Personalized Meta-Federated Learning for IoT-Enabled Health Monitoring

Zhenge Jia, *Member, IEEE*, Tianren Zhou, *Member, IEEE*, Zheyu Yan, *Member, IEEE*, Jingtong Hu, *Senior Member, IEEE*, and Yiyu Shi, *Senior Member, IEEE*

***Abstract*—Federated learning (FL) has been widely adopted in IoT-enabled health monitoring on biosignals thanks to its advantages in data privacy preservation. However, the global model trained from FL generally performs unevenly across subjects since biosignal data is inherent with complex temporal dynamics. The morphological characteristics of biosignals with the same label can vary significantly among different subjects (i.e., intersubject variability) while biosignals with varied temporal patterns can be collected on the same subject (i.e., intrasubject variability). To address the challenges, we present the personalized meta-federated learning (PMFed) framework for personalized IoT-enabled health monitoring. Specifically, in the FL stage, a novel momentum-based model aggregating strategy is introduced to aggregate clients' models based on domain similarity in the meta-FL paradigm to obtain a well-generalized global model while speeding up the convergence. In the model personalizing stage, an adaptive model personalization mechanism is devised to adaptively tailor the global model based on the subject-specific biosignal features while preserving the learned cross-subject representations. We develop an IoT-enabled computing framework to evaluate the effectiveness of PMFed over three real-world health monitoring tasks. Experimental results show that the PMFed excels at detection performances in terms of F1 and accuracy by up to 9.4% and 8.7%, and reduces training overhead and throughput by up to 56.3% and 63.4% when compared with the SOTA FL algorithms.**

***Index Terms*—Embedded system, federated learning (FL), personal.**

I. Introduction

IOT-ENABLED health monitoring, the health monitoring system with the capability of IoT [1], has been considered to be a promising solution to out-of-hospital healthcare applications, such as arrhythmias detection [2] and activities monitoring [3]. Recently, deep learning (DL) has gained growing attention in IoT-enabled health monitoring thanks to its advantages in feature extraction automation. DL-based methods have been applied in various applications, including arrhythmias detection [4], human activity recognition [5], etc. In the healthcare field, however, the data privacy constraint restricts data from being aggregated online because uploading personal health data to a cloud server is prohibitive in certain application scenarios [6]. To address this issue, federated learning (FL) is proposed and its key idea is to collaboratively train a global model without uploading personal health data by aggregating clients' models with a weighted average on the server. While FL has shown great promise in healthcare applications, inter- and intra-subject variability of biosignals makes it challenging to ensure the optimal detection performances across health monitor recipients [7].

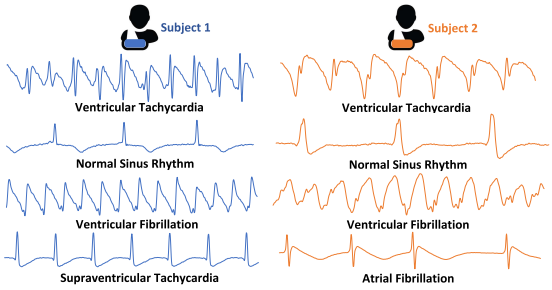


Fig. 1. Inter- and intra-subject variability reflected on IEGMs of two subjects.

Intersubject variability caused by individual differences can lead to slight or significant variation over biosignals (with the same label) in terms of morphological characteristics among subjects [7]. As shown in Fig. 1, the intracardiac electrograms (IEGMs) segments on the same row are with the same arrhythmia label [e.g., ventricular tachycardia (VT)] but from two subjects. The morphological characteristics of these segments on the same row demonstrate intersubject variable patterns. Another variability, which is intrasubject variability, can lead to a nonstationarity of biosignals on the same subject. As shown by IEGMs segments on the same column of Fig. 1, segments with different types of arrhythmias can be collected from one subject. Different types of arrhythmias present significantly different morphological characteristics on the same subject. Therefore, the biosignal data patterns of each individual subject are highly personalized and heterogeneous.

To tackle this heterogeneity problem in FL, FedProx [8] is proposed to extend FedAvg [9] by adding a proximal point update for local optimization. Local fine-tuning using local data is another key solution to adapt the global model to the individual [10], [11]. To further improve the generalization of the global model, meta-learning is embedded into FL to obtain a well-generalized global model [12], [13]. To avoid

Manuscript received 30 August 2023; revised 4 February 2024; accepted 7 April 2024. Date of publication 15 April 2024; date of current version 20 September 2024. This work was supported in part by National Science Foundation under Grant CNS-2122320, Grant CCF-2324937, and Grant CNS-2235364; and in part by NIH under Grant R01EB033387. This article was recommended by Associate Editor T. Mitra. *(Corresponding authors: Zhenge Jia; Yiyu Shi.)*

Zhenge Jia, Zheyu Yan, and Yiyu Shi are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: zjia2@nd.edu; zyan2@nd.edu; yshi4@nd.edu).

Tianren Zhou is with the School of Computer Science and Technology, Shandong University, Qingdao 266200, China (e-mail: trzhou@mail.sdu.edu.cn).

Jingtong Hu is with the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15261 USA (e-mail: jthu@pitt.edu).

Digital Object Identifier 10.1109/TCAD.2024.3388908

domain shifting in local model personalization via fine-tuning, FedBN [14], and SiloBN [15] are proposed to keep local batchnorm (BN) statistics from aggregation.

