# FedL2T: Personalized Federated Learning with Two-Teacher Distillation for Seizure Prediction

Jionghao Lou[1], Jian Zhang[2], Zhongmei Li[1], Lanlan Chen[1], Enbo Feng[1,*]
[1]Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai, China
[2]School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China
jhlou@mail.ecust.edu.cn, zhjmaster@sina.com, {lizhongmei, llchen}@ecust.edu.cn, enbo.feng@outlook.com

***Abstract*—The training of deep learning models in seizure prediction requires large amounts of Electroencephalogram (EEG) data. However, acquiring sufficient labeled EEG data is difficult due to annotation costs and privacy constraints. Federated Learning (FL) enables privacy-preserving collaborative training by sharing model updates instead of raw data. However, due to the inherent inter-patient variability in real-world scenarios, existing FL-based seizure prediction methods struggle to achieve robust performance under heterogeneous client settings. To address this challenge, we propose FedL2T, a personalized federated learning framework that leverages a novel two-teacher knowledge distillation strategy to generate superior personalized models for each client. Specifically, each client simultaneously learns from a globally aggregated model and a dynamically assigned peer model, promoting more direct and enriched knowledge exchange. To ensure reliable knowledge transfer, FedL2T employs an adaptive multi-level distillation strategy that aligns both prediction outputs and intermediate feature representations based on task confidence. In addition, a proximal regularization term is introduced to constrain personalized model updates, thereby enhancing training stability. Extensive experiments on two EEG datasets demonstrate that FedL2T consistently outperforms state-of-the-art FL methods, particularly under low-label conditions. Moreover, FedL2T exhibits rapid and stable convergence toward optimal performance, thereby reducing the number of communication rounds and associated overhead. These results underscore the potential of FedL2T as a reliable and personalized solution for seizure prediction in privacy-sensitive healthcare scenarios.**

***Index Terms*—EEG, seizure prediction, federated learning, knowledge distillation**

## I. Introduction

Epilepsy is a chronic neurological disorder caused by abnormal synchronous discharges in the brain, affecting over 65 million people worldwide [1]. Electroencephalography (EEG), which records brain activity through electrodes placed on the scalp, has become the primary technique for epilepsy analysis due to its non-invasiveness and high temporal resolution. Clinical studies have shown that a pre-ictal state, occurring several minutes before a seizure, is distinguishable from both seizure (ictal) and normal (inter-ictal) states in EEG recordings [2]. This distinction between pre-ictal and inter-ictal states makes seizure prediction possible.

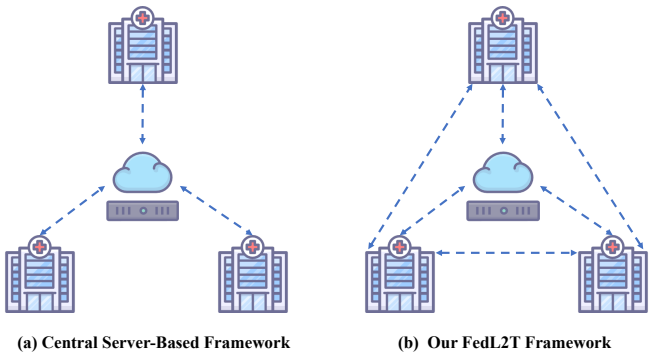


Fig. 1: Comparison of traditional FL and the proposed FedL2T framework. **(a)** Traditional FL relies on a central server for model aggregation, limiting personalization under heterogeneous data. **(b)** FedL2T introduces an additional peer teacher for distillation, enabling both global and cross-client knowledge transfer to promote learning diversity.

In recent years, deep learning has demonstrated remarkable success in predicting seizures [3]. However, this success largely depends on the availability of large-scale datasets. However, collecting sufficient EEG data for seizure prediction remains challenging due to the infrequent nature of seizures and the fact that EEG acquisition is typically confined to clinical settings. The scarcity of data underscores the importance of cross-institutional data sharing.

Nevertheless, privacy and ethical concerns, such as those mandated by the GDPR [4], severely restrict centralized data collection. Federated Learning (FL) [5] provides a promising solution by enabling collaborative model training without sharing raw data. A representative approach is the classical FedAvg algorithm [5], which enables data owners to conduct local training and iteratively aggregate their models on a central server. However, directly substituting local models with the aggregated global model in each round may degrade performance, especially under non-independent and identically

This work was supported by National Key Research and Development Program of China (2022YFB3305900), National Natural Science Foundation of China (62394343), Fundamental Research Funds for the Central Universities (ICT2024A01), Shanghai Rising-Star Program (24QA2706100), National Natural Science Foundation of China under Grant 62376095.

distributed (non-IID) data settings. To mitigate this issue, personalized Federated Learning (pFL) [6] allows each client to maintain a personalized model tailored to its local data, unlike traditional FL which uses a shared global model. For instance, Ditto [7] introduces a proximal regularization term to stabilize local updates, while FedBN [8] and FedRep [9] adopt personalized batch normalization and prediction layers, respectively, to better align feature representations across heterogeneous clients. In addition, knowledge distillation (KD) techniques [10] have been employed to prevent local models from being overwritten by the global model. In KD-based FL methods, the global model acts as a teacher that provides soft labels to guide each client's model. This allows clients to absorb global knowledge while retaining personalization, thereby preventing the direct replacement of local model parameters. However, approaches [11] that rely solely on soft labels from the averaged global model often suffer from anisotropic drift, especially in highly heterogeneous scenarios such as seizure prediction, where EEG distributions vary widely across subjects.

To address these limitations, we propose FedL2T, a novel personalized federated learning framework based on a two-teacher knowledge distillation paradigm (L2T). FedL2T enables each client to learn from both a global teacher and a dynamically assigned peer teacher. Specifically, each client acts not only as a student but also as a potential teacher for others, facilitating direct client-to-client knowledge transfer. This design reduces reliance on a single global teacher and enhances representation diversity. By offering diverse guidance, the peer teacher enables exploration beyond local optima, akin to mutation in genetic algorithms, which helps prevent premature convergence [12]. Meanwhile, the global teacher provides stable supervision and helps prevent overfitting. In addition, FedL2T performs hybrid distillation using soft labels and intermediate feature representations. The distillation strength is adaptively adjusted by prediction confidence, emphasizing reliable samples while down-weighting uncertain ones. This collaborative framework supports mutual learning between each client and the global teacher, as well as one-way knowledge transfer from peer clients. To further improve training stability under heterogeneous conditions, a regularization mechanism is introduced to constrain personalized model updates within a controlled range relative to the global model. Finally, we evaluate FedL2T on the public CHB-MIT dataset and a private real-world EEG dataset. Experimental results show that FedL2T consistently outperforms existing methods with faster and more stable convergence under fully supervised settings. Furthermore, it maintains high accuracy across low-label conditions, demonstrating strong robustness to label scarcity and underscoring its practical applicability and potential for real-world clinical deployment.

