practices where multi-modal data are available for model training and only MRI data are used for testing;
3. We present a multi-instance-based knowledge distillation scheme, and the experiments on three independent datasets demonstrate the advantage our proposed scheme;

4. Our framework can generate heat-maps on the basis of image patches for visual interpretation.

The rest of the paper is organized as follows. In Section 2, we review related works. In Section 3, we introduce the studied datasets, and our new method. In Section 4, we evaluate the proposed method. In Section 5, we compare our method with previous methods and discuss the limitations of the current study. In Section 6, we conclude this work.

2. Related work

2.1. Knowledge distillation

Knowledge distillation aims to train a student network under the guidance of a trained teacher network. With additional knowledge beyond the usual data supervision (*e.g.,* annotations), the student can be better optimized. The pioneering knowledge distillation work proposed by Hinton et al. (2015) minimizes the Kullback-Leibler (KL) divergence between output probabilities of a teacher and a student network. The relative probabilities of incorrect classes convey useful information about how a teacher tends to generalize. This distillation strategy is simple yet effective, and has been used in various applications (Gou et al., 2020). Some other methods transfer knowledge by representation mimicking. FitNets (Romero et al., 2014) matches the full feature maps between a teacher and a student. Attention transfer (Zagoruyko and Komodakis, 2017) attempts to match the spatial attention maps. Apart from directly matching a teacher's representation, similarity preserving (Tung and Mori, 2019) encourages a student to preserve the pairwise activation similarities derived by a teacher network. Representation-based distillation methods bypass the problem of few output classes conveying limited knowledge. Although these representation-based methods have achieved reasonable performance, their effectiveness has not been validated when representations are generated from data of very different modalities.

Cross-modal knowledge transfer is an arising research topic of knowledge distillation. Given a teacher network trained on the source modality, the teacher's knowledge is transferred to a student network applied on a different modality (Gupta et al., 2016; Passalis and Tefas, 2018). Knowledge transfer among different modalities enables effective utilizing of complementary information provided by multi-modal data. Do et al. (2019) propose a visual question answering method, in which a teacher model learns trilinear interactions from image-question-answer inputs, then transfers knowledge to a student model that learns bilinear interactions from image-question inputs. Cross-modal distillation fits the scenarios where certain modalities of data or labels are not available during the training or test stage. Dou et al. (2020) leverage cross-modal distillation to address the unpaired multi-modal segmentation task. Li et al. (2020) exploit the modality-shared knowledge to facilitate image segmentation of the target-modality. Based on the above studies, we investigate the knowledge transfer from multi-modal data to a single-modal disease prediction model.

2.2. Deep learning in AD diagnosis

Recently, many deep learning-based methods are proposed for AD diagnosis. MRI has the advantages of non-invasively capturing disease-related pathological patterns and common availability. Thus, a relatively large amount of MRI data can be used for training deep learning-based AD diagnosis models. Liu et al. (2018) exploit anatomical landmarks to develop an MRI-based deep learning model for AD classification and MCI conversion prediction. Lian et al. (2018) propose an AD diagnosis network with integration of discriminative atrophy localization, feature extraction and classifier construction. Lee et al. (2019) aim to build an interpretable diagnosis model by extracting regional abnormality representation from MRI. Despite all these efforts, multi-modal methods have always demonstrated a better diagnostic performance than the

single-modal MRI-based methods. Spasov et al. (2019) devise a light-weight network that takes a combination of MRI, cognitive measurements, genetic, and demographic data as input, and achieved impressive results on MCI conversion prediction. Qiu et al. (2020) utilize multi-modal inputs of MRI, age, gender, and Mini-Mental State Examination score for training a deep learning model, and demonstrate the model's effectiveness of AD diagnosis on several datasets. Pan et al. (2020) propose a diagnosis framework utilizing MRI and PET data, and achieve promising results. However, these multi-modal methods seem to under-utilize the MRI data, since the performance with the data other than MRI is already high. In practice, multi-modal medical data are hard to get, leaving room for further improvement using only MRI data.

2.3. Multi-instance learning

Multi-instance learning (MIL) tries to solve a weakly supervised learning problem that each labelled bag in a training set contains multiple unlabelled instances (Amores, 2013; Carbonneau et al., 2018). The goal of MIL is to predict the class labels of unseen bags. Under the MIL setting, a positive bag contains at least one positive instance, while all instances in a negative bag are negative. Compared with fully supervised learning methods, MIL reduces the annotation cost and fits many real-life applications. For example, by leveraging weakly labelled training videos, a deep multiple instance ranking framework is proposed for anomaly detection in surveillance videos (Sultani et al., 2018). A multi-instance deep learning method is developed to discover discriminative local anatomies for bodypart recognition (Yan et al., 2016). In the domain of medical image analysis, obtaining pixel-level annotation costs very much. With MIL and the weak supervision of per image diagnosis, a large quantity of medical images can be used for model training (Kandemir and Hamprecht, 2015). An attention-based MIL method presented in (Ilse et al., 2018) achieves promising results on cancer diagnosis via histology image. A landmark-based deep MIL framework use MRI data (Liu et al., 2018) for brain disease diagnosis. These studies suggests that multi-instance representations leverage rich information derived from data. In the present study, instead of using MIL for producing the final diagnosis, we leverage multi-instance probabilities as the medium for knowledge distillation.

3. Material and methods

In this section, we first introduce the datasets and MRI pre-processing pipeline used in this study (Section 3.1). Then we present the proposed multi-modal multi-instance distillation scheme including the mathematical formulation of model (Section 3.2), the network architectures of the proposed multi-modal teacher and the MRI-based student (Section 3.3), the novel multi-instance distillation method (Section 3.4), and the distillation loss and implementation details (Section 3.5–3.6).

3.1. Datasets and image pre-processing

We collected data from three public datasets, ADNI-1, ADNI-2 (Jack Jr. et al., 2008), and the Australian Imaging, Biomarkers and Lifestyle (AIBL) database (Ellis et al., 2009)[1]. The ADNI-3 dataset was not used because it hadn't release long enough follow-ups. It is worth noting that all subjects in ADNI-2 were newly enrolled, so there are no subjects appeared in both ADNI datasets. Only baseline data were used in this study.

---

[1] Data used in this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (adni.loni.usc.edu) and the Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL) database (www.aibl.csiro.au). As such, the investigators within the ADNI and AIBL contributed to the design and implementation of ADNI and AIBL 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: online

Table 1
Demographic characteristics of the studied subjects.

Dataset	Category	No. of subjects	Age Range	Females/Males
ADNI-1	sMCI	100	57.8–87.9	34/66
	pMCI	164	55.2–88.3	66/98
	NC	226	59.9–89.6	109/117
	AD	179	55.1–90.9	87/92
ADNI-2	sMCI	115	55.0–88.6	48/67
	pMCI	76	56.5–85.9	36/40
	NC	186	56.2–89.0	96/90
	AD	140	55.6–90.3	56/84

All T1-weighted images were pre-processed using FSL's anatomical processing pipeline[2] (Jenkinson et al., 2012). First, all images were re-oriented to the standard space (Montreal Neurological Institute, MNI) and automatically cropped before bias-field correction. Next, the corrected images were registered non-linearly using the MNI T1 Template (MNI152_T1_1mm) (Grabner et al., 2006). Then, the skull was stripped using the FNIRT-based approach (Jenkinson et al., 2012), and the brain-extracted images were regenerated to a standard space of size $182 \times 218 \times 182$. Processed images had an identical spatial resolution ($1mm \times 1mm \times 1mm$). To further reduce the redundant computational cost, each image was cropped to $144 \times 176 \times 144$ (see Fig. S4). The 3D images were standardized to zero mean and unit variance before feeding into the network. Pre-processed images were manually checked, and the images with insufficient stereotaxic registration and insufficient skull stripping were excluded.

Clinical data employed in this study consisted of 13 scalar variables: age, gender, ethnic category, racial category, education years, Clinical Dementia Rating Sum of Boxes (1 variable), Alzheimer's disease assessment scale (2 variables: ADAS11, ADAS13), episodic memory evaluations in the Rey Auditory Verbal Learning Test (4 variables: immediate, learning, forgetting, percent forgetting), as well as APOe4 genotyping (1 variable). All clinical data used in this study were from baseline assessments. The clinical data were standardized to zero mean and unit variance before feeding into the network. The mean and standard deviation on the training set were used to apply the standardization on the validation and test set. The detailed clinical information of ADNI subjects was reported in Table S1 of the *Supplementary Materials*.

ADNI-1: The baseline ADNI-1 dataset used in this study consisted of 1.5T T1-weighted MRI images scanned from 669 subjects. According to standard clinical criteria, subjects were divided into three groups: normal control (NC), MCI, and AD. According to whether MCI subjects converted to AD within 36 months, MCI subjects were further categorized into: 1) sMCI (stable MCI), if diagnosis was MCI at all available time points (0–96 months), and had diagnosis records for at least 36 months; 2) pMCI (progressive MCI), if diagnosis was MCI at baseline but the conversion to AD was reported after baseline within 3 years, and without a reversion to MCI or NC at any available follow-up (0–96 months). The other MCI subjects who had the last diagnosis reported after baseline within 2 years, or had unstable diagnoses (reversion, or conversion after 3 years) were not included in this study. The dataset contained 100 sMCI, 164 pMCI, 226 NC, and 179 CE subjects The demographic information of the ADNI-1 subjects is presented in Table 1.

ADNI-2: The baseline ADNI-2 dataset used in this study contained 3T T1-weighted MRI images scanned from 517 subjects. The same clinical criteria as in ADNI-1 were used to separate the subjects into 115 sMCI, 76 pMCI, 186 NC, and 140 CE subjects (see Table 1).

AIBL: The AIBL dataset used in this study contained 3T T1-weighted MRI images scanned from 37 subjects, which consists of 25 sMCI and 12 pMCI. The demographic characters of the AIBL subjects were presented in Table S1 of the *Supplementary Materials*.

