Heterogeneous Structured Federated Learning with Graph Convolutional Aggregation for MRI-Based Mental Disorder Diagnosis

1st Yao Hu
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong SAR
City University of Hong Kong
Shenzhen Research Institute
Shenzhen, China
y.hu@my.cityu.edu.hk

2nd Rui Liu
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Hong Kong SAR
ruiliu@polyu.edu.hk

3rd Jiaqi Zhang
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Hong Kong SAR
jqzhang927@gmail.com

4thZhi-An Huang*
Research Office
City University of Hong Kong (Dongguan)
Dongguan, China
City University of Hong Kong
Shenzhen Research Institute
Shenzhen, China
huang.za@cityu-dg.edu.cn

5th Linqi Song*
Department of Computer Science
City University of Hong Kong
Kowloon Tong, Hong Kong SAR
City University of Hong Kong
Shenzhen Research Institute
Shenzhen, China
linqi.song@cityu.edu.hk

6th Kay Chen Tan*
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Hong Kong SAR
kctan@polyu.edu.hk

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10651026

***Abstract*—To relieve the growing burden of mental disorders, deep learning techniques have emerged as a promising tool to aid clinicians by detecting abnormal patterns in neuroimaging data. However, the efficacy of such models is contingent upon access to vast pools of patient data, which is impractical for individual healthcare institutions. Moreover, the privacy-preserving policy regulations governing medical images further complicate the pooling of information necessary for training robust models. Federated Learning (FL) offers a solution to this dilemma by aggregating the local model updates without compromising patient privacy. However, current studies fail to adequately account for the need to personalize models according to the diverse structures of local data. In this work, an effective heterogeneous structured FL framework using graph convolutional aggregation dubbed GAHFL is proposed to diagnose mental disorders on functional magnetic resonance imaging data. In addition, we propose to perform the global model self-evaluation to enable the training to emphasize the samples that are difficult to classify. To solve the catastrophic forgetting problem, we build a historical logit pool to awaken the global model's recognition ability by performing a server knowledge self-distillation. Empirical evaluations demonstrate that the proposed framework achieves averaged diagnosis AUC values of 69.01% and 69.04% with different sizes of public datasets of ABIDE-I and ADHD-200 datasets, respectively. The ablation studies and robustness validation test further demonstrate the superior performance of our framework.**

***Index Terms*—Federated learning, heterogeneous structured local model, graph convolutional network, autism spectrum disorder (ASD), attention deficit/hyperactivity disorder (ADHD), functional magnetic resonance imaging (fMRI).**

I. Introduction

Mental disorders, such as autism spectrum disorder (ASD) and attention deficit/hyperactivity disorder (ADHD), have become a growing global public health concern. Their high prevalence gradually poses a huge pressure on the health center services [1], [2]. In the past decades, computer-aided diagnosis (CAD) approaches are developed to address the psychiatrist shortage by automatically analyzing high-resolution medical images, e.g., functional magnetic resonance imaging (fMRI) [3], [4]. fMRI can investigate aberrant neurobiological functions in mental disorders by detecting tiny changes in blood flow [5], [6]. Recently, deep learning-based CAD approaches (DL-CAD), e.g., long short-term memory network (LSTM) [7], gated recurrent units (GRU) [8], and hopfield neural network [9] et al., achieved decent performance in mental disorder diagnosis. However, the successful training of deep learning models tends to require sufficient training samples.

In the real world, it is impractical to harvest sufficiently large amounts of fMRI training samples. Data sharing strategy naturally becomes a promising strategy to increase the available training data. However, its practical implementation is hampered by the high risk of privacy leakage and legal issues. To address these problems, federated learning (FL) [10]–[12]

*Corresponding author.

emerges as a decentralized collaborative paradigm to locally train models in individual data owners (i.e., clients) and iteratively aggregates the uploaded model parameters on the central server. In this way, the global models aggregated with decentralized knowledge can achieve superior performance over the local models, without the sharing of raw samples. As FL has significant potential in medical institution collaboration, many researchers developed different applications and technical improvements [13]–[15].

Current medical FL studies follow a common assumption that every participating client employs an identical structured local model. In practice, however, different medical institutions may gather disparate types and volumes of data, stemming from their distinct patient populations, imaging methodologies, and diagnostic procedures. The uniform structured local model may fail to capture the specific data attributes and variations prevalent across all institutions. Furthermore, the availability of computational resources may influence the clients' capacity to train the local models. Therefore, it is implausible to presume that all medical institutions employ an identical structured local model.

Heterogeneous structured FL (HFL) has also been an active area of research in view of natural images by allowing every client to build different structured local models. The most popular stream to integrate knowledge from uploaded heterogeneous models relies on a central public dataset. Then, knowledge distillation is executed on the public server dataset to ensemble knowledge from heterogeneous models [16]–[18]. For instance, FedGEMS [17] randomly and equally splits the samples from each dataset into public and private datasets. Then, the knowledge is selectively distilled on the public dataset according to the entropy values of predicted logits. However, the logit entropy may fail to effectively represent the model reliance and capture the relevance among all clients. That is to say, how to judge the 'expert' clients for each sample in the extensive collection of uploaded heterogeneous structured local models still remains an unsolvable problem.