## II. Related Work

Federated learning has emerged as a promising approach for EEG-based health monitoring, enabling collaborative training of seizure detection and prediction models across distributed hospitals without sharing patient data. Baghersalimi et al. [13] applied the standard FedAvg algorithm to train the ECG-based seizure detection model on edge devices, demonstrating improved accuracy and energy efficiency. Saemaldahr and Ilyas [14] extended FedAvg within a three-tier architecture, integrating Spiking-GCNN and fuzzy inference to model the patient-specific seizure prediction task. To reduce communication overhead and central dependency, Ding et al. [15] proposed Fed-ESD for seizure detection, using fog nodes as dynamic aggregators in a hierarchical FL framework. Baghersalimi et al. [16] further extended this direction with a fully decentralized FL approach, combining adaptive model assembly and knowledge distillation to address non-IID challenges. Most recently, Ding et al. [17] presented FedMWAD, which uses module-wise weighted aggregation and personalized local updates to improve generalization while preserving individual adaptability across heterogeneous EEG sources for seizure prediction.

While most existing studies have concentrated on seizure detection, research on federated seizure prediction remains relatively limited. Moreover, the effectiveness of personalized patient-specific modeling still leaves room for improvement. To advance this line of research, we propose FedL2T, which, to our knowledge, is the first framework to introduce a two-teacher knowledge distillation strategy for personalized seizure prediction in a federated setting. By allowing each client to learn jointly from a global model and a dynamically assigned peer model, FedL2T improves model diversity and adaptability without sacrificing privacy or convergence stability. Extensive experiments demonstrate that FedL2T achieves superior classification performance compared to state-of-the-art FL methods.

## III. Methods

### A. Problem Definition

Given $K$ clients, each with a private labeled dataset $D_k = \left\{\left(x_i^k, y_i^k\right)\right\}_{i=1}^{n_k}$, which remains locally stored and is never shared. In the FedL2T framework, each client maintains two structurally identical models: a personalized model $\boldsymbol{P}$ tailored to its own data, and a transfer model $\boldsymbol{T}$, periodically synchronized with the global model $\boldsymbol{T}_G^r$ aggregated from all client transfer models. The goal of FedL2T is to collaboratively train robust personalized models across clients while preserving data privacy.

### B. Overview of the FedL2T Framework

Traditional KD-based FL methods typically rely on a single global teacher aggregated by averaging model parameters across clients. While distillation helps prevent the complete overwriting of local knowledge, it often falls short in highly heterogeneous tasks such as personalized EEG-based seizure prediction, where the global model fails to capture patient-specific characteristics. To address this, we propose the FedL2T framework, inspired by the mutation mechanism in genetic algorithms [12], which improves population diversity to escape local optima. FedL2T supplements the global teacher with a dynamically assigned peer teacher from another client

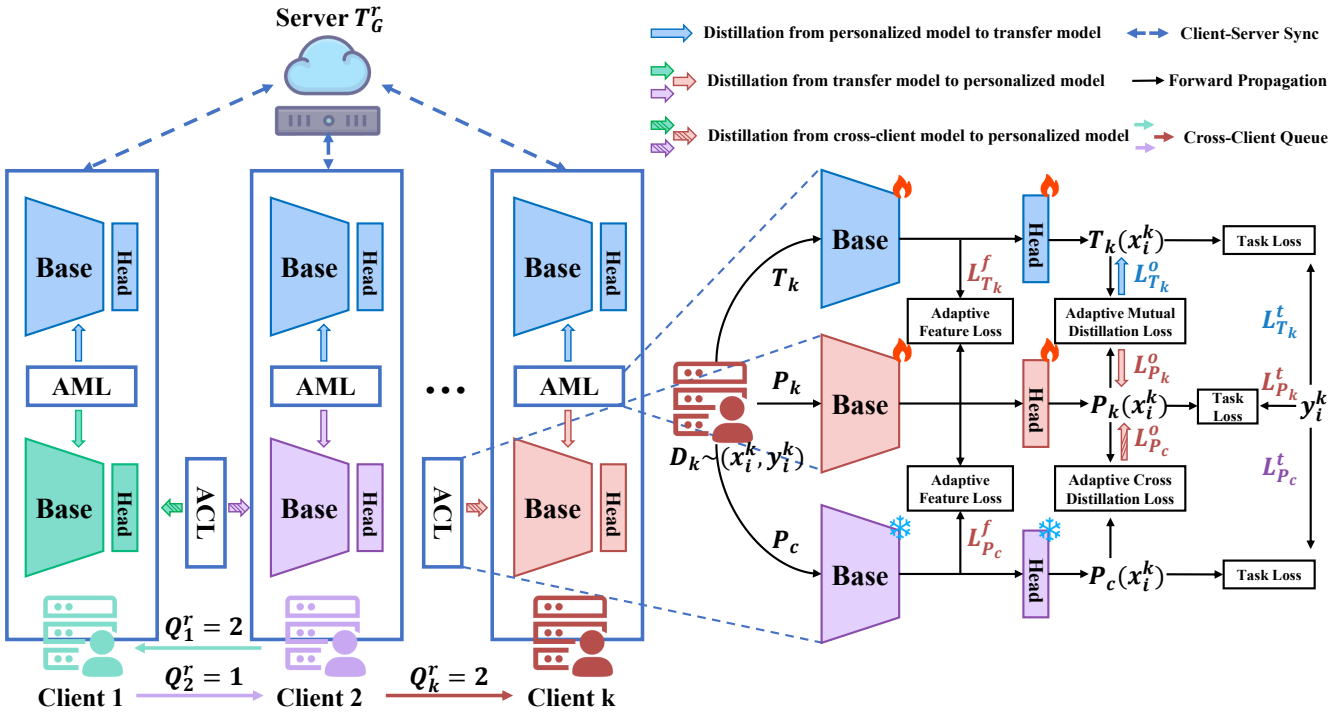


Fig. 2: Overview of the proposed FedL2T framework. Each client maintains a personalized model $\boldsymbol{P}_k$ and a transfer model $\boldsymbol{T}_k$. During local training, FedL2T performs two complementary knowledge distillation processes: Adaptive Mutual Learning (AML) between $\boldsymbol{P}_k$ and $\boldsymbol{T}_k$, and Adaptive Cross Learning (ACL) from a peer model $\boldsymbol{P}_c$ assigned via a cross-client queue $\mathcal{Q}^r$. Both soft predictions and intermediate features are leveraged for multi-level distillation. The transfer models $\boldsymbol{T}_k$ are periodically synchronized with the global model $\boldsymbol{T}_G^r$ aggregated on the server, enabling global coordination without sharing raw data. Colored arrows indicate the direction of knowledge transfer. Flame and snowflake icons respectively indicate updated and frozen parameters, with the latter excluded from gradient updates during training.

during training, promoting richer and more diverse knowledge transfer. An overview of the FedL2T framework is illustrated in Fig. 2.

