2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-1179-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/EMBC46164.2021.9630382

Federated Learning via Conditional Mutual Learning for Alzheimer's Disease Classification on T1w MRI

Ya-Lin Huang[1,2], Hao-Chun Yang[1,2], Chi-Chun Lee[1,2]

Abstract— **Data-driven deep learning has been considered a promising method for building powerful models for medical data, which often requires a large amount of diverse data to be sufficiently effective. However, the expensive cost of collecting and the privacy constraints lead to the fact that existing medical datasets are small-scale and distributed. Federated learning via model distillation is a data-private collaborative learning where the model can leverage all available data without direct sharing. The data knowledge is shared by distillation through the multi-site average prediction scores on the public dataset. However, the average consensus is suboptimal to individual client due to data domain shift in MRI data caused by acquisition protocols, recruitment criteria, etc. In this work, we propose a federated conditional mutual learning (FedCM) to improve the performance by considering the clients' local performance and the similarity between clients. This work is the first federated learning on multi-dataset Alzheimer's disease classification by 3DCNN using T1w MRI. Our method achieves the best recognition rates comparing with FedMD and other frameworks. Further visualization and relevance ranking on the region of interests (ROI) in human brains implies that the left hemisphere may have greater relevance than the right hemisphere does. Several potential regions are listed for future investigation.**

I. INTRODUCTION

With an aging global population, age-related disorders have become a major health problem. Alzheimer's disease (AD) is the most common form of dementia, of which neurodegenerative conditions are considered the most dreaded disease to the elderly. The prevalence of AD is around 3% at the age of 65 and is astonishing 33% for those that are 80 years old and more [1]. Mild cognitive impairment (MCI) has been generally taken as an intermediate state between normal aging and the onset of AD; hence the recognition of MCI is important for prophylactic treatment of AD. The application of deep learning for early detection and stage classifying in AD by using non-invasive data has been considered as a promising clinically assistive method. However, modern DL models require a large amount of data to achieve clinical-grade accuracy, while the privacy concerns restrict access or data sharing in the healthcare domain [2].

Federated Learning (FL) is a learning paradigm that holds great promise on distributed learning by training the algorithm collaboratively across sites without exchanging the data itself [3]. In 2017, McMahan et.al proposed Federated Averaging(FedAvg)[4] following a server-client setup with repeated steps: (1) the local clients train its model for several epochs, and transmit the model weights to the center; (2) the center server collects and aggregate a global model by weight averaging; (3) the clients obtain the global model as the initial model for next local training. Recent studies have shown that models trained under the FL framework have better performance than models that only see isolated site-specific data. Compared with centralized data training, comparable performance can also be achieved[5][6].

Nevertheless, this collaborative framework, while designed for privacy, is still vulnerable to inference attacks by the malicious server or clients through model weight sharing. Malicious members can update the fallacious weight to mislead the update direction or reconstruct other members' data through the pattern encoded in the parameters of the shared model[7]. FL scheme also limits the diversity of the model structure to handle non-IID clients. In addition, the centralized model may be not well-adapted to individual client due to the data size variability where clients have a varying data amount with distinct data distributions[8]. Inspired by the knowledge transfer algorithms, Federated Learning via Model Distillation (FedMD)[9] leverages knowledge distillation[10] to achieve model heterogeneity by sharing the prediction on public data to obtain an average consensus as teacher prediction without sharing private data or model structure. In the absence of direct contact with model weights, this method also prevents reconstruction attacks. However, the average consensus lacks the ability to handle site-wise heterogeneity. In MRI data, the variability of scanners and sites are confounds that hinder the direct pooling of data collected from different sites due to domain shift results from a range of issues, e.g., MRI acquisition protocols, recruitment criteria, different machines, etc.[11].

In this work, we propose FedCM, a novel federated mutual distillation framework with a conditioning mechanism on site-wise performance and probability distribution similarity. We verified the effectiveness of our approach in Alzheimer's disease binary and three-class classification using structural magnetic resonance image (sMRI). In order to simulate the scenario where the clients are composed of different collection sites, we used 3 public datasets (including ADNI, OASIS, and AIBL) and split them into sub-datasets by sites as local clients. In a follow-up analysis, we demonstrate the model attention visualization on the region of interests(ROIs) of sMRI, and discuss the attention differences between 8 different model training frameworks.