However, almost all previous methods are proposed only for image data. The morphological characteristics heterogeneity problem is more severe for biosignal data, which gives rise to the challenges as follows: 1) the global model could be easily skewed by the subjects with unique morphological characteristics of biosignal data due to intersubject variability. The global model with poor generalization could result in a highly biased classification across testing subjects; 2) fine-tuning the global model with local data could lead to overfitting or domain shifting, and therefore result in a poor adaptation to the individual. It is because the testing subject's data collected for local fine-tuning can be very different from the future sensed data due to intrasubject variability; and 3) existing FL methods generally focus on boosting performances for clients (subjects) participating in federated training, without considering the model generalization where the deep model would be applied on the unseen testing subjects.

To address the aforementioned challenges, in this article, we propose personalized meta-federated learning (PMFed) framework for IoT-enabled health monitoring. PMFed is conducted in a manner of 1) meta-federated training to obtain a well-generalized global model and 2) subject-specific model personalization to properly personalize the global model to fit the unseen testing subject. Specifically, in the federated training stage, the clients would apply the meta-FL paradigm with the client's data to train the local model. Once the clients' models are uploaded on the server, a novel momentum-based model aggregating strategy based on clients' domain similarity is proposed. In this way, the global model could unbiasedly learn the representations of the biosignal data across subjects and finish with well-generalized model parameters while facilitating faster model convergence. In the subject-specific model personalization stage, an adaptive model personalization mechanism is proposed to adaptively select personalized nontrainable parameters and learning rate of fine-tuning based on domain similarity. The mechanism could prevent the well-generalized model from overfitting in local fine-tuning for unseen testing subjects. We further implement an IoT computing platform to evaluate the detection and practical performances of PMFed. Experimental results show that PMFed outperforms SOTA FL algorithms in terms of detection and practical performances over three real-world health monitoring tasks. The main contributions of this article are as follows:

1) We propose PMFed framework that conducts subject-specific health monitoring in IoT.
2) We introduce a novel momentum-based model aggregation strategy based on training subjects' biosignal data domain similarity to obtain a well-generalized global model while facilitating faster model convergence in federated training.
3) We devise an adaptive model personalization mechanism that effectively personalizes the global model for unseen testing subjects by adaptively setting parameters.
4) Experimental results demonstrate that PMFed outperforms SOTA FL algorithms in terms of F1 and accuracy by up to 9.4% and 8.7%, respectively. PMFed also reduces training overhead and communication throughput by up to 56.3% and 63.4% when compared to existing FL algorithms.

II. Background

In this section, we introduce the background information about IoT-enabled health monitoring, DL-based health monitoring, and FL designs for health monitoring, respectively.

A. IoT-Enabled Health Monitoring

IoT-enabled health monitoring has gained growing attention in recent years with the rapid development of biomedical sensors and wireless transmitters. When compared with conventional monitoring, connectivity is the main advantage in IoT-enabled health monitoring, especially for out-of-hospital health conditions surveillance and treatment [16]. There are various IoT-enabled health monitoring applications. For example, implantable cardioverter-defibrillators (ICDs) are manufactured to provide in-time defibrillation on the detected ventricular arrhythmias (VAs) [17]. The remote monitoring function in modern ICDs is established with the integration of IoT capability. Atrial fibrillation (AF) detection on monitors, such as an insertable cardiac monitor (ICM) [2] and cardiac patch [18], have been greatly integrated with the capability of IoT. While providing the AF detection function on the device, those IoT-enabled monitors could upload the sensed rhythm data to the server for further diagnosis and provide professional medical recommendations by doctors. Furthermore, IoT-enabled health monitors are widely adopted in the general health monitoring field. Smartwatches (e.g., Fitbit watch [19] and Apple watch [20]) provide a wide range of health monitoring functions.

B. Deep Learning-Based Health Monitoring

In current computer-aided methods design, essential features and detection criteria are first derived from clinical trials and then transformed into a program that is runnable on the IoT monitors [17]. Considerable expertise is required to optimize the extracted feature set, detection criteria, and programmable parameters. DL provides an alternative solution to address the shortage of expertise. DL could automatically learn to extract essential features and perform classification via self-training. These distinctive advantages are driving the utilization of DL in health monitoring on biosignals. DL-based methods have achieved outstanding performance in a variety of health monitoring tasks. For example, DL-based arrhythmia detection on 12-lead electrocardiogram (ECG) has achieved cardiologist-level performance in terms of accuracy on twelve-class arrhythmia classification [21]. An automated detection system for Parkinson's disease (PD) is proposed to detect PD using a convolutional neural network (CNN) based on sensed electroencephalogram (EEG) signal [22].

C. Federated Learning in Health Monitoring

FL enables user-end devices to collaborate with a server to train a global model without data sharing. Specifically, the training paradigm of FL is to aggregate local model updates without accessing the personal data on the user end. The classic FL algorithm, FedAvg [9], distributes the global model to all clients at the beginning of each training round. Once the

