Federated Multi-Task Learning for Joint Diagnosis of Multiple Mental Disorders on MRI Scans

Zhi-An Huang, Yao Hu, Rui Liu, Xiaoming Xue, Zexuan Zhu, *Senior Member, IEEE*, Linqi Song, *Senior Member, IEEE*, and Kay Chen Tan, *Fellow, IEEE*

Abstract*—*Objective: **Deep learning (DL) techniques have been introduced to assist doctors in the interpretation of medical images by detecting image-derived phenotype abnormality. Yet the privacy-preserving policy of medical images disables the effective training of DL model using sufficiently large datasets. As a decentralized computing paradigm to address this issue, federated learning (FL) allows the training process to occur in individual institutions with local datasets, and then aggregates the resultant weights without risk of privacy leakage.** ***Methods:*** **We propose an effective federated multi-task learning (MTL) framework to jointly identify multiple related mental disorders based on functional magnetic resonance imaging data. A federated contrastive learning-based feature extractor is developed to extract high-level features across client models. To ease the optimization conflicts of updating shared parameters in MTL, we present a federated multi-gate mixture of expert classifier for the joint classification. The proposed framework also provides practical modules, including personalized model learning, privacy protection, and federated biomarker interpretation.** ***Results:*** **On real-world datasets, the proposed framework achieves robust diagnosis accuracies of 69.48 ± 1.6%, 71.44 ± 3.2%, and 83.29 ± 3.2% in autism spectrum disorder, attention deficit/hyperactivity disorder, and schizophrenia, respectively.** ***Conclusion:*** **The proposed framework can effectively ease the domain shift between clients via federated MTL.** ***Significance:*** **The current work provides insights into exploiting the advantageous knowledge shared in related mental disorders for improving the generalization capability of computer-aided detection approaches.**

***Index Terms*—Federated multi-task learning, functional magnetic resonance imaging, domain shift, joint diagnosis, autism spectrum disorder, attention deficit/hyperactivity disorder, schizophrenia.**

I. Introduction

MENTAL healthy issues have become a growing global public health concern. It is reported that 17.6% of the adult population worldwide are suffering from common mental disorders [1]. In the past decade, computation-aided detection approaches have emerged to address an escalating psychiatrist shortage through analysis of high-resolution magnetic resonance imaging (MRI) [2], [3]. Particularly, deep learning (DL) techniques can achieve outstanding diagnostic performance when sufficient training data are available [4]. DL models require sufficiently vast training data to ensure proper generalization of the trained model.

However, it is unpractical to harvest large amounts of neuroimaging data at a single site in practice. Data sharing strategy naturally becomes a promising solution by increasing both data size and diversity. This strategy holds a significant potential but faces several problems. Although data anonymization seemingly bypass privacy protection rules, it cannot technically fully eliminate privacy leakage [5]. To reduce the risk of data sharing, data owners can specify their own application policies and license restrictions for access control. Nevertheless, once the data is released, it is impossible to manage data transmission because recalling data from approved users is practically unenforceable. Furthermore, data ownership also stifles the opportunities for data sharing because of legal access, technological, and behavioural barriers.

Despite these concerns, federated learning (FL) [6], a decentralized computing paradigm has been introduced to address the

Manuscript received 19 January 2022; revised 1 July 2022 and 28 August 2022; accepted 21 September 2022. Date of publication 30 September 2022; date of current version 21 March 2023. This work was supported in part by the National Key Research and Development Project under Grant 2019YFE0109600, in part by the National Natural Science Foundation of China under Grants 62202399, 61871272, U21A20512, and 61876162, in part by the Research Grants Council of the Hong Kong SAR under Grant PolyU11211521, in part by the open Project of BGIShenzhen under Grant BGIRSZ20200002, and in part by the City University of Hong Kong Dongguan Research Institute. *(Corresponding author: Yao Hu.)*

Zhi-An Huang is with the Center for Computer Science and Information Technology, City University of Hong Kong Dongguan Research Institute, China.

Yao Hu is with the Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong, SAR, and also with the City University of Hong Kong, Shenzhen Research Institute, Shenzhen 518060, China (e-mail: yaohu4-c@my.cityu.edu.hk).

Rui Liu, Xiaoming Xue, and Linqi Song are with the Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong, SAR, and also with the City University of Hong Kong, Shenzhen Research Institute, China.

Zexuan Zhu is with the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, China, and also with the College of Computer Science and Software Engineering, Shenzhen University, China.

Kay Chen Tan is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, SAR.

This article has supplementary downloadable material available at https://doi.org/10.1109/TBME.2022.3210940, provided by the authors.

Digital Object Identifier 10.1109/TBME.2022.3210940

aforementioned issues by enabling the collaborative training of distributed data from multiple institutions. Instead of moving medical image data beyond the firewalls of archived institutions, FL enables the training process of DL models to locally occur at each participating institution in parallel. Only the trainable parameters of local models (e.g., gradients and/or weights) are transferred for aggregation update and redistribution. Another significant merit of FL is to overcome the space-sensitive limitations. Since the high-dimensional medical images are always storage-intense, in a FL setup, it is not necessary to duplicate the data from local institutions to a centralized pool (e.g. cloud server) where data need to be duplicated again by users for training their local models. Consequently, facing the exponential growth of global data, users can be exempt from the great burden of hardware upgrade for data storage and high-performance computing. Sheller et al. [7] compared FL with two collaborative learning methods on brain tumor segmentation and demonstrated its superiority in terms of accuracy. They further concluded that the model trained via FL can achieve similar performance as the model achieved with centralized data [8]. A brief review of multi-institutional collaboration methods can be found in Fig. 1 of the Supplemental Materials.

Despite the merits of FL, data divergence poses a challenge for applying FL in digital health. Medical images distributed from multiple institutions are particularly diverse. The participating institutions are prone to follow their specific criteria of sampling protocols, recruitment strategies, and participant instructions [9]. Since sample data are assumed to be independent and identically distributed (IID) across participants in both FL and non-FL, such data divergence has to be solved for data sharing strategies. Although FL benefits greatly from potentially increased diversity of data sources by reducing certain effects of bias, data divergence often leads to a dilemma where the global optimal solution of central model (a.k.a. source domain) is not optimal for an individual collaborating local model (a.k.a. target domain). In other words, not all the knowledge from source domain is advantageous for the training in target domain. The common FL technique (e.g., federated averaging dubbed FedAvg [6]) and other collaborative learning techniques struggle to address this problem. Their learning processes lack personalized model targeted at each client. The gap of data distribution between source domain and target domain (i.e., domain shift [10], [11]) becomes an obstacle to the success of FL.

The common knowledge shared in related mental disorders should be taken into consideration while deploying a FL framework across institutions in real-world applications. The participating institutions (e.g. healthcare centers and hospitals) tend to collect, curate, and maintain an image data pool involving various mental disorders. Inspired by multi-task learning (MTL) [12], it is effective to obtain a general representation by jointly learning a number of related tasks (i.e. simultaneously diagnosing multiple related disorders) from similar data distribution. In digital health, combining FL and MTL is a natural fit to model the optimal correlation structure for each diagnostic task with highly consistent feature patterns from other tasks. Therefore, it avoids the need to diagnose a new disorder from scratch and eases the domain shifts. In addition to MTL, parameter tuning and transfer learning (TL) are commonly leveraged to address domain shift in FL. To properly verify the performance of federated MTL in healthcare, autism spectrum disorder (ASD), attention deficit/hyperactivity disorder (ADHD), and schizophrenia (SCZ) are selected as three representative case studies in this work. Since they have been found to share common traits in genetic roots and neuroanatomical phenotypes, there is some overlap between the clinical symptoms of these mental disorders [13]. Therefore, the joint diagnosis of these mental disorders can be regarded as multiple related tasks.

In this paper, a Federated Multi-Task Learning framework for Joint Diagnosis of multiple mental disorders dubbed FMTLJD is developed based on functional MRI (fMRI) scans. fMRI can perform the noninvasive, safe measure for brain activity by detecting tiny changes in blood flow. It can be used to examine the brain functional connectivities (FCs) among functionally interconnected regions of interests (ROIs) [14]. The technical novelty in this work mainly concentrates on the potential exploration of both diagnostic performance and model interpretation by combing FL and MTL. The proposed FMTLJD framework is comprised of federated contrastive learning-based feature extractor (FCLFE), federated multi-gate mixture of expert classifier (FMMoE), and three functional modules including federated positive TL with personalization layers (FTLPL), privacy protection, and federated biomarker interpretation analysis. FCLFE extracts highly relevant features across client models and then FMMoE jointly classifies multiple related disorders via federated MTL. On real-world fMRI datasets, the experimental results indicate that FMTLJD can achieve promising accuracies of 69.48 $\pm$ 1.6%, 71.44 $\pm$ 3.2%, and 83.29$\pm$3.2% in ASD, ADHD, and SCZ, respectively. To the best of our knowledge, it is the first attempt to propose a federated MTL framework in digital health. We aim to address the challenges posed by such a novel paradigm based on the existing well-performed algorithms. The main contributions of this paper can be summarized as follows:

1) A federated MTL paradigm is first proposed to jointly classify multiple mental disorders by sharing fruitful information associated with related tasks. We provide effective solutions to address the challenges posed by this paradigm. The proposed federated MTL paradigm is more appropriate in real-world applications.
2) In FMTLJD, a series of modules are developed to address domain shift from different perspectives effectively. The simulation results suggest that FMTLJD can benefit more from a larger size of participating institutions, holding promise in future popularization.
3) The proposed federated biomarker interpretation analysis provides two visualization tools to help medical professionals improve clinical decision-making in terms of global levels, client levels, and individual levels.