[1] Ya-Lin Haung (yalinn@gapp.nthu.edu.tw), Hao-Chun Yang (chadyang.hc@gmail.com) and Chi-Chun Lee (cclee@ee.nthu.edu.tw) are with the Department of Electrical Engineering, National Tsing Hua University (NTHU), Taiwan.

[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

TABLE I: Demographic information

Dataset	Label	Subjects	Sessions	Male	Female	Age
ADNI	CN	169	628	306	322	76.2±5.2
	MCI	301	1142	689	453	74.4±7.2
	AD	139	366	187	179	75.9±7.3
OASIS	CN	316	316	119	197	45.1±23.9
	MCI	70	70	31	39	76.2±7.2
	AD	30	30	10	20	78.0±6.9
AIBL-1	CN	317	555	249	306	74.8±6.6
	MCI	77	111	59	52	77.1±6.7
	AD	69	102	43	59	76.7±7.4
AIBL-2	CN	168	303	141	162	72.4±6.0
	MCI	47	61	40	21	75.0±7.4
	AD	29	37	19	18	74.0±8.6

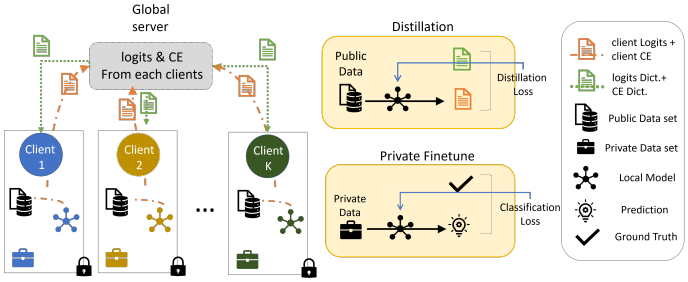


Fig. 1: A schematic of our proposed federated conditional mutual learning framework.

II. RESEARCH METHODOLOGY

A. MRI Data Description and Preprocessing

Data used in the preparation of this work were obtained from three open datasets: the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [1] [12], the Open Access Series of Imaging Studies(OASIS-1) [13], the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) Study [14].

In this work, we used probability maps of GM from T1-weighted (T1w) MRI as our image input. The MRI scans were preprocessed with the Clinica software platform [2] [15]. First, the raw datasets were converted into BIDS format, and the t1-volume preprocessing pipeline of clinica was then applied to the images [16]. In this pipeline, tissue segmentation, bias correction, and spatial normalization were performed simultaneously onto the input image using the unified segmentation approach of SPM12[3]. Next, a group template is created using DARTEL[17] to map the subject's tissue probability to the native space. Lastly, the DARTEL to MNI space method is applied to retrieve the images for classification and analysis. The products of the pipeline are the probability maps of gray matter (GM) and white matter (WM). The subjects are categorized into three groups:

- **CN**: subjects who were diagnosed as cognitively normal.
- **MCI**: subjects who were diagnosed as mild cognitive impairment (MCI), early MCI (EMCI), or late MCI (LMCI).
- **AD**: subjects who were diagnosed as Alzheimer's Disease.

Noted that in OASIS, the patients with clinical dementia rating (CDR) score 0 were labeled as CN, while patients scoring with 0.5 or greater were labeled as AD, even for those with CDR 0.5 who elsewhere may be considered to be diagnosed as MCI. For correction, we re-assigned the label to those in OASIS with CDR 0.5 as MCI, since this work involves multiple data sites, different criteria of labeling may mislead the interpretation.

The demographics of ADNI, OASIS, and AIBL are described in Table I. The subsets of AIBL are split into AIBL-1 AIBL-2 depending on the the client data centers.

[1] adni.loni.usc.edu
[2] www.clinica.cun
[3] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

B. Task Definition

Given a large public dataset $D_0 := \{(X_i^0, y_i^0)\}_{i=1}^{N_0}$ and K small client datasets splitted into training set $D_{k,tr} := \{(X_i^{k,tr}, y_i^{k,tr})\}_{i=1}^{N_{k,tr}}$ and validation set $D_{k,val} := \{(X_i^{k,val}, y_i^{k,val})\}_{i=1}^{N_{k,val}}$, our objective is to find the optimal f_k for each client k that could perform beyond using each individual site's training data while only allow to access public D_0 and D_k. On the other words, we propose an improved federated learning framework that utilize the public dataset as an information sharing medium and benefit each client's learning without compromising their privacy. We will detail our learning framework in the following section.