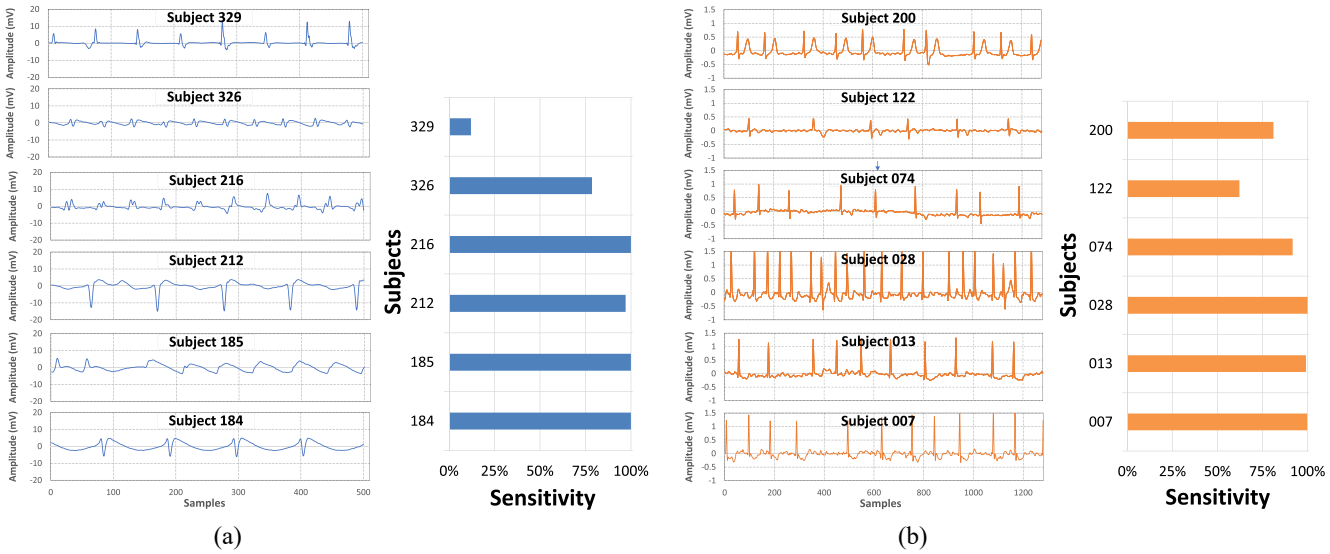


Fig. 2. Biosignal segments with the same label and the corresponding sensitivity performances over six subjects. (a) IEGMs and sensitivity. (b) ECG and sensitivity.

server aggregates the updated neural networks updated with the local data of each client, it averages the parameters of all models with weights to obtain a new global model for the next training round. The global model would be finally distributed to the user for accurate detection.

To address the privacy concerns of the conventional DL training paradigm (i.e., data aggregation in a server), FL has been actively explored in health monitoring. Chen et al. [23] devised an FL scheme for wearable health monitoring where a group of smartphones collaborates to train a shared CNN model with a cloud server for human activity recognition. Warnat-Herresthal et al. [24] proposed a FL paradigm that units edge computing and blockchain techniques to conduct accurate disease classifications while maintaining high confidentiality. Additionally, Tan et al. [25] proposed a tree-based FL approach for personalized treatment with electronic health records from different hospitals.

III. Motivations

In this section, we demonstrate the intrinsic characteristics of inter- and intra-subject variability, and present preliminary experimental results to illustrate their effects on conventional FL algorithms. The biosignal data used in the experiments is with the type of IEGMs and ECG. The detailed experimental setup is introduced in Section V-A.

A. Intersubject Variability and the Effects on FL

It is challenging to accurately detect events or diseases on biosignal due to its complex temporal dynamics. The complex patterns of biosignal are generally caused by intersubject variability. Such variability, reflected on biosignal, is the varied temporal patterns of the signal with the same label over different individuals. As shown in Fig. 2(a), there are IEGM segments with the same arrhythmia label (i.e., VT) retrieved from six different subjects from Ann Arbor Electrogram Libraries (AAEL) [26]. In the same manner, Fig. 2(b) presents six ECG segments with the same arrhythmia label (i.e., AF) from long-term AF dataset (LTAFDB) [27], [28]. The detailed description of the dataset is introduced in Section V-A.

As demonstrated by the figures, the biosignal segments with the same type of arrhythmia or action can show slightly or even significantly different temporal patterns (i.e., morphological characteristics) of various subjects. For example, in Fig. 2(b), subject 122 experiences a much lower QRS amplitude and longer QRS interval when compared with the ECG segment of other subjects. Meanwhile, the other subjects' ECG segments demonstrate a slight variation in morphological characteristics in terms of QRS-peak and QRS intervals. The same phenomenon also appears in IEGM segments shown in Fig. 2(a). As a result, there is a group of subjects with major morphological characteristics while there is also a portion of subjects with unique characteristics. It indicates that the biosignal patterns of subjects are not always uniformly distributed but naturally personalized with feature distribution skew.

In the FL paradigm, intersubject variability is even more severe since the data cannot be aggregated and each client is treated equally in the learning process. A global model in a conventional FL algorithm may not effectively learn cross-subject representations and could only adapt well to the subject with major features. Fig. 2 demonstrates the detection performance of CNNs trained with FedAvg [9] for VA and AF detection. The performances are reported in terms of sensitivity on each selected subject's biosignal segments with VA or AF labels. As shown in Fig. 2, the global model performs poorly on some individuals (e.g., subject 329 in VA detection and subject 122 in AF detection) due to the unique morphological characteristics caused by intersubject variability.

Though there are methods, such as FedProx [8] and PerFedAvg [12], to obtain a better global model, these methods are proposed only for image data and cannot fully address the negative effect of intersubject variability to the global model. Therefore, to perform subject-specific detection in health monitoring on biosignals, it is demanding to devise a