The rest of this paper is organized as follows. Section II illustrates the model design in detail. Section III

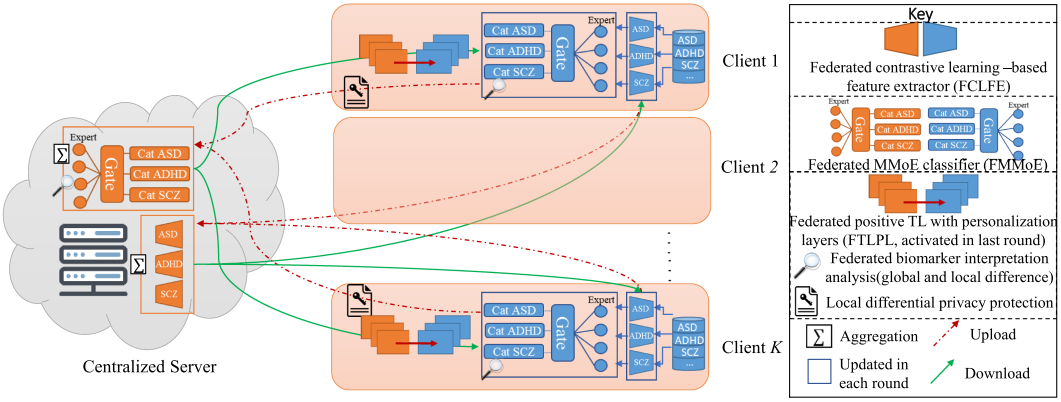


Fig. 1. The flowchart of the proposed FMTLJD framework. Firstly, the local model on each client extracts the domain-specific features and then trains their local MMoE classifier based on their datasets. Secondly, the parameters of local MMoE and feature extractor are uploaded to the centralized server for updating the global MMoE and feature extractor. Thirdly, the aggregated MMoE and feature extractor are downloaded to each client for next communication round of training. In the last communication round, FTLPL is performed to augment personalized model learning.

conducts a series of simulation experiments to evaluate the proposed framework. Finally, Section IV concludes this paper.

II. Model Design

In this section, the proposed methods and procedures are elaborated by following the flowchart shown in Fig. 1. The proposed FMTLJD framework is consisted of FCLFE, FMMoE, and three functional modules including FTLPL, privacy protection and federated biomarker interpretation analysis.

A. Problem Formulation

The primary goal of this work is to collaboratively train multiple client models of K data owners without the need to consolidate or exchange their respective private datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$. The overall loss function of a federated MTL paradigm $\mathcal{L}_{fl}$ can be derived from the all combination of M related diagnostic tasks $\{\mathcal{T}_i\}_{i=1}^{M}$ and the corresponding K local losses $\{\mathcal{L}_{ij}\}_{j=1}^{K}$ as follows

$$\mathcal{L}_{fl}(\mathcal{D}; \Theta) = \sum_{i=1}^{M} \sum_{j=1}^{K} \mathcal{L}_{ij}(D_j; \Theta) \tag{1}$$

where Θ is the set of model parameters to be learned. Let $\mathcal{L}_{con}$ represent the total learning loss of $\{\mathcal{T}_i\}_{i=1}^{M}$ for the conventional mode in which all datasets $\mathcal{D} = \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_K$ are centralized to train a global model. Ideally, $\mathcal{L}_{fl}$ should be very close to $\mathcal{L}_{con}$. Given a non-negative real number δ, the optimization framework can be reached as

$$|\mathcal{L}_{fl} - \mathcal{L}_{con}| < \delta \tag{2}$$

$$\Theta^* = \arg\min |\mathcal{L}_{fl} - \mathcal{L}_{con}|. \tag{3}$$

To analyze the effectiveness of a FL system, we can say that δ-performance loss can be observed as compared with the conventional mode with a single pool of all centralized data.

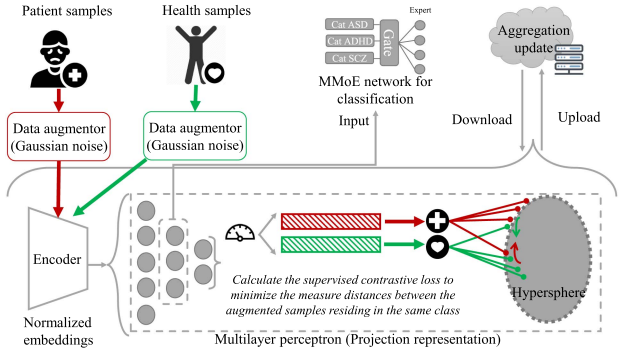


Fig. 2. The workflow of FCLFE. The input patient samples and health samples refer to the real fMRI datasets of case group (e.g., ASD, ADHD, and SCZ) and control group (i.e., typical control), respectively.

B. Federated Contrastive Learning Based Feature Extractor

Inspired by the merits of supervised contrastive learning, feature extraction for high-dimensional FC features can be beneficial from a better representation learning based on limited labeled samples in FL. Based on supervised contrastive learning [15], FCLFE is proposed to ease data divergence across clients for federated feature extraction by extending the similar structure of [16]. In this work, Pearson correlation coefficient (PCC) is used to calculate the FC features as the inputs of each subject based on the mean time series of ROIs.

Fig. 2 illustrates the workflow of FCLFE. For a set of n feature/label pairs on a client model denoted as $\{x_i, y_i\}_{i=1,\ldots,n}$, a data augmentation method with Gaussian noise is first conducted on each sample once, resulting in $2n$ pairs, i.e., $\{\tilde{x}_i, \tilde{y}_i\}_{i=1,\ldots,2n}$ [15]. This step of data augmentation is intended to reduce the risk of overfitting. Then we apply an encoder network to match the $2n$ augmented samples to the corresponding representation vectors through the same normalized embedding. A multilayer perceptron (MLP) [17] is deployed to extract the higher levels of abstract representations with nonlinear transformations. Again,

the output of MLP is normalized to calculate the overall supervised contrastive loss $\mathcal{L}_{sc}$ by distributing on the unit hypersphere z. By leveraging the label information, we hope to minimize the measured distances between the augmented samples belonging to the same class. As such, the $\mathcal{L}_i^{sc}$ of the i-th augmented sample can be formulated as follows:

$$\mathcal{L}_i^{sc} = \frac{-1}{|P(i)|} \sum_{j=1}^{2n} 1_{i \neq j} \cdot 1_{\tilde{y}_i = \tilde{y}_j} \cdot \log \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{k=1}^{2n} 1_{i \neq k} \cdot \exp(z_i \cdot z_k / \tau)} \tag{4}$$

where $P(i)$ represents the set of indices of the i-th augmented sample sharing the same label, and $|P(i)|$ is the cardinality of $P(i)$. The scalar temperature parameter $\tau \in \mathbb{R}^+$ is set to 0.1 following [15]. The objective is to iteratively minimize the overall $\mathcal{L}^{sc} = \sum_{i=1}^{2n} \mathcal{L}_i^{sc}$ so as to obtain the well-trained feature extraction results embedding in the middle layer of MLP, which are fed to the forthcoming classification task via MMoE network.

To improve the generalizability of feature extraction, the parameters of both encoder network and MLP are uploaded to the centralized server for aggregation update. One diagnostic task should correspond to one global aggregation model to extract the task-specific relevant feature representation. For a single diagnostic task, let N be the total amount of real samples aggregated from the fixed K clients, namely $N = \sum_{k=1}^{K} n_k$. Based on the current model parameters $W^{(t)} = \{W_1^{(t)}, \ldots, W_K^{(t)}\}$, a typical global aggregation [6] is adopted to derive the next model parameters $W^{(t+1)}$ as

$$W^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} W_k^{(t)}. \tag{5}$$

The aggregated model parameters $W^{(t+1)}$ are distributed back to all clients for the next communication round of training. This procedure is repeated until it reaches up to the maximum iterations.

C. Federated Multi-Gate Mixture of Experts Classifier

For exploring the subtle but significant FC patterns, MTL can capture large variability to provide a comprehensive picture of the functional image-derived phenotypes through implicit data augmentation. Concerning the highly complex networks involved in multiple diagnostic tasks, federated MTL allows "eavesdropping" to mediate the positive and negative knowledge between different tasks [18]. In federated MTL, domain shifts can be viewed from two perspectives, i.e., the gap between different tasks and between different sites. For the former, the trainable gating network within FMMoE dynamically assigns different gated weights to different classification tasks by mixing the outputs of the expert networks. For the latter, the aggregated updated expert network enables FMMoE to obtain a collaboratively-learned general model by lessening the site-specific dependence.

As a basic building block of FMMoE, MMoE is first introduced on principle and architecture. To better illustrate the characteristics of MMoE, we compare the general architecture between MMoE and the other two extensively used model, i.e., the single task model (STM in Fig. 3(a)) and shared-bottom model (SBM in Fig. 3(b)) [19]. In STM, each task is learned in isolation without knowledge transfer from other related tasks, while SBM leverages the shared-bottom layer to lap up all transferable knowledge for learning the related tasks simultaneously. SBM could struggle to effectively update the shared parameters because the complex heterogeneity of multiple related tasks could lead to an inconsistent optimization problem. In comparison, as shown in Fig. 3(c), MMoE can adjust the shared parameters in expert networks by managing the trade-offs between inter-task relatedness and task-specific variations.