C. Federated Learning via Model Distillation (FedMD)

Our model is primarily motivated as an extension of Federated Learning via Model Distillation (FedMD)[9] that enables federated learning for individual client models through mutual consensus knowledge distillation while preserving data privacy. Each private model D_k was initially trained on the huge public dataset D_0 which shares a similar task with the private dataset. Then each client will then train on its private training set $D_{k,tr}$ for client-wise adaptation. To perform the cross-clients knowledge distillation, each client would then interpret the knowledge in the private model by computing the prediction logits on the public dataset D_0. Finally, these logits from each of the clients would be gathered and averaged as teacher logits then further send back to each client and fine-tune again as knowledge distillation for each client. Note that the logits predicted on the public dataset was the only information shared throughout the framework to guarantee the privacy preservation.

D. Federated Conditional Mutual Learning (FedCM)

While the original FedMD has achieved great success on federated learning particularly on the synthesized datasets[9], there exist limitations in transferring framework into the real-world medical application. The major constraint is that the whole learning process relies on each client's distilled knowledge, but neglect the fact that there could exist a large heterogeneity among clients and directly average the logits as teacher knowledge could harm the individual site's model performance. To deal with this limitation, we propose the Federated Conditional Mutual Learning (FedCM) that enables the framework with client-aware Mutual Learning[18].

TABLE II: A summary table for experimental results. Mean1: Mean AUC over 3 tasks in binary classification; *: where ds15 is higher than transfer; **: where only predict on a specific label, is excluded when comparing scores ; bold: the highest value.

OASIS

	CN-AD				CN-MCI				MCI-AD				Binary	CN-MCI-AD			
	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE	Mean1	ACC	AUC	SEN	SPE
UB	0.827	0.931	0.733	0.835	0.767	0.873	0.886	0.741	0.680	0.590	0.600	0.714	0.798	0.721	0.794	0.621	0.621
DS15	0.838	0.906*	0.933	0.829	0.808	0.836*	0.686	0.835	0.540	0.577	0.400	0.600	0.773	0.678	0.754*	0.577	0.843
TFL	0.809	0.863	0.867	0.804	0.352	0.515	0.686	0.278	0.560	0.589	0.467	0.600	0.655	0.389	0.601	0.427	0.427
FedAvg[4]	0.890	0.947	0.800	0.899	0.632	0.823	0.943	0.563	0.300	0.773	1.000**	0.000**	0.848	0.168	0.727	0.333**	0.667
FedMD[9]	0.925	**0.972**	0.800	0.937	0.653	0.899	0.971	0.582	0.640	0.564	0.200	0.829	0.812	0.702	0.816	0.574	0.873
$FedCM^E$	0.850	0.962	1.000	0.835	0.762	0.903	0.914	0.728	0.660	0.665	0.533	0.714	0.843	0.721	**0.838**	0.629	**0.870**
$FedCM^J$	0.931	0.942	0.400	0.981	0.845	**0.905**	0.771	0.861	0.700	**0.726**	0.533	0.771	**0.858**	0.705	0.705	0.333**	0.667
$FedCM^{E+J}$	0.919	0.963	1.000	0.911	0.824	**0.905**	0.857	0.816	0.680	0.636	0.400	0.800	0.835	**0.745**	0.816	**0.647**	**0.874**

AIBL-1

	CN-AD				CN-MCI				MCI-AD				Binary	CN-MCI-AD			
	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE	Mean1	ACC	AUC	SEN	SPE
UB	0.853	0.907	0.796	0.863	0.670	0.666	0.607	0.684	0.636	0.705	0.755	0.541	0.759	0.544	0.688	0.495	0.495
DS15	0.734	0.732	0.653	0.749	0.537	0.567	0.590	0.525	0.536	0.619*	0.531	0.541	0.639	0.536	0.652	0.483	0.777
TFL	0.817	0.919	0.857	0.810	0.478	0.714	0.836	0.395	0.573	0.575	0.408	0.705	0.736	0.416	0.595	0.408	0.408
FedAvg[4]	0.606	0.862	0.939	0.544	0.340	0.592	0.934	0.202	0.555	0.684	0.000**	1.000**	0.713	0.164	0.605	0.333**	0.667
FedMD[9]	0.904	0.928	0.796	0.924	0.713	0.700	0.607	0.738	0.664	0.679	0.408	0.869	0.769	0.598	0.723	0.564	0.792
$FedCM^E$	0.862	0.908	0.735	0.886	0.725	0.685	0.541	0.768	0.645	0.723	0.673	0.623	0.772	0.643	0.722	**0.587**	**0.811**
$FedCM^J$	0.891	0.913	0.735	0.920	0.756	**0.713**	0.541	0.806	0.664	**0.724**	0.633	0.689	**0.784**	0.760	**0.775**	0.333**	0.667
$FedCM^{E+J}$	0.888	**0.934**	0.388	0.981	0.809	0.709	0.508	0.878	0.655	0.672	0.531	0.754	0.772	0.660	**0.732**	0.555	**0.799**