3.2. Mathematical formulation of model

Let $\{(\boldsymbol{X}_n, \boldsymbol{y}_n)\}_{n=1}^N$ be a training set containing N samples, where $\boldsymbol{X}_n$ denotes input data from the nth subject, and $\boldsymbol{y}_n$ denotes the corresponding class label (0 for sMCI, 1 for pMCI). The MCI conversion prediction task is formed as a classification problem, which can be formulated by estimating conditional probabilities of the target variable $\boldsymbol{y}$, given the input data $\boldsymbol{X}$:

$$p(\boldsymbol{y}|\boldsymbol{X};\theta), \tag{1}$$

where θ represents the parameters of the neural network. The goal is to learn the network parameter so that the true conditional probabilities can be approximated.

3.3. Multi-modal teacher & single-modal student

Given the scarcity and high cost of multi-modal data, we propose to improve an MRI-based prediction model by leveraging knowledge distillation. The complementary knowledge provided by multi-modal data is extracted by a teacher network. Then knowledge distillation is performed between the teacher network and a student network. After training, the student network is expected to effectively make disease prediction with only MRI data. Our prediction framework consists of a multi-modal attention network as the teacher, and an MRI-based network as the student. The details of the network architectures are given as follows.

3.3.1. Student network

The student network is based on ResNet (He et al., 2016a) using residual block proposed in (He et al., 2016b). The network consists of one root convolutional layer and three residual blocks (Fig. 2). Feature down-sampling is achieved via strided convolutions. After three residual blocks, the feature map is down-sampled for 8 times. Then two double attention block (A^2N) (Chen et al., 2018) are cascaded on top of the network to efficiently model the long-range inter-dependencies (Burkov and Lempitsky, 2018). In the task of MRI-based MCI conversion prediction, the long-range inter-dependencies could be the spatial relationships between different areas of the brain. The following layers consist of a global average pooling layer, a dropout layer, and two fully connected layers. As shown in Fig. 2, the MRI-based student network takes a whole MRI scan as input, and aims to predict whether an MCI patient would convert to AD. Although existing MRI-based methods have achieved reasonable prediction results, there is still a big performance gap between them and multi-modal methods. Next we introduce a teacher network that uses multi-modal data.

3.3.2. Teacher network

The teacher network utilizes an MRI feature extractor with the same architecture as the student network. In addition, a clinical feature extractor and a multi-modal attention module are introduced to build the whole teacher network, *i.e.,* a multi-modal attention network. The learned clinical features are utilized to guide the generation of attention maps. Next, the learned attention maps are used to weightedly average the visual features and generate the aggregated features. The network architecture is inspired by (You et al., 2019). The diagram of the teacher network is shown in Fig. 3. The detailed definitions are given as follows.

The extracted visual features $\boldsymbol{I}$ are with spatial resolution of $H \times W \times D$ and C channels. Firstly, we use two $1 \times 1 \times 1$ convolutions to project the visual features to the clinical space. The projected features are denoted as $\boldsymbol{I}_e \in \mathbb{R}^{H \times W \times D \times E}$, where E represents the dimension of the clinical features. Following (Qiu et al., 2020; Spasov et al., 2019), the clinical data are passed through a sequence of two fully-connected layers to produce clinical features $\boldsymbol{J} \in \mathbb{R}^E$. The multi-modal attention map is generated by computing the cosine similarity between the projected
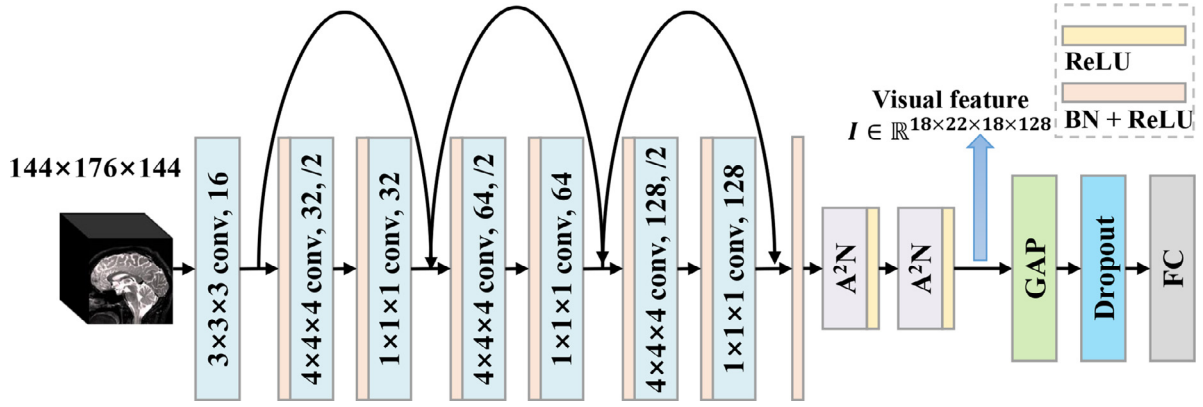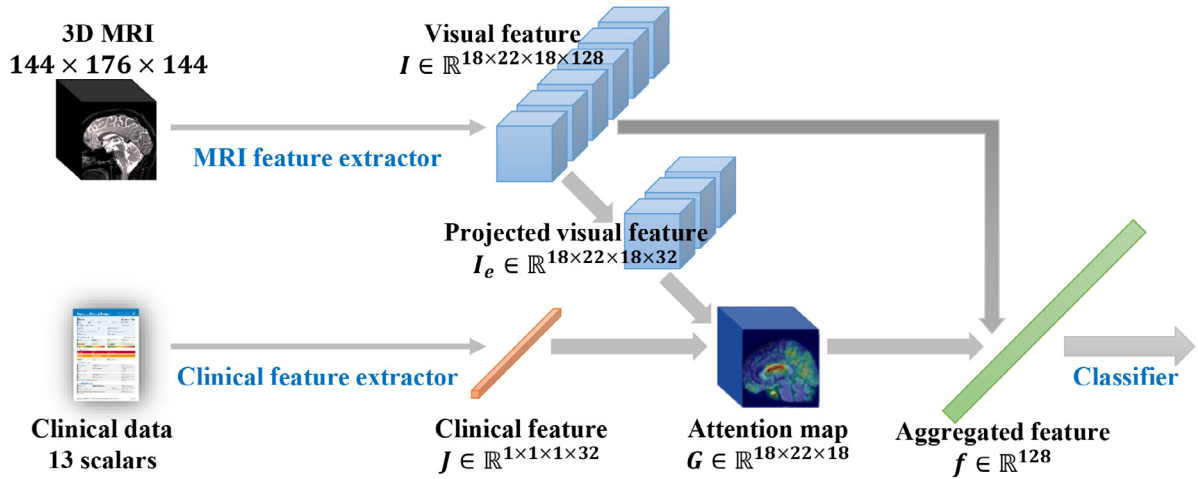
Fig. 2. Architecture of the MRI-based student network.

Fig. 3. Diagram of the multi-modal attention network, *i.e.,* the teacher network. The visual features are extracted and projected to the clinical space. The learned clinical features and the projected visual features are used to generate the attention maps. Then the visual features are weightedly averaged using the attention maps to generate the aggregated features.

visual feature and I_e^l and clinical feature J at location l.

$$g^l = ReLU \frac{I_e^{lT} J}{\left\| I_e^l \right\| \left\| J \right\|}. \qquad (2)$$

The attention map $G = [g^1, \ldots, g^{HWD}]$ is normalized by dividing the sum of attentive values across all space locations. The normalized attention map $U = [u^1, \ldots, u^{HWD}]$ is used to combine visual features, *i.e.,*

$$f = \sum_{l=1}^{HWD} u^l \mathbf{I}^l. \qquad (3)$$

The attention-weighted features f are passed to two fully-connected layers for generating predictions. The teacher network is able to extract the complementary information and cross-modal dependencies from multi-modal data for better prediction. We choose to use the multi-modal attention network as the teacher network, because it is light-weight and effective, and has an architecture similar to the student network. The teacher network is a combination of the student network and the multi-modal attention module, which fuses multi-modal features and helps the teacher network achieve promising performance. The multi-modal attention network is designed to be similar to the student network, since the architecture difference between student and teacher network might degrade the performance of knowledge transfer (Gou et al., 2020).

3.4. Multi-instance knowledge distillation

As aforementioned, the basic approach of knowledge distillation is to enforce a student to match the teacher's probabilistic outputs (Hinton et al., 2015). However, a binary classification model's outputs are not sufficiently informative to carry useful knowledge. A previous study has demonstrated that directly matching output probabilities does not offer enough guidance to a student with two output classes (Shen et al., 2016). To overcome this problem, we leverage multi-instance probabilities as a medium for knowledge distillation. A prediction network trained in a supervised manner is expected to learn a mapping from an input MRI to an output diagnostic label. When the network takes MRI patches as input, the outputs can be considered as the disease scores, which reflect the regional severities of the disease. The disease scores contain the information of how the network encodes the local regions.

Here, we employ patch-level disease scores to perform knowledge distillation. Specifically, we first partition the MRI into $M = N \times N \times N$ patches of size $x \times y \times z$, where x, y and z are the coronal, sagittal, and axial sizes of the patches, respectively. The patches partitioned from the image X make the individual instances in a bag. A network generates a disease score v ranging from 0 to 1 for each MRI patch (*i.e.,* instance). Then, softmax function is applied over a bag of disease scores $(v^1, v^2, \ldots, v^M)$ to produce the probabilistic estimates of multi-instances. The probability of instance m of image X given by the student and

teacher are computed as:

$$p_S^m = \frac{exp(v_S^m)}{\sum_{m=1}^M exp(v_S^m)}, \quad p_T^m = \frac{exp(v_T^m)}{\sum_{m=1}^M exp(v_T^m)}, \tag{4}$$

where p_S^m denotes the student's prediction, while p_T^m denotes the teacher's prediction. A bag of disease scores represent the disease-related status distributed across the brain. With softmax activation, the multi-instance probabilities can convey useful information about the relationship between patches. When enforcing the student to mimic the teacher's probabilistic estimates of multi-instances, the teacher guides the student to better localize and grasp the disease-related information within MRI data. Intuitively, the proposed distillation method works like data augmentation, while the targets of patches are provided by a teacher rather than the human-annotated ground truth.