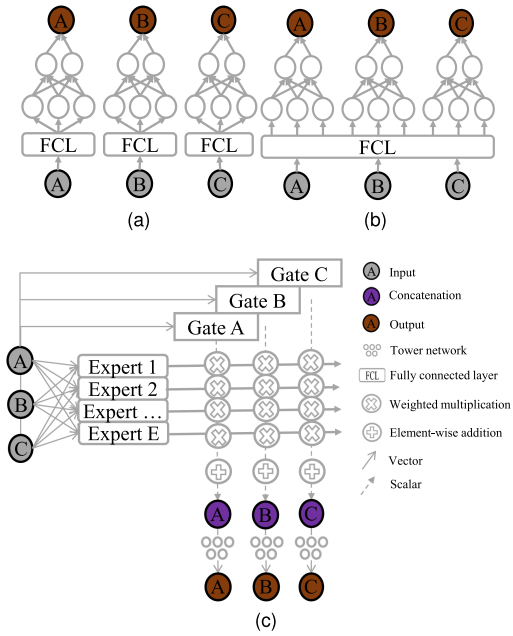


Fig. 3. Architecture showcases of STM, SBM, and MMoE. Since the inputs of multiple tasks (denoted as "A," "B," and "C") have the same FC feature dimension, each expert network can directly accept the inputs to train their weights alternatively for multi-task learning. (a) STM. (b) SBM. (c) MMoE.

The adopted MMoE model [20] is composed of expert networks and gating networks (each corresponds to a particular task). As a group of stacking neural networks, the expert networks substitute the shared-bottom layer to learn different feature representations given the inputs of multiple tasks. Since some tasks could be less related, knowledge transfer between them should be penalized accordingly. Each gating network learns to obtain an optimal mixture pattern by assembling these expert networks with different learned weights. Let g_i be the gating network with a trainable matrix $\omega_i \in \mathbb{R}^{E \times D}$ for the i-th task, where E and D represent the amount of assigned expert networks and feature dimension of input x, respectively. For deriving g_i, a *softmax* function is performed with a linear transformation for input x as follows.

$$g_i(x) = \text{softmax}(\omega_i x) \quad \text{s.t.} \sum_{e=1}^{E} g_i^e(x) = 1 \tag{6}$$

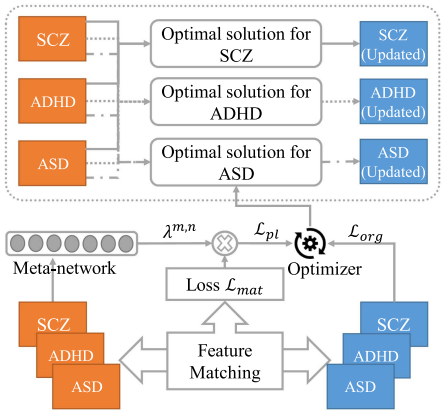


Fig. 4. The proposed federated positive TL method with personalization layers for adaptively selective knowledge transfer.

After that, g_i can yield a weight distribution over all expert networks $\{f^e\}_{e=1}^{E}$, representing a weighted sum of outputs. The output of MMoE for the i-th task (marked as F_i) can be infused by the concatenation of g and f as

$$F_i(x) = \sum_{e=1}^{E} g_i^e(x) f^e(x). \tag{7}$$

Finally, an MLP built upon the output of MMoE for each task serves as a tower network to further refine the task-specific representation and make predictions. Given K clients in the federated setup, combining their model parameters in the centralized server not only can blend the "expert" knowledge distilled from different clients but also address the biases caused by domain shift. The aggregation update of FMMoE is basically consistent with the procedure shown in Eq. (5), except for the definition of n_k. In the manner of MTL, here n_k refers to the k-th client's sample number involving all diagnostic tasks, instead a single one.

D. Important Functional Modules of FMTLJD

1) Personalized Model Learning: Due to the domain gap, it is widely recognized that not all the knowledge of source domain can be useful for target models. Personalization [21] is an important factor to determine what knowledge to be transferred and how to achieve accurate adaptive knowledge transfer. Especially in federated MTL, the complicated relationships between multiple diagnostic tasks make personalization become more critical. To address this issue, as we can see in Fig. 4, a novel personalized model learning method FTLPL is proposed to adaptively transfer advantageous knowledge based on the intermediate feature maps in the concatenation layers of FMMoE.

In the general settings of TL, the first step is to estimate the feature gap between the source network S and a target network T. Following the steps of minimizing ℓ_2 objective in FitNet [22], we can define the distance measure of intermediate feature maps between the m^{th} source task $S^m(x)$ and the n^{th} target task $T_\theta^n(x)$ as

$$\|S^m(x) - r_\theta(T_\theta^n(x))\|_2^2 \tag{8}$$

where r_θ is used as a pointwise convolution with a linear transformation controlled by θ [23]. According to [24], the feature matching objective can be calculated as follows:

$$\mathcal{L}_{mat}^{m,n} = \frac{1}{BC}\sum_{i}^{B}\sum_{j}^{C}(S^m(x)_{i,j} - r_\theta(T_\theta^n(x))_{i,j})^2 \tag{9}$$

where B and C represent the batch size and spatial size of concatenation layers, respectively. By taking the feature information of source model into consideration, FTLPL utilizes a meta-network f_ϕ to generate the weights as a function, $\lambda^{m,n} = \text{ReLU6}(f_\phi^{m,n}(S^m(x)))$, where ϕ is a trainable parameter set of the meta-network and ReLU6 is a modification of ReLU activation function with a maximum size of 6. As such, we can derive the loss of the generated personalization layers for all target models as

$$\mathcal{L}_{pl}(\theta|X^T, \phi) = \sum_{(m,n)\in H} \lambda^{m,n}\mathcal{L}_{mat}^{m,n}, \tag{10}$$

where H is a set of candidate task pairs. Finally, let $\mathcal{L}_{org}$ be the original loss of target models for classification (e.g., cross entropy), and the total loss $\mathcal{L}_{total}$ is reached as follows:

$$\mathcal{L}_{total}(\theta|X^T, \phi) = \mathcal{L}_{org}(\theta|X^T, Y^T) + \delta\mathcal{L}_{pl}(\theta|X^T, \phi) \tag{11}$$

where the hyper-parameter $\delta > 0$ regulates the effect of $\mathcal{L}_{pl}$. Based on the optimization process of [23], ϕ and θ are alternatively trained to minimize $\mathcal{L}_{total}$.

2) Privacy Protection: As shown in Fig. 5, we showcase the preprocessed FC matrix of one subject randomly sampled from each dataset. Since individual FC matrix has the potential to draw inferences about single subjects and could act as a fingerprint identifying subjects from a large group [25], privacy-preserving technique should be taken into consideration in this framework. To this end, differentially private stochastic gradient descent (DP-SGD) [26] is introduced to offer a strong standard for privacy guarantee. Through modifying the mini-batch SGD optimization process, DP-SGD can realize the privacy preservation of distributed data processing systems. In the proposed FMTLJD framework, DP-SGD is applied on the private local datasets of client models.

Given a random subset of samples $x_i \in \{x_1, \ldots, x_N\}$ and the corresponding loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i^N \mathcal{L}(\theta, x_i)$, the gradient $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$ is calculated at each SGD step of epoch t. In order to mask the privacy information of $\mathbf{g}_t$, we clip the ℓ_2 norm of $\mathbf{g}_t$ and add noise on it according to the following two equations as

$$\bar{\mathbf{g}}_t(x_i) = \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{Z}\right) \tag{12}$$

$$\tilde{\mathbf{g}}_t = \frac{1}{G}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 Z^2 \mathbf{I})\right) \tag{13}$$

where σ, G, and Z represent the noise scale, group size, gradient norm bound, respectively. $\mathcal{N}(0, \sigma^2 Z^2 \mathbf{I})$ refers to the Gaussian distribution with mean 0 and standard deviation $\sigma^2 Z^2$. The optimization steps of DP-SGD are consistent with that of typical

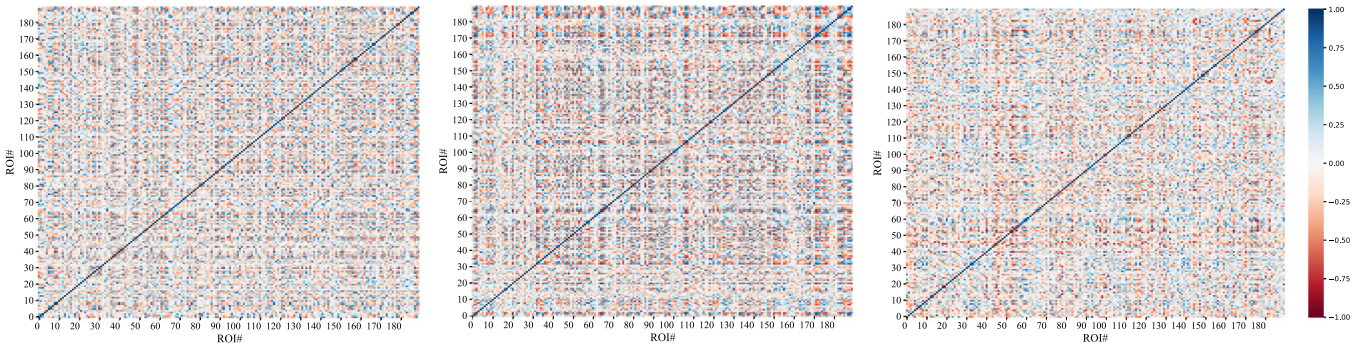


Fig. 5. Showcase of preprocessed FC matrix from each dataset.

SGD methods. Provided here is the basic rationale of DP-SGD and more detailed information can be found in [26].