AIBL-2

	CN-AD				CN-MCI				MCI-AD				Binary	CN-MCI-AD			
	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE	Mean1	ACC	AUC	SEN	SPE
UB	0.821	0.946	0.938	0.808	0.602	0.593	0.400	0.644	0.717	0.754	0.500	0.833	0.765	0.521	0.647	0.439	0.439
DS15	0.809	0.895	0.875	0.801	0.170	0.500	1.000**	0.000**	0.543	0.563	0.125	0.767	0.652	0.760	0.499	0.333**	0.667
TFL	0.722	0.933	1.000	0.692	0.455	0.559	0.667	0.411	0.674	0.681	0.250	0.900	0.724	0.484	0.651	0.427	0.427
FedAvg[4]	0.099	0.915	1.000**	0.000**	0.210	0.582	0.900	0.068	0.652	0.821	0.000**	1.000**	0.772	0.141	0.682	0.339	0.663
FedMD[9]	0.901	0.942	0.813	0.911	0.653	0.679	0.533	0.678	0.761	0.821	0.563	0.867	0.814	0.464	0.767	0.530	0.733
$FedCM^E$	0.827	**0.973**	1.000	0.808	0.602	**0.703**	0.633	0.596	0.761	0.840	0.625	0.833	**0.838**	0.391	**0.817**	0.643	0.751
$FedCM^J$	0.938	0.919	0.750	0.959	0.619	0.670	0.567	0.630	0.739	**0.848**	0.688	0.767	0.812	0.760	0.686	0.333**	0.667
$FedCM^{E+J}$	0.877	0.961	0.875	0.877	0.545	0.629	0.667	0.521	0.652	0.742	0.625	0.667	0.777	0.422	0.796	**0.649**	**0.760**

Algorithm 1: FedCM

Input : Public dataset D_0, private dataset $D_{k,tr}, D_{k,val}$, private model f_k, for k=1...m

Output: Trained model f_k

Initialize: Pre-train f_k of each client to convergence on the public D_0 . Transfer Learning: Fine tune f_k with private $D_{k,tr}$.

while *Collaboration training* **do**

  Evaluate: Each client evaluate the model on private $D_{k,ts}$ by cross-entropy loss H_k .

  Communicate: Each client compute the class scores on public data D_0, and transmits it to the server along with the last performance H_k. The server collects the record from every clients, then returns the collecting result to each client.

  Mutual learn: Each client updates its model f_k by conditioned mutual loss (Eq.3).

  Revisit: Each client trains its model f_k on own private $D_{k,tr}$ for few epochs.

Each client in FedCM periodically uploads the *predicted logits on the public data* and the *cross-entropy (CE) loss on the private test data* to a server. The server sends the set of logits and CE loss of all clients except the receiving client to the client. Then, each clients updates its knowledge by mutual distillation, and then fine-tune on its private dataset for personalization. The said operation of FedCM is depicted in Figure 1.

Specifically, two additional client-wise mutual conditions are explicitly incorporated into the framework:

- **Entropy Ratio Conditioning (E)**: First, we want to ensure that only reliable knowledge gets distilled and shared. To evaluate a client model's reliability, we calculate the cross-entropy $H(g) = -\sum_{c=1}^{C} y_c \, log \, g_c$ from the former evaluation step, where g is the predicted class probability. Then client j's model reliability term α_j could be formulated as:

$$\alpha_j = 1 + e^{-(\frac{H(g)_j}{H(g)_k})^{\gamma_1}} \tag{1}$$

  where the larger α_j refers that client j's model is more reliable during knowledge distillation. The difference enlarges hyperparameter γ_1 is set as 2 in this work.