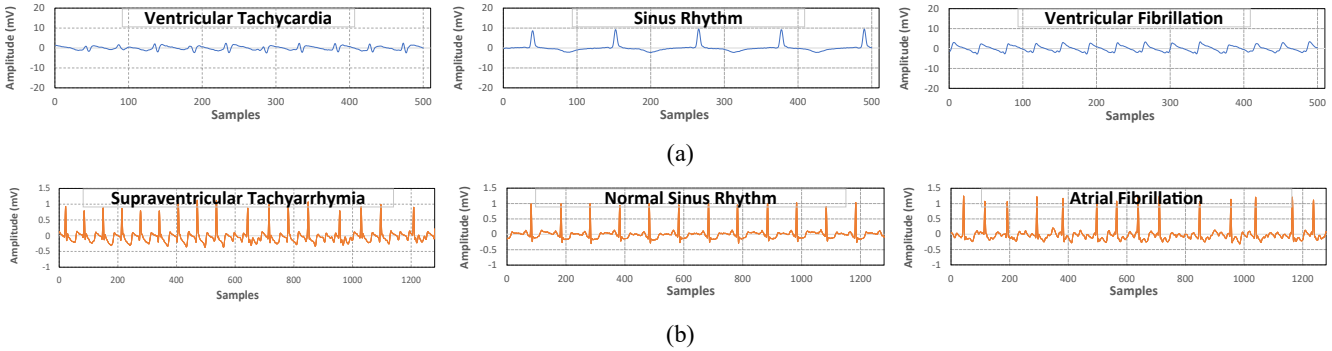


Fig. 3. Biosignal segments with different labels over the same subject. (a) IEGMs segments with different labels of subject 326. (b) ECG segments with different labels of subject 07.

FL algorithm that effectively learns cross-subject data representations by considering the subjects with either major or unique biosignals' morphological characteristics.

B. Intrasubject Variability and the Effects on FL

Intrasubject variability, reflected on biosignals, is the varied temporal patterns of the time-series data over the same subject. As shown in Fig. 3, there are three segments with three different types of arrhythmia of IEGMs and ECG, respectively. Fig. 3 indicates that different types of arrhythmia lead to significantly different biosignals' morphological characteristics on the same subject. Additionally, it is hard for the subject to experience each specific type of arrhythmia, and gather the corresponding signal segments. In health monitoring tasks, multiple types of arrhythmia may be concluded as one label for classification purposes. As a result, intrasubject variability would lead to label and quantity distribution skew in model personalization.

To address the performance degradation caused by intrasubject variability, a practical way in FL is to perform local fine-tuning on the global model with the targeting subject's data [10], [11]. However, intrasubject variability hinders performance improvement through the simple fine-tuning strategy. Fig. 4 presents the detection performances of the fine-tuned models in terms of accuracy over six subjects in VA detection and AF detection, respectively. The data utilized in fine-tuning is a small group of the targeting subject's biosignal segments that are extracted randomly (The data extraction process is introduced in Section V-A1). These subjects are defined as unseen testing subjects, who choose not to participate in federated training to avoid data breaches. As shown in Fig. 4, most of the models fine-tuned with the targeting subjects' biosignal data gain a performance improvement in terms of accuracy. However, the models of subject 326 in Fig. 4(a) and subject 074 in Fig. 4(b) experience performance degradation after being fine-tuned with the subject-specific biosignal data.

The performances shown in Fig. 4 indicate that intrasubject variability poses a challenge to the model personalization in health monitoring. Though there are methods, such as FedBN [14] and SiloBN [15], proposed to utilize local BN statistics during fine-tuning, these methods cannot fully address the impact of intrasubject variability coming from unseen subjects. It is therefore expected to propose a model personalization method that properly personalizes the model

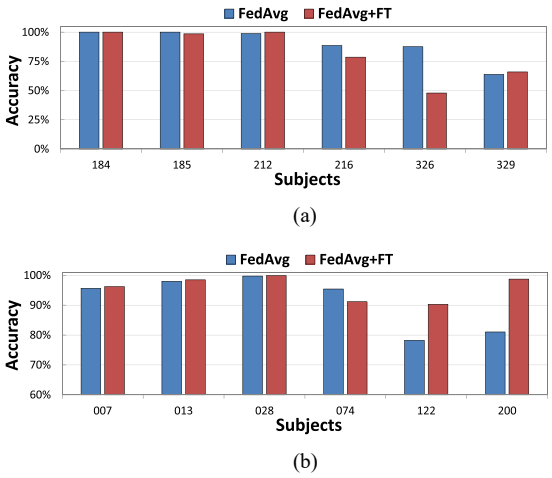


Fig. 4. Individual detection accuracy over FedAvg and FedAvg with simple fine-tuning strategy. (a) Individual accuracy in VA detection. (b) Individual accuracy in AF detection.

with limited but skewed labeled biosignal data of unseen subjects.

IV. Personalized Meta-Federated Learning Framework

In this section, we first present the system overview of the proposed PMFed framework and then introduce two essential processes (i.e., federated training and local personalization). The methodologies, including momentum-based model aggregation strategy and adaptive model personalization mechanism, are presented along with the introduction of these processes.

A. System Overview

Fig. 5 shows the system overview of the PMFed framework. We first develop a computing framework consisting of a server and a line of IoT health monitors as clients. There are two essential processes of PMFed: 1) meta-federated training conducted on participant clients and the server to obtain a well-generalized global model and 2) subject-specific model personalization conducted on the local client to generate a personalized model.