3) Federated Biomarker Interpretation Analysis: As a de facto knowledge discovery model, FMTLJD benefits from the data-driven principle and thereby allows for detecting reproducible neurological "fingerprints" as interpretive biomarkers associated with specific mental disorders. Thanks to a central trusted server to coordinate FL in this work, federated biomarker interpretation analysis can diagnose patients across different demographics, enabling high-quality clinical decision making.

To this end, a novel federated biomarker interpretation method is presented based on the assumption that those remarkable biomarkers with larger discriminative power for classification should be more associated with the unique neural patterns of mental disorder. The guided back-propagation method [27] is used to compute the sub-gradient as $\partial f/\partial x_k$ for the k-th layer input x_k, where f is the current layer output. Then the gradient-based score indicates the relative importance of each FC feature input based on the probability output of corresponding correct class. To concentrate on the significant changes of ROIs, we aggregate scores from all associated FC features to calculate the weighted mean of each ROI. Two visualization tools i.e., saliency map and circular graph are used to analyze remarkable disorder-related ROIs and FCs, respectively. Since the bias and specificity of target domain are inevitable, there is still some inconsistency between global model and client model in search of remarkable disorder-related ROIs and FCs. Based on data-driven principle, the remarkable biomarkers selected by global model and client model represent the global-level and client-level findings, respectively. Furthermore, the individual-specific remarkable biomarkers can also be identified for a given subject and then represented as the individual-level finding. Therefore, the federated biomarker interpretation analysis can be conducted separately in terms of global levels, client levels, and individual levels. To ease the understanding of FMTLJD, the pseudo-code is presented in Algorithm 1.

Algorithm 1: The Pseudo-Code of FMTLJD.

Input: Preprocessed data X, phenotypic traits X_p, label Y, client number K, max_federated_rounds T.
Output: Predicted probabilities of test set Y_{te}^{pred}.

Init global & local models: $W^{(0)}$ & $\{W^{1,(0)}, \ldots, W^{K,(0)}\}$
for $k \leftarrow 1, \ldots, K$ **do** ▷ split data for each client
  $[X_{tr}^k, X_{te}^k, X_{p\_tr}^k, X_{p\_te}^k, Y_{tr}^k, Y_{te}^k] \leftarrow$ Split(X^k, X_p^k, Y^k)
end for
for $t \leftarrow 1, \ldots, T$ **do**
  for $k \leftarrow 1, \ldots, K$ **do** ▷ run in parallel
    Receive $W_{fe}^{(t)}, W_{clf}^{(t)}$ from server and update local model
    $\triangle W_{fe}^{k,(t)}, X_{tr}^{k'}, X_{te}^{k'} \leftarrow$FCLFE($W_{fe}^{(t)}, X_{tr}^k, Y_{tr}^k$, DP-SGD)
    $\triangle W_{clf}^{k,(t)} \leftarrow$FMMoE($W_{clf}^{(t)}, X_{tr}^{k'}, X_{p\_tr}^k, Y_{tr}^k$, DP-SGD)
    Send $\triangle W_{fe}^{k,(t)}, \triangle W_{clf}^{k,(t)}, n_k$ to server for aggregation
  end for
  $\{W_{fe}^{(t+1)}, W_{clf}^{(t+1)}\} \leftarrow$Global_update via Eq.(5) ▷ server
  Return $\{W_{fe}^{(t+1)}, W_{clf}^{(t+1)}\}$ to each client
  if $t == T$ **then** ▷ last round
    for $k \leftarrow 1, \ldots, K$ **do** ▷ run in parallel
      for $t' \leftarrow 1, \ldots, T'$ **do** ▷ epoch of FTLPL
        $W_{clf}^{k,(t'+1)} \leftarrow$FTLPL.fit($W_{clf}^{(T)}, W_{clf}^{k,(t')}$) via Eq. (11)
      end for
    $Y_{te}^{pred(k)} \leftarrow$FMMoE.pred($W_{clf}^{k^*}, X_{te}^k, X_{p\_te}^k$)
    end for
  end if
end for

III. Experiments and Results

To thoroughly evaluate the proposed FMTLJD framework, a series of simulation experiments are conducted in this section. First of all, the adopted datasets and preprocessing pipelines are introduced. The experimental configurations and evaluation metrics are provided. Then the proposed framework is compared with the state-of-the-art FL frameworks in digital health. Ablation studies are executed to further validate the effectiveness of major components within FMTLJD. In addition,

TABLE I
PERFORMANCE COMPARISON ON ABIDE, ADHD-200, AND COBRE

Method	ABIDE(ASD)					ADHD-200(ADHD)					COBRE(SCZ)
	Client1	Client2	Client3	Client4	*Avg%[ACC/SPE/SEN]*	Client1	Client2	Client3	Client4	*Avg%[ACC/SPE/SEN]*	*Avg%[ACC/SPE/SEN]*
Section A: Comparison with the state-of-the-art FL frameworks in digital health											
FMTLJD	70.12	68.87	69.99	68.95	**69.48±1.6/69.65±3.8/73.38±4.8**	69.93	76.91	70.64	68.29	**71.44±3.2/75.68±2.8/65.55±6.3**	83.29±3.2/82.52±3.4/86.08±4.2
Fed-MoE	63.02	61.45	66.14	66.3	64.23±3.7/57.56±7.7/75.67±9.1	57.81	67.71	72.91	63.02	65.36±2.7/67.25±4.7/46.15±12.2	72.22±9.7/69.82±9.7/80.55±13.3
Fed-Align	57.12	60.93	61.79	61.09	60.32±4.1/52.43±14.3/63.31±12.6	48.06	68.23	61.3	57.29	58.72±4.7/64.29±8.2/28.26±10.6	69.44±4.3/59.26±3.9/70.83±7.8
FedHealth†	67.06	67.86	70.81	71.86	69.39±6.4/68.46±13.4/78.06±17.1	65.38	73.58	72.84	68.17	69.99±4.1/73.88±6.3/60.58±6.2	**83.33±9.9/80.47±5.3/87.44±11.2**
Fed-DAN	68.69	68.46	67.32	68.41	68.22±2.9/66.35±6.3/72.27±6.9	69.96	71.33	65.87	70.31	69.37±2.4/71.36±7.1/63.91±10.3	78.65±2.5/81.61±5.9/76.48±11.5
Fed-DANN	67.69	69.93	69.44	66.61	68.42±2.4/67.67±2.9/70.91±5.7	68.70	70.54	68.90	67.64	68.94±2.7/72.41±3.4/66.38±2.6	79.17±2.5/77.51±4.4/83.21±6.0
Summary(w/t/l)					(4-1-0)/(5-0-0)/(3-2-0)					(5-0-0)/(5-0-0)/(4-1-0)	(4-1-0)/(4-1-0)/(4-1-0)
Section B: Comparison with the different common classification strategies shown in Fig. 6											
FMTLJD	70.12	68.87	69.99	68.95	**69.48±1.6/69.65±3.8/73.38±4.8**	69.93	76.91	70.64	68.29	**71.44±3.2/75.68±2.8/65.55±6.3**	83.29±3.2/82.52±3.4/86.08±4.2
FMTLJD_Single	62.05	61.56	61.42	61.86	61.72±2.6/60.74±3.3/66.67±6.0	57.87	68.87	64.45	62.71	63.47±10.8/66.87±9.2/59.0±10.3	69.63±3.5/66.59±5.5/77.27±11.3
FMTLJD_Mix†	69.37	70.63	64.38	68.9	68.32±1.6/73.77±2.4/67.69±1.3	69.28	72.66	73.95	67.86	70.93±2.6/77.18±1.3/57.14±3.5	**85.24±3.6/92.09±7.5/80.51±4.1**
FMTLJD_Fed	66.87	66.84	64.73	69.29	66.93±3.0/66.13±3.2/70.54±5.1	56.25	69.38	65.84	63.57	63.76±4.3/66.04±3.2/55.97±6.8	80.16±4.6/84.55±5.6/80.03±4.6
Summary(w/t/l)					(3-0-0)/(2-0-1)/(3-0-0)					(2-1-0)/(2-0-1)/(3-0-0)	(2-0-1)/(1-1-1)(3-0-0)
Section C: Comparison with the different common classification strategies shown in Fig. 6 using DP-SGD ($\epsilon = 4.5$)											
FMTLJD_DP	67.31	66.75	67.32	68.50	**67.47±2.3/66.68±2.2/72.33±4.7** ↓2.9%/↓4.3%/↓1.4%	65.99	69.77	68.03	66.98	**67.69±2.5/72.17±3.4/45.96±7.1** ↓5.2%/↓4.6%/↓29.9%	**79.56±6.1**/84.28±7.3/77.20±9.1 ↓4.5%/↑2.1%/↓10.3%
Fed-MoE_DP	61.46	59.89	62.50	63.43	61.82±1.5/58.42±6.1/69.85±11.1 ↓3.7%/↑1.5%/↓7.7%	54.17	61.98	66.92	62.31	61.34±5.3/67.29±7.2/37.52±13.7 ↓6.2%/↑0.06%/↓18.7%	68.94±4.5/67.66±4.5/74.03±8.5 ↓4.5%/↓3.1%/↓8.1%
Fed-Align_DP	53.80	59.50	57.83	57.71	57.21±2.4/44.35±6.5/64.06±8.9 ↓5.2%/↓15.4%/↑1.2%	42.25	64.58	58.68	55.17	55.17±9.4/58.05±9.9/32.67±12.3 ↓6.0%/↓9.7%/↑15.6	65.42±3.4/66.45±4.2/58.23±12.1 ↓5.8%/↑12.1%/↓17.8%
FMTLJD_Single_DP	61.64	58.78	61.17	62.77	61.09±2.4/61.38±2.7/59.79±6.5 ↓1.0%/↑0.1%/↓10.3%	57.56	64.53	62.41	58.40	60.73±1.7/65.61±2.0/39.64±7.7 ↓4.3%/↓1.9%/↓32.8%	69.44±5.8/67.13±5.6/**79.11±14.2** ↓0.2%/↑0.08%/↓2.4%
FMTLJD_Fed_DP	61.97	63.63	63.06	64.61	63.32±2.7/63.17±2.7/63.49±4.2 ↓5.4%/↓4.5%/↓9.9%	55.84	65.67	62.25	57.27	60.25±2.5/64.72±1.7/41.79±6.9 ↓5.5%/↓2.0%/↓25.3%	76.04±7.7/**89.57±6.7**/69.90±8.7 ↓5.14%/↑5.9%/↓12.7%
Summary(w/t/l)					(5-0-0)/(5-0-0)/(5-0-0)					(5-0-0)/(5-0-0)/(5-0-0)	(5-0-0)/(4-0-1)/(3-2-0)