*1) Cross-Client Distillation Pathway:* To enable client-to-client knowledge sharing, FedL2T establishes a cross-client distillation pathway in each communication round. The cross-client queue is denoted as $\mathcal{Q}^r = \{\mathcal{Q}_1^r, \mathcal{Q}_2^r, \ldots, \mathcal{Q}_k^r\}_{k=1}^K$, where $\mathcal{Q}_k^r$ represents the peer teacher for client $k$ in round $r$. It is generated randomly each round under the constraint $\mathcal{Q}_k^r \neq k$, allowing each client to learn from a different peer and introducing representational diversity into the training process. An example queue $\{2, 1, \ldots, 2\}$ in Fig. 2 illustrates a scenario with three clients. Once $\mathcal{Q}^r$ is determined, all clients update in parallel. The update includes two components: Adaptive Cross Learning (ACL) and Adaptive Mutual Learning (AML), which are described in the following subsections.

*2) Adaptive Cross Learning:* Given $\mathcal{Q}_k^r = c$, client $k$ receives the $\boldsymbol{P}_c$ from client $c$ and performs one-way knowledge distillation from $\boldsymbol{P}_c$ to its personalized model $\boldsymbol{P}_k$. Acting as a cross-client teacher, $\boldsymbol{P}_c$ provides external knowledge through multi-level distillation, thereby enhancing the generalization capability of $\boldsymbol{P}_k$. In addition to distillation, $\boldsymbol{P}_k$ is optimized on its private dataset $D_k$ using cross-entropy loss to fit the ground truth labels. This local supervision ensures that $\boldsymbol{P}_k$ retains its base performance and mitigates the risk of client drift. Meanwhile, $\boldsymbol{P}_c$ performs forward inference on $D_k$ with frozen parameters and remains unchanged. For each training sample $(x_i^k, y_i^k) \in D_k$, let $\boldsymbol{P}_k(x_i^k)$ and $\boldsymbol{P}_c(x_i^k)$ denote the soft probabilities predicted by the personalized and cross-client models, respectively. The task losses $\mathcal{L}_{P_k}^t$ and $\mathcal{L}_{P_c}^t$ are computed as:

$$\mathcal{L}_{P_k}^t = \mathcal{L}_{CE}\left(\boldsymbol{P}_k(x_i^k), y_i^k\right), \ \mathcal{L}_{P_c}^t = \mathcal{L}_{CE}\left(\boldsymbol{P}_c(x_i^k), y_i^k\right) \quad (1)$$

where $\mathcal{L}_{CE}(p, y) = -[y \log p + (1-y)\log(1-p)]$ denotes the binary cross-entropy loss, which is widely used in seizure prediction to quantify the discrepancy between predicted probabilities and binary ground-truth labels.

Previous studies [18], [19] have shown that incorporation of knowledge distillation at the feature level can provide a finer-grained and more comprehensive transfer of information between models. Thus, we adopt a multi-level distillation strategy comprising both output-level and feature-level alignment to enhance cross-client knowledge transfer. Specifically, the output-level loss $\mathcal{L}_{P_c}^o$ corresponds to the Adaptive Cross

Distillation Loss module in Fig. 2, while the feature-level loss $\mathcal{L}^f_{P_c}$ corresponds to the Adaptive Feature Loss module. These losses are defined as follows:

$$\mathcal{L}^o_{P_c} = \mathcal{L}_{KL}\left(\boldsymbol{P}_c(x_i^k) \,\|\, \boldsymbol{P}_k(x_i^k)\right) \;/\; (\mathcal{L}^t_{P_c} + \mathcal{L}^t_{P_k}) \quad (2)$$

$$\mathcal{L}^f_{P_c} = \mathcal{L}_{MSE}\left(\boldsymbol{P}_{cb}(x_i^k), \boldsymbol{P}_{kb}(x_i^k)\right) \;/\; (\mathcal{L}^t_{P_c} + \mathcal{L}^t_{P_k}) \quad (3)$$

where $\mathcal{L}_{KL}$ denotes the Kullback–Leibler divergence between the predicted soft probabilities of the teacher $\boldsymbol{P}_c$ and the student $\boldsymbol{P}_k$, and $\mathcal{L}_{MSE}$ computes the mean squared error between their intermediate feature representations extracted by the base block ($b$) of model $\boldsymbol{P}$ (detailed in Section IV-B). Furthermore, to achieve adaptive regulation of knowledge transfer, we adopt the strategy proposed in FedKD [20], where both distillation losses are normalized by the sum of task losses. This design ensures that, when task losses are large, indicating less reliable predictions, the impact of distillation is reduced to prevent the propagation of low-confidence knowledge. In contrast, when both models achieve low task losses, the normalized distillation loss increases accordingly, which enhances the transfer of reliable knowledge and helps mitigate overfitting.

*3) Adaptive Mutual Learning:* Complementary to ACL, which introduces diversity via cross-client knowledge transfer, AML facilitates mutual knowledge exchange [11], [21] between the local transfer model $\boldsymbol{T}_k$ and the personalized model $\boldsymbol{P}_k$. This bidirectional interaction enables $\boldsymbol{P}_k$ to absorb global knowledge from $\boldsymbol{T}_k$, while $\boldsymbol{T}_k$, enriched with local task-specific information, is periodically uploaded to the server and aggregated into the global model $\boldsymbol{T}^r_G$. Specifically, during each local epoch, both $\boldsymbol{P}_k$ and $\boldsymbol{T}_k$ are updated using local data $D_k$ while engaging in an adaptive mutual distillation process to exchange knowledge bidirectionally. The output-level distillation loss from $\boldsymbol{P}_k$ to $\boldsymbol{T}_k$, corresponding to the Adaptive Mutual Distillation Loss module in Fig. 2, is defined as:

$$\mathcal{L}^o_{T_k} = \mathcal{L}_{KL}\left(\boldsymbol{P}_k(x_i^k) \,\|\, \boldsymbol{T}_k(x_i^k)\right) / \left(\mathcal{L}^t_{P_k} + \mathcal{L}^t_{T_k}\right) \quad (4)$$

where $\mathcal{L}^t_{P_k}$ and $\mathcal{L}^t_{T_k}$ represent the task losses of $\boldsymbol{P}_k$ and $\boldsymbol{T}_k$, respectively, both computed using the formulation in Eq. (1) with the corresponding model substituted. The remaining losses, including the output-level distillation loss $\mathcal{L}^o_{P_k}$ (from $\boldsymbol{T}_k$ to $\boldsymbol{P}_k$) and the feature-level loss $\mathcal{L}^f_{T_k}$, are computed following Eqs. (2) and (3), with $\boldsymbol{T}_k$ substituted for $\boldsymbol{P}_c$.