In meta-federated training, as shown in Fig. 5, each client who participated in training would perform meta-learning

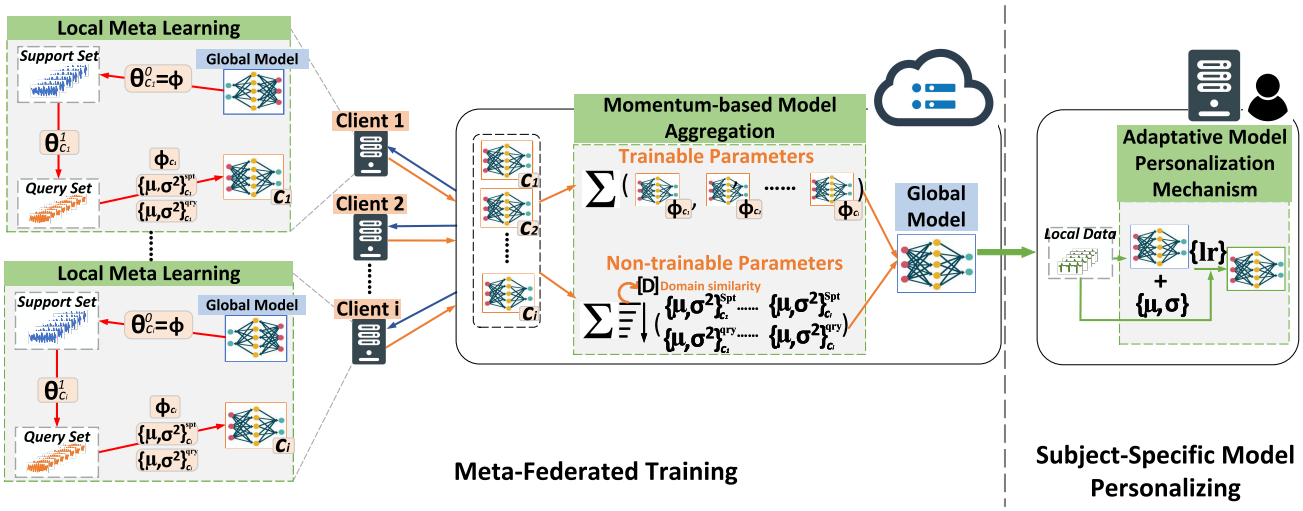


Fig. 5. System overview of PMFed.

on the received model and upload it to the server. Once the uploaded models are received, the server would conduct momentum-based model aggregation based on the domain similarity between clients and distribute the updated global model to clients. The process would be executed iteratively until the model parameters converge. Once the meta-federated training is completed, we would obtain a global model that unbiasedly learns cross-subject representations.

Subject-specific model personalization, as shown in Fig. 5, is conducted on the IoT monitor of the testing (unseen) subject. The process aims to personalize the global model downloaded from the server by fine-tuning the model with a limited amount of local biosignal data from the testing subject. The proposed adaptive model personalization mechanism would adaptively set the hyperparameters of the fine-tuning process and model parameters.

B. Meta-Federated Training

To improve model generalization, we propose local meta-learning and momentum-based model aggregation strategy to enable the global model to unbiasedly learn cross-subject representations.

1) Local Meta-Learning on Client: We first introduce the corresponding definitions and notations. The meta-federated training starts with the initial meta-model ϕ initialized on the server. The purpose of the model is to classify the biosignal segments into different classes *cls*. In each round, the participated clients (denoted as $\mathcal{C}$) would start the FL process with the server.

As shown in Fig. 5, the meta-model (global model) ϕ would be distributed to each client in $\mathcal{C}$. Once the meta-model ϕ is received by the client $c_i \in \mathcal{C}$, the client starts to formulate a *taskset* T_{c_i} for meta-learning. The taskset T_{c_i} contains a *support set* $\tau^s_{c_i}$ and a *query set* $\tau^q_{c_i}$. The support set $\tau^s_{c_i}$ contains p number of data points of each targeting class while the support set $\tau^q_{c_i}$ contains q number of data points of each targeting class. Therefore, the support set $\tau^s_{c_i}$ is

$$\tau^s_{c_i} = \{(x_j, y_j)\}_{j \in M^{\text{spt}}_{c_i}} \text{ for } j = 1, \ldots, cls \cdot p \tag{1}$$

where (x_j, y_j) is the data-label pairs and $M^{\text{spt}}_{c_i}$ is the set containing the indices of data-label pairs of c_i for the support set. The query set $\tau^q_{c_i}$ is defined as follows:

$$\tau^q_{c_i} = \{(x'_j, y'_j)\}_{j' \in M^{\text{qry}}_{c_i}} \text{ for } j = 1, \ldots, cls \cdot q \tag{2}$$

where (x'_j, y'_j) is the data-label pairs, and $M^{\text{qry}}_{c_i}$ is the set containing the indices of c_i in $M^{\text{qry}}_{c_i}$. Note that the indices in $M^{\text{spt}}_{c_i}$ and $M^{\text{qry}}_{c_i}$ are mutually exclusive.

With the preparation of $\tau^s_{c_i}$ and $\tau^q_{c_i}$, the local meta-learning process is conducted on the client c_i. The first step is *inner update* [29], where the received meta-model ϕ is updated over the support set $\tau^s_{c_i}$. The loss of model θ_{c_i} (i.e., $\theta_{c_i} \leftarrow \phi$ at the initial step as shown in Fig. 5) on $\tau^s_{c_i}$ is calculated as follows:

$$\mathcal{L}_{\tau^s_{c_i}}(\theta_{c_i}) = \frac{1}{|\tau^s_{c_i}|} \sum_{(x,y) \in \tau^s_{c_i}} \mathcal{L}\left(f_{\theta_{c_i}}(x), y\right) \tag{3}$$

where $f_{\theta_{c_i}}(x)$ represents the model inference conducted on the input x with the model parameters θ_{c_i}. The loss function $\mathcal{L}$ can be cross-entropy loss for the classification task and other types of loss functions targeting different tasks. Next, the meta-model is updated by calculating the gradient on the support set for one step

$$\theta_{c_i} = \theta_{c_i} - \alpha \nabla_{\theta_{c_i}} \mathcal{L}_{\tau^s_{c_i}}(\theta_{c_i}) \tag{4}$$

where α is the inner-update learning rate. The distribution statistics $\{\boldsymbol{\mu}, \boldsymbol{\sigma}\}^{\text{spt}}_{c_i}$ of each BN layer over the support set is recorded as well.