- **Jensen Shannon Conditioning (J)** [19]: Second, we presume that a client should learn more from the clients with similar statistical distributions to reduce client bias. To measure the heterogeneity of two clients' samples, we utilize JS Divergence , $D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)$ for computing the probability distribution divergence, where $M = \frac{P+Q}{2}$. Jensen Shannon (JS) Divergence is a symmetrized and smoothed version of Kullback-Leibler (KL) divergence $D_{KL}(P||Q) = \sum_{x \subseteq X} P(x) \log_2 \frac{P(x)}{Q(x)}$, where P and Q are probability distributions. The value of JS divergence is bounded by 0 and 1, and is 0 for identical distributions.

$$\beta_j = (1 - D_{JS}(\, f_k(x_0) \,||\, f_j(x_0)\,))^{\gamma_2} \tag{2}$$

  where the larger β_j refers that the data of client j and client k are relatively homogeneous. The difference enlarges hyperparameter γ_2 is set as 2 in this work.

Finally, the conditioned mutual distillation loss can be formulated as:

$$L_{distil}^{k} = \frac{1}{N_0} \sum_{j \neq k}^{m} \sum_{i}^{N_0} \alpha_j \, \beta_j \, \| f_k(x_i^0)) - f_j(x_i^0) \| \quad (3)$$

III. EXPERIMENTAL SETUP AND RESULTS

A. Experimental Setup

We carried out the classification as binary class and three-class task of CN, MCI and AD for full comparison. Because of the scarcity of AD label, we split each dataset into train and test set with a ratio of 50%-50%. The performance was evaluated in accuracy (ACC), area under curve (AUC), sensitivity (SEN) and specificity (SPE). The metrics are averaged in 3-class classification.

1) Comparison Models: We conduct our experiment with 8 models for comprehensive comparison. ADNI dataset was assigned as public dataset and OASIS, AIBL-1, AIBL-2 as each client's private dataset.

The base 3DCNN architecture based on VGG7 net that consists of six 3D-convolutional layers, which kernel size is 3x3x3 and stride is 1, with numbers of channels 8-8-16-16-32-32, four max-pooling layers and two fully-connected layers. The private train set was down-sampled by 15 for each label to simulate the extremely small dataset.

- **Upperbound(UB)**: The pre-trained 3DCNN is fine-tuned on the aggregation of DS15 samples from all client.
- **DS15**: The base 3DCNN is trained on the down-sampled private set (DS15 sample) with max epoch = 100. The patience is set as 5 for early stopping.
- **Transfer Learning(TFL)**: The base 3DCNN is pre-trained on public dataset, then fine-tuned on private DS15 sample dataset.
- **FedAvg[4]**: The knowledge is exchanged by periodically updating the model initialization weights with the average model weights from every clients.
- **FedMD[9]**: The knowledge is exchanged by distillation through the average logits of each clients on public dataset. Details are depicted in Sec. II-C.
- **FedCM**: To compare the effectiveness of two conditioning terms, we construct $FedCM^E$ as FedCM with entropy ratio conditioning; $FedCM^J$ as FedCM with JS conditioning; $FedCM^{E+J}$ as FedCM with both conditioning.

The batch size of each model was set as 16. The 3DCNN was optimized using Adam optimizer with learning rate = 0.001.

B. Experimental Results and Discussion

Table II. summarizes our complete experimental results. Our proposed method achieves the best recognition rates in most of the tasks among 8 comparison models. Several observations can be summarized. In binary classification, AUC is sufficient for evaluating a model's performance, while all metrics should be considered for a full comparison in a 3-class classification.

In general, the recognition ability of DS15 is low, as expected, in both binary or 3-class tasks. However, we found that the TFL functions is worse than DS15 in most of the tasks trained on OASIS. The phenomenon could be related to the differences between ADNI and OASIS caused by the severe age distributional difference, and it is known that the brain structure varies significantly with age [20].

Then, we observed that the AUC is enhanced in binary tasks after applying FedAvg comparing with DS15, while the 3-class task's result is still unsatisfying. Also, we noticed that both $FedCM^J$ and FedAvg consistently predict on a specific label in 3-class tasks (demonstated by a value of 0.333 SEN). This phenomenon could probably be related to the poor performance of initial transfer. In FedAvg, an originally well-performed model may be degraded significantly by another sub-optimal model through averaging. This procedure continuously leads FedAvg to a worse recognition ability. In $FedCM^J$, the models tend to learn from the model with similar probability distribution which could be detrimental when both models are starting at a wrong initial state.