† denotes the framework exists a risk of privacy leakage; The results of each client are represented in terms of accuracy. ↓/↑ represents the worse/better difference compared to model without DP privacy protocol.
The number of wins(w), ties(t), and losses(l) are statistically summarized to depict the comparison between FMTLJD/FMTLJD_DP and other competitors.

the effect analysis of varying client numbers is used to evaluate the reliability and effectiveness of practical applications. Finally, remarkable ROIs and FCs of each mental disorder are visualized to provide federated biomarker interpretation analysis.

A. Data Acquisition and Processing

In this work, all resting state fMRI sample data along with their phenotypical information (e.g., age, gender, and handedness) are gathered from three publicly released aggregation databases, namely the Autism Brain Imaging Data Exchange I (ABIDE for ASD) [28], the ADHD-200 Competition (ADHD-200 for ADHD), and the Center for Biomedical Research Excellence (COBRE for SCZ).

ABIDE [1]: This database archives 1035 valid samples aged 7 to 64, including 505 ASD patients and 530 typical controls (TCs). To reduce the variance between preprocessing pipelines, five different teams implemented fairly similar preprocessing steps using their preferred tools. We selected the data preprocessed by the Configurable Pipeline for the Analysis of Connectomes [29].

ADHD-200[2]: This database was originally created for the diagnostic classification in a global competition. ADHD-200 archives 939 valid datasets aged 7 to 21 from 358 ADHD patients and 581 TCs. Similarly, the preprocessed data of ADHD-200 can be downloaded from the preprocessed pipeline of Athena.

COBRE[3]: This database contributes 146 valid samples aged 18 to 65 from 72 SCZ patients and 74 TCs. The preprocessed pipeline of [30] is performed for COBRE including a series of steps: volume removal, slice timing correction and realignment, motion correction, spatial normalization, bandpass filtering, normalization by the MNI template, and smoothing with a 6-mm Gaussian kernel of full width at half maximum.

In the experiment, we selected the Craddock 200 (CC200) functional parcellation atlas [31] to partition the connectivity graphs of each individual sample into 200 representative ROIs. The specific parameters of the preprocessing pipeline are summarized in Table I of the Supplemental Materials. In order to coordinate the settings of MTL, the final 190 ROIs are taken into consideration in accordance with the setup of Athena pipeline. By retrieving from the upper triangular elements of a correlation matrix, each sample data is converted into a feature vector of size 17,955, representing all pairwise FC features/connections between the ROIs.

B. Experimental Configuration

Since ABIDE and ADHD-200 are multi-site collections with collaborations of 17 and 7 international imaging sites, respectively, we equally divided both databases into four groups as different clients for FL simulation as shown in Table II of the Supplemental Materials. As a single-site database, COBRE was randomly split into four group accordingly. The model parameter settings of each component within FMTLJD are tabulated in Table III of the Supplemental Materials. Three-fold cross validation is adopted to evaluate the effectiveness of proposed framework. Since the multiple diagnostic tasks jointly solved in this work can be considered as multiple binary classification problems, accuracy (ACC), specificity (SPE), and sensitivity (SEN) are used to assess the performance by following [2]. In Table I and Table II, the Wilcoxon rank-sum test is also utilized to verify the performance change of FMTLJD compared with the other competing approaches to indicate which metric is statistically significant at the level of $\boldsymbol{p \leq 0.05}$.

[1]http://fcon_{1}000.projects.nitrc.org/indi/abide
[2]http://fcon_{1}000.projects.nitrc.org/indi/adhd200
[3]http://fcon_{1}000.projects.nitrc.org/indi/retro/cobre.html

TABLE II
ABLATION STUDIES BASED ON THE PROPOSED FMTLJD FRAMEWORK

Variants	ABIDE(ASD)					ADHD-200(ADHD)					COBRE(SCZ)
	Client1	Client2	Client3	Client4	*Avg%[ACC/SPE/SEN]*	Client1	Client2	Client3	Client4	*Avg%[ACC/SPE/SEN]*	*Avg%[ACC/SPE/SEN]*
ST w/ SVM	61.10	64.87	65.35	63.46	63.69±2.4/62.77±2.2/70.27±13.9 ↓8.3%/↓9.9%/↓4.2%	58.13	63.3	67.76	57.36	61.64±3.1/68.22±1.8/40.46±13.3 ↓13.7%/↓9.9%/↓38.3%	76.67±6.0/76.46±8.5/**88.83±10.5** ↓7.9%/↓7.3%/↑3.2%
ST w/ MMoE	60.10	60.21	61.80	61.69	60.96±3.5/62.67±3.6/59.93±6.7 ↓12.3%/↓10.0%/↓18.3%	43.80	67.24	73.48	59.02	60.88±3.1/64.25±1.2/27.59±9.7 ↓14.8%/↓15.1%/↓57.9%	69.44±10.5/57.07±10.2/72.87±10.9 ↓16.6%/↓30.8%/↓15.3%
ST w/ STM	57.86	61.29	60.00	60.00	59.79±1.9/59.24±6.1/62.36±15.1 ↓13.9%/↓14.9%/↓15.0%	51.62	67.29	72.56	57.38	62.19±6.7/64.72±3.3/39.26±11.0 ↓12.9%/↓14.5%/↓40.1%	68.05±5.0/69.44±16.3/33.33±11.6 ↓18.3%/↓15.9%/↓61.3%
MT w/ SBM	64.79	66.45	65.98	68.18	66.41±3.1/**71.82±3.8**/66.09±5.9 ↓4.4%/↑2.2%/↓9.9%	46.51	71.20	83.23	59.92	62.21±3.4/66.38±2.5/40.96±9.4 ↓12.9%/↓12.3%/↓24.6%	68.40±8.4/70.04±9.3/36.31±10.3 ↓17.9%/↓15.1%/↓57.8%
MT w/o FCLFE	66.99	66.00	65.21	65.40	65.90±2.1/65.58±5.7/70.30±5.2 ↓5.2%/↓5.8%/↓4.2%	62.69	75.12	71.10	63.77	68.17±2.5/68.21±3.5/59.60±6.1 ↓4.6%/↓9.9%/↓9.1%	81.36±6.1/**83.45±8.1**/81.25±9.1 ↓2.3%/↑1.1%/↓5.6%
FMTLJD	70.12	68,87	69.99	68.95	**69.48±1.6**/69.65±3.8/**73.38±4.8**	69.93	76.91	70.64	68.29	**71.44±3.2/75.68±2.8/65.55±6.3**	**83.29±3.2**/82.52±3.4/86.08±4.2
Summary(w/t/l)					(5-0-0)/(4-0-1)/(5-0-0)					(5-0-0)/(5-0-0)/(5-0-0)	(4-1-0)/(4-0-1)/(4-0-1)

The results of each client are represented in terms of accuracy; ST: Single-task pattern, MT: Multi-task pattern, w/: with, w/o: without; ↓/↑ represents the worse/better difference compared to FMTLJD (the best model).
The number of wins(w), ties(t), and losses(l) are statistically summarized to depict the comparison between FMTLJD and other competitors.

TABLE III
EFFECT ANALYSIS OF FCLFE

FMTLJD	w/o FCLFE	w/ FCLFE
Trainable parameters	86.91M	5.04M (down by 94.2%)
Average elapsed time (per federated round)	8.16s	1.87s (down by 77.1%)
Uploaded model size (.pkl)	331.0 MB	19.2 MB (down by 94.2%)

Runtime environment: One Intel Core i7-8700K@3.70GHz & one NVIDIA GeForce RTX 2080 Ti GPU

C. Performance Comparison

This subsection explores and analyzes the overall performance of FMTLJD based on three different comparison experiments. For a fair comparison, all compared frameworks are configured with the best model parameters. Note that all compared frameworks take turns to test the performance of each client as target domain while other clients are combined as source domain for global training. FMTLJD is first compared with three state-of-the-art FL frameworks in digital health, i.e., Fed-MoE [32], Fed-Align [32], and FedHealth [33]. Fed-MoE and Fed-Align are two privacy-preserving FL frameworks with domain adaptation to classify ASD and TC on ABIDE database. For addressing domain shift between sites, Fed-MoE leverages an MoE layer to dynamically assigns gated weights so as to merge global models and private models. It is noted that the purpose of MoE layer in [32] is significantly different from our FMTLJD. Our FMMoE module aims to explore the common knowledge shared among related disorders rather than mitigating the domain shift between the global model and local model. Fed-Align employs the local feature extractor and global discriminator to generalize the source domain into a common space of target data. FedHealth is the first federated TL framework of wearable healthcare to centralize local data for training in a centralized server and then build personalized models by TL. Therefore, FedHealth benefits more from direct data aggregation instead of compromising on private security. In addition the state-of-the-art FL frameworks, two representative domain adaptation techniques, i.e., deep adaptation networks (DAN) [34] and domain adaptive neural networks (DANN) [35], are also compared to represent benchmark performance for applying existing domain adaptation methods to federated MTL. DAN enhances the feature transferability from task-specific layers of the neural network by mean-embedding matching of the multi-layer representation across domains in a reproducing kernel Hilbert space. DANN incorporates the maximum mean discrepancy measure as a regularizer to reduce the distribution mismatch between source domain and target domain with a gradient reversal layer. For a fair comparison, we approximately modified their source codes to fit for the federated MTL setting. The variants of DAN and DANN we devised are termed by Fed-DAN and Fed-DANN, respectively. To investigate the risk of privacy leakage for these compared FL frameworks, more analysis discussion is provided in Table IV of the Supplemental Materials.