In this study, a Graph convolutional network-based Aggregation method for Heterogeneous Structured Federated Learning (GAHFL) framework is developed for mental disorder diagnosis using fMRI datasets. In GAHFL, within each communication round, every client trains the unique structured local model on its private data and uploads the trained model to the central server. On the central server, to judge the 'expert' clients, we propose a graph convolutional network (GCN)-based knowledge aggregation method (GCNA). Specifically, GCNA formulates the connection between uploaded local models and the global one as graphs. Subsequently, a GCN is developed to dynamically generate a weight for each node in the graph. These weights serve as the model-level weights for facilitating knowledge transfer from heterogeneous structured local models to the global model. Finally, every client downloads the global model and transfers the global knowledge to local model using its private samples. On real-world fMRI datasets, the experimental results indicate that GAHFL can achieve promising averaged AUC values of 69.01% (ACC: 69.57%, SPE: 71.03%, and SEN: 69.02%) and 69.04% (ACC: 68.28%; SPE: 76.70%; SEN: 62.94%) on ABIDE-I and ADHD-200, respectively. We sum up the main contributions of our work into three-folds. First, we propose a novel heterogeneous structured federated learning framework dubbed GAHFL for mental disorder diagnosis on MRI scans. To the best of our knowledge, the proposed GAHFL is the first HFL framework developed for CAD-based mental disorder diagnosis. Second, we propose a GCN-based aggregation method, which formulates the connection between local and global models as graphs and uses the GCN to update the weight value of every node. Third, we showcase the superiority and robustness of the proposed GAHFL on two real world fMRI datasets.

The remainder of this paper is organized as follows. In Section II, we illustrate the proposed framework in detail. Section III conducts a series of simulation experiments to demonstrate the superiority of GAHFL. Finally, we conclude this paper in Section IV.

II. Methodology

In this section, the proposed GAHFL framework is elaborated by following the flowchart shown in Fig. 1, which mainly consists of the local training phase and the global training stage.

A. Problem Formulation

The primary goal of this work is to train a centralized global DL-CAD model θ using the heterogeneous structured local models and the public fMRI samples without implicitly accessing the private local data. We assume there are K clients in FL process and each client $k \in [K]$ processes its private dataset $\mathcal{D}_k = \left\{ \left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)} \right) \right\}_{i=1}^{n_k}$, where n_k denotes the size of corresponding training samples. Each client can train its own heterogeneous structured model (θ_k) using $\mathcal{D}_k$. The objective in client k can be reached as follows,

$$f_k(\theta_k) = \mathcal{L}\left(\theta_k; \mathcal{D}_k\right), \tag{1}$$

where $\mathcal{L}$ denotes the utilized loss function between the predictions and actual annotations. On the central server, we train the centralized global model (θ) using the public dataset $\mathcal{D}_{\mathrm{p}} = \left\{ \left(\mathbf{x}_p^{(i)}, \mathbf{y}_p^{(i)} \right) \right\}_{i=1}^{n_p}$. The optimization objective can be formulated as,

$$\begin{aligned} F(\boldsymbol{\theta}) &\overset{\text{def}}{=} \min_{\boldsymbol{\theta}} \sum_{k=1}^{K} \frac{n_k}{\sum_{k=1}^{K} n_k} \cdot f_k(\boldsymbol{\theta}), \\ \text{s.t. } \boldsymbol{\theta} &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{\theta}; \mathcal{D}_{\mathrm{p}}, \theta_1, ..., \theta_K\right). \end{aligned} \tag{2}$$

where the subjection denotes the loss of knowledge transfer from heterogeneous structured models to the global model based on the public data.

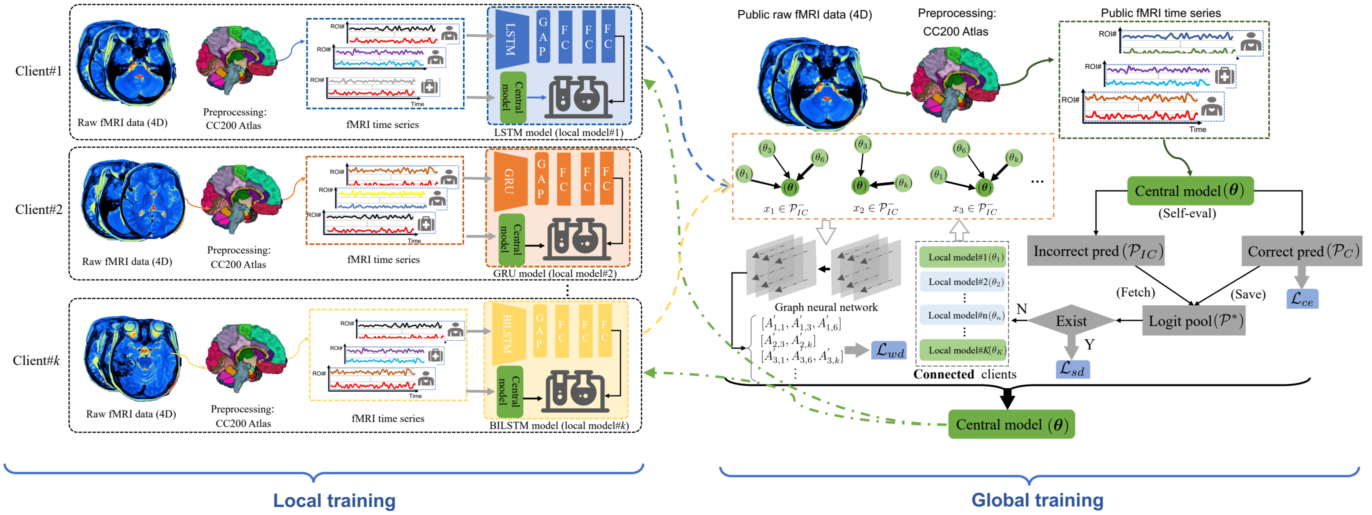
Client#1
Raw fMRI data (4D)
Preprocessing: CC200 Atlas
ROI#
ROI#
ROI#
Time
fMRI time series
LSTM
GAP
FC
FC
FC
Central model
LSTM model (local model#1)
Client#2
Raw fMRI data (4D)
Preprocessing: CC200 Atlas
ROI#
ROI#
ROI#
Time
fMRI time series
GRU
GAP
FC
FC
FC
Central model
GRU model (local model#2)
Client#k
Raw fMRI data (4D)
Preprocessing: CC200 Atlas
ROI#
ROI#
ROI#
Time
fMRI time series
BILSTM
GAP
FC
FC
FC
Central model
BILSTM model (local model#k)
Local training
Public raw fMRI data (4D)
Preprocessing: CC200 Atlas
Public fMRI time series
ROI#
ROI#
ROI#
Time
(θ_3) (θ_6) (θ_1) $(\boldsymbol{\theta})$
$x_1 \in \mathcal{P}^-_{IC}$
(θ_3) $(\boldsymbol{\theta})$ (θ_k)
$x_2 \in \mathcal{P}^-_{IC}$
(θ_6) (θ_k) (θ_1) $(\boldsymbol{\theta})$
$x_3 \in \mathcal{P}^-_{IC}$
...
Graph neural network
$[A'_{1,1}, A'_{1,3}, A'_{1,6}]$
$[A'_{2,3}, A'_{2,k}]$
$[A'_{3,1}, A'_{3,6}, A'_{3,k}]$
$\mathcal{L}_{wd}$
Local model#1 (θ_1)
Local model#2 (θ_2)
Local model#n (θ_n)
Local model#$K(\theta_K)$
Connected clients
Central model $(\boldsymbol{\theta})$
(Self-eval)
Incorrect pred $(\mathcal{P}_{IC})$
Correct pred $(\mathcal{P}_C)$
(Fetch)
(Save)
$\mathcal{L}_{ce}$
Logit pool $(\mathcal{P}^*)$
Exist
N
Y
$\mathcal{L}_{sd}$
Central model $(\boldsymbol{\theta})$
Global training