In viewing all of the results, our proposed methods outperform the former frameworks in our experiment. In binary classification, by averaging the AUC results over the binary tasks (**Mean1**), $FedCM^J$ has the highest average AUC 0.858 in OASIS and 0.784 in AIBL-1, and $FedCM^E$ has the highest value of 0.838 in AIBL-2. In 3-class tasks, $FedCM^E$ and Fed^{E+J} are relatively better than FedMD considering the comprehensive performance of AUC, SEN, and SPE. This indicates that our proposed conditioning distills the model with more effective knowledge than FedMD does.

Finally, we noticed that the model using the distillation framework usually outperforms the upperbound. This may imply that the model learns from the latent knowledge which is less affected by irrelevant noises.

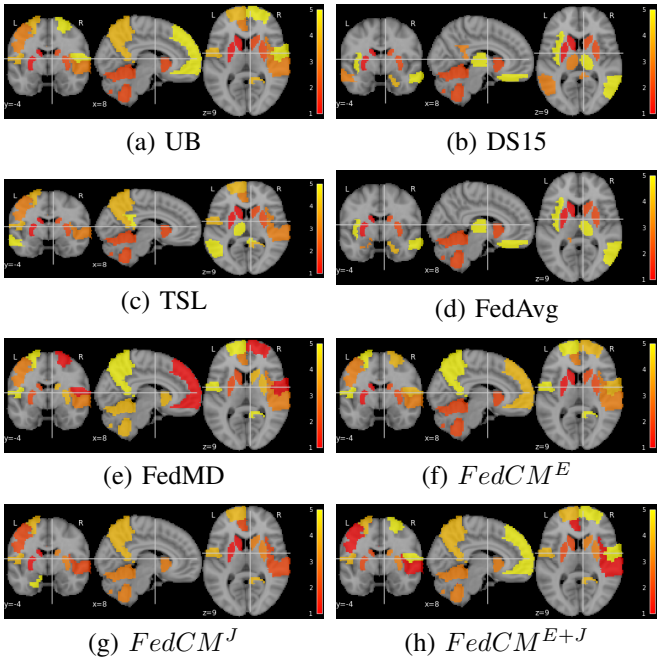
To conclude, our proposed method efficiently improves the model recognition rate on both binary classification and three class classification, while two conditioning methods show different importance under different initial recognition abilities of the participants. The combination between CE ratio conditioning and JS divergence conditioning is the key to obtaining the best performing model.

IV. ANALYSIS

To further interpret the convolutional network for both understandings the model's operation and draw potential clinical insights. We visualize the interpretation of models through computing the gradient of the network's output class score with respect to the image [21]. We take the absolute value of the computed gradient as relevance scores to every pixel on the image. For better interpretation, the relevance scores are sum-up by region of interest (ROI) for ranking the importance of the anatomical regions of the brain to the model. Here, We focus on discussing the relevance analysis on recognizing AD patients by 3-class-task models trained on OASIS dataset due to the page limits.

TABLE III: A summary of the top 15 relevant ROI to 8 model framework predicting on AD patients of OASIS dataset. L: left hemisphere, R: right hemisphere.