The comparison results are shown in the Section A of Table I. The proposed FMTLJD framework achieves promising accuracies of 69.48 ± 1.6% (SPE: 69.65%, SEN: 73.38%), 71.44 ± 3.2% (SPE: 75.68%, SEN: 65.55%), and 83.29±3.2% (SPE: 82.52%, SEN: 86.08%) on ABIDE, ADHD-200, and COBRE, respectively. The Wilcoxon rank-sum test statistically demonstrates the performance improvement achieved by FMTLJD over the compared state-of-the-art frameworks. The performance of Fed-MoE and Fed-Align is inferior to that of others. This is because their adopted domain adaption techniques do not work on all sites. Their simulation studies [32] also showed that only two out of four sites got improved in accuracy. It is also noted that their classification performance is skewed by the unbalance between specificity and sensitivity. For intelligent auxiliary diagnosis analysis, missed diagnosis can generate a more serious consequence than misdiagnosis does. That is to say, sensitivity is more important to specificity in most cases. Fed-MoE and Fed-Align struggle to maintain adequate sensitivity on ADHD-200 and then achieve inferior reliability for the diagnosis of ADHD. Fed-DAN and Fed-DANN achieve decent classification performance fairly close to FedHealth. Yet, it could be challenging for them to solve various challenges posed by such a federated MTL paradigm. The complicated relationship between different diagnostic tasks inevitably limits their classification effectiveness.

FedHealth can reach relatively decent results comparable to FMTLJD. However, FedHealth needs to upload 70% data of each client to form public datasets for building an initial cloud model. The data centralization of FedHealth could result in a high risk of privacy disclosure. In this work, there is no need to incorporate the privacy protection of homomorphic encryption used in FedHealth for comparison. FedHealth can be considered as a combination of centralized data sharing and FL. In comparison with FedHealth, FMTLJD shows a substantial improvement of classification performance, very close to the

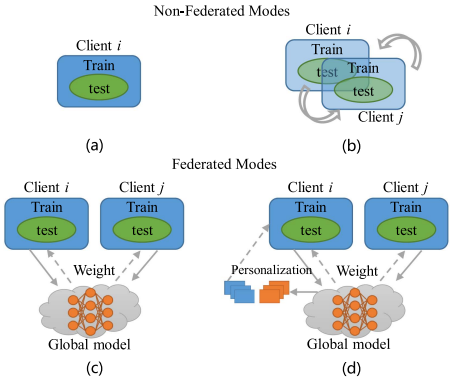


Fig. 6. Comparison with different common classification strategies. Transparent colors represent no privacy-preserving capability. (a) Single. (b) Mix. (c) Fed. (d) FMTLJD.

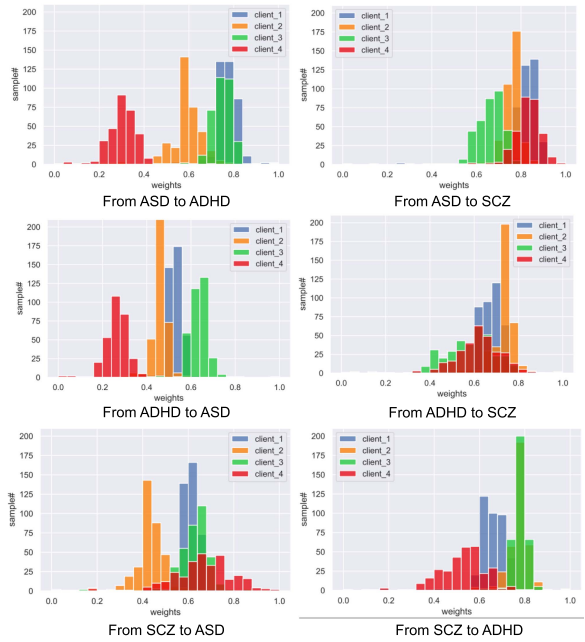


Fig. 7. Histogram visualization to interpret the learned weights transferred between different tasks for testing samples.

levels of centralized data sharing architecture. This can be attributed to the fact that as a first federated MTL framework FMTLJD can learn common knowledge shared in all or part of related tasks by introducing an inductive bias from more demographic data of multiple related disorders.

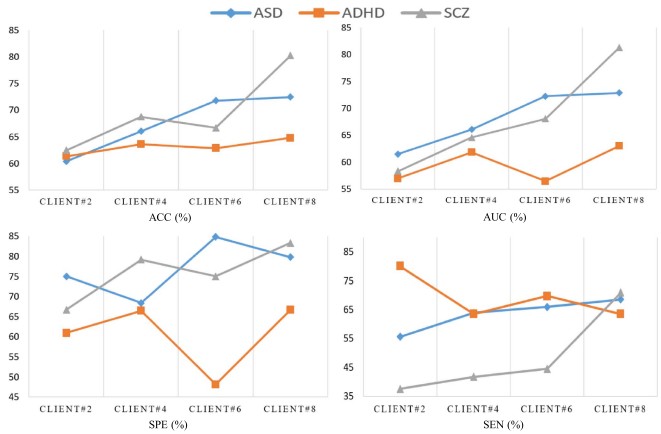
To further evaluate FMTLJD from multiple perspectives in MTL paradigm, four common classification strategies [32] are compared in terms of non-federated modes and federated modes as shown in Fig. 6. Non-federated modes include independent sets of training/testing within a single site (abbr. *Single*) and gathering available data from all clients for training a global model like centralized data sharing (abbr. *Mix*). In federated modes, traditional FL without personalization like FedAvg [6] (abbr. *Fed*) and FL with personalization adopted in this work (abbr. *FMTLJD*) are included. As shown in the Section B of Table I, FMTLJD_Single achieves the baseline accuracy levels of non-federated modes for ASD (61.72 $\pm$ 2.6%), ADHD (63.47 $\pm$ 10.8%), and SCZ (69.63 $\pm$ 3.5%) by only using the local data. FMTLJD_Fed represents the baseline levels of federated modes, clearly surpassing the performance of FMTLJD_Single on ABIDE and COBRE databases. It supports the idea that FL is helpful to address the issues of data isolation. As expected, FMTLJD_Mix outperforms FMTLJD_Fed by directly incorporating all available data from clients. To our surprise, FMTLJD is superior to FMTLJD_Mix in ABIDE and ADHD-200 databases. This result also demonstrates that without the reduced risk of privacy leakage, FMTLJD enables a reliable diagnostic detection which is competitive to the ideal scenario of gathering all multi-site data for training.

It is well-known that encryption functions can lead to the decline in accuracy [36]. We also conducted the effect analysis of DP-SGD on Fed-MoE, Fed-Align, and these common classification strategies of FMTLJD (except FMTLJD_Mix). For the variants of FMTLJD, about 0.2%-5.5% of reductions can be seen in accuracy with relatively higher standard deviation. The main reason is that the adverse effects of differential privacy mechanisms are borne unequally by clients and the tail or underrepresented participants often suffer most [36]. DP-SGD was demonstrated to provide a plausible privacy guarantee when the privacy budget ε should range from 0.1 (high privacy regime) to $ln|x|$ (low privacy regime), where x represents the sample number [37]. In this work, the default value of ε is set to 4.5, within the reasonable privacy range of ε. Theoretically, DP-SGD with reasonable setting of ε can strongly prevent different types of privacy leakage [38].

It is shown that Fed-MoE_DP and Fed-Align_DP also suffer from deteriorated performance by using DP-SGD with $\varepsilon = 4.5$. As compared to the baseline version, the accuracy of the encrypted versions (Fed-MoE_DP vs. Fed-Align_DP) decreased by 3.7%/5.2%, 6.2%/6.0%, and 4.5%/5.8% in ABIDE, ADHD-200, and COBRE datasets, respectively. We can observe that, under the same privacy regime, the performance adverse effect brought by DP-SGD can be more significant on both Fed-MoE and Fed-Align than FMTLJD. The observation indicates that FMTLJD could be more tolerable to the use of DP-SGD as compared with Fed-MoE and Fed-Align. Additionally, we also conduct an experiment to evaluate the performance change of FMTLJD under different settings of ε. As shown in Fig. 2 of the Supplemental Materials, varying ε from low privacy regime to high privacy regime generally results in various degrees of accuracy degradation by 1.74%, 2.95%, and 5.39% in ABIDE, ADHD-200, and COBRE datasets, respectively. The results suggest that the larger sample number x in datasets could be less sensitive to the performance change of FMTLJD under different settings of ε, especially in terms of SPE and SEN. Generally speaking, the cost of DP-SGD is limited and acceptable. FMTLJD with DP-SGD still achieves notable advantages compared with Fed-MoE and Fed-Align.