Fig. 1. The flowchart of the proposed GAHFL framework. The framework consists of the local training phase and the global training stage. In the local training phase, every client trains its own DL-CAD model for mental disorder diagnosis and uploads the model to the central server. On the central server, GAHFL adaptively leverages the uploaded local models to train the global model using the public fMRI samples. After training, every client downloads the global model as aggregated knowledge for knowledge distillation in the local training stage.

B. Heterogeneous Structured Local Model Training

For each communication round in GAHFL, every client first trains its own heterogeneous structured local model using private data. As shown in the left part of Fig. 1, each client first utilizes the Craddock 200 (CC200) functional parcellation atlas [19] to partition the brain into distinct regions of interest (ROIs). Then, the average blood oxygen level-dependent signal is extracted from each ROI to transform the raw fMRI data into fMRI time series, which serve as training samples. GAHFL leverages different local training methods on various communication rounds. Since clients do not receive a knowledgeable global model in the first communication round, the binary cross-entropy loss ($\mathcal{L}_{bce}$) is directly used on the private samples to train local models. In the following rounds, after downloading the global model, each client leverages it as the aggregated knowledge to guide the local training. Concretely, we distill the knowledge from the global model to the local one using the Kullback Leibler (KL)-divergence function ($\mathcal{L}_{KL}$). Accordingly, the training objective for client k can be formulated as follows,

$$\mathcal{L}_k\left(\theta_k; \theta, \mathcal{D}_k\right) = \mathcal{L}_{bce}\left(\theta_k; \mathcal{D}_k\right) + \alpha \mathcal{L}_{KL}\left(lg_{\theta_k}(\mathcal{D}_k) | lg_{\theta}(\mathcal{D}_k)\right), \tag{3}$$

where the hyper-parameter $\alpha > 0$ regulates the effect of $\mathcal{L}_{KL}$. The notions $lg_{\theta_k}(\cdot)$ and $lg_{\theta}(\cdot)$ represent the output logit values of θ_k and θ, respectively. After finishing the local training, the trained heterogeneous structured local models are uploaded to the central server.

C. GCNA-based Global Model Training

Due to the issue of model heterogeneity, the central server fails to directly aggregate local knowledge via the parameter aggregation method. To solve this problem, FedGEMS [17] and FedET [18] leverage the public data as an intermediate to transfer knowledge from local models to the central one. However, facing the extensive collection of uploaded models, the determination of 'expert' local models for achieving an effective knowledge transfer for global model training still remains an unsolvable problem. To this end, we develop the GCNA method in the global training stage to select 'expert' clients by determining their individual model transfer weights.

As illustrated in the right part of Fig. 1, the global training begins with a self-evaluation of the global model on the public dataset $\mathcal{D}_{\text{p}}$. Based on the evaluation results, $\mathcal{D}_{\text{p}}$ can be divided into correct predictions ($\mathcal{P}_C$) and incorrect predictions ($\mathcal{P}_{IC}$). Then, a server knowledge self-distillation strategy with $\mathcal{L}_{sd}$ and the GCNA for weighted knowledge distillation using $\mathcal{L}_{wd}$ to enable the global model to learn from $\mathcal{P}_C$ and $\mathcal{P}_{IC}$, respectively. Accordingly, the total loss of global training can be achieved as follows,

$$\mathcal{L}_{gt} = \mathcal{L}_{bce} + \beta \mathcal{L}_{sd} + \gamma \mathcal{L}_{wd}, \tag{4}$$

where $\beta > 0$ and $\gamma > 0$ are used to regulate the effect of corresponding loss functions. Since the θ has mastered knowledge for the samples in $\mathcal{P}_C$, we simply leverage $\mathcal{L}_{bce}$ to train the global model from $\mathcal{P}_C$. However, the catastrophic forgetting problem may make the trained model lose previously learned knowledge, leading to incorrect predictions for data it had learned. To solve this problem, we save the correct logits into a logit pool $\mathcal{P}^*$ and use the temporally averaged method [20] to update the stored logits. We denote the temporally averaged logit values of x_i at t-th training iteration as $E^{(t)}_{lg}(x_i)$, which is calculated as $E^{(t)}_{lg}(x_i) = \phi E^{(t-1)}_{lg}(x_i) + (1-\phi) lg(x_i)$, where $E^{(t-1)}_{lg}(x_i)$ is the temporally averaged parameters in previous $(t-1)$ iterations and ϕ represents an ensembling momentum parameter. Specially, $E^{0}_{lg}(x_i) = lg_{(x_i)}$. Subsequently, the established summary knowledge is used to enable the global model to recall samples that were accurately predicted in the past. Therefore, to learn knowledge from incorrect predicted data $\mathcal{P}_{IC}$, we first need to check if they exist in $\mathcal{P}^*$.