No.	UB	DS15	TSL	FedAvg	FedMD	$FedCM^{E}$	$FedCM^{J}$	$FedCM^{E+J}$
1	Cerebelum (L)	Angular gyrus (L)	Cerebelum (L)	Cerebelum (L)	Cerebelum (L)	Cerebelum (L)	Cerebelum (L)	Cerebelum (L)
2	Cerebelum (R)	Angular gyrus (R)	Cerebelum (R)	Cerebelum (R)	Cerebelum (R)	Cerebelum (R)	Cerebelum (R)	Cerebelum (R)
3	Frontal gyrus (L)	Cerebelum (L)	Cuneus (L)	Cingulate gyrus (L)	Cingulate gyrus (L)	Frontal gyrus (L)	Cingulate gyrus (L)	Cingulate gyrus (L)
4	Frontal gyrus (R)	Cerebelum (R)	Frontal gyrus (L)	Cuneus (L)	Cuneus (L)	Frontal gyrus (R)	Cuneus (L)	Cuneus (L)
5	Fusiform gyrus (L)	Cingulate gyrus (L)	Frontal gyrus (R)	Frontal gyrus (L)	Frontal gyrus (L)	Fusiform gyrus (L)	Frontal gyrus (L)	Frontal gyrus (L)
6	Hippocampus (L)	Frontal gyrus (L)	Fusiform gyrus (L)	Frontal gyrus (R)	Frontal gyrus (R)	Hippocampus (L)	Frontal gyrus (R)	Frontal gyrus (R)
7	Hippocampus (R)	Frontal gyrus (R)	Hippocampus (L)	Heschl's gyrus (L)	Fusiform gyrus (L)	Hippocampus (R)	Fusiform gyrus (L)	Fusiform gyrus (L)
8	Lateral orbital gyrus (L)	Fusifo m gyrus (L)	Hippocampus (R)	Hippocampus (L)	Lingual gyrus (L)	Lateral orbital gyrus (L)	Hippocampus (L)	Lingual gyrus (L)
9	Parahippocampal Gyrus (L)	Lingual gyrus (L)	Parahippocampal gyrus (L)	Hippocampus (R)	Parahippocampal gyrus (L)	Parahippocampal gyrus (L)	Lingual gyrus (L)	Anterior orbital gyrus (R)
10	Parahippocampal Gyrus (R)	Parahippocampal gyrus (L)	Parahippocampal gyrus (R)	Parahippocampal gyrus (L)	Parahippocampal gyrus (R)	Parahippocampal gyrus (R)	Lingual gyrus (R)	Lateral orbital gyrus (L)
11	Parietal gyrus (R)	Parahippocampal gyrus (R)	Parietal gyrus (R)	Parahippocampal gyrus (R)	Parietal gyrus (R)	Parietal gyrus (R)	Anterior orbital gyrus (R)	Lateral orbital gyrus (R)
12	Precentral gyrus (R)	Precentral gyrus (R)	Precuneus (L)	Parietal gyrus (R)	Precentral gyrus (R)	Precentral gyrus (R)	Parahippocampal gyrus (L)	Parahippocampal Gyrus (L)
13	Temporal gyrus (L)	Temporal gyrus (L)	Temporal gyrus (L)	Temporal gyrus (L)	Temporal gyrus (L)	Temporal gyrus (L)	Temporal gyrus (L)	Precentral gyrus (R)
14	Temporal gyrus (R)	Temporal gyrus (R)	Temporal gyrus (R)	Temporal gyrus (R)	Temporal gyrus (R)	Temporal gyrus (R)	Temporal gyrus ®	Temporal gyrus (L)
15	Vermis	Vermis	Vermis	Vermis	Vermis	Vermis	Vermis	Vermis



(a) UB (b) DS15 (c) TSL (d) FedAvg (e) FedMD (f) $FedCM^{E}$ (g) $FedCM^{J}$ (h) $FedCM^{E+J}$

Fig. 2: The Visualization of the top 5 relevant ROI of 3DCNN predicting AD patients on OASIS. The models are trained on private OASIS dataset for 3-class tasks. The color stands for the priority of relevance, where the color red has the highest relevance value while yellow has the lowest.

In figure 2., we observed that all models share several overlapping ROIs. Eight ROIs are attended by all the models (see table III), including of cerebellum (L, R), frontal gyrus (L, R), parahippocampal gyrus (R), temporal gyrus (L), and vermis, where these regions are proved to have structural or volume change in AD patients [22]. However, we noticed that there still exists significant differences among the models: (1)attention difference on ROI, (2)models have different priority to ROIs. Subsequent observations confirm the significance of these two factors.

First, we focus on DS15 and UB. One is only allowed to use a private DS15 sample, while the other has direct access to all data samples. Interestingly, they all have a unique viewpoint of ROIs: only DS15 values the left and right angular gyrus, and UB additionally sees left Lateral orbital gyrus comparing with other models. More interestingly, studies have shown that the angular gyrus syndrome shares many clinical features with Alzheimer's disease, and these two conditions are easily confused[23]. Secondly, we studied the difference between UB and FedMD and noticed that compared with UB, FedMD also identifies the left cingulate gyrus, left lingual gyrus, and left cuneus. Then, we compare our proposed $FedCM^{E+J}$, which has a better recognition rates, and FedMD to see if there exists any distinct ROI difference. The top 8 relevant ROIs of the two models are the same. The disjoint sets between FedMD and $FedCM^{E+J}$ are listed: from FedMD, right parahippocampal gyrus, right parietal gyrus, right temporal gyrus; from $FedCM^{E+J}$, right anterior orbital gyrus, left and right lateral orbital gyrus. The brain regions in the disjoint set of FedMD are indicated to have changes in AD [22]. As for the $FedCM^{E+J}$, orbital gyrus is examined to have widespread damages from the viewpoint of neurofibrillary tangle (NFT) pathology [24]. Finally, we noticed that the left hemisphere always has higher priority than right hemisphere, which may imply that left hemisphere has more significance in recognizing Alzheimer's disease.