The second step is *outer update* [29], where the support set-specific model θ_{c_i} is evaluated on the query set $\tau^q_{c_i}$. The purpose of this step is to evaluate the generalization and training ability of θ_{c_i} and provide the updating direction for the meta-model. The loss of $\theta^m_{c_i}$ over the query set $\tau^q_{c_i}$ is calculated as follows:

$$\mathcal{L}_{\tau^q_{c_i}}(\theta_{c_i}) = \frac{1}{|\tau^q_{c_i}|} \sum_{(x,y) \in \tau^q_{c_i}} \mathcal{L}\left(f_{\theta_{c_i}}(x), y\right). \tag{5}$$

Next, the gradient of the loss in (5) over the meta-model ϕ is defined as follows:

$$\nabla_{\phi} \mathcal{L}_{\tau^q_{c_i}}(\theta_{c_i}) = \left(I - \alpha \nabla^2_{\phi} \mathcal{L}_{\tau^q_{c_i}}(\phi)\right) \nabla_{\theta_{c_i}} \mathcal{L}_{\tau^q_{c_i}}(\theta_{c_i}) \tag{6}$$

where the second-derivative is based on the acquisition of θ_{c_i} derived from the meta-model ϕ using the support set $\tau^s_{c_i}$ [29]. Since the second-derivative part $I - \alpha\nabla^2_\phi \mathcal{L}_{\tau^q_{c_i}}(\phi)$ is generally small in terms of value but with high-computational complexity, the gradient in (6) can be further calculated with first-order approximation [29] as follows:

$$\nabla_\phi \mathcal{L}_{\tau^q_{c_i}}(\theta_{c_i}) \approx \nabla_{\theta_{c_i}} \mathcal{L}_{\tau^q_{c_i}}(\theta_{c_i}). \tag{7}$$

Based on the approximated gradient, the meta-model on the client c_i is updated as follows:

$$\phi_{c_i} = \phi - \beta\nabla_{\theta_{c_i}} \mathcal{L}_{\tau^q_{c_i}}(\theta_{c_i}) \tag{8}$$

where β is the outer-update learning rate. The distribution statistics $\{\boldsymbol{\mu}, \boldsymbol{\sigma}\}^{\text{qry}}_{c_i}$ on query set is calculated as well.

The intuition of local meta-learning is to enable the local model to learn to generalize to subject-specific data. Local meta-learning involves a two-level training process based on the support set and query set. The support set and query set serve two distinct roles in local meta-learning [29]. The main benefit of having a support set is that it allows the model to adapt quickly to subject-specific data with very few examples. This is crucial in situations where we do not have a large amount of data available for each subject, a common scenario in real-world healthcare applications. The query set, on the other hand, is used to evaluate how well the model has adapted to the subject-specific data based on the fine-tuned model on the support set. By having a separate query set of the same subject, we can obtain an unbiased estimate of the model's performance on the subject with intrasubject variability, as it has not seen these examples during the fine-tuning phase on the support set.

2) Momentum-Based Model Aggregation on Server: Once ϕ_{c_i} together with distribution statistics ($\{\boldsymbol{\mu}, \boldsymbol{\sigma}\}^{\text{spt}}_{c_i}$, $\{\boldsymbol{\mu}, \boldsymbol{\sigma}\}^{\text{qry}}_{c_i}$) of all $c_i \in \mathcal{C}$ are received by the server, the model aggregation is conducted to obtain the new meta-model ϕ for the next training iteration. Fig. 2 shows that simply averaging the parameters of the uploaded models could result in a global model with highly biased classification due to intersubject variability. Existing optimization methods [8], [12], [14] cannot generate a well-generalized global model (shown in Section V-B1) since they are not specifically designed for unseen clients.

To address the issue, we conduct a series of preliminary experiments by setting various rules on modifying or freezing different components of the neural network during model aggregation. Based on the performance of the aggregated global model under various strategies, we empirically find that the distribution statistics of BN play an essential role in model generalization by calculating the detection performances of individual subjects. The performances show that the quality of the global model can be easily affected by the subjects' biosignal data with unique temporal patterns. In this article, the distribution statistics (i.e., mean and variance) of BN are defined as nontrainable BN parameters. Therefore, instead of being averaged at server [8] or preserved at local [14], these nontrainable BN parameters should be carefully aggregated in each iteration to reduce the negative impact on the generalization of the global model.

We first propose a novel similarity measurement to calculate the domain similarity between two clients based on cosine similarity. The mean and variance of each BN layer from every client's query set are utilized in the similarity calculation as follows:

$$S_{\cos}(c_i, c_j) = \sum_{l \in L_{bn}} \frac{\boldsymbol{\mu}_{c_i,l} \cdot \boldsymbol{\mu}_{c_j,l}}{\|\boldsymbol{\mu}_{c_i,l}\| \|\boldsymbol{\mu}_{c_j,l}\|} + \frac{\boldsymbol{\sigma}_{c_i,l} \cdot \boldsymbol{\sigma}_{c_j,l}}{\|\boldsymbol{\sigma}_{c_i,l}\| \|\boldsymbol{\sigma}_{c_j,l}\|} \tag{9}$$

where L_{bn} is the set of BN used in all layers. $\boldsymbol{\mu}_{c_i,l}$ and $\boldsymbol{\sigma}_{c_j,l}$ represent the vectors of mean and variance of all channels on the certain layer's BN of $\{\boldsymbol{\mu}, \boldsymbol{\sigma}\}^{\text{qry}}_{c_i}$, respectively. We then construct the similarity matrix D with the size of $|\mathcal{C}| \times |\mathcal{C}|$ where D_{ij} represents the domain similarity between the client c_i and c_j. The clustering algorithm DBSCAN [30] is then invoked to find the outliers O (i.e., the subjects with unique biosignal morphological characteristics) and the main cluster X of clients based on D.