Although the two-teacher strategy comprising ACL and AML facilitates both diversity and global knowledge integration, data heterogeneity may cause personalized models to deviate considerably from the global model. This deviation often leads to severe fluctuations during distillation, resulting in unstable training and poor generalization. To mitigate this issue, we introduce a regularization mechanism inspired by [7], which penalizes the divergence of $\boldsymbol{P}_k$ from the global model $\boldsymbol{T}^r_G$. The proximal term is formulated as:

$$\mathcal{L}_{\text{prox}} = \frac{\mu}{2}\left\|\boldsymbol{P}_k - \boldsymbol{T}^r_G\right\|_2^2 \quad (5)$$

where $\mu$ is a tunable hyperparameter that controls the strength of regularization.

**Algorithm 1:** The FedL2T Algorithm

**Input:** Total communication rounds $R$, Local epochs $E$
Total number of clients $K$
The network structure for $\boldsymbol{P}$, $\boldsymbol{T}$
**Output:** The trained personalized local models $\{\boldsymbol{P}_k\}_{k=1}^K$
**Server executes:**

Initialize the model parameters $\boldsymbol{P}$, $\boldsymbol{T}$ for all clients
**for** each round $r$ in $R$ **do**
  **for** each client $k$ **in parallel do**
    Assign random cross-client teacher $\boldsymbol{P}_c$
    $\boldsymbol{T}_k \leftarrow$ **ClientUpdate**$(k, \boldsymbol{P}_k, \boldsymbol{P}_c)$
  **end for**
  Update global model: $\boldsymbol{T}^{r+1}_G \leftarrow \sum_{k=1}^K \frac{n_k}{n}\boldsymbol{T}_k$
  Broadcast $\boldsymbol{T}^{r+1}_G$ to all $\boldsymbol{T}_k$
**end for**

**ClientUpdate**$(k, \boldsymbol{P}_k, \boldsymbol{P}_c)$**:**

**for** each local epoch $e$ in $E$ **do**
  Compute task losses $\mathcal{L}^t_{P_k}$, $\mathcal{L}^t_{T_k}$ and $\mathcal{L}^t_{P_c}$
  Compute output distillation losses $\mathcal{L}^o_{P_k}$, $\mathcal{L}^o_{T_k}$ and $\mathcal{L}^o_{P_c}$
  Compute feature distillation losses $\mathcal{L}^f_{T_k}$ and $\mathcal{L}^f_{P_c}$
  Compute proximal term $\mathcal{L}_{prox}$
  $\mathcal{L}_{T_k} \leftarrow \mathcal{L}^t_{T_k} + (\mathcal{L}^o_{T_k} + \mathcal{L}^f_{T_k})$
  $\mathcal{L}_{P_k} \leftarrow \mathcal{L}^t_{P_k} + (\mathcal{L}^o_{P_k} + \mathcal{L}^f_{T_k}) + \lambda_c(\mathcal{L}^o_{P_c} + \mathcal{L}^f_{P_c}) + \mathcal{L}_{prox}$
  $\boldsymbol{T}_k \leftarrow \boldsymbol{T}_k - \eta\nabla_{T_k}\mathcal{L}_{T_k}$
  $\boldsymbol{P}_k \leftarrow \boldsymbol{P}_k - \eta\nabla_{P_k}\mathcal{L}_{P_k}$
**end for**
**return** $\boldsymbol{T}_k$

*4) Client Update and Server Aggregation:* Finally, the total loss of client $k$ in FedL2T at each epoch is formulated as:

$$\mathcal{L}_{T_k} = \mathcal{L}^t_{T_k} + (\mathcal{L}^o_{T_k} + \mathcal{L}^f_{T_k}) \quad (6)$$

$$\mathcal{L}_{P_k} = \mathcal{L}^t_{P_k} + (\mathcal{L}^o_{P_k} + \mathcal{L}^f_{T_k}) + \lambda_c(\mathcal{L}^o_{P_c} + \mathcal{L}^f_{P_c}) + \mathcal{L}_{prox} \quad (7)$$

where $\boldsymbol{T}_k$ is optimized using task supervision and distillation from $\boldsymbol{P}_k$, while $\boldsymbol{P}_k$ incorporates an additional proximal term and cross-client distillation from $\boldsymbol{P}_c$, weighted by $\lambda_c$ to balance contributions from both teachers. Subsequently, $\boldsymbol{T}_k$ and $\boldsymbol{P}_k$ are updated by gradient descent:

$$\boldsymbol{T}_k = \boldsymbol{T}_k - \eta\nabla_{T_k}\mathcal{L}_{T_k},\ \boldsymbol{P}_k = \boldsymbol{P}_k - \eta\nabla_{P_k}\mathcal{L}_{P_k} \quad (8)$$

where $\eta$ denotes the learning rate. The gradients $\nabla_{T_k}\mathcal{L}_{T_k}$ and $\nabla_{P_k}\mathcal{L}_{P_k}$ are computed with respect to the total loss of $\boldsymbol{T}_k$ and $\boldsymbol{P}_k$, respectively.

After $E$ epochs of local updates, each client uploads its optimized $\boldsymbol{T}_k$ to the server, which aggregates all transfer models into a global model $\boldsymbol{T}^{r+1}_G$ by weighted averaging:

$$\boldsymbol{T}^{r+1}_G = \sum_{k=1}^K \frac{n_k}{n}\boldsymbol{T}_k, \quad \text{where } n = \sum_{k=1}^K n_k \quad (9)$$

This aggregated model is then broadcast as the new initialization for $\boldsymbol{T}_k$ in the next round. Through global updates of

$\boldsymbol{T}_k$ and local training of $\boldsymbol{P}_k$, FedL2T enables personalized optimization while preserving data privacy. The detailed steps of the proposed FedL2T algorithm are presented in Algorithm 1.

## IV. Experiments

This section presents the experimental setup to evaluate our proposed FedL2T framework for epileptic seizure prediction.

### A. EEG Datasets

To evaluate the proposed method, experiments are conducted on both a public EEG dataset and a private EEG dataset, with all recordings following the international 10–20 electrode placement system. The public CHB-MIT dataset [22] includes long-term scalp EEG recordings from 23 pediatric subjects sampled at 256 Hz, with 21 commonly available electrodes selected as input. The private dataset consists of EEG recordings from 18 adult patients acquired at 200 Hz, using 16 channel configurations. All raw EEG data are preprocessed with notch filtering to suppress power-line noise, followed by z-score normalization. For the CHB-MIT dataset, the continuous signals are segmented into 7-second windows, while the private data is segmented using the same number of sampling points with 50% overlap to augment data samples. Each segment is denoted as $x_i \in \mathbb{R}^{chs \times T}$, where $chs$ is the number of EEG channels and $T$ is the number of sampling points. In the FL setting, each subject is treated as an individual client with a private dataset $D_k = (x_i^k, y_i^k)_{i=1}^{n_k}$.