In summary, a model's performance could be strongly differed by a few attention differences or the slight change in ROIs importance rankings. This phenomenon is consistent with the difficulty in diagnosing neurological disorders considering the entangled and subtle connections among multiple brain regions. Moreover, we draw on the gradients of the neural network to demonstrate that these ROIs could have an influence in distinguishing CN, MCI, and AD. In our future work, we could extend the statistical investigations on the topography of pathological changes from a network-based viewpoint.

V. Conclusion and Future Work

In this work, we introduced a novel federal mutual knowledge distillation framework and validated the framework for classifying CN, MCI, and AD through 3DCNN prediction training on a very small brain imaging data set. Experimental results show that the proposed method is a promising for multi-site FL to improve MR Image classification without compromising privacy. In 3-class classification, our method achieved an accuracy rate of 74.5%, 76.0%, 76.0% in OASIS, AIBL-1, and AIBL-2 respectively, which improved the result of FedMD by 4.3%, 16.2%, 29.7%. Compared with FedMD, FedCM can help improve performance by 6.8%, 2.4%, and -0.6% in the unweighted accuracy in the 2-class classification.

In the future work, we will investigate the interaction between two conditioning mechanisms for optimal parameter setting. Lastly, we will continue to experiment using unlabeled data or synthetic data (such as generative adversarial network (GAN)) for better distillation.

REFERENCES

[1] A. Association *et al.*, "2018 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[2] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[3] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, 2021.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[5] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 92–104.

[6] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results," *Medical Image Analysis*, vol. 65, p. 101765, 2020.

[7] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 2512–2520.

[8] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[9] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," *arXiv preprint arXiv:1910.03581*, 2019.

[10] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1607–1616.

[11] E. Kondrateva, M. Pominova, E. Popova, M. Sharaev, A. Bernstein, and E. Burnaev, "Domain shift in computer vision models for mri data analysis: an overview," in *Thirteenth International Conference on Machine Vision*, vol. 11605. International Society for Optics and Photonics, 2021, p. 116050H.

[12] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685–691, 2008.

[13] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[14] K. Ellis, A. Bush, D. Darby, D. De Fazio, J. Foster, P. Hudson, N. Lautenschlager, N. Lenzo, R. Martins, P. Maruff *et al.*, "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," 2009.

[15] A. Routier, N. Burgos, J. Guillon, J. Samper-González, J. Wen, S. Bottani, A. Marcoux, M. Bacci, S. Fontanella, T. Jacquemont *et al.*, "Clinica: an open source software platform for reproducible clinical neuroscience studies," 2019.

[16] J. Samper-González, N. Burgos, S. Bottani, S. Fontanella, P. Lu, A. Marcoux, A. Routier, J. Guillon, M. Bacci, J. Wen *et al.*, "Reproducible evaluation of classification methods in alzheimer's disease: Framework and application to mri and pet data," *NeuroImage*, vol. 183, pp. 504–521, 2018.

[17] J. Ashburner, "A fast diffeomorphic image registration algorithm," *Neuroimage*, vol. 38, no. 1, pp. 95–113, 2007.

[18] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

[19] L. Lee, "Measures of distributional similarity," *arXiv preprint cs/0001012*, 2000.

[20] C. A. Raji, O. Lopez, L. Kuller, O. Carmichael, and J. Becker, "Age, alzheimer disease, and brain structure," *Neurology*, vol. 73, no. 22, pp. 1899–1905, 2009.

[21] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[22] E. Canu, D. G. McLaren, M. E. Fitzgerald, B. B. Bendlin, G. Zoccatelli, F. Alessandrini, F. B. Pizzini, G. K. Ricciardi, A. Beltramello, S. C. Johnson *et al.*, "Mapping the structural brain changes in alzheimer's disease: The independent contribution of two imaging modalities," *Journal of Alzheimer's Disease*, vol. 26, no. s3, pp. 263–274, 2011.

[23] N. Nagaratnam, T. A. Phan, C. Barnett, and N. Ibrahim, "Angular gyrus syndrome mimicking depressive pseudodementia," *Journal of Psychiatry and Neuroscience*, vol. 27, no. 5, p. 364, 2002.

[24] G. W. Van Hoesen, J. Parvizi, and C.-C. Chu, "Orbitofrontal cortex pathology in alzheimer's disease," *Cerebral Cortex*, vol. 10, no. 3, pp. 243–251, 2000.