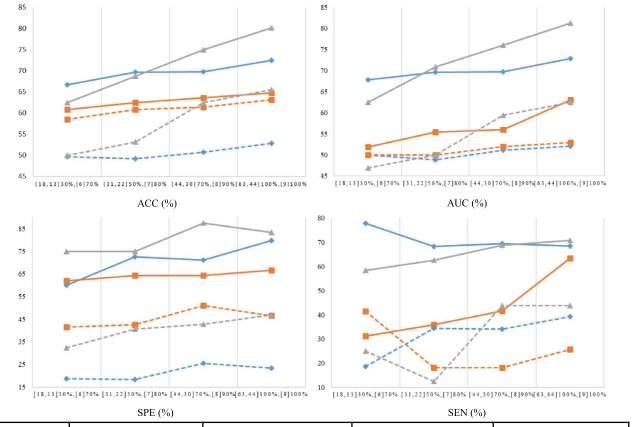
D. Ablation Studies

FCLFE, FMMoE and federated MTL paradigm are three important contributors in this work. We conducted ablation studies to evaluate their effectiveness based on single-task pattern and multi-task pattern. As one of the most popular classifiers in medical image analysis, support vector machine (SVM) [39]



FMTLJD	Avg ACC (%)			Avg AUC (%)			Avg SPE (%)			Avg SEN (%)		
	ASD	ADHD	SCZ	ASD	ADHD	SCZ	ASD	ADHD	SCZ	ASD	ADHD	SCZ
Client#2 (1/12)	60.4	61.4	62.5	61.5	57.0	58.3	75.0	60.9	66.7	55.6	**80.0**	37.5
Client#4 (0/12)	66.0	63.6	68.8	66.1	61.8	64.6	68.4	66.5	79.2	63.9	63.5	41.7
Client#6 (1/12)	71.8	62.9	66.7	72.2	56.4	68.1	**84.8**	48.1	75.0	65.9	69.7	44.4
Client#8 (10/12)	**72.4**	**64.8**	**80.2**	**72.9**	**63.0**	**81.3**	79.8	**66.7**	**83.3**	**68.5**	63.4	**70.8**

(a)



FMTLJD (each of Client#8)	Avg ACC (%)			Avg AUC (%)			Avg SPE (%)			Avg SEN (%)		
	ASD	ADHD	SCZ	ASD	ADHD	SCZ	ASD	ADHD	SCZ	ASD	ADHD	SCZ
30%, 70%(1/12)	66.7	60.8	62.5	67.8	51.9	62.5	60.0	61.9	75.0	**77.9**	31.3	58.3
50%, 80%(0/12)	69.7	62.5	68.8	69.7	55.4	70.8	72.7	64.3	75.0	68.3	35.9	62.5
70%, 90%(2/12)	69.8	63.6	75.0	69.8	**63.6**	76.0	71.2	64.3	**87.5**	69.4	41.7	68.8
100%, 100%(9/12)	**72.4**	**64.8**	**80.2**	**72.9**	63.0	**81.3**	**79.8**	**66.7**	83.3	68.5	**63.4**	**70.8**

(b)

Fig. 8. Performance change of FMTLJD by increasing client numbers (a) and the proportions of sample sizes within small institutions (b). In the diagram of the right column, [18,13]30%,[6]70% indicates there are 30% proportions of ASD and ADHD datasets with size of 18 and 14 as well as 70% proportion of SCZ dataset with size of 6. The dashed lines represent the performance levels of *Single* mode via MTL. (a) Increasing the client numbers from 2 to 8. (b) Each of the eight participating clients takes turns to increase theproportions of sample sizes.

is employed to test the benchmark performance of traditional machine learning in single-task pattern. Three architectures shown in Fig. 3, i.e., STM, SBM, and MMoE, are compared based on the similar network settings. FMTLJD without FCLFE is also compared for the effect analysis. As shown in Table II, the variants in multi-task pattern outperform those in single-task pattern. It is demonstrated that the proposed federated MTL paradigm can be a good recipe to enable more effective learning by accumulating larger subject populations with more general demographic distributions. Since each diagnostic task has its own unique noise patterns, federated MTL paradigm can ideally average such data-dependent noise to reduce the risk of overfitting.

Amongst the variants of single-task pattern, SVM achieves the best overall performance. Given the insufficient data within local clients, SVM is more effective in handling small, complex datasets through the appropriate linear transformation. Back to the comparison of STM, SBM, and MMoE, SBM as a variant of multi-task patterns can only show a significant improvement on ABIDE. Indeed, SBM can reduce the risk of overfitting based on the shared parameters. However, the resistance of updating shared parameters also grows exponentially because each diagnostic task is assigned the same weight. The optimization conflicts between diagnostic tasks could overwhelm the learning process of SBM. Therefore, the diagnostic task with more sample data (i.e., ABIDE) can dominate the total loss. FMMoE is proposed to replace SBM in order to learn the flexible parameter sharing patterns for each diagnostic task. As described in Table II, FMTLJD with FMMoE shows different levels of performance improvement on these three databases. We can observe 12.9% and 17.9% of accuracy promotions on ADHD-200 and COBRE, respectively.

The meta-network within FTLPL allows the MMoE model to target at each diagnostic task by adaptively transferring advantageous knowledge from other diagnostic tasks with different weights. In order to interpret how the targeted task can be affected by other tasks in each client, Fig. 7 shows the learned gating weights transferred between different tasks for each testing sample. Generally, the transferable effects between different tasks are quite notable and almost distributed in the adjacent values, which range from 0.4 to 0.8. This suggests that the knowledge transfer between different tasks indeed functions as an effective inter-medium to improve the generalization to a target domain. Additionally, through the histogram visualization, we also note that the model of Client 4 appears to have lower gating weights to transfer knowledge from other tasks, especially for the cases from ASD to ADHD and from ADHD to ASD. Based on the results shown in Table I, the model performance of Client 4 is below the average on ABIDE and ADHD-200 datasets. Therefore, in this work, such transferable effects could be positively correlated with the classification performance for each client.

As a federated supervised feature extractor, FCLFE is developed to reduce feature dimension by 94.3% (from 17,955 to 1,024). It is shown that FMTLJD with FCLFE results in 5.2%, 4.6%, and 2.3% of accuracy increases on ABIDE, ADHD-200, and COBRE, respectively. Moreover, Table III illustrates that FCLFE can reduce trainable parameters, elapsed time, and uploaded model size by 94.2%, 77.1%, and 94.2%, respectively. The results demonstrate that FMTLJD with FCLFE can be competent to practical applications in rigorous conditions, e.g., limited computational resources, communication-intense environment, and real-time analysis. Furthermore, as shown in Fig. 3 and Table VI of the Supplemental Materials, we also

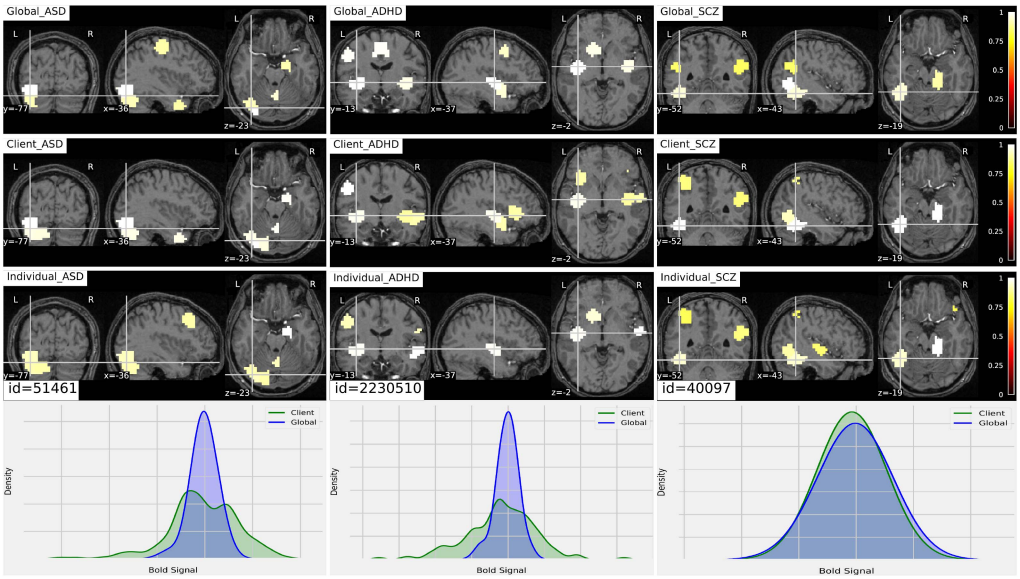


Fig. 9. The saliency map visualization to showcase the ten top remarkable disorder-related ROIs in terms of global levels, client levels, and individual levels. In the last row, kernel density estimator [42] is used to analyze the divergence of fMRI data distribution between global levels and client levels.

conducted t-distributed stochastic neighbor embedding (t-SNE) data analysis [40] and cross validation to compare supervised contrastive learning (represented by FCLFE) with two typical methods belonging to standard end-to-end supervised feature learning, i.e., MLP and autoencoder (AE) [41]. The experimental results shown in the Supplemental Materials demonstrate that, given the limited labeled samples in FL, supervised contrastive learning has an advantage over standard end-to-end supervised learning for high-dimensional FC feature extraction. Due to the small amount of training samples, we also evaluate the performance of FCLFE on avoiding overfitting and comparison with different data augmentation methods in Fig. 4 and Table VII of the Supplemental Materials, respectively.