The Seizure Occurrence Period (SOP) and Seizure Prediction Horizon (SPH) are set to 30 and 5 minutes, respectively, balancing timely intervention with reduced patient anxiety. Inter-ictal segments are defined as EEG windows occurring at least 4 hours from any seizure to ensure clear separation from ictal states. To address class imbalance, inter-ictal samples are undersampled. To reflect realistic clinical settings where only past data can inform future predictions, a leave-last-event-out strategy is adopted: for each patient with $N$ seizures, the last is used for testing, and the remaining $(N-1)$ seizures are used for training. The detailed subject information for both the CHB-MIT and private datasets is summarized in Table I. As the experimental results report subject-wise performance, segment-based accuracy (Acc) is adopted as the primary evaluation metric for a concise comparison. Accuracy, defined as the proportion of correctly classified segments among all segments, offers a reliable measure of model performance under class-balanced conditions.

### B. Implementation Details

The model architecture employed in our experiments is summarized in Table II. All experiments are implemented in PyTorch and run on a system with an NVIDIA GeForce RTX 4090 GPU and Ubuntu OS. The models are optimized using stochastic gradient descent (SGD) with a learning rate of $\eta = 0.01$. The regularization coefficient $\mu$ is set to 0.2, and the cross-client distillation weight $\lambda_c$ is set to 0.5. A total of $R = 100$ communication rounds are performed, with each client conducting $E = 1$ local epoch per round.

TABLE I: Detailed Description of Subjects Included in the CHB-MIT and Private Datasets.

| CHB-MIT | | | | | | Private | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Subject | Age | Seizures | Subject | Age | Seizures | Subject | Age | Seizures | Subject | Age | Seizures |
| 1 | 11 | 6 | 14 | 9 | 4 | S01 | 24 | 4 | S10 | 32 | 5 |
| 2 | 11 | 3 | 16 | 7 | 5 | S02 | 21 | 4 | S11 | 48 | 4 |
| 3 | 14 | 6 | 17 | 12 | 3 | S03 | 85 | 3 | S12 | 60 | 4 |
| 5 | 7 | 5 | 18 | 18 | 6 | S04 | 46 | 4 | S13 | 41 | 4 |
| 6 | 1.5 | 7 | 19 | 19 | 3 | S05 | 48 | 3 | S14 | 21 | 4 |
| 7 | 14.5 | 3 | 20 | 6 | 5 | S06 | 17 | 5 | S15 | 49 | 3 |
| 8 | 3.5 | 4 | 21 | 13 | 4 | S07 | 35 | 4 | S16 | 87 | 3 |
| 9 | 10 | 4 | 22 | 9 | 3 | S08 | 26 | 4 | S17 | 39 | 3 |
| 10 | 3 | 6 | 23 | 6 | 4 | S09 | 59 | 3 | S18 | 25 | 4 |

TABLE II: Details of Model Architecture.

| Layer | Input | Type | Kernel | Stride | Pool | Stride | Output |
|---|---|---|---|---|---|---|---|
| Conv1 | $[chs,\ T]$ | Conv1d | 5 | 2 | 2 | 2 | $[64,\ T//4]$ |
| Conv2 | $[64,\ T//4]$ | Conv1d | 3 | 2 | 2 | 2 | $[64,\ T//16]$ |
| Conv3 | $[64,\ T//16]$ | Conv1d | 3 | 1 | 2 | 2 | $[64,\ T//32]$ |
| Conv4 | $[64,\ T//32]$ | Conv1d | 2 | 1 | 2 | 2 | $[128,\ T//64]$ |
| GRU | $[128,\ T//64]^\top$ | GRU | – | – | – | – | [128] |
| FC | [128] | Linear | – | – | – | – | [2] |

Note: // denotes integer division; $^\top$ indicates transposition. The GRU contains 128 hidden units, with only the final hidden state used as output.

## V. Results and Discussion

### A. Comparison with State-of-the-art Methods

To evaluate the effectiveness of the proposed FedL2T framework, we conduct extensive experiments against several representative FL methods, including FedAvg [5], FedBN [8], FedRep [9], FedMRL [23], Ditto [7], and FML [11], as well as the Local baseline where each client is trained independently without collaboration, i.e., patient-specific training strategy commonly adopted in the seizure prediction task. As shown in Table III, our proposed FedL2T achieves the highest average accuracy of 98.00% on the CHB-MIT dataset, significantly outperforming all competing methods. Notably, it surpasses the Local baseline by 5.56%, demonstrating its ability to retain personalization while benefiting from cross-client knowledge transfer. Similar advantages are observed on the private dataset in Table IV, where FedL2T reaches 96.40% average accuracy, exceeding the Local baseline by 5.72% and consistently outperforming all other methods. These results validate the robustness and generalizability of the proposed two-teacher learning paradigm strategy across heterogeneous datasets.

Serving as a lower bound for FL performance, FedAvg achieves average accuracies of 69.09% and 78.77% on CHB-MIT and the private datasets, respectively. In contrast, all five personalized FL methods exhibit notable improvements by incorporating different personalization strategies. FedBN applies local batch normalization to mitigate feature shift, FedRep decouples shared representation learning from personalized classification heads, and FedMRL introduces a nested model design to accommodate structural heterogeneity. Moreover, Ditto incorporates a regularization-based dual model strategy to stabilize local adaptation, while FML employs mutual knowledge distillation to transfer global knowledge. However, their performance relative to the Local baseline is inconsistent.

TABLE III: Comparison with State-of-the-art and Ablation Variants on CHB-MIT Dataset.

| Subject | Local | FedAvg [5] | FedBN [8] | FedRep [9] | FedMRL [23] | Ditto [7] | FML [11] | L2T-C | L2T-CG | FedL2T (Ours) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 96.89 | 88.91 | 98.44 | 97.28 | 96.50 | 96.89 | 94.94 | 95.91 | 97.47 | 100.00 |
| 2 | 70.04 | 57.98 | 99.61 | 98.05 | 70.43 | 96.89 | 83.66 | 82.88 | 66.54 | 99.61 |
| 3 | 95.91 | 85.41 | 84.63 | 94.75 | 98.83 | 92.80 | 97.28 | 98.64 | 99.61 | 97.67 |
| 5 | 99.03 | 95.33 | 94.75 | 98.25 | 99.22 | 99.81 | 97.08 | 99.42 | 100.00 | 99.81 |
| 6 | 77.24 | 79.77 | 95.33 | 64.79 | 66.54 | 99.81 | 58.95 | 65.56 | 100.00 | 99.81 |
| 7 | 93.39 | 38.13 | 96.11 | 99.81 | 80.16 | 90.47 | 94.94 | 98.25 | 93.00 | 98.83 |
| 8 | 96.69 | 89.69 | 99.42 | 100.00 | 98.44 | 99.42 | 97.86 | 99.22 | 100.00 | 100.00 |
| 9 | 100.00 | 70.04 | 99.61 | 96.11 | 99.81 | 100.00 | 99.03 | 100.00 | 100.00 | 100.00 |
| 10 | 96.50 | 55.45 | 91.25 | 99.61 | 100.00 | 99.81 | 99.61 | 100.00 | 100.00 | 99.81 |
| 14 | 70.04 | 56.23 | 97.67 | 92.22 | 92.02 | 100.00 | 91.25 | 98.64 | 100.00 | 100.00 |
| 16 | 92.41 | 93.77 | 98.64 | 93.19 | 98.83 | 96.69 | 93.97 | 98.25 | 100.00 | 94.36 |
| 17 | 93.97 | 49.22 | 69.84 | 97.08 | 89.69 | 58.17 | 92.02 | 92.80 | 97.47 | 81.13 |
| 18 | 99.42 | 71.79 | 81.52 | 99.42 | 99.61 | 99.22 | 98.05 | 99.81 | 100.00 | 100.00 |
| 19 | 99.22 | 50.00 | 89.11 | 100.00 | 99.03 | 93.39 | 99.03 | 99.42 | 99.81 | 100.00 |
| 20 | 100.00 | 59.14 | 96.89 | 99.81 | 97.67 | 96.30 | 100.00 | 99.81 | 100.00 | 100.00 |
| 21 | 94.36 | 55.06 | 50.00 | 54.28 | 93.00 | 90.47 | 87.94 | 84.44 | 98.05 | 93.00 |
| 22 | 88.72 | 84.63 | 99.42 | 97.67 | 93.77 | 99.81 | 99.42 | 98.83 | 97.08 | 100.00 |
| 23 | 100.00 | 63.04 | 99.81 | 100.00 | 99.81 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Avg | 92.44 | 69.09↓23.35 | 91.23↓1.21 | 93.46↑1.02 | 92.96↑0.52 | 95.00↑2.56 | 93.61↑1.17 | 95.10↑2.66 | 97.17↑4.73 | **98.00**↑5.56 |