In contrast to the conventional model aggregation strategy by averaging each uploaded model's parameters, for nontrainable BN parameters (i.e., mean and variance of BN), we first filter out the outlier clients in O. For the clients in X, we find the central point of their mean and variance of BN. The central point is defined as follows:

$$\{\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\sigma}}\} = \sum_{i \in X} \frac{1}{|X|} \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}^{\text{qry}}_{c_i}. \tag{10}$$

Next, we sort the clients in X based on the distance between each client's distribution statistics and the central point $\{\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\sigma}}\}$ in ascending order, and obtain the sorted list L. The *momentum-based model aggregation strategy* on nontrainable BN parameters is conducted for each client $c_i \in L$ in order as follows:

$$\begin{aligned} \boldsymbol{\mu} &= (1-\gamma) \cdot \left((1-\gamma)\boldsymbol{\mu} + \gamma \boldsymbol{\mu}^{\text{spt}}_{c_i}\right) + \gamma \cdot \boldsymbol{\mu}^{\text{qry}}_{c_i} \\ \boldsymbol{\sigma} &= (1-\gamma) \cdot \left((1-\gamma)\boldsymbol{\sigma} + \gamma \boldsymbol{\sigma}^{\text{spt}}_{c_i}\right) + \gamma \cdot \boldsymbol{\sigma}^{\text{qry}}_{c_i} \end{aligned} \tag{11}$$

where γ is the momentum parameter and the later client's distribution statistics weigh higher. The trainable parameters of the model of all clients in $\mathcal{C}$ would be aggregated by averaging the corresponding parameter as follows:

$$\phi_{\text{tr}} = \sum_{i \in \mathcal{C}} \frac{1}{|\mathcal{C}|} \phi_{c_i}. \tag{12}$$

In the end, the meta-model ϕ would be obtained on the server by integrating the trainable and nontrainable BN parameters.

Algorithm 1 illustrates the process of the proposed meta-federated training. The proposed meta-federated training starts at the server, where all clients are selected at the beginning of each round (line 4). Once the meta-model ϕ is distributed and received by the client, the client would perform local meta-learning in parallel. For each client, it first formulates the support set and the query set with local data (line 15), and then conducts the cross-subject learning process. The inner update is conducted on the support set (line 16) and the outer update is conducted on the query set (line 17). The generated ϕ_{c_i} along with the mean and variance of BN statistics would be finally uploaded to the server for model aggregation (line 18). Once the models of all clients are collected, the server starts to construct the affinity matrix with the proposed similarity measurement defined in (9) over all models of selected clients in C_M (lines 6–8). Next, the set X of clients in each cluster and the set O of outlier clients are obtained by applying the DBSCAN algorithm on the affinity matrix (line 9). The mean and variance of BN with momentum-based aggregation would

Algorithm 1: Meta-Federated Training

Given ϕ: deep model initial parameters.
Given R: the number of total rounds.
Given $\mathcal{C}$: the set of clients participated in training.
1 **RunServer**($\phi, R, \mathcal{C}$):
2 **for** $r = 1, 2, \ldots, R$ **do**
3 **for** *each client* $c_i \in \mathcal{C}$ ***in parallel*** **do**
4 $\phi_{c_i} \leftarrow$ **RunClient**(ϕ)
5 **end**
6 **for** *each* $(c_i, c_j) \in \mathcal{C} \times \mathcal{C}$ **do**
7 $D_{ij} \leftarrow S_{\cos}(c_i, c_j)$ by Eqn. (9)
8 **end**
9 $X, O \leftarrow$ DBSCAN(D)
10 $\boldsymbol{\mu}, \boldsymbol{\sigma} \leftarrow$ by Eqn. (10) & (11)
11 $\phi_{\text{tr}} \leftarrow \sum_{i \in \mathcal{C}} \frac{1}{|\mathcal{C}|} \phi_{c_i}$
12 $\phi \leftarrow \{\phi_{\text{tr}}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$ *#integrating trainable and non-trainable BN parameters*
13 **end**
14 **RunClient**(ϕ):
15 Formulate support set $\tau_{c_i}^s$ and query set $\tau_{c_i}^q$, $\theta_{c_i} \leftarrow \phi$
16 $\theta_{c_i} \leftarrow$ by Eqn. (3) & (4)
17 $\phi_{c_i} \leftarrow$ by Eqn. (5) & (8)
18 Upload ϕ_{c_i}, $\{\boldsymbol{\mu}, \boldsymbol{\sigma}\}_{c_i}^{\text{spt}}$, $\{\boldsymbol{\mu}, \boldsymbol{\sigma}\}_{c_i}^{\text{qry}}$

then be obtained as nontrainable BN parameters (line 10). The trainable model parameters would be aggregated by averaging the parameter (line 11). In the end, the new meta-model ϕ could be obtained by integrating trainable and nontrainable BN parameters (line 12).

The intuition of the strategy is to reduce the negative impact caused by outlier training clients on the global model generalization. Our observations show that good nontrainable BN parameters could greatly affect the global model's generalization. The proposed strategy helps to alleviate the domain shifting by putting less weight on the training clients classified as outliers during model aggregation on nontrainable BN parameters.