For the samples existing in $\mathcal{P}^*$, we term them as $\mathcal{P}_{IC}^*$ and believe the global model once had a command of their beneficial knowledge. Thus, the reserved $\mathcal{P}^*$ is used to recover the global model's previous knowledge by performing the self-knowledge distillation. Specifically, for $x_i \in \mathcal{P}_{IC}^*$, the KL-divergence between output logit $lg(x_i)$ and stored logit $\mathcal{P}^*(x_i)$ along with the $\mathcal{L}_{bce}$ between $lg(x_i)$ and the ground truth label are formulated as the self-distillation function ($\mathcal{L}_{sd}$), which is shown as follows,

$$\mathcal{L}_{sd} = \epsilon\mathcal{L}_{bce} + (1-\epsilon)\mathcal{L}_{KL}\left(lg(x_i)|E_{lg}^{(t)}(x_i)\right), \tag{5}$$

where ϵ represents a positive hyper-parameter to balance the effects of $\mathcal{L}_{bce}$ and $\mathcal{L}_{KL}$.

For the remaining samples within $\mathcal{P}_{IC}$, called $\mathcal{P}_{IC}^-$, these samples have never been accurately predicted. Therefore, global models are considered incapable of learning them by itself. To this end, we leverage the proposed GCNA method to extract helpful knowledge from the uploaded heterogeneous structured local models to refine the global model. For $x_i \in \mathcal{P}_{IC}^-$, we select the uploaded models that can make accurate predictions as beneficial models. However, different local models may contribute varying degrees of valuable knowledge. To assess the contribution degree among beneficial models, we formulate the link between the helpful and global models as graphs and use a GCN to update the relevance information. For $x_i \in \mathcal{P}_{IC}^-$, the connections between helpful heterogeneous structured models and the global model can be formulated in a graph as $G_i = (V_i, E_i)$, in which the n_i beneficial local models and global model are considered as the vertices V_i while their connections are termed as the edges E_i, respectively. The predicted logit sets from helpful local and global models are used as embedding ($Z_i \in \mathbb{R}^{(n_i+1)\times 2}$) to represent the node. Then, we use the entropy of logit values as represent the relationships between nodes to calcualte the adjacency matrix ($A_i \in \mathbb{R}^{(n_i+1)\times(n_i+1)}$) of G_i. Inspired by the remarkable ability of GCN to capture intricate and non-linear relationships among nodes, we propose to use GCN to update the node relationship for selecting 'expert' clients.

Algorithm 1 The pseudo-code of GAHFL. T is the number of communication rounds; G and L represent the global and local model training epoch, respectively; $\mathcal{D}_k$ and $\mathcal{D}_\text{p}$ are private dataset in client k and the public dataset on the server.

```
function SERVEREXECUTION($\mathcal{D}_\text{p}$, $G$)    ▷Run on the server
    Initialize $\theta$ as global model
    for each round $t \in 1, 2, \ldots, T$ do
        for each connected client $k$ in parallel do
            $\theta_k \leftarrow ClientUpdate\ (\theta, t, L)$
        end for
        for $g \leftarrow 1, ..., G$ do # Start global training iteration
            for $x, y \in \mathcal{D}_p$ do
                Divide $\mathcal{P}_C$, $\mathcal{P}_{IC}^*$, $\mathcal{P}_{IC}^-$ based on $\theta(x, y)$;
                if $x, y \in \mathcal{P}_C$ then
                    Update $\theta$ via $\mathcal{L}_{bce}$;
                else if $x, y \in \mathcal{P}_{IC}^*$ then
                    Update $\theta$ via Eq. (5);
                else if $x, y \in \mathcal{P}_{IC}^-$ then
                    Update $\theta$ via $[\theta_1, \ldots, \theta_k]$ and Eq. (7);
                end if
            end for
        end for
    end for
    return $\theta$
end function

function CLIENTUPDATE($\theta$, $t$, $L$)    ▷Run on client $k$
    Initialize $\theta_k$ as local model
    for $l \leftarrow 1, ..., L$ do # Start local training iteration
        for $x, y \in \mathcal{D}_k$ do
            if $t$ = 1 then
                Update $\theta_k$ via $\mathcal{L}_{bce}$;
            else if t >1 then
                Update $\theta_k$ via $\theta$ and Eq. (3);
            end if
        end for
    end for
    return $\theta_k$
end function
```

With the node embedding set $Z = \left\{Z_1, \ldots, Z_{|\mathcal{P}_{IC}^-|}\right\}$ and the adjacency matrix set $A = \left\{A_1, \ldots, A_{|\mathcal{P}_{IC}^-|}\right\}$, the formulation of l-th layer of GCN can be calculated as

$$H^{(l)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l-1)}W^{(l-1)} + b^{(l-1)}\right), \tag{6}$$

where $\tilde{D}$ is the degree matrix of $\tilde{A}$ and $\tilde{A}$ denotes the adjacency matrix A with self-connecting edges, respectively. The weight matrix $W^{(l-1)}$ and bias term $b^{(l-1)}$ represent learnable parameters and σ is the employed activation function. We denote the extracted features from $l-1$ graph convolution layer as $H^{(l-1)}$, specifically $H^0 = Z$. Using the output node embedding set H' from the last layer, the adjacency matrix A' between all nodes is updated by calculating the dot product, i.e., $A' = Z' \cdot Z'^\top$. As such, for $x_i \in \mathcal{P}_{IC}^-$, we can determine the knowledge transfer weights of beneficial heterogeneous structured local models ($A'_{i,1}, \ldots, A'_{i,n_i}$), which is employed as intensities on the knowledge distillation. Accordingly, the weighted knowledge distillation $\mathcal{L}_{wd}$ can be written as:

$$\mathcal{L}_{wd} = \xi\mathcal{L}_{bce} + (1-\xi)\sum_{i=1}^{|\mathcal{P}_{IC}^-|}\sum_{k=1}^{n_i} A'_{i,k}\,\mathcal{L}_{KL}\left(lg_\theta(x_i)|lg_{\theta_k}(x_i)\right), \tag{7}$$

where ξ functions as a positive hyperparameter to modulate the impact of the given function. Upon completing the global training, each client downloads the global model as aggregated knowledge to guide the local training using Eq. (3). To ease the understanding of the GAHFL framework, the pseudo-code is shown in Algorithm 1.