Note: All values represent classification accuracy (%) for each subject. The "Avg" row reports the average accuracy across all subjects, where **Bold** denotes the highest averaged accuracy. The up↑ and down↓ arrows represent the improvement or decline relative to the Local baseline.

TABLE IV: Comparison with State-of-the-art and Ablation Variants on Private Dataset.

| Subject | Local | FedAvg [5] | FedBN [8] | FedRep [9] | FedMRL [23] | Ditto [7] | FML [11] | L2T-C | L2T-CG | FedL2T (Ours) |
|---|---|---|---|---|---|---|---|---|---|---|
| S01 | 92.98 | 67.17 | 85.34 | 73.81 | 88.85 | 93.36 | 97.99 | 96.62 | 97.74 | 93.61 |
| S02 | 99.37 | 75.69 | 100.00 | 100.00 | 99.75 | 92.48 | 99.00 | 97.24 | 99.00 | 99.50 |
| S03 | 96.74 | 95.99 | 99.75 | 100.00 | 100.00 | 99.37 | 97.49 | 97.37 | 99.12 | 99.50 |
| S04 | 51.88 | 67.29 | 55.64 | 50.13 | 83.46 | 66.79 | 88.60 | 84.46 | 85.09 | 82.32 |
| S05 | 97.87 | 69.80 | 99.50 | 99.75 | 100.00 | 97.62 | 99.37 | 99.25 | 99.12 | 98.87 |
| S06 | 98.50 | 86.59 | 99.75 | 100.00 | 99.87 | 78.95 | 100.00 | 99.75 | 99.75 | 99.87 |
| S07 | 91.48 | 57.64 | 86.47 | 91.73 | 93.23 | 79.32 | 68.42 | 83.83 | 87.72 | 91.85 |
| S08 | 75.94 | 46.62 | 65.16 | 56.89 | 84.71 | 81.33 | 93.98 | 94.74 | 91.85 | 96.92 |
| S09 | 92.98 | 88.22 | 93.73 | 95.74 | 94.74 | 94.36 | 96.24 | 97.62 | 97.99 | 98.25 |
| S10 | 92.11 | 76.32 | 96.49 | 77.32 | 86.22 | 98.62 | 50.50 | 93.48 | 94.99 | 93.11 |
| S11 | 97.49 | 91.23 | 99.25 | 92.61 | 95.74 | 99.75 | 95.11 | 97.49 | 98.50 | 99.25 |
| S12 | 87.72 | 81.08 | 97.87 | 99.87 | 97.74 | 89.10 | 98.25 | 93.48 | 94.61 | 99.50 |
| S13 | 91.48 | 92.48 | 92.23 | 99.50 | 92.36 | 95.36 | 95.99 | 99.00 | 99.25 | 97.12 |
| S14 | 84.71 | 70.80 | 94.61 | 93.23 | 91.48 | 82.46 | 92.98 | 98.12 | 95.49 | 88.10 |
| S15 | 96.49 | 72.56 | 70.68 | 87.72 | 93.11 | 98.25 | 97.99 | 96.12 | 99.62 | 99.12 |
| S16 | 96.74 | 81.70 | 88.97 | 98.87 | 97.87 | 91.35 | 100.00 | 98.12 | 98.75 | 99.00 |
| S17 | 88.47 | 96.74 | 99.37 | 70.93 | 79.57 | 99.62 | 99.62 | 68.92 | 86.97 | 99.50 |
| S18 | 99.25 | 99.87 | 99.75 | 89.85 | 89.97 | 99.87 | 90.10 | 99.75 | 99.75 | 99.75 |
| Avg | 90.68 | 78.77↓11.91 | 90.25↓0.43 | 87.66↓3.02 | 92.70↑2.02 | 91.00↑0.32 | 92.31↑1.63 | 94.19↑3.51 | 95.85↑5.17 | **96.40**↑5.72 |

Note: All values represent classification accuracy (%) for each subject. The "Avg" row reports the average accuracy across all subjects, where **Bold** denotes the highest averaged accuracy. The up↑ and down↓ arrows represent the improvement or decline relative to the Local baseline.

While FedMRL, Ditto and FML generally achieve slight improvements over Local on both datasets, FedBN and FedRep occasionally underperform. This inconsistency highlights the challenge of stable personalization and further emphasizes the effectiveness of FedL2T in handling data heterogeneity.

### *B. Ablation Study*

To assess the contribution of each component within the proposed FedL2T framework, we conduct ablation studies on both the CHB-MIT and private datasets. Specifically, we evaluate two simplified variants: L2T-C, which utilizes only cross-client collaboration via a single peer teacher without any involvement from the central server, and L2T-CG, which incorporates both global and peer supervision but omits the proximal term used in the full FedL2T.