3.5. Knowledge transfer and overall loss function

In our framework, the teacher's knowledge is encoded in the form of probabilities of multi-instances. We transfer the knowledge by minimizing the KL divergence between the two networks' probabilistic estimates of multi-instances, p_T and p_S. The knowledge distillation (KD) loss is defined as follows:

$$\mathcal{L}_{KD} = D_{KL}(p_T \| p_S) = \sum_{i=1}^N \sum_{m=1}^M p_T^m \log \frac{p_T^m}{p_S^m}. \tag{5}$$

In addition, a binary cross-entropy (BCE) loss is computed using the whole MRI as input and the corresponding conversion label as the target:

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{n=1}^N y_n \cdot \log p(\hat{y}_n | X_n; \theta) + \tag{6}$$
$$(1 - y_n) \cdot \log(1 - p(\hat{y}_n | X_n; \theta)),$$

where θ represents the parameters of the neural network, $\hat{y}_n$ represents the estimated target.

Then, the total loss for training the student network is defined as a summary of the BCE-loss and the KD-loss weighted by hyperparameter α:

$$\mathcal{L} = \mathcal{L}_{BCE} + \alpha \mathcal{L}_{KD}. \tag{7}$$

During the initial training process, the probability estimates of multi-instances of student are handicapped, since the student hasn't built a robust mapping from the input to the output. The distillation loss may not be very useful or even be counterproductive (Anil et al., 2018). Thus, we adopt the gradually increased supervision from the teacher network by changing α in Eq. 7,

$$\alpha_t = \frac{1}{2}\alpha(1 + \cos(\frac{T_{cur}}{T_{max}}\pi + \pi)), \tag{8}$$

where T_{cur} denotes the current number of epochs, and the T_{max} denotes the max training epochs. The subscript t means changing with time. By doing this, the student network could gradually receive guidance from the teacher network.

In addition to training the teacher and student sequentially, we investigate the Deep Mutual Learning (DML) (Zhang et al., 2018) setting where the teacher and the student learn collaboratively and teach each other throughout the training process. Mutually matching the probability estimates can increase each network's posterior entropy, and help them achieve better generalization (Zhang et al., 2018). With deep mutual learning, multi-modal knowledge is expected to be thoroughly exploited by the student network.

3.6. Implementation details

In the student network, the visual feature maps ($I \in \mathbb{R}^{18 \times 22 \times 18 \times 128}$) were processed with global average pooling to produce 128-dimensional features. The teacher network generated the 128-dimensional features with attention weighted combining. The next two fully-connected (FC)

layers had 128 and 16 units, respectively. The output layer had one unit for binary classification. Dropout with rate 0.4 was activated for the first FC layer. We used ELU (Clevert et al., 2016) activation for the FC layers, and Sigmoid activation for the output layer to generate disease scores. We evenly divided each MRI image into 8 non-overlapping patches of size $72 \times 88 \times 72$. Note that the patch size of each axis was set as multiples of 8 to ensure that no border pixel was discarded during the convolution process.

Our models were implemented with PyTorch (https://pytorch.org/). In addition, we defined our model in mixed-precision using the NVIDIA APEX library (https://github.com/NVIDIA/apex) to reduce the GPU memory. We used the batch size of 16 samples. Each mini-batch consisted of balanced samples of each category. In order to obtain more reliable gradient estimates, we aggregated the gradients for a total of 4 mini-batches before performing a training step. We used the ADAM optimizer (Kingma and Ba, 2015) to perform the update steps for all experiments. The learning rate was set as 2e-4. A constant learning rate was used since we did not found much improvement with learning rate decay. The training of an independent student or teacher took 30 or 20 epochs, respectively. The training of a student using KD-loss took 30 epochs with $\alpha = 20$. In addition, the training under the DML setting took 30 epochs with $\alpha = 150$.

Random flipping along the coronal axis and Gaussian noise are used for data augmentation. In addition, we use mixup (Zhang et al., 2017) to generate convex combinations of pairs of examples and their labels for training the networks. Considering the small amounts of MCI subjects, we use AD and NC subjects as auxiliary data during the training process. Previous studies have shown the prediction of MCI conversion can be improved by the auxiliary knowledge learned from AD and NC (Lian et al., 2018; Moradi et al., 2015). We combine AD and pMCI as the positive group, since pMCI has similar atrophy patterns as AD. Meanwhile, the NC and sMCI subjects are combined as the negative group, as sMCI has less disease-related anomalous. With the help of mixup generating convex combinations of pairs of examples and their labels, the network is regularized with better generalizability to differentiate pMCI from sMCI. Note that the AD and NC subject are only used in the training process, both the validation and test set consist of only MCI subjects.

4. Experiments and analysis

In this section, we first present the experimental settings and results. Next, we investigate the influences of the number of image patches, and the cumulative strategy of the KD-loss on prediction performance. Finally, we present the visual interpretation analysis.

4.1. Experimental settings

The proposed distillation approach was compared with similarity preserving (SP) (Tung and Mori, 2019) and attention transfer (AT) (Zagoruyko and Komodakis, 2017). SP losses were computed using the features of the first FC layer, since we found the losses computed using the feature maps of the last convolution layer performed poorly. AT losses were computed using the feature maps of the ResNet and each of the two A^2N-blocks. We set the weight of the distillation loss for attention transfer as $\beta_{AT} = 1000$, and for similarity-preserving as $\beta_{SP} = 100$, respectively. Specifically, both SP and AT losses were computed with the features generated using whole MRI images as input. All the methods used the same auxiliary samples and the same data augmentation procedure.

We used four metrics to evaluate the prediction performance: accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under receiver operating characteristic curve (AUC). Before defining these metrics, we use TP, TN, FP, and FN to denote the true positive, true negative, false positive, and false negative values, respectively. Then we get ACC=(TP+TN)/(TP+TN+FP+FN), SEN=TP/(FN+TP), and SPE=TN/(TN+FP). The AUC is computed based on all possible pairs

Table 2
Results of MCI conversion prediction on the ADNI-2 dataset.

Method	Student	AT	SP	M3ID (Ours)	M3ID+DML (Ours)	Clinical SVM	Teacher
AUC	0.822*†	0.838*†	0.844*†	0.856*	**0.871**	0.903	0.912
(std.)	(0.004)	(0.002)	(0.002)	(0.006)	**(0.004)**	(0.002)	(0.002)
ACC	0.752*†	0.764*†	0.772*	0.780*	**0.800**	0.821	0.832
(std.)	(0.027)	(0.012)	(0.015)	(0.008)	**(0.016)**	(0.008)	(0.017)
SEN	0.766	0.737*	0.774	0.745*	**0.784**	0.805	0.829
(std.)	(0.031)	(0.030)	(0.025)	(0.034)	**(0.007)**	(0.052)	(0.021)
SPE	0.743*†	0.783*	0.770*†	0.804	**0.810**	0.831	0.835
(std.)	(0.057)	(0.014)	(0.013)	(0.033)	**(0.028)**	(0.038)	(0.017)

The best result is bolded and the second best is underlined. AT: attention transfer (Zagoruyko and Komodakis, 2017). SP: similarity preserving (Tung and Mori, 2019). Clinical SVM: a SVM classifier trained with clinical data. *: significantly different from the results with M3ID + DML (t-test, $p < 0.05$). †: significantly different from the results with M3ID (t-test, $p < 0.05$).

of SEN and 1-SPE, which are obtained by changing the thresholds performed on the prediction probabilities. AUC is a better metric than accuracy in imbalanced datasets and real-world applications, especially in clinical practice (Bekkar et al., 2013; Hajian-Tilaki, 2013; Huang and Ling, 2005). We compared the results across different methods with two-sample t-test, using the one-tailed distribution.

We first evaluated the models' prediction performance on the ADNI-2 dataset. The models were trained with MCI data in ADNI-1 and the auxiliary AD/NC data (Section 4.2). Then, we exchanged the training and test dataset, and performed evaluations on the ADNI-1 dataset (Section 4.3). Next, we conducted experiments to study the influences of the number of image patches (Section 4.4), and the effectiveness of the cumulative strategy of the KD-loss (Section 4.5). In the above experiments, 10% of the training MCI samples were randomly selected for held-out validation. The model with the highest AUC value on the hold-out validation dataset was saved for evaluation.

4.2. Results of MCI conversion prediction

Table 2 compares the prediction results obtained by the student network, teacher network, and the students trained with various knowledge distillation methods. The models were evaluated on MCI subjects of ADNI-2 dataset. The results evaluated on AIBL were shown in Table S5 in the *Supplementary Materials*. We reported means and standard deviations of the results over the five model runs. The proposed M3ID method was used in two settings: 1) the teacher and student networks were trained sequentially (M3ID in Table 2); 2) the teacher and student networks were trained simultaneously with DML (M3ID + DML in Table 2).

As shown in Table 2, the multi-modal teacher network achieved the highest prediction AUC, suggesting the effectiveness of complementary information provided by multi-modal data. We also trained a SVM classifier with the clinical data, which consisted of 13 scalars. The SVM classifier achieved an AUC of 0.903, which was much higher than the MRI-based student network. The high prediction performance using clinical data suggests the superior predictive value of clinical data over MRI data. On the other hand, this implies the demand of reducing the performance gap between using multi-modal data and single-modal MRI data, since the gathering of clinical data requires much human efforts and costs.