TABLE I
DATA DISTRIBUTION FOR EACH CLIENT

ABIDE-I	Client #1	Client #2	Client #3	Client #4

Sites	UM_1(106)
	UM_2(34)
	UCLA_1(72)
	UCLA_2(26)
	OHSU(26)
	CALTECH(37)
		USM(71)
		TRINITY(47)
		MAX_MUM(52)
		STANFORD(39)
		SDSU(36)
		OLIN(34)
			LEUVEN_1(29)
			LEEUVEN_2(34)
			PITT(56)
			YALE(56)
			KKI(48)
			SBL(30)
			CMU(27)
				NYU(175)

Total #	301	279	280	175
ADHD-200	Client #1	Client #2	Client #3	Client #4

Sites	OHSU(258)	NEURO(94)
		NYU(73)
		WASHU(59)
			PEKING(113)
			PITT(98)
				BROWN(244)

Total #	258	226	211	244

III. EXPERIMENTS AND RESULTS

A. Configuring Simulation Settings

In this work, we evaluate the effectiveness of our proposed GAHFL framework on two public resting-state fMRI aggregation datasets, i.e., Autism Brain Imaging Data Exchange I[1] (ABIDE-I for ASD) and the ADHD-200 Competition[2] (ADHD-200 for ADHD). Specifically, the ABIDE-I dataset collects 1035 valid samples (aged 7 to 64) from 17 different international sites with 505 ASD patients and 530 health controls (HCs). We use the Configurable Pipeline for the Analysis of Connectomes (CPAC)[3] [21] to preprocess ABIDE-I. The ADHD-200 dataset archives 358 ADHD patients and 581 HCs (aged 7 to 21) from 7 international sites. The ADHD-200 samples are processed via the pipeline of Athena[4]. In our study, the Craddock 200 (CC200) atlas is adopted to extract the mean time-series for a set of 200 ROIs. To simulate the experimental settings of FL, we divide both databases into four groups as different clients as presented in Table I following [13]. In order to keep the time course dimensions of split samples identical, we perform random cropping on the time-series of each subject [22]. Concretely, we fix the sequence length $T = 90$ and randomly crop 10 sequences for each sample. As such, we can crop each sampled sequence into a space-time matrix with the size of 90×200. The cropped samples are used as training sets for model training.

To learn from the time-series, we pre-defined four different four commonly used models, including the LSTM, GRU, bidirectional LSTM (BILSTM), and a shallow LSTM (SLSTM) as heterogeneous structured local models. Specific parameter settings and model configurations of GAHFL are tabulated in Table II. We use five-fold cross validation (CV) to assess the effectiveness of the proposed framework, in which all datasets are divided into five folds. Then, each fold serves as testing in turns while the rest folds are used as training sets. The classification performance is evaluated via four different evaluation metrics, including accuracy (ACC), specificity (SPE), sensitivity (SEN), and the area under the receiver operating characteristic curve (AUC).

TABLE II
PARAMETER SETTINGS OF GAHFL

For GAHFL			
# of client	4	Communication rounds	40
Configuration of GRU	GRU & GAP& F#1-F#2-F#3-F#4		
Configuration of LSTM	LSTM & GAP & F#1-F#2-F#3-F#4		
Configuration of SLSTM	LSTM & GAP & F#1-F#2-F#4		
Configuration of BILSTM	BILSTM & GAP & F#1-F#2-F#3-F#4		
LSTM hidden size	32	GRU hidden size	32
BILSTM hidden size	32	GAP size	90
F#1 configuration	10	F#2 configuration	20
F#3 configuration	10	F#4 configuration	2
Batch_first	True	Activation function	Softmax
For local training			
Training epoch	35	Batch size	256
Learning rate	0.001	Optimizer	Adam
Dropout rate	0.25	Hyper-parameter α	5
For global training			
Training epoch	50	Batch size	256
Learning rate	0.001	Optimizer	Adam
Dropout rate	0.25	GCN hidden dim	20
GCN layer	2	Hyper-parameter β	1
Hyper-parameter γ	10	Hyper-parameter ϕ	0.8
Hyper-parameter ϵ	0.5	Hyper-parameter ξ	0.4

Note: F denotes fully connected layer

B. Performance Comparison

The FedGEMS [17] and FedET [18] are used for comparison, representing the state-of-the-art performance in HFL. FedGEMS randomly and equally splits local samples from every client into a public and a private dataset and then performs knowledge distillation on the public dataset to transfer heterogeneous structured local models' knowledge to the global model. The entropy values of predicted logits are taken as weights to aggregate knowledge. Similarly, FedET leverages the variance within logit vectors as weights to derive the consensus out of the ensemble of models. In addition, FedET also encourages diversity across models to improve the generalization performance of global model. The FedET is properly modified to leverage the supervisory information from the public dataset by replacing the generated pseudo-labels with the actual annotations. The ratio of the public dataset occupying the training set (R) is an important hyper-parameter to balance the effectiveness and privacy degree, e.g., $R = 50\%$ in FedGEMS. Here, we compare the model performance under different ratios, i.e., $R \in \{20\%, 30\%, 40\%\}$. The specific results are tabulated in Table III.