As shown in Table III and Table IV, the cross-client variant L2T-C improves accuracy by 2.66% and 3.51% over the Local baseline on the CHB-MIT and private datasets, respectively. These gains already outperform several existing baselines, including FML, which also adopts teacher-based distillation. This highlights the effectiveness of cross-client collaboration over centralized global supervision in highly heterogeneous settings. Building upon L2T-C, the L2T-CG variant introduces global model guidance, forming a two-teacher framework with adaptive distillation strength control. This leads to performance improvements of 4.73% and 5.17% on the respective datasets, demonstrating the benefit of combining global and cross-client knowledge sources. Finally, by incorporating a proximal term into L2T-CG, the full FedL2T framework further stabilizes local updates, yielding the highest performance

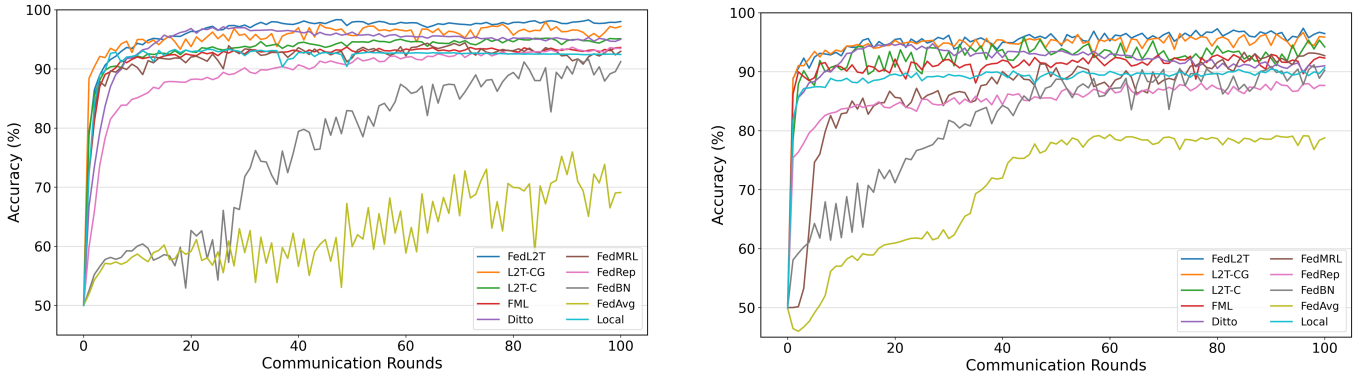


Fig. 3: Comparison of classification accuracy (%) across communication rounds on the CHB-MIT (left) and Private (right) datasets. Each curve represents the average test accuracy of a federated method across clients.

TABLE V: Accuracy (%) Comparison under Varying Labeled Data Ratios (%) on Two Datasets. "Mean ± Std" Represents the Average Accuracy and Standard Deviation Across 18 Subjects.

| Methods | CHB-MIT | | | | Private | | | |
|---|---|---|---|---|---|---|---|---|
| | 1% Labels | 5% Labels | 10% Labels | 25% Labels | 1% Labels | 5% Labels | 10% Labels | 25% Labels |
| Local | 73.99 ± 17.46 | 86.10 ± 13.31 | 89.66 ± 13.95 | 90.43 ± 13.57 | 81.60 ± 12.83 | 85.90 ± 10.60 | 87.52 ± 10.57 | 88.62 ± 10.88 |
| FedAvg [5] | 52.84 ± 12.42 | 54.17 ± 8.36 | 63.80 ± 10.09 | 66.75 ± 22.95 | 49.65 ± 2.54 | 50.49 ± 15.12 | 55.53 ± 10.26 | 61.48 ± 23.52 |
| FedBN [8] | 58.84 ± 13.76 | 71.66 ± 10.43 | 82.77 ± 12.36 | 88.63 ± 13.31 | 79.35 ± 12.18 | 84.45 ± 12.21 | 87.01 ± 11.78 | 89.55 ± 12.85 |
| FedRep [9] | 69.53 ± 15.16 | 79.83 ± 16.47 | 86.42 ± 14.05 | 90.24 ± 13.08 | 81.84 ± 14.99 | 83.88 ± 14.85 | 85.55 ± 15.15 | 86.32 ± 15.68 |
| FedMRL [23] | 72.56 ± 9.05 | 82.34 ± 15.56 | 90.30 ± 12.10 | 91.06 ± 11.63 | 82.05 ± 15.45 | 85.89 ± 14.00 | 88.43 ± 13.32 | 91.73 ± 8.00 |
| Ditto [7] | 80.56 ± 12.70 | 86.15 ± 11.25 | 89.20 ± 10.13 | 91.49 ± 9.16 | 83.10 ± 11.33 | 86.58 ± 11.45 | 87.80 ± 11.75 | 90.59 ± 10.52 |
| FML [11] | 82.22 ± 11.75 | 88.29 ± 10.22 | 91.07 ± 9.30 | 92.38 ± 9.60 | 85.53 ± 13.11 | 89.31 ± 11.89 | 91.07 ± 10.83 | 91.77 ± 9.76 |
| FedL2T (Ours) | **87.33 ± 10.37** | **92.06 ± 8.65** | **95.21 ± 4.76** | **97.09 ± 4.76** | **89.79 ± 8.26** | **93.94 ± 5.48** | **94.40 ± 4.95** | **95.72 ± 3.40** |

across both datasets. These results validate the necessity and complementarity of each component in the FedL2T design.

### C. Comparison of Convergence Rate

We further analyze the convergence behaviors of all FL methods, including seven baselines and three ablation variants, as depicted in Fig. 3. FedL2T consistently achieves the highest accuracy in the final communication rounds on both datasets. Notably, it exhibits a relatively fast convergence rate, particularly in the early stage of training, second only to the L2T-CG variant. This rapid progress can be attributed the efficient escape from local optima enabled by cross-client collaboration. However, the dynamic nature of peer interactions in L2T-C and L2T-CG introduces larger fluctuations. Interestingly, Ditto demonstrates the most stable convergence owing to the proximal term that constrains local updates. Nevertheless, its performance declines in later stages due to its exclusive dependence on global model knowledge.

Building on these insights, the full FedL2T framework combines the stability of Ditto with the collaborative advantages of L2T-CG. As a result, it achieves a favorable balance between convergence speed, training stability, and final accuracy, reaching high performance with minimal oscillations.

### D. Evaluation under Limited Label Supervision

To simulate practical scenarios with limited labeled data, we evaluate the proposed FedL2T framework against several representative FL methods on both the CHB-MIT and private EEG datasets. The experiments are conducted under varying proportions of labeled data (1%, 5%, 10%, and 25%) to assess performance under different levels of supervision.

TABLE VI: Impact of $\lambda_c$ on Accuracy (%) for Two Datasets.

| $\lambda c$ | 0 | 0.25 | 0.50 | 0.75 | 1.00 | 1.25 | 1.50 |
|---|---|---|---|---|---|---|---|
| CHB-MIT | 96.35 | 97.61 | **98.00** | 97.58 | 97.06 | 96.97 | 96.49 |
| Private | 93.26 | 95.46 | **95.75** | 95.27 | 95.22 | 95.02 | 94.86 |

As shown in Table V, FedL2T consistently outperforms all baseline methods across both datasets and all labeling levels, demonstrating strong adaptability to limited supervision. Notably, when the labeled data ratio drops below 10%, most existing methods fall short of the Local baseline. This performance degradation is particularly evident for methods such as FedMRL and Ditto, which, despite their advantages under fully supervised conditions, struggle to maintain effectiveness with reduced labeled data. The scarcity of labeled data may amplify inter-client heterogeneity, increasing variance among local models and hindering the quality of global aggregation. Consequently, methods relying heavily on global model guidance tend to be more vulnerable under such conditions. While FML also leverages global knowledge, it mitigates this issue through knowledge distillation, allowing more flexible and localized adaptation. Building on this idea, FedL2T further improves robustness by integrating global and cross-client supervision in a balanced manner and introducing adaptive weighting to ensure effective knowledge transfer under low-

label settings. This design contributes to its consistently superior performance in low-resource federated learning scenarios.