In Table 2, all knowledge distillation methods improved the MRI-based student network's prediction performance. Knowledge transferred from the multi-modal attention network helped the student network grasp more informative features. Compared with other distillation methods, our M3ID approach improved the student's prediction AUC by 1.8% than using AT, by 1.2% than using SP. This demonstrates the effectiveness of the proposed method for multi-modal knowledge transferring. When training the student in DML setting, the AUC was further

improved to 0.871. This suggests that DML helps the student to better exploit the multi-modal knowledge during the training process, and generalize better to test data.

As we have discussed that the binary class outputs of a teacher network are not sufficiently informative to teach a student. We validated this by experiments. The results were presented in Table S3 and Table S4 of the *Supplementary Materials*. Compared with our method, the vanilla KD using the whole image as a single patch achieved poorer results. We also conducted additional experiments using SP under the DML setting. The results in Table S7 show that our method achieved better results.

4.3. Influence of data partition

In the above experiments, we evaluated different methods on the baseline ADNI-2 dataset. To study the generalization of the proposed framework, we evaluated the models' prediction performance on baseline ADNI-1 dataset (trained with MCI in ADNI-2 and additional AD/NC). The prediction results were reported in Table 3. The results evaluated on AIBL were shown in Table S6 in the *Supplementary Materials*. We reported means and standard deviations of the results over five model runs.

The multi-modal teacher network again achieved the highest prediction AUC. The student trained with only data supervision (*i.e.,* conversion labels) achieved an AUC of 0.777. With various multi-modal knowledge distillation methods, the student network's prediction AUC was increased by 1.1% to 3.1%, respectively. Our proposed distillation method again brought bigger AUC increases to the student network than the other methods. The student trained in DML setting achieved a higher prediction performance compared with the student trained with traditional setting of distillation. Although the AT and SP produced higher accuracy and sensitivity values, our methods achieved better AUC scores, which were of great importance in clinic and imbalanced datasets. The results in Table 3 demonstrates the generalizability of the proposed framework.

By comparing the results in Table 2 and Table 3, it can be seen that the prediction performance on the ADNI-2 are better than the prediction performance on the ADNI-1. There might be two possible reasons: one is the insufficient training sample size, since there are more MCI samples in ADNI-1 than those in ADNI-2 (264 scans vs. 191 scans); the other is that the MCI subjects of these two datasets have different disease statuses (see Table S1 and Table S2 in the *Supplementary Materials*), and the dataset itself could be the bottleneck of performance improvement. Meanwhile, our setting of using two independent dataset is closer to the real-world clinic situation, because a large existing dataset can be used for training models to be evaluated on newly enrolled patients.

Table 3
Results of MCI conversion prediction on the ADNI-1 dataset.

Method	Student	AT	SP	M3ID (Ours)	M3ID+DML (Ours)	Clinical SVM	Teacher
AUC	0.777*†	0.788*†	0.794*†	0.805*	**0.808**	0.821	0.854
(std.)	(0.003)	(0.002)	(0.003)	(0.002)	**(0.002)**	(0.003)	(0.008)
ACC	0.722*†	**0.742**	0.736	0.733	0.741	0.753	0.776
(std.)	(0.008)	**(0.015)**	(0.014)	(0.007)	(0.020)	(0.009)	(0.022)
SEN	0.710 †	**0.778**	0.754	0.730	0.738	0.728	0.760
(std.)	(0.007)	**(0.054)**	(0.021)	(0.022)	(0.034)	(0.050)	(0.042)
SPE	0.744	0.682*†	0.706*†	0.738	**0.746**	0.786	0.804
(std.)	(0.022)	(0.049)	(0.021)	(0.029)	**(0.019)**	(0.044)	(0.047)

The best result is bolded and the second best is underlined. AT: attention transfer (Zagoruyko and Komodakis, 2017). SP: similarity preserving (Tung and Mori, 2019). Clinical SVM: a SVM classifier trained with clinical data. *: significantly different from the results with M3ID + DML (t-test, $p < 0.05$). †: significantly different from the results with M3ID (t-test, $p < 0.05$).
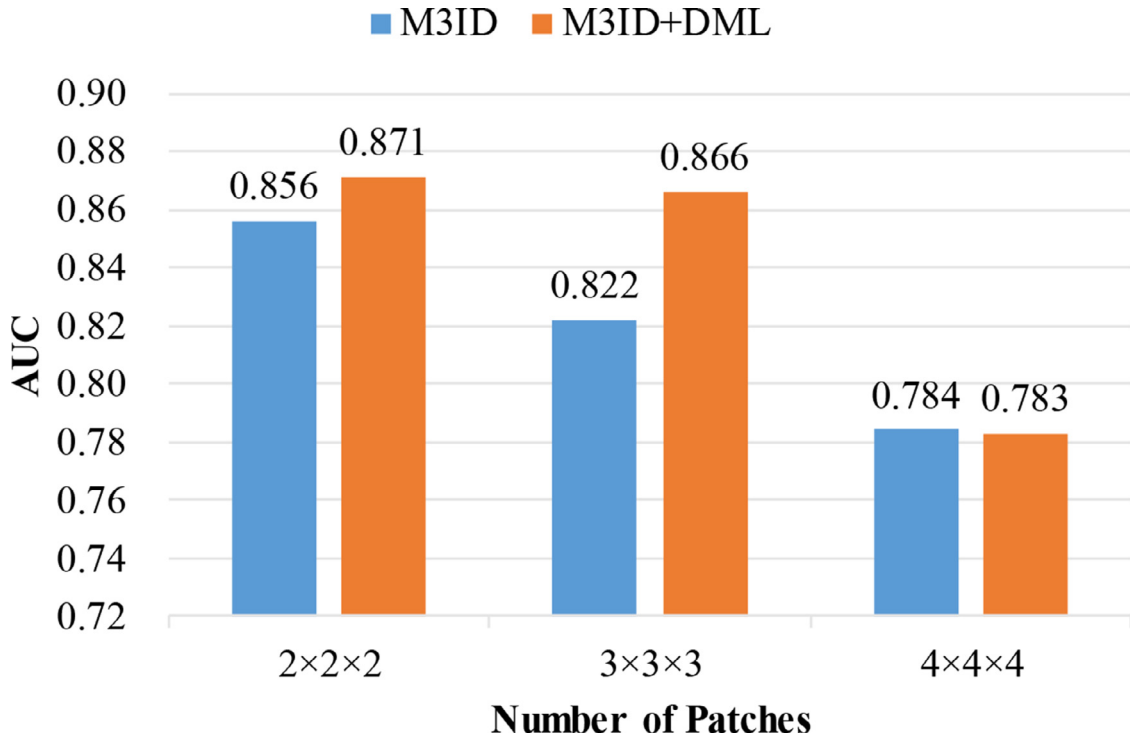

Fig. 4. Influence of the number of image patches on the performance.

4.4. Influence of the number of image patches

In the above experiments, we adopted a fixed number of patches (*i.e.,* $2 \times 2 \times 2$) with size of $72 \times 88 \times 72$ for our proposed distillation method. We then explored the influence of the number of patches on the performance of knowledge distillation. Along each axis, we divided the whole image into 3 and 4 equal parts, producing 27 and 64 patches, respectively. Likewise, the patch size of each axis was multiples of 8 to ensure that no border pixel was discarded during the convolution process. When dividing an MRI into 27 patches, each patch is of size $48 \times 64 \times 48$. When dividing an MRI into 64 patches, each patch is of size $40 \times 48 \times 40$.

Fig. 4 shows the AUC values of MCI conversion prediction on the ADNI-2 dataset. The best results were obtained by using 8 patches (*i.e.,* $2 \times 2 \times 2$). When increasing the number of patches, the corresponding patch size decreases. Fig. 4 suggests that the patches with much smaller sizes than the whole image are not suitable for transferring knowledge, since the structural information might be lost during the network's mapping process. In addition, using a large number of patches significantly

increase the computational burden, which limits the proposed method's practical applications.

4.5. Influence of the cumulative strategy of the KD-loss

In this group of experiments, we investigated the influence of different cumulative strategies of the KD-loss on the performance. The prediction results on the ADNI-2 dataset were reported in Table 4: increasing the hyperparameter α from zero following the cosine function (Eq. 8) consistently brought performance improvements; the prediction performance with a constant α was comparatively poor. This suggests that the student network could learn better by gradually receiving more guidance from the teacher network. At the initial stage of training, a high knowledge distillation loss might hamper the accurate learning of a student (Anil et al., 2018). We conducted additional experiments with the linear cumulative strategy ($\alpha_t = \alpha \times (1/T_{max}) * T_{cur}$), and provided results in Table 4 for comparison. The linear function also worked and produced comparable results. This suggests the effectiveness of our distillation method. The AUCs using cosine function (Eq. 8) are slightly

Table 4

Influence of the cumulative strategy of the KD-loss on the performance.

Strategy	Constant α		Linear α		Cosine α	
	M3ID	M3ID +DML	M3ID	M3ID +DML	M3ID	M3ID +DML
AUC	0.824	0.845	0.850	<u>0.861</u>	0.856	**0.871**
ACC	0.759	0.754	0.770	**0.801**	0.791	<u>0.796</u>
SEN	0.697	**0.803**	0.776	0.776	0.711	<u>0.789</u>
SPE	0.800	0.722	0.765	<u>0.817</u>	**0.843**	0.800

The best result is bolded and the second best is underlined.

higher than those using linear function. We conjecture this could result from the growth trend of the cosine function: starts with a smaller value when the network is handicapped; produces a higher value as the network getting robust.

To better understand the influence of α_t, we provided change curves of the total loss, the KD-loss and the α_t-weighted KD-loss during the training processes in Fig. S1. In addition, we presented the performance with different α_t in Fig. S2 of the *Supplementary Materials*.