As we can see using 30% training data as public samples, GAHFL achieves the promising AUC results of 69.28% (ACC: 69.80%, SPE: 72.50%, and SEN: 68.14%) and 69.43% (ACC: 68.55%; SPE: 76.83%; SEN: 63.41%) on ABIDE-I and ADHD-200, respectively. The performance of the FedGEMS and FedET frameworks is inferior to that of our GAHFL. This is mainly because the FedGEMS fails to capture the relevance among all clients and effectively determine the model reliance value. In addition, FedET does not additionally emphasize the training on the samples that the global model cannot correctly classify. By contrast, the GCNA formulates the model connection as graphs and leverages GCN to capture the client reliance so as to update the model reliability. Further-

[1] http://fcon_1000.projects.nitrc.org/indi/abide

[2] http://fcon_1000.projects.nitrc.org/indi/adhd200/

[3] https://fcp-indi.github.io/

[4] https://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline

TABLE III
PERFORMANCE COMPARISON ON ABIDE-I AND ADHD-200 (%).

	ABIDE-I (ASD)				ADHD-200 (ADHD)			
Methods	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC
Public data ratio R = 20%								
FedGEMS	65.83(↓1.7)	67.61(↓2.1)	64.11(↓2.0)	65.46(↓1.4)	65.43(↓0.9)	69.82(↓6.0)	61.65(↑0.6)	65.60(↓1.7)
FedET	64.89(↓2.6)	**72.34(↑2.6)**	60.28(↓5.8)	65.50(↓1.4)	63.66(↓2.7)	71.12(↓4.2)	58.85(↓2.2)	64.33(↓3.0)
GAHFL	**67.51**	69.75	**66.11**	**66.90**	**66.34**	**75.77**	**61.01**	**67.29**
Public data ratio R = 30%								
FedGEMS	67.99(↓1.8)	64.58(↓7.9)	**75.12(↑7.0)**	67.59(↓1.7)	67.22(↓1.3)	75.07(↓1.8)	62.10(↓1.3)	67.94(↓1.5)
FedET	68.97(↓0.8)	66.23(↓6.3)	73.69(↑5.6)	68.46(↓0.8)	66.82(↓1.7)	73.47(↓3.4)	61.95(↓1.5)	67.21(↓2.2)
GAHFL	**69.80**	**72.50**	68.14	**69.28**	**68.55**	**76.83**	**63.41**	**69.43**
Public data ratio R = 40%								
FedGEMS	69.34(↓2.1)	70.48(↓0.3)	68.23(↓4.6)	69.13(↓1.7)	68.56(↓1.4)	74.42(↓3.1)	63.95(↓0.5)	68.93(↓1.5)
FedET	69.50(↓1.9)	69.38(↓1.5)	70.33(↓2.5)	69.22(↓1.6)	68.06(↓1.9)	75.28(↓2.2)	62.96(↓1.5)	68.70(↓1.7)
GAHFL	**71.40**	**70.83**	**72.81**	**70.84**	**69.96**	**77.51**	**64.41**	**70.41**

↑ and ↓ represent the better and worse difference compared to the GAHFL, respectively.

more, the utilization of historical logit information to awaken the global model's memory for addressing the catastrophic forgetting problem could further improve the generalization ability. It is observed that FedGEMS struggles to manage a trade-off between SPE and SEN in ADHD-200 dataset due to the skewed data distribution (358 vs 581). As indicated by the experimental results, the GAHFL can alleviate this issue to some extent. As expected, the proposed framework gets performance improvement with the increasing of available public samples. Specifically, both ACC and AUC can steadily increase despite some variances shown in SPE and SEN. The improvements could be attributed to the ability of GCNA to capture model relevance and determine the knowledge transfer weights. Therefore, even when only 20% of the training set consists of public data, GAHFL achieves an average ACC and AUC of 66.93% and 67.10%, respectively. The results are still superior to those of FedGEMS and FedET with averaged ACC and AUC improvements of 1.57% and 2.18%, demonstrating the effectiveness of GCNA in easing data scarcity. Furthermore, when R increases from 30% to 40%, the averaged AUC of FedGEMS and FedET are improved by 1.35% and 0.89%, respectively, whereas those of GAHFL are improved by 1.51%. It turns out that the GAHFL can obtain a more obvious promotion with the increasingly available public dataset. Based on these results, we conclude that the GAHFL is more effective and efficient than FedGEMS and FedET in extracting knowledge from heterogeneous structured local models.

C. Ablation Studies

Since $\mathcal{L}_{bce}$, $\mathcal{L}_{sd}$, and $\mathcal{L}_{wd}$ are three key components in the proposed GCNA, we conduct an ablation study to evaluate their contributions to the performance improvement with R = 30%. The experimental results are shown in Table IV.

First, the $\mathcal{L}_{bce}$ is excluded to investigate its influence. We can observe that there are 3.24% and 3.09% of AUC degradation on the ABIDE-I and ADHD datasets, respectively. The degradation is mainly because the removal of $\mathcal{L}_{bce}$ disables the GCNA to further learn the samples that were once correctly predicted. The immediate data exclusion may make the model fail to learn the intrinsic data knowledge, even though it can make the correct prediction once. In addition, the catastrophic forgetting problem may cause the model to gradually lose its diagnosis ability with increasing training iterations. Subsequently, the proposed framework without $\mathcal{L}_{sd}$ is evaluated. We note that the declines of ACC and AUC reach a ratio of 2.00% and 1.73% on average in the two datasets. The absence of $\mathcal{L}_{sd}$ diminishes the capability of global model to leverage the historical information to purify its knowledge of these incorrect predictions. Finally, we evaluate the exclusion of $\mathcal{L}_{wd}$ from the proposed GAHFL, which results in a noticeable performance drop of 2.81% in terms of AUC averaged across two datasets. This is mainly because the removal of $\mathcal{L}_{wd}$ disables the enhanced training on $\mathcal{P}_{IC}$. Furthermore, these results demonstrate the effectiveness of leveraging the supervision from beneficial local models to help the global model refine the decision boundary. Summarily, $\mathcal{L}_{bce}$, $\mathcal{L}_{sd}$, and $\mathcal{L}_{wd}$ are demonstrated to make great contributions to the performance improvement of GAHFL.