E. Effect Analysis of Incremental Clients

Sufficiently large curated datasets are vital to the effectiveness of federated MTL. The nature of FL encourages participating institutions to promote the scale effects of potentially accessing to more training datasets. We evaluated the performance change of FMTLJD by increasing the number of clients while keeping the same sample size for different mental states in each client. To reduce the disturbance effect of datasets, the data samples of each client are equally distributed with the same proportion of positive and negative samples from ABIDE, ADHD-200, and COBRE. According to the situation with maximum client number (i.e., eight), each client is consistently allocated with the minimum fixed sample numbers of 63, 44, 9, and 148 in ASD, ADHD, SCZ, and TC, respectively. The area under the receiver operating characteristic curve (AUC) is also used to estimate the probability that FMTLJD ranks a randomly selected individual with mental disorder higher than a randomly selected individual from TCs. As can be seen in Fig. 8(a), there is an upward trend in four metrics for these mental disorders as expected. Specifically, both accuracy and AUC can steadily increase despite some variances shown in specificity and sensitivity.

Based on the productive global model, those institutions with relatively small datasets could still benefit from the collaborative learning in FL. For simulating the performance change of FMTLJD in such a scenario, each of the eight participating clients was selected in turns to change the proportions of sample sizes at four levels as shown in Fig. 8(b). Since the sample number of SCZ is extremely deficient, the proportions are changed from 70% to 100%. The results demonstrate that FMTLJD can maintain decent performance at different levels of data deficiency. To evaluate the baseline performance, we also tested each client model in *Single* mode via MTL (i.e., without using FL). As a result, these client models suffer from significant performance degradation despite increasing available datasets. In our simulation experiments, the majority of client models fail to converge during the training process. There are five, six, and four out of the total eight client models which cannot make prediction for ASD, ADHD, and SCZ, respectively. Therefore, FMTLJD also enables effective learning for those participating institutions with relatively small datasets.

F. Federated Biomarker Interpretation Analysis

In addition to computer-aided detection, FMTLJD can also provide federated biomarker interpretation analysis to help medical professionals improve clinical decision-making. The identified remarkable disorder-related biomarkers generate plausible explanations for the image-derived phenotype changes between various pathologies and healthy states. Thanks to the guided back-propagation method, the ten top remarkable disorder-related ROIs and FCs are visually interpreted by the saliency maps and circular graphs shown in Fig. 9 and Fig. 5 of the Supplemental Materials, respectively. As shown in Table VIII

of the Supplemental Materials, through manual validation, 90%, 80%, and 100% of the ten top detected ROIs have been supported to have associations with ASD, ADHD, and SCZ, respectively. Generally, the global levels and client levels are fairly similar on each disorder. For the saliency map visualization, 80%, 50%, and 80% of the ten top remarkable ROIs have overlap between both levels in ASD, ADHD, and SCZ, respectively. The divergence of fMRI data distribution between both levels does not have a distinct effect on the biomarker analysis. For circular graph visualization, given the vast amount of FCs (17,955), the included ROIs and their K-nearest neighbors are considered as a group to share the same remarkable FC (K is set to 5 in this work). Accordingly, 50%, 60%, and 30% of the ten top remarkable FCs have overlap between both levels on ASD, ADHD, and SCZ, respectively. Based on the data-driven outcomes, the reproducible disorder-related biomarkers are characterized to offer deeper insights into the pathophysiological mechanism of mental disorders. From the perspectives of individual levels, we can quickly capture those personal unique biomarkers which are not present in global levels and client levels. The federated biomarker interpretation analysis is expected to provide a clue to advance the future of individualized healthcare.

IV. Conclusion

In this work, an effective federated MTL framework FMTLJD has been presented for the joint diagnosis of multiple mental disorders using fMRI data. The major innovation lies in exploiting the highly consistent feature patterns from other related tasks in a federated MTL fashion. For addressing domain shift in FL, FCLFE, FMMoE, and FTLPL are proposed to improve the performance of FMTLJD from different perspectives. FCLFE extracts the disorder-related features to ease data divergence across client models. FMMoE mediates the heterogeneity of multiple related tasks for classification by optimizing the mixture patterns of expert networks. FTLPL performs advantageous knowledge transfer from source domain to target domain. FMTLJD can be equipped with DP-SGD to offer a strong standard of privacy guarantee. FMTLJD also provides federated biomarker interpretation analysis in terms of global levels, client levels, and individual levels. Extensive simulation experiments have demonstrated the effectiveness and reliability of FMTLJD. Through the effect analysis of incremental clients, more participating institutions can be more beneficial for the performance of FMTLJD. The current work is anticipated to help medical professionals set up early detection and personalized treatment in FL.

References

[1] R. V. Rodrigues et al., "Screening for common mental disorders using the SRQ-20 in medical students from Porto Velho-RO, Brazil," *J. Adv. Med. Pharmaceut. Sci.*, vol. 40, no. 2, pp. 33–45, 2021.

[2] Z.-A. Huang et al., "Identifying autism spectrum disorder from resting-state fMRI using deep belief network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2847–2861, Jul. 2021.

[3] Z. Yang et al., "Functional informed fiber tracking using combination of diffusion and functional MRI," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 3, pp. 794–801, Mar. 2019.

[4] Z.-A. Huang et al., "Identification of autistic risk candidate genes and toxic chemicals via multilabel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3971–3984, Sep. 2021.

[5] Y. Chen, X. Sun, and Y. Jin, "Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 4229–4238, Oct. 2020.

[6] B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[7] M. J. Sheller et al., "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*, Springer, 2018, pp. 92–104.

[8] M. Sheller et al., "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.

[9] A. Messé, H. Benali, and G. Marrelec, "Relating structural and functional connectivity in MRI: A simple model for a complex brain," *IEEE Trans. Med. Imag.*, vol. 34, no. 1, pp. 27–37, Jan. 2015.

[10] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, Mar. 2022.

[11] M. Ingalhalikar et al., "Functional connectivity-based prediction of Autism on site harmonized ABIDE dataset," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 12, pp. 3628–3637, Dec. 2021.

[12] Y.-A. Huang et al., "Predicting microRNA–disease associations from lncRNA-microRNA interactions via multiview multitask learning," *Brief. Bioinf.*, vol. 22, no. 3, 2021, Art. no. bbaa133.

[13] M. T. M. Park et al., "Neuroanatomical phenotypes in mental illness: Identifying convergent and divergent cortical phenotypes across autism, ADHD and schizophrenia," *J. Psychiatry Neurosci.: JPN*, vol. 43, no. 3, 2018, Art. no. 201.

[14] D. Sahoo, T. D. Satterthwaite, and C. Davatzikos, "Hierarchical extraction of functional connectivity components in human brain using resting-state fMRI," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 940–950, Mar. 2021.

[15] P. Khosla et al., "Supervised contrastive learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[16] T. Chen et al., "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[17] N. Stamatis, D. Parthimos, and T. M. Griffith, "Forecasting chaotic cardiovascular time series with an adaptive slope multilayer perceptron neural network," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 12, pp. 1441–1453, Dec. 1999.

[18] Y. Zhang and Q. Yang, "An overview of multi-task learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 30–43, 2018.

[19] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[20] J. Ma et al., "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1930–1939.

[21] V. Kulkarni et al., "Survey of personalization techniques for federated learning," in *Proc. 4th World Conf. Smart Trends Syst., Secur. Sustain.*, 2020, pp. 794–797.

[22] Y. Cheng et al., "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 126–136, 2018.

[23] Y. Jang et al., "Learning what and where to transfer," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3030–3039.

[24] Y. LeCun et al., "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[25] E. S. Finn et al., "Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity," *Nature Neurosci.*, vol. 18, no. 11, pp. 1664–1671, 2015.

[26] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[27] R. R. Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[28] A. Di Martino et al., "The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," *Mol. Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[29] C. Craddock et al., "The neuro bureau preprocessing initiative: Open sharing of preprocessed neuroimaging data and derivatives," *Front. Neuroinform.*, vol. 7, 2013, Art. no. 27.

[30] C.-G. Yan et al., "DPABI: Data processing & analysis for (resting-state) brain imaging," *Neuroinformatics*, vol. 14, no. 3, pp. 339–351, 2016.

[31] R. C. Craddock et al., "A whole brain fMRI atlas generated via spatially constrained spectral clustering," *Hum. Brain Mapping*, vol. 33, no. 8, pp. 1914–1928, 2012.

[32] X. Li et al., "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Med. Image Anal.*, vol. 65, 2020, Art. no. 101765.

[33] Y. Chen et al., "FedHealth: A federated transfer learning framework for wearable healthcare," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 83–93, Jul./Aug. 2020.
[34] M. Long et al., "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.
[35] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[36] E. Bagdasaryan et al., "Differential privacy has disparate impact on model accuracy," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 15479–15488.
[37] T. Murakami and Y. Kawamoto, "Utility-optimized local differential privacy mechanisms for distribution estimation," in *Proc. 28th USENIX Secur. Symp.*, 2019, pp. 1877–1894.
[38] C. Dwork et al., "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptography Conf.*, Springer, 2006, pp. 265–284.
[39] X. Tang et al., "A real-time arrhythmia heartbeats classification algorithm using parallel delta modulations and rotated linear-kernel support vector machines," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 4, pp. 978–986, Apr. 2020.
[40] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
[41] S. Mostafa et al., "Autoencoder based methods for diagnosis of autism spectrum disorder," in *Proc. Int. Conf. Comput. Adv. Bio Med. Sci.*, Springer, 2019, pp. 39–51.
[42] M. Kerpicci, H. Ozkan, and S. S. Kozat, "Online anomaly detection with bandwidth optimized hierarchical kernel density estimators," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 4253–4266, Sep. 2021.