### E. Analysis of Hyperparameter

A key innovation of the proposed FedL2T framework lies in enabling each personalized model $\boldsymbol{P}_k$ to distill knowledge from two sources: a local transfer model $\boldsymbol{T}_k$ and a peer model $\boldsymbol{P}_c$. Each distillation path is adaptively weighted via task-loss-based normalization, reflecting the reliability of supervision. In addition, the hyperparameter $\lambda_c$ explicitly modulates the contribution from the peer model.

To assess the sensitivity of $\lambda_c$, we evaluate model performance on both datasets with $\lambda_c$ ranging from 0 to 1.5. As shown in Table VI, accuracy increases with larger $\lambda_c$, reaching its peak at $\lambda_c = 0.5$, and declines thereafter. Specifically, the case of $\lambda_c = 0$ corresponds to a variant of FML augmented with adaptive distillation loss and proximal regularization. Its performance exceeds that of both FML and Ditto, demonstrating the value of these enhancements. Overall, the results validate the importance of balancing complementary knowledge sources, with $\lambda_c = 0.5$ serving as an effective default.

## VI. Conclusions

This paper presents FedL2T, a personalized federated learning framework with a two-teacher distillation strategy for seizure prediction. By allowing each client to learn from both a global model and a dynamically assigned peer, FedL2T facilitates more effective and diverse knowledge transfer. The framework further integrates adaptive multi-level distillation and proximal regularization to enhance training stability and personalization under non-IID conditions. Extensive experiments on both public and private EEG datasets demonstrate that FedL2T consistently outperforms state-of-the-art federated learning methods in terms of accuracy, convergence speed, and robustness under limited supervision. These results highlight the potential of FedL2T for deployment in real-world clinical settings and offer a promising direction for scalable personalized FL in healthcare. While the current framework focuses on homogeneous model architectures and random peer assignment for simplicity, it does not explicitly compare communication cost. However, the fast and stable convergence of FedL2T implicitly reduces the number of communication rounds. Future work will explore support for heterogeneous models, similarity-based peer selection, low-cost communication extensions, and broader biomedical applications.

## References


[1] J. Werner, B. Kohli, P. P. Bernardo, C. Gerum, and O. Bringmann, "Energy-Efficient Seizure Detection Suitable for Low-Power Applications," in *2024 International Joint Conference on Neural Networks (IJCNN)*, Jun. 2024, pp. 1–8.

[2] G. Segal, N. Keidar, M. Herskovitz, and Y. Yaniv, "Personalized preictal EEG pattern characterization: Do timing and localization matter?" *Frontiers in Neuroscience*, vol. 19, May 2025.

[3] Y. Li, Y. Liu, Y.-Z. Guo, X.-F. Liao, B. Hu, and T. Yu, "Spatio-Temporal-Spectral Hierarchical Graph Convolutional Network With Semisupervised Active Learning for Patient-Specific Seizure Prediction," *IEEE Transactions on Cybernetics*, vol. 52, no. 11, pp. 12 189–12 204, Nov. 2022.

[4] J. P. Albrecht, "How the GDPR will change the world," *European Data Protection Law Review*, vol. 2, no. 3, 2016.

[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 54, 2017, pp. 1273–1282.

[6] J. Zhang, Y. Liu, Y. Hua, H. Wang, T. Song, Z. Xue, R. Ma, and J. Cao, "Pfllib: A beginner-friendly and comprehensive personalized federated learning library and benchmark," *Journal of Machine Learning Research*, vol. 26, no. 50, pp. 1–10, 2025.

[7] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and Robust Federated Learning Through Personalization," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Jul. 2021, pp. 6357–6368.

[8] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated Learning on Non-IID Features via Local Batch Normalization," in *International Conference on Learning Representations (ICLR)*, Oct. 2020.

[9] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting Shared Representations for Personalized Federated Learning," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Jul. 2021, pp. 2089–2099.

[10] A. Mora, I. Tenison, P. Bellavista, and I. Rish, "Knowledge Distillation in Federated Learning: A Practical Guide," in *Proceedings of the 33th International Joint Conference on Artificial Intelligence (IJCAI)*, Aug. 2024.

[11] T. Shen, J. Zhang, X. Jia, F. Zhang, Z. Lv, K. Kuang, C. Wu, and F. Wu, "Federated mutual learning: A collaborative machine learning method for heterogeneous data, models, and objectives," *Frontiers of Information Technology & Electronic Engineering*, vol. 24, no. 10, pp. 1390–1402, Oct. 2023.

[12] K. Sastry, D. Goldberg, and G. Kendall, "Genetic Algorithms," in *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*, 2005, pp. 97–125.

[13] S. Baghersalimi, T. Teijeiro, D. Atienza, and A. Aminifar, "Personalized Real-Time Federated Learning for Epileptic Seizure Detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 2, pp. 898–909, Feb. 2022.

[14] R. Saemaldahr and M. Ilyas, "Patient-Specific Preictal Pattern-Aware Epileptic Seizure Prediction with Federated Learning," *Sensors*, vol. 23, no. 14, p. 6578, Jan. 2023.

[15] W. Ding, M. Abdel-Basset, H. Hawash, S. Abdel-Razek, and C. Liu, "Fed-ESD: Federated learning for efficient epileptic seizure detection in the fog-assisted internet of medical things," *Information Sciences*, vol. 630, pp. 403–419, Jun. 2023.

[16] S. Baghersalimi, T. Teijeiro, A. Aminifar, and D. Atienza, "Decentralized Federated Learning for Epileptic Seizures Detection in Low-Power Wearable Systems," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 6392–6407, May 2024.

[17] Y. Ding, W. Zhao, and K. Huang, "FedMWAD: Module-wise weighted aggregation federated learning combined with Ditto for patient-independent seizure prediction," *Information Processing & Management*, vol. 62, no. 6, p. 104253, Nov. 2025.

[18] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, and Y. Bengio, "FitNets: Hints for Thin Deep Nets," in *International Conference on Learning Representations (ICLR)*, 2015.

[19] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 5776–5788.

[20] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, "Communication-efficient federated learning via knowledge distillation," *Nature Communications*, vol. 13, no. 1, p. 2032, Apr. 2022.

[21] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep Mutual Learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018, pp. 4320–4328.

[22] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Harvard-MIT Health Sciences and Technology, Massachusetts Institute of Technology, 2009.

[23] L. Yi, H. Yu, C. Ren, G. Wang, X. Liu, and X. Li, "Federated model heterogeneous matryoshka representation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.