C. Subject-Specific Model Personalization

The next essential step is to personalize the meta-model ϕ to adapt to the testing subject c's biosignal data domain and obtain the subject-specific detection model. Fine-tuning the global model with local data is an effective and practical way to perform model personalization [10]. In our scenarios, to perform model personalization, the testing subject c is required to formulate the personalizing set τ_c^p, which contains a limited number of biosignal data segments with all available targeting classes. The simple fine-tuning process starts with calculating the loss on the data-label pairs in τ_c^p as follows:

$$\mathcal{L}_{\tau_c^p}\left(\phi_c^{p(m-1)}\right) = \frac{1}{|\tau_c^p|} \sum_{(x,y) \in \tau_c^p} \mathcal{L}\left(f_{\phi_c^{p(m-1)}}(x), y\right) \tag{13}$$

where m indicates the current fine-tuning step and $\phi_c^{p(0)}$ is the received global model ϕ^*. The loss function can be cross-entropy loss on the data-label pairs in τ_c^p for the classification task. One or multiple steps of update with gradient can be conducted as follows:

$$\phi_c^{p(m)} = \phi_c^{p(m-1)} - \alpha \nabla_{\phi_c^{p(m-1)}} \mathcal{L}_{\tau_c^p}\left(\phi_c^{p(m-1)}\right) \tag{14}$$

where α is the learning rate in the fine-tuning process. However, as introduced in Section III, the intrasubject variability greatly hinders performance improvement of the model personalized by the local fine-tuning. To address the issues, we propose an *adaptive model personalization mechanism*. The core tenet of the proposed mechanism is to adaptively set distribution statistics of BN and the learning rate of local fine-tuning.

We first calculate the cosine similarity $S_{\cos}(\phi^*, \phi_c^p)$ between the distribution statistics of the meta-model and local model as follows:

$$S\left(\phi^*, \phi_c^p\right) = \sum_{l \in L_{bn}} \frac{\boldsymbol{\mu}_{\phi^*} \cdot \boldsymbol{\mu}_{\phi_c^{p(1)}}}{\|\boldsymbol{\mu}_{\phi^*}\| \|\boldsymbol{\mu}_{\phi_c^{p(1)}}\|} + \frac{\boldsymbol{\sigma}_{\phi^*} \cdot \boldsymbol{\sigma}_{\phi_c^{p(1)}}}{\|\boldsymbol{\sigma}_{\phi^*}\| \|\boldsymbol{\sigma}_{\phi_c^{p(1)}}\|} \tag{15}$$

where $\phi_c^{p(1)}$ is the local model obtained by conducting one-step local fine-tuning on τ_c^p. To set the distribution statistics of BN for the targeting client, we set the rule as follows:

$$S_{\cos}\left(\phi^*, \phi_c^p\right) < S_{\text{lower}} \tag{16}$$

where S_{lower} is the preset similarity hurdle. If the condition in (16) is stratified, the mean and variance of ϕ^* would be integrated into those of $\phi_c^{p(1)}$ with a portion ratio ϵ as follows:

$$\begin{aligned} \boldsymbol{\mu}_{\phi^*} &= (1-\epsilon) \cdot \boldsymbol{\mu}_{\phi^*} + \epsilon \cdot \boldsymbol{\mu}_{\phi_c^{p(1)}} \\ \boldsymbol{\sigma}_{\phi^*} &= (1-\epsilon) \cdot \boldsymbol{\sigma}_{\phi^*} + \epsilon \cdot \boldsymbol{\sigma}_{\phi_c^{p(1)}} \end{aligned} \tag{17}$$

where both mean and variance would be frozen during local fine-tuning. Otherwise, the mean and variance of the global model would be updated in a general manner during fine-tuning. To set the learning rate of local fine-tuning, we further calculate the first-step loss $\mathcal{L}_{\tau_c^p}(\phi^*)$ over τ_c^p. We then set the rule as follows:

$$\mathcal{L}_{\tau_c^p}\left(\phi^*\right) > L_{\text{upper}} \tag{18}$$

where L_{upper} is the upper bound of the loss. If the condition in (19) is stratified, the learning rate of the local fine-tuning would be modified as follows:

$$\beta = \alpha \cdot n \tag{19}$$

where β is the new learning rate and n is a preset value. Otherwise, the learning rate β would keep the same as the original learning rate α. With $\phi_c^{p(0)} \leftarrow \phi^*$, the local fine-tuning with the new BN statistics and learning rate can be conducted as follows:

$$\phi_c^{p(m)} = \phi_c^{p(m-1)} - \beta \nabla_{\phi_c^{p(m-1)}} \mathcal{L}_{\tau_c^p}\left(\phi_c^{p(m-1)}\right) \tag{20}$$

where m is the fine-tuning steps.

The intuition of the mechanism is to alleviate the negative impact of intrasubject variability on the performance of the model personalized by local fine-tuning. Different from the simple fine-tuning strategy, the two main factors (i.e., BN parameters affecting domain adaptation and learning rate affecting model personalization speed) are adaptively set based on the similarity and loss between the global and local model tailored to the subject-specific data.

V. Experiments

A. Experimental Setup

1) Dataset and Data Preparation: To evaluate the effectiveness of the proposed method, we utilize three datasets with various types of biosignal data as different health monitoring

applications. The datasets utilized in the experiments are shown as follows:

AAEL: The first application is VAs detection over the AAEL dataset [26], which is one of the largest IEGMs datasets utilized by implantable device manufacturers to evaluate their algorithms. The data preprocessing scheme is: 1) We utilize IEGM recordings from the RVA-Bi