TABLE IV
ABLATION STUDIES ON THE PROPOSED GAHFL FRAMEWORK (%).

Methods	ABIDE-I (ASD)			
	ACC	SPE	SEN	AUC
w/o $\mathcal{L}_{bce}$	66.53(↓3.3)	64.01(↓8.5)	71.01(↑2.9)	66.04(↓3.2)
w/o $\mathcal{L}_{sd}$	67.44(↓2.4)	69.59(↓2.9)	65.80(↓2.3)	67.44(↓1.8)
w/o $\mathcal{L}_{wd}$	66.62(↓3.2)	67.90(↓4.6)	65.60(↓2.5)	66.49(↓2.8)
Full	69.80	72.50	68.14	69.28
Methods	ADHD-200 (ADHD)			
	ACC	SPE	SEN	AUC
w/o $\mathcal{L}_{bce}$	65.40(↓3.1)	73.94(↓2.9)	60.00(↓3.4)	66.34(↓3.1)
w/o $\mathcal{L}_{sd}$	66.91(↓1.6)	74.74(↓2.1)	61.93(↓1.5)	67.82(↓1.6)
w/o $\mathcal{L}_{wd}$	66.30(↓2.3)	71.38(↓5.5)	62.28(↓1.1)	66.61(↓2.8)
Full	68.55	76.83	63.41	69.43

Notes: w/o is the abbreviation of without; ↑ and ↓ represent the better and worse difference compared to the full GAHFL, respectively.

D. Robustness Validation

Besides training a powerful global model, it is also important to ensure the framework is resistant to different types

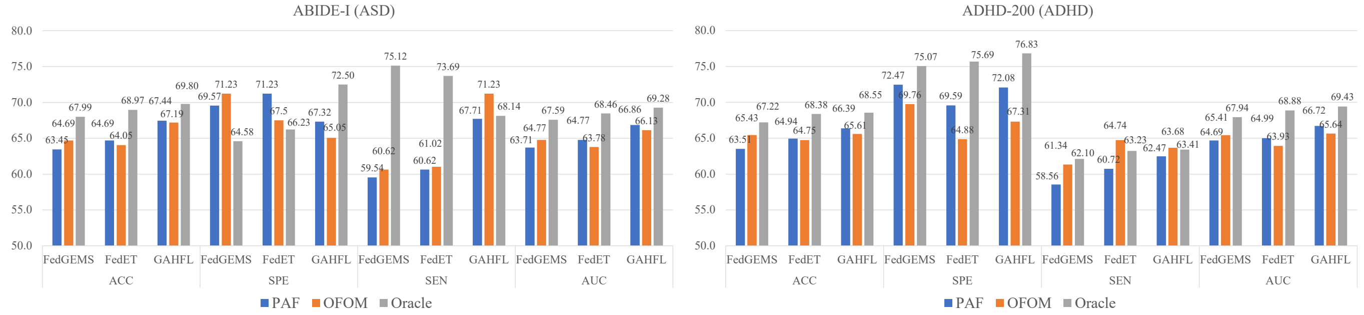

Fig. 2. Robustness validation of different HFL frameworks by adding different attacks. The oracle denotes the model performance without any attacks.

of adversaries, i.e., the framework's robustness. To validate the robustness of the proposed GAHFL framework, we add two different attacks, including naive poisoning (PAF) [23] and one far one mean (OFOM) [23], on the uploaded heterogeneous structured models with $R = 30\%$. Specifically, PAF can poison the δ portion of clients by crafting the malicious update θ_m to be arbitrarily far from the mean of the benign update as follows:

$$\theta_m = \frac{\sum_{k=1}^{(1-\delta)n} \theta_k}{(1-\delta)n} + \theta', \tag{8}$$

where θ' is an arbitrarily large vector that has the same size as $|\theta_k|$. The OFOM attack generates two malicious updates as follows:

$$\theta_1^m = \frac{\sum_{k=1}^{n} \theta_k}{n} + \theta', \quad \theta_2^m = \frac{\sum_{k=1}^{n} \theta_k + \theta_1^m}{n+1}, \tag{9}$$

where θ_1^m is arbitrarily far away from the true mean and the θ_2^m represents the empirical mean of benign updates and θ_1^m, respectively. The poisoning attacks are added following the setting of [17]. The results are presented in Fig. 2.

Based on the results, we can find all three frameworks suffer from the added attacks. Specifically, the added PAF and OFOM attacks lead to the accuracy degradation of 2.49% and 2.55% on ABIDE-I and AHDH-200 datasets, respectively. In contrast, the accuracy performance of FedGEMS and FedET has an average drop of 3.34% and 4.18%, respectively. The results demonstrate the distinct advantage of GAHFL in defending the poisoning attacks over FedET and FedGEMS. The reasons could be attributed to the following reasons: i) FedET does not perform a sample division, thereby struggling to select helpful clients so as to emphasize the incorrect predictions. ii) Without effectively capturing the relevance of various local models, FedGEMS fails to assign proper weights for the helpful local models, thus aggravating the negative influence of the added attacks. In addition, we observe that FedGEMS outperforms the FedET under both poisoning attacks, illustrating the effectiveness of the helpful local model selection strategy in improving the model robustness. Furthermore, increasing the available training samples may also increase the model's robustness against poisoning attacks, which is validated by the small AUC gap on the larger ABIDE-I dataset.