4.6. Visual interpretation analysis

In this group of experiments, we visualized the multi-modal attention maps and the patch scores for further analysis. Fig. 5 shows the multi-modal attention maps generated by the teacher network. In Fig. 5, the hippocampus, ventricle, temporal and parietal lobe are highlighted. These areas have been previously reported in studies concerning the brain atrophy in AD (Weiner et al., 2012). This suggests that the multi-modal attention helps the teacher network to extract features from the discriminative areas.

The multi-instance probabilities are encoded to represent the complicated atrophy distributions of brains. With the multi-instance probabilities as extra guidance, the teacher network teaches the student network to reweight patches across the brain. Then the student has a better capacity of discriminative feature extraction and localization, which can be demonstrated in the visualizations of patch scores. Fig. 6 visualizes the patch scores generated by the student network after distillation. For pMCI cases, the non-overlapping patches have different scores, and the patches near temporal lobe and hippocampus have higher scores. The disease-related areas like hippocampus and amygdala are located in the inner region of the temporal lobe. This demonstrates the model's capacity of discriminative feature extraction and localization. For sMCI subjects, patch scores are relatively low, and the highlighted areas are around the lateral ventricle. The visualization suggests that the network assesses regional severities of the diseases in MRI, and leverages the patch relationships for prediction.

In addition, Fig. 5 and Fig. 6 show that the individual cases and the average results are largely overlapped. This suggests that the network can focus on the disease-related areas. By comparing Fig. 5 and Fig. 6, we can see that the teacher's attention covers more areas than the student highlights. This could result from the multi-modal data providing complementary information. The teacher network uses an additional attention module to model the multi-modal correlations, for example, those between image and cognition. Thus the teacher network can captures more subtle patterns such as the structural changes of sulci (highlighted border areas of brains in Fig. 5), which are affected by AD and correlated to cognition (Bertoux et al., 2019). Although the guidance from teacher helps the student to focus more on the disease-related areas, some subtle patterns might be overly difficult for a student with less parameters and access to only MRI data.

We compared the heat-maps generated by a student network trained without knowledge distillation (KD) and with KD in Fig. 7. The visualization results suggest that multi-instance knowledge distillation helps the network to reweight the importance of each patch, and focus on the disease-related areas. Then the student has a better capacity of discrim-

inative feature extraction and localization, and achieves better performance (see Table 2 and Table 3). We presented additional figures in the *Supplementary Materials* to further illustrate the effectiveness of our distillation method: Fig. S3 shows that the teacher's multi-instance probabilities reveals the patch relationships; Fig. S5 visualizes the slices best showing the highlighted areas of sMCI. With our multi-instance distillation scheme, the network is capable of providing visual interpretations, which could be useful for medical decision makings.

5. Discussion

In this study, we present a novel knowledge distillation method for MCI conversion prediction. The multi-modal knowledge is transferred from a teacher network to a student network that only has access to MRI data. With the multi-modal knowledge as the extra supervision, the prediction performance of the MRI-based student network is effectively improved.

5.1. Comparison with previous works

In Table 5, we briefly summarized several state-of-the-art results on MCI conversion prediction using the ADNI dataset. We listed these method's results achieved with only MRI data, and the results using multi-modal data. The direct comparison may be affected because of the different number of subjects and the image pre-processing step used in these studies. Most previous studies (as shown in Table 5) used 36 months as the decision threshold between sMCI and pMCI. For a fair comparison, we also adopted 36 months to categorize sMCI and pMCI. Compared with the MRI-based prediction methods, our framework significantly increased the MRI-based prediction model's performance (see Table 5). Some previous studies' results (Moradi et al., 2015; Spasov et al., 2019) using a combination of clinical data and MRI were much better than the performance achieved with only MRI data. Although this suggests the predictive values of clinical data, the disadvantages are obvious: the extra clinical data are necessary, and the highly accessible MRI data only play an auxiliary role. Our framework reduces the performance gap between using multi-modal data and using only MRI. Thus, our framework has promising potential application. Further, our experiments were conducted in a strict setting. The prediction model was trained and evaluated on three datasets (ADNI-1, ADNI-2, AIBL). With the more challenging evaluation protocol, the effectiveness of the proposed framework was ensured.

Many previous studies leverage the neurological expertise to define disease-related regions-of-interests (ROI), and extract features from these ROIs or nearby image patches for building diagnosis and prediction models (Bron et al., 2015; Eskildsen et al., 2013; Lian et al., 2018; Liu et al., 2018; Moradi et al., 2015). These studies demonstrate that the local image patches convey disease-related information, and the ROI priors are useful for better feature extraction. This is similar to our scheme that estimates probabilities of multi-instances. In our proposed scheme, the multi-instance probabilities are encoded to represent the complicated atrophy distributions. In conventional multi-instance learning (Amores, 2013), a bag is represented by the most discriminative instance or the average output of all the instances in the bag. By doing this, the distribution of atrophy patterns might be lost. In contrast, our method does not directly use multi-instance representations or outputs for classification, but considers the probability distribution of multi-instances as the extra supervision. With a better capacity of representing the complicated and distributed atrophy patterns, the multi-instance probabilities implicitly guide the student network to reweight the patches across a brain and achieve better performance. In addition, the multi-instance prediction models in (Lian et al., 2018; Liu et al., 2018) require pre-defined landmarks for building an extra location proposal module, then perform feature extraction according to the proposed locations. In contrast, we use evenly partitioned image patches as input
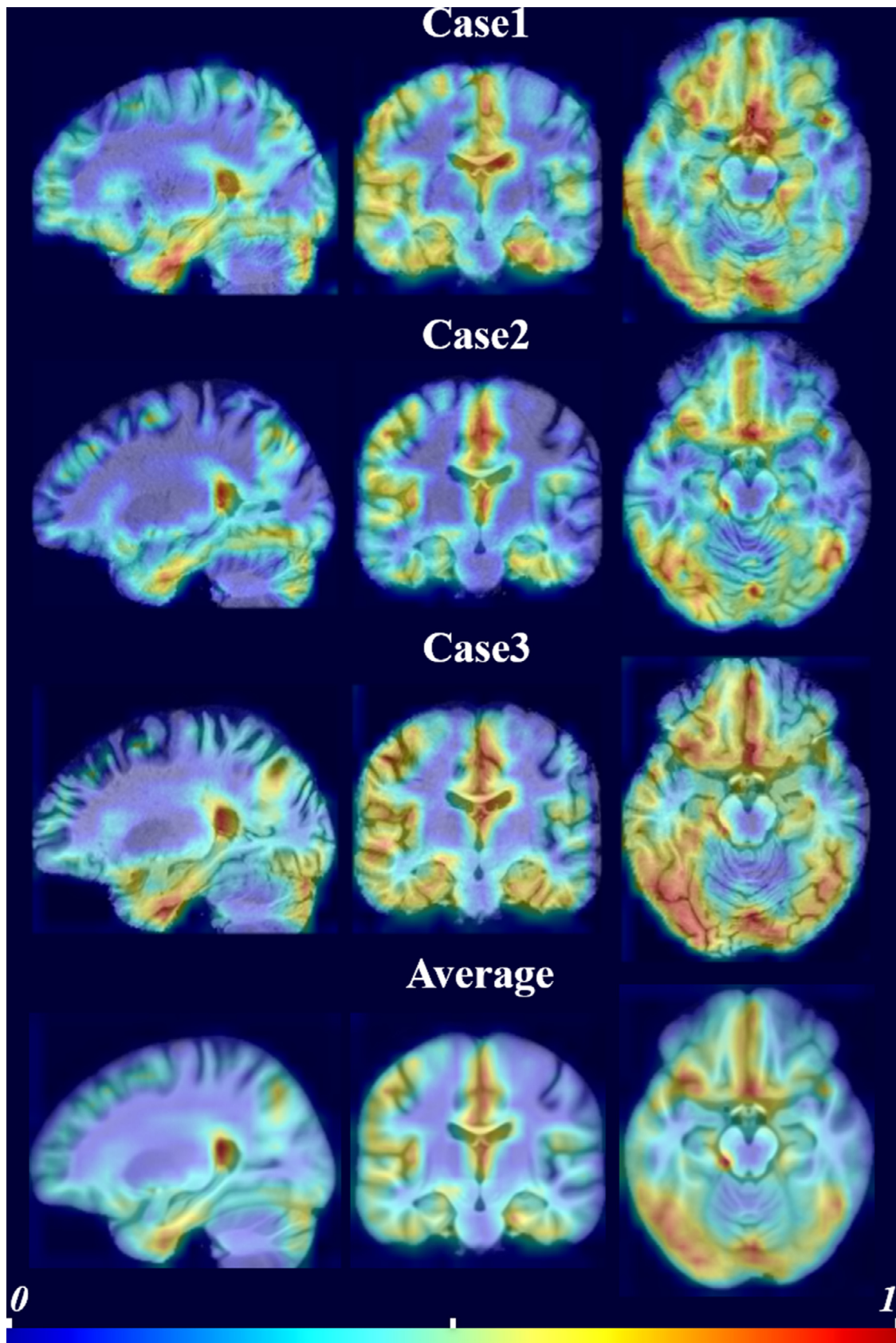
Fig. 5. Visualization of multi-modal attention maps. The attention maps are upscaled to the same size as the input MRI image, and overlayed onto the MRI image. The columns denote the views in three anatomical planes. The average results are computed using all correctly predicted test cases.
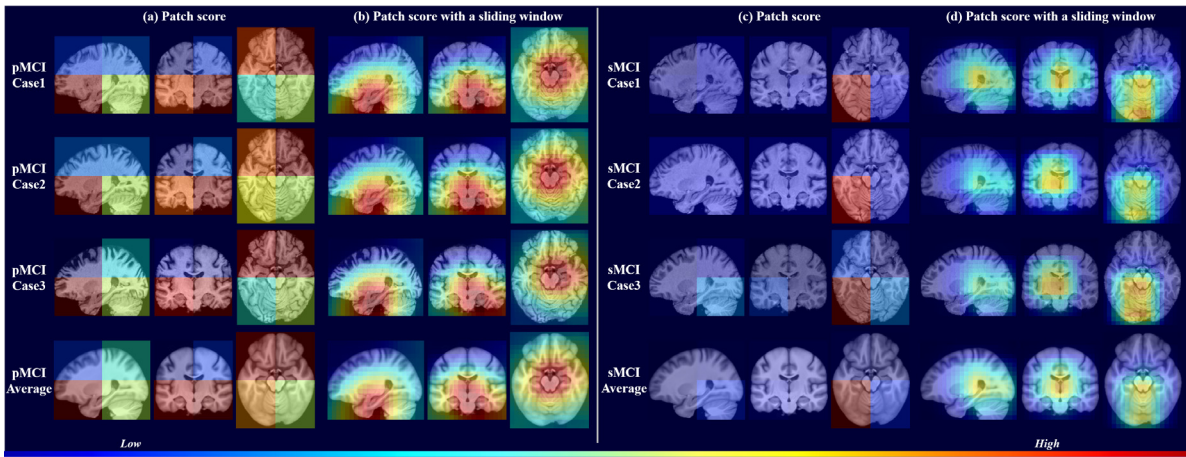
Fig. 6. Visualization of patch scores of pMCI and sMCI subjects. Panel (a) and (c): 8 non-overlapping patches evenly divided from a whole MRI; panel (b) and (d): patches extracted with a sliding window running across the whole MRI. All of the patches are of size $72 \times 88 \times 72$. The average results are computed using all correctly predicted test cases.
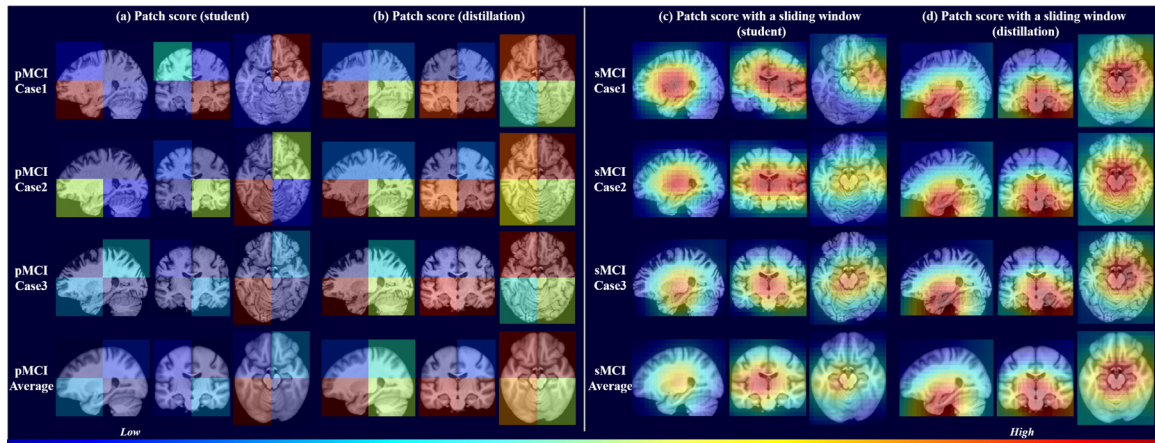
Fig. 7. Visualization of patch scores generated without and with distillation. The student trained with distillation loss focus more on hippocampus and amygdala, while the vanilla student highlights diffuse areas across the brain. The 8 non-overlapping patches (of size $72 \times 88 \times 72$) are evenly divided from a whole MRI. The average results are computed using all correctly predicted test cases.

Table 5

A summary of MRI-based studies for predicting MCI-to-AD conversion using the ADNI dataset.

Study	Number of pMCI/sMCI	Method	Conversion time	AUC (MRI)		AUC (MRI+)	
				ADNI-1	ADNI-2	ADNI-1	ADNI-2
Moradi et al. (2015)	164/100	Gray matter density map + low density separation	36 months	0.766	-	0.903 (+ age & clinical data)	-
Liu et al. (2016)	117/117	Gray matter density map + Ensemble SVM	24 months	0.834	-	-	-
Liu et al. (2018)	205/465	MRI patches + CNN	36 months	-	0.776	-	-
Lian et al. (2018)	205/465	MRI patches + CNN	36 months	-	0.781	-	-
Lin et al. (2018)	169/139	MRI patches + CNN + PCA + Lasso + ELM	36 months	0.829	-	0.861 (+ age)	-
Spasov et al. (2019)	181/228	Whole brain MRI + CNN	36 months	0.790	-	0.925 (+ age & clinical data)	-
Pan et al. (2020)	192/502	Whole brain MRI + CNN with Fisher representation	36 months	-	0.818	-	0.839 (+PET)
Ours	240/215	Whole brain MRI + CNN	36 months	0.808	0.871	0.854	0.912
						(+ age & clinical data)	

"CNN" refers to convolution neural network; "MRI + " refers to using multi-modal data in addition to MRI. "ADNI-1" and "ADNI-2" refer the corresponding evaluation datasets.

without the need of the localization processes or landmarks. Therefore, our deep learning framework is more annotation-efficient.

Compared with the vanilla distillation method (Hinton et al., 2015) that enforces the matching of class probabilities, our method enforce a matching of the probability estimates of multi-instances. The additional experiments demonstrate our method's advantage over the vanilla distillation method (Hinton et al., 2015) (Table S3 and Table S4) in the *Supplementary Materials*. Compared with the representation-based distillation methods (including SP (Tung and Mori, 2019) and AT (Zagoruyko and Komodakis, 2017)) that encourage a student to mimic a teacher's representation space in different aspects, our method aims to match the probabilities of the output space. Considering the student network and the teacher network building with different inductive biases, their solution paths of expressing the representation space could be different. In addition, large gaps may exist between different sources of data, and lead to various responses in the intermediate layers. In the

above situations, mimicking of representation space may not convey accurate knowledge. However, the differences in the representation space do not prevent the teacher and the student from independently making the right predictions. The experiments also demonstrate our method's superiority over the representation-based methods.

Our method lies between the probability-based method and representation-based distillation method: we use the multi-instance probabilities in the output space; the multi-instances probabilities work like a kind of representation rather than directly produce prediction outputs. In addition, our method could also be counted as multi-task learning: the student network is trained with the data supervision and the extra supervision distilled from the teacher network. Different from the conventional multi-task learning that uses clinical data as additional targets, we use the multi-modal teacher network's predictions as the additional targets. We provide additional experimental analysis with respect to the effectiveness of temperature scaling and A^2N-block in the *Supplementary Materials*. The results in Table S8 show that the temperature scaling brought no performance improvement. The prediction performance (see Table S9) using the network without A^2N-block is poor, and our method still outperforms the other distillation methods.

5.2. Limitations and future work

Although the proposed framework has achieved competitive results for MRI-based conversion prediction, there is still room for improvement. Firstly, considering the medical data of high costs are used only during the training process, all available modalities of data can be leveraged to improve the current framework. For example, functional imaging, FDG-PET, and CSF data are useful for disease diagnosis and biomarker discovery (Ding et al., 2019; Li et al., 2019; Wang et al., 2020; 2019). We could further the study by using all available imaging and non-imaging clinical data for optimized knowledge extraction and distillation. Secondly, our teacher network only utilized the patients with complete multi-modal data. The patients missing some modalities of data were excluded. While in practice, modalities missing of medical data is usual due to the cost, data quality, and patient dropouts. Therefore, combining the methods of effectively utilizing incomplete multi-modal data into the proposed framework would be a promising direction. Thirdly, we processed the 3D patches without leveraging their location information, which could be another source of supervision. Explicitly using location information of these patches might further improve the model's ability for feature extraction. Finally, although we evaluated our frameworks on three datasets, further experiments in a larger population are needed to evaluate and improve the framework.

6. Conclusion

In this study, we developed a multi-modal multi-instance distillation (M3ID) method for MRI-based MCI conversion prediction. We enforced an MRI-based student network to mimic a multi-modal teacher's probability estimates for 3D patches, and guide the student network to better model the subtle features in MRI. The proposed framework was evaluated on three public datasets. The experimental results shows that our M3ID approach effectively improved the prediction performance of the MRI-based student model. Our proposed framework enables better disease prediction using less medical examinations.

Credit authorship contribution statement

Hao Guan: Conceptualization, Data curation, Methodology, Validation, Visualization, Writing – original draft. **Chaoyue Wang:** Supervision, Project administration, Writing – review & editing. **Dacheng Tao:** Supervision, Funding acquisition, Writing – review & editing.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118586.

References

Amores, J., 2013. Multiple instance classification: review, taxonomy and comparative study. Artif Intell 201, 81–105. doi:10.1016/j.artint.2013.06.003. http://www.sciencedirect.com/science/article/pii/S0004370213000581

Anil, R., Pereyra, G., Passos, A., Ormándi, R., Dahl, G., Hinton, G.E., 2018. Large scale distributed neural network training through online distillation. ArXiv 1505–1512. doi:10.1109/tvlsi.2020.2981443. arXiv:1804.0.

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., Jones, E., 2011. Alzheimer's disease. The Lancet 377 (9770), 1019–1031. doi:10.1016/S0140-6736(10)61349-9.

Bekkar, M., Djemaa, H., Alitouche, T.A., 2013. Evaluation measures for models assessment over imbalanced data sets. Journal of Information Engineering and Applications 3, 27–38. doi:10.5121/ijdkp.2013.3402.

Bertoux, M., Lagarde, J., Corlier, F., Hamelin, L., Mangin, J.-F., Colliot, O., Chupin, M., Braskie, M.N., Thompson, P.M., Bottlaender, M., Sarazin, M., 2019. Sulcal morphology in Alzheimer's disease: an effective marker of diagnosis and cognition. Neurobiol. Aging 84, 41–49. doi:10.1016/j.neurobiolaging.2019.07.015. http://www.sciencedirect.com/science/article/pii/S0197458019302192

Braskie, M.N., Thompson, P.M., 2013. Understanding cognitive deficits in Alzheimer's disease based on neuroimaging findings. Trends Cogn Sci 17 (10), 510–516. doi:10.1016/j.tics.2013.08.007.

Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Mendez Orellana, C., Meijboom, R., Pinto, M., Meireles, J.R., Garrett, C., Bastos-Leite, A.J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cardenas-Pena, D., Alvarez-Meza, A.M., Dolph, C.V., Iftekharuddin, K.M., Eskildsen, S.F., Coupe, P., Fonov, V.S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K.R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G.M., Stoyanov, Z.V., Sorensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J.C., Niessen, W.J., Klein, S., Alzheimer's Disease Neuroimaging, I., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. Neuroimage 111, 562–579. doi:10.1016/j.neuroimage.2015.01.048. https://www.ncbi.nlm.nih.gov/pubmed/25652394

Burkov, E., Lempitsky, V., 2018. Deep Neural Networks with Box Convolutions. In: Adv. Neural Inf. Process. Syst., pp. 6211–6221. http://papers.nips.cc/paper/7859-deep-neural-networks-with-box-convolutions.pdf

Carbonneau, M.-A., Cheplygina, V., Granger, E., Gagnon, G., 2018. Multiple instance learning: a survey of problem characteristics and applications. Pattern Recognit 77, 329–353. doi:10.1016/j.patcog.2017.10.009. http://www.sciencedirect.com/science/article/pii/S0031320317304065

Carretti, B., Borella, E., Fostinelli, S., Zavagnin, M., 2013. Benefits of training working memory in amnestic mild cognitive impairment: specific and transfer effects. Int Psychogeriatr 25 (4), 617–626. doi:10.1017/s1041610212002177.

Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J., 2018. A2-nets: Double attention networks. In: Adv. Neural Inf. Process. Syst., pp. 352–361. doi:10.1007/978-3-030-01246-5_22.

Clevert, D.-A., Unterthiner, T., Hochreiter, S., 2016. Fast and accurate deep network learning by exponential linear units (ELUs). CoRR 5441–5451. doi:10.1039/c8sc00148k. arXiv:1511.0.

Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L., Pericak-Vance, M.A., 1993. Gene dose of apolipoprotein e type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261 (5123), 921–923. doi:10.1126/science.8346443.

Ding, Y., Sohn, J.H., Kawczynski, M.G., Trivedi, H., Harnish, R., Jenkins, N.W., Lituiev, D., Copeland, T.P., Aboian, M.S., Mari Aparici, C., Behr, S.C., Flavell, R.R., Huang, S.Y., Zalocusky, K.A., Nardo, L., Seo, Y., Hawkins, R.A., Hernandez Pampaloni, M., Hadley, D., Franc, B.L., 2019. A deep learning model to predict a diagnosis of alzheimer disease by using (18)F-FDG PET of the brain. Radiology 290 (2), 456–464. doi:10.1148/radiol.2018180958. https://www.ncbi.nlm.nih.gov/pubmed/30398430

Do, T.K.L., Do, T.-T., Tran, H., Tjiputra, E., Tran, Q.D., 2019. Compact Trilinear Interaction for Visual Question Answering. In: IEEE/CVF International Conference on Computer Vision (ICCV) doi:10.1109/iccv.2019.00048.

Dou, Q., Liu, Q., Heng, P.A., Glocker, B., 2020. Unpaired multi-modal segmentation via knowledge distillation. IEEE Trans. Med. Imag. 39 (7), 2415–2425. doi:10.1109/tmi.2019.2963882.

Driscoll, I., Davatzikos, C., An, Y., Wu, X., Shen, D., Kraut, M., Resnick, S.M., 2009. Longitudinal pattern of regional brain volume change differentiates normal aging from MCI. Neurology 72 (22), 1906–1913. doi:10.1212/WNL.0b013e3181a82634. https://pubmed.ncbi.nlm.nih.gov/19487648, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2690968/

Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., Ames, D., 2009. The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 21 (4), 672–687. doi:10.1017/s1041610209009405.

Eskildsen, S.F., Coupe, P., Garcia-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L., Alzheimer's Disease Neuroimaging, I., 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. Neuroimage 65, 511–521. doi:10.1016/j.neuroimage.2012.09.058. https://www.ncbi.nlm.nih.gov/pubmed/23036450

Ewers, M., Sperling, R.A., Klunk, W.E., Weiner, M.W., Hampel, H., 2011. Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. Trends Neurosci 34 (8), 430–442. doi:10.1016/j.tins.2011.05.005. https://www.ncbi.nlm.nih.gov/pubmed/21696834

Genin, E., Hannequin, D., Wallon, D., Sleegers, K., Hiltunen, M., Combarros, O., Bullido, M.J., Engelborghs, S., De Deyn, P., Berr, C., Pasquier, F., Dubois, B., Tognoni, G., Fiévet, N., Brouwers, N., Bettens, K., Arosio, B., Coto, E., Del Zompo, M., Mateo, I., Epelbaum, J., Frank-Garcia, A., Helisalmi, S., Porcellini, E., Pilotto, A., Forti, P., Ferri, R., Scarpini, E., Siciliano, G., Solfrizzi, V., Sorbi, S., Spalletta, G., Valdivieso, F., Vepsäläinen, S., Alvarez, V., Bosco, P., Mancuso, M., Panza, F., Nacmias, B., Bossù, P., Hanon, O., Piccardi, P., Annoni, G., Seripa, D., Galimberti, D., Licastro, F., Soininen, H., Dartigues, J.F., Kamboh, M.I., Van Broeckhoven, C., Lambert, J.C., Amouyel, P., Campion, D., 2011. APOE And alzheimer disease: a major gene with semi-dominant inheritance. Mol Psychiatry 16 (9), 903–907. doi:10.1038/mp.2011.52.

Gou, J., Yu, B., Maybank, S.J., Tao, D., 2020. Knowledge distillation: A Survey. ArXiv. arXiv:2006.05525. https://ui.adsabs.harvard.edu/abs/2020arXiv200605525G

Grabner, G., Janke, A.L., Budge, M.M., Smith, D., Pruessner, J., Collins, D.L., 2006. Symmetric Atlasing and Model Based Segmentation: An Application to the Hippocampus in Older Adults. In: Med Image Comput Comput Assist Interv (MICCAI), Vol. 9, pp. 58–66. doi:10.1007/11866763_8.

Gupta, S., Hoffman, J., Malik, J., 2016. Cross Modal Distillation for Supervision Transfer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2827–2836. doi:10.1109/cvpr.2016.309.

Hajian-Tilaki, K., 2013. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med 4 (2), 627–635. https://pubmed.ncbi.nlm.nih.gov/24009950 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/

Hampel, H., Toschi, N., Baldacci, F., Zetterberg, H., Blennow, K., Kilimann, I., Teipel, S.J., Cavedo, E., dos Santos, A.M., Epelbaum, S., Lamari, F., Genthon, R., Dubois, B., Floris, R., Garaci, F., Lista, S., 2018. Alzheimer'S disease biomarker-guided diagnostic workflow using the added value of six combined cerebrospinal fluid candidates: $a\beta1$-42, total-tau, phosphorylated-tau, NFL, neurogranin, and YKL-40. Alzheimer's & Dementia 14 (4), 492–501. doi:10.1016/j.jalz.2017.11.015. http://www.sciencedirect.com/science/article/pii/S1552526017338724

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778. doi:10.1109/cvpr.2016.90.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. ArXiv 630–645. doi:10.1007/978-3-319-46493-0_38. arXiv:1804.01661. https://ui.adsabs.harvard.edu/abs/2018arXiv180401661Y

Hinton, G.E., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. ArXiv abs/1503.0. doi:10.1109/icassp.2016.7472621.

Huang, J., Ling, C., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 17, 299–310. doi:10.1109/tkde.2005.50.

Ilse, M., Tomczak, J.M., Welling, M., 2018. Attention-based deep multiple instance learning. ArXiv. arXiv:1802.

Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J Magn Reson Imaging 27 (4), 685–691. doi:10.1002/jmri.21049. https://www.ncbi.nlm.nih.gov/pubmed/18302232

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. Neuroimage 62 (2), 782–790. doi:10.1016/j.neuroimage.2011.09.015. https://www.ncbi.nlm.nih.gov/pubmed/21979382

Kandemir, M., Hamprecht, F.A., 2015. Computer-aided diagnosis from weak supervision: abenchmarking study. Computerized Medical Imaging and Graphics 42, 44–50. doi:10.1016/j.compmedimag.2014.11.010. http://www.sciencedirect.com/science/article/pii/S0895611114001852

Kingma, D.P., Ba, J., 2015. Adam: A Method for stochastic optimization. CoRR 148–151. doi:10.1134/s1064562416020071. arXiv:1412.6980. https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6980K

Lee, E., Choi, J.-S., Kim, M., Suk, H.-I., 2019. Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning. Neuroimage 202, 116113. doi:10.1016/j.neuroimage.2019.116113. http://www.sciencedirect.com/science/article/pii/S1053811919307049

Li, K., Yu, L., Wang, S., Heng, P.-A., 2020. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, pp. 775–783. doi:10.1609/aaai.v34i01.5421.

Li, Y., Yang, H., Lei, B., Liu, J., Wee, C.Y., 2019. Novel effective connectivity inference using ultra-Group constrained orthogonal forward regression and elastic multilayer perceptron classifier for MCI identification. IEEE Trans. Med. Imag. 38 (5), 1227–1239. doi:10.1109/TMI.2018.2882189. https://www.ncbi.nlm.nih.gov/pubmed/30475714

Lian, C., Liu, M., Zhang, J., Shen, D., 2018. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE Trans. Pattern Anal. Mach. Intell. 1. doi:10.1109/TPAMI.2018.2889096. https://www.ncbi.nlm.nih.gov/pubmed/30582529

Lin, W., Tong, T., Gao, Q., Guo, D., Du, X., Yang, Y., Guo, G., Xiao, M., Du, M., Qu, X., Alzheimer's Disease Neuroimaging, I., 2018. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. Front Neurosci 12, 777. doi:10.3389/fnins.2018.00777. https://www.ncbi.nlm.nih.gov/pubmed/30455622

Liu, M., Zhang, D., Shen, D., 2016. Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. IEEE Trans. Med. Imag. 35 (6), 1463–1474. doi:10.1109/TMI.2016.2515021. https://www.ncbi.nlm.nih.gov/pubmed/26742127

Liu, M., Zhang, J., Adeli, E., Shen, D., 2018. Landmark-based deep multi-instance learning for brain disease diagnosis. Med. Image Anal. 43, 157–168. doi:10.1016/j.media.2017.10.005. https://www.ncbi.nlm.nih.gov/pubmed/29107865

Liu, M., Zhang, J., Yap, P.T., Shen, D., 2017. View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data. Med. Image Anal. 36, 123–134. doi:10.1016/j.media.2016.11.002. http://www.ncbi.nlm.nih.gov/pubmed/27898305

Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V., 2016. Unifying distillation and privileged information. CoRR doi:10.1109/pac.2017.13. arXiv:1511.0.

Mitchell, A.J., Shiri-Feshki, M., 2008. Temporal trends in the long term risk of progression of mild cognitive impairment: a pooled analysis. Journal of Neurology, Neurosurgery & Psychiatry 79 (12), 1386–1391. doi:10.1136/jnnp.2007.142679. https://jnnp.bmj.com/content/79/12/1386

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Alzheimer's Disease Neuroimaging, I., 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398–412. doi:10.1016/j.neuroimage.2014.10.002. http://www.ncbi.nlm.nih.gov/pubmed/25312773

Pan, Y., Liu, M., Lian, C., Xia, Y., Shen, D., 2020. Spatially-Constrained fisher representation for brain disease identification with incomplete multi-Modal neuroimages. IEEE Trans Med Imaging doi:10.1109/TMI.2020.2983085. https://www.ncbi.nlm.nih.gov/pubmed/32217472

Pandya, S.Y., Clem, M.A., Silva, L.M., Woon, F.L., 2016. Does mild cognitive impairment always lead to dementia? a review. J Neurol Sci 369, 57–62. doi:10.1016/j.jns.2016.07.055. http://www.ncbi.nlm.nih.gov/pubmed/27653867

Passalis, N., Tefas, A., 2018. Learning deep representations with probabilistic knowledge transfer. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 268–284. doi:10.1007/978-3-030-01252-6_17.

Petersen, R.C., 2004. Mild cognitive impairment as a diagnostic entity. J. Intern. Med. 256 (3), 183–194. doi:10.1111/j.1365-2796.2004.01388.x.

Qiu, S., Joshi, P.S., Miller, M.I., Xue, C., Zhou, X., Karjadi, C., Chang, G.H., Joshi, A.S., Dwyer, B., Zhu, S., Kaku, M., Zhou, Y., Alderazi, Y.J., Swaminathan, A., Kedar, S., Saint-Hilaire, M.H., Auerbach, S.H., Yuan, J., Sartor, E.A., Au, R., Kolachalama, V.B., 2020. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain doi:10.1093/brain/awaa137. https://www.ncbi.nlm.nih.gov/pubmed/32357201

Reiman, E.M., Langbaum, J.B.S., Tariot, P.N., 2010. Alzheimer's prevention initiative: a proposal to evaluate presymptomatic treatments as quickly as possible. Biomark Med 4 (1), 3–14. doi:10.2217/bmm.09.91.

Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y., 2014. Fitnets: hints for thin deep nets. ArXiv doi:10.1109/acpr.2015.7486451. arXiv:1707.

Roses, A.D., 1996. Apolipoprotein E alleles as risk factors in Alzheimer's disease. Annu Rev Med 47, 387–400. doi:10.1146/annurev.med.47.1.387.

Ruffy, F., Chahal, K., 2019. The state of knowledge distillation for classification. ArXiv. arXiv:1912.1.

Shen, J., Vesdapunt, N., Boddeti, V.N., Kitani, K.M., 2016. In teacher we trust: learning compressed models for pedestrian detection. ArXiv 126. doi:10.1007/s11263-018-1077-3. arXiv:1612.00478.

Shen, L., Kim, S., Risacher, S.L., Nho, K., Swaminathan, S., West, J.D., Foroud, T., Pankratz, N., Moore, J.H., Sloan, C.D., Huentelman, M.J., Craig, D.W., Dechairo, B.M., Potkin, S.G., Jack, C.R., Weiner, M.W., Saykin, A.J., 2010. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort. Neuroimage 53 (3), 1051–1063. doi:10.1016/j.neuroimage.2010.01.042.

Sherman, D.S., Mauser, J., Nuno, M., Sherzai, D., 2017. The efficacy of cognitive intervention in mild cognitive impairment (MCI): a meta-analysis of outcomes on neuropsychological measures. Neuropsychol Rev 27 (4), 440–484. doi:10.1007/s11065-017-9363-3.

Spasov, S., Passamonti, L., Duggento, A., Lio, P., Toschi, N., Alzheimer's Disease Neuroimaging, I., 2019. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. Neuroimage 189, 276–287. doi:10.1016/j.neuroimage.2019.01.031. https://www.ncbi.nlm.nih.gov/pubmed/30654174

Sultani, W., Chen, C., Shah, M., 2018. Real-World Anomaly Detection in Surveillance Videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/cvpr.2018.00678.

Tang, J., Shivanna, R., Zhao, Z., Lin, D., Singh, A., Chi, E.H.-h., Jain, S., 2020. Understanding and improving knowledge distillation. ArXiv. arXiv:2002.0.

Teipel, S.J., Cavedo, E., Lista, S., Habert, M.-O., Potier, M.-C., Grothe, M.J., Epelbaum, S., Sambati, L., Gagliardi, G., Toschi, N., Greicius, M.D., Dubois, B., Hampel, H., group, f.t.I.-p.s., (APMI), A.P.M.I., 2018. Effect of Alzheimer's disease risk and protective factors on cognitive trajectories in subjective memory complainers: an INSIGHT-preAD study. Alzheimer's & Dementia 14 (9), 1126–1136. doi:10.1016/j.jalz.2018.04.004.

Tung, F., Mori, G., 2019. Similarity-Preserving Knowledge Distillation. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1365–1374. doi:10.1109/iccv.2019.00145.

Vapnik, V., Izmailov, R., 2015. Learning using privileged information: similarity control and knowledge transfer. J. Mach. Learn. Res. 16, 2023–2049. doi:10.1007/978-3-319-17091-6_1.

Wang, M., Lian, C., Yao, D., Zhang, D., Liu, M., Shen, D., 2020. Spatial-temporal dependency modeling and network hub detection for functional MRI analysis via convolutional-Recurrent network. IEEE Trans Biomed Eng 67 (8), 2241–2252. doi:10.1109/TBME.2019.2957921. https://www.ncbi.nlm.nih.gov/pubmed/31825859

Wang, P., Liu, Y., Shen, D., 2019. Flexible locally weighted penalized regression with applications on prediction of Alzheimer's disease neuroimaging Initiative's clinical scores. IEEE Trans. Med. Imag. 38 (6), 1398–1408. doi:10.1109/TMI.2018.2884943. https://www.ncbi.nlm.nih.gov/pubmed/30530315

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., Morris, J.C., Petersen, R.C., Saykin, A.J., Schmidt, M.E., Shaw, L., Siuciak, J.A., Soares, H., Toga, A.W., Trojanowski, J.Q., 2012. The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. Alzheimers Dement 8 (1 Suppl), S1–68. doi:10.1016/j.jalz.2011.09.172.

Yan, Z., Zhan, Y., Peng, Z., Liao, S., Shinagawa, Y., Zhang, S., Metaxas, D.N., Zhou, X.S., 2016. Multi-Instance deep learning: discover discriminative local anatomies for bodypart recognition. IEEE Trans. Med. Imag. 35 (5), 1332–1343. doi:10.1109/TMI.2016.2524985.

Yao, S., Liu, Y., Zheng, X., Zhang, Y., Cui, S., Tang, C., Lu, L., Xu, N., 2020. Do non-pharmacological interventions prevent cognitive decline? a systematic review and meta-analysis. Transl Psychiatry 10 (1), 19. doi:10.1038/s41398-020-0690-4.

You, R., Guo, Z., Cui, L., Long, X., Bao, S.Y.-Z., Wen, S., 2019. Cross-Modality attention with semantic graph embedding for multi-label classification. ArXiv doi:10.1109/jurse.2019.8808940. arXiv:1912.07872. https://ui.adsabs.harvard.edu/abs/2019arXiv191207872Y

Zagoruyko, S., Komodakis, N., 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. ArXiv 102786. doi:10.1016/j.cviu.2019.07.006. arXiv:1612.0.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. Mixup: beyond empirical risk minimization. ArXiv 504–519. doi:10.1007/978-3-030-01231-1_31.

Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H., 2018. Deep Mutual Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4320–4328. doi:10.1109/cvpr.2018.00454.

Zhou, T., Thung, K.H., Zhu, X., Shen, D., 2019. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Hum Brain Mapp 40 (3), 1001–1016. doi:10.1002/hbm.24428. https://www.ncbi.nlm.nih.gov/pubmed/30381863

Zhu, Y., Zhu, X., Kim, M., Yan, J., Kaufer, D., Wu, G., 2019. Dynamic hyper-graph inference framework for computer-assisted diagnosis of neurodegenerative diseases. IEEE Trans. Med. Imag. 38 (2), 608–616. doi:10.1109/TMI.2018.2868086. https://www.ncbi.nlm.nih.gov/pubmed/30183622