IV. Conclusions and Future Work

In this study, an effective heterogeneous structured FL framework named GAHFL has been first proposed for the diagnosis of various mental disorders using fMRI data. This framework leverages the graph convolutional network-based knowledge aggregation method to dynamically update the distributed model weights. Specifically, it formulates the connections between valuable local models and the central model through graph representation by refining the features to identify 'expert' clients. The effectiveness of GAHFL is demonstrated by 5-fold cross-validation. The ablation studies demonstrate the important role of each component in enhancing overall performance. In addition, the robustness of GAHFL is validated by its resilience against various adversarial attacks. In the future, we intend to refine the knowledge aggregation method for heterogeneous local models by eliminating the dependency on the construction of public datasets. Furthermore, we also expect to apply GAHFL to a broader spectrum of deep learning-based downstream analysis tasks, such as brain tumor segmentation and localization.

V. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62202399, 62371411, 62306259, and U21A20512, in part by the Research Grants Council of the Hong Kong SAR under Grants GRF 11217823, PolyU11211521, PolyU15218622, PolyU15215623, and PolyU25216423, and in part by the Hong Kong Polytechnic University (Project IDs: P0039734, P0035379, P0043563, and P0046094), and in part by InnoHK initiative, the Government of the HKSAR, and in part by the City University of Hong Kong (Dongguan).

References

[1] Y. Hu, Z.-A. Huang, R. Liu, X. Xue, X. Sun, L. Song, and K. C. Tan, "Source free semi-supervised transfer learning for diagnosis of mental disorders on fMRI scans," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 778–13 795, 2023.

[2] Y. Hu, Z.-A. Huang, R. Liu, X. Xue, L. Song, and K. C. Tan, "A dual-stage pseudo-labeling method for the diagnosis of mental disorder on MRI scans," in *Proceedings of International Joint Conference on Neural Networks*. IEEE, 2022, pp. 1–8.

[3] Z.-A. Huang, Z. Zhu, C. H. Yau, and K. C. Tan, “Identifying autism spectrum disorder from resting-state fMRI using deep belief network,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 2847–2861, 2021.

[4] Z.-A. Huang, R. Liu, Z. Zhu, and K. C. Tan, “Multitask learning for joint diagnosis of multiple mental disorders in resting-state fMRI,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022, early access, doi = 10.1109/TNNLS.2022.3225179.

[5] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, “Spatial–temporal co-attention learning for diagnosis of mental disorders from resting-state fMRI data,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023, early access, doi = 10.1109/TNNLS.2023.3243000.

[6] R. Liu, Z-A Huang, Y. Hu, Z. Zhu, K-C. Wong, and K. C. Tan, “Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022, early access, doi=10.1109/TNNLS.2022.3219551.

[7] Y. Hu, X. Sun, X. Nie, Y. Li, and L. Liu, “An enhanced LSTM for trend following of time series,” *IEEE Access*, vol. 7, pp. 34 020–34 030, 2019.

[8] T. Wadhera, J. Bedi, and S. Sharma, “Autism spectrum disorder prediction using bidirectional stacked gated recurrent unit with time-distributor wrapper: an eeg study,” *Neural Computing and Applications*, vol. 35, no. 13, pp. 9803–9818, 2023.

[9] K. C. Tan, H. Tang, and S. S. Ge, “On parameter settings of hopfield networks applied to traveling salesman problems,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, no. 5, pp. 994–1002, 2005.

[10] Y. Hu, X. Sun, Y. Tian, L. Song, and K. C. Tan, “Communication efficient federated learning with heterogeneous structured client models,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 3, pp. 753–767, 2023.

[11] Y. Hu, X. Sun, Y. Chen, and Z. Lu, “Model and feature aggregation based federated learning for multi-sensor time series trend following,” in *Proceedings ofInternational Work-Conference on Artificial Neural Networks*. Springer, 2019, pp. 233–246.

[12] Y. Chen, X. Sun, and Y. Hu, “Federated learning assisted interactive eda with dual probabilistic models for personalized search,” in *International conference on swarm intelligence*. Springer, 2019, pp. 374–383.

[13] Z.-A. Huang, Y. Hu, R. Liu, X. Xue, Z. Zhu, L. Song, and K. C. Tan, “Federated multi-task learning for joint diagnosis of multiple mental disorders on MRI scans,” *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 4, pp. 1137–1149, 2023.

[14] 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.

[15] L. Peng, N. Wang, N. Dvornek, X. Zhu, and X. Li, “Fedni: Federated graph learning with network inpainting for population-based disease prediction,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, pp. 2032–2043, 2022.

[16] D. Li and J. Wang, “Fedmd: Heterogenous federated learning via model distillation,” in *NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019.

[17] S. Cheng, J. Wu, Y. Xiao, and Y. Liu, “Fedgems: Federated learning of larger server models via selective knowledge fusion,” *arXiv preprint arXiv:2110.11027*, 2021.

[18] Y. J. Cho, A. Manoel, G. Joshi, R. Sim, and D. Dimitriadis, “Heterogeneous ensemble knowledge transfer for training large models in federated learning,” in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022.

[19] R. C. Craddock, G. A. James, P. E. Holtzheimer III, X. P. Hu, and H. S. Mayberg, “A whole brain fMRI atlas generated via spatially constrained spectral clustering,” *Human Brain Mapping*, vol. 33, no. 8, pp. 1914–1928, 2012.

[20] Y. Ge, D. Chen, and H. Li, “Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification,” *arXiv preprint arXiv:2001.01526*, 2020.

[21] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. S. Khundrakpam, J. D. Lewis, Q. Li, M. Milham *et al.*, “The neuro bureau preprocessing initiative: Open sharing of preprocessed neuroimaging data and derivatives,” *Frontiers in Neuroinformatics*, vol. 7, 2013.

[22] N. C. Dvornek, P. Ventola, and J. S. Duncan, “Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks,” in *Proceedings of International Symposium on Biomedical Imaging*. IEEE, 2018, pp. 725–728.

[23] H. Chang, V. Shejwalkar, R. Shokri, and A. Houmansadr, “Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer,” in *Proceedings of Advances in Neural Information Processing Systems*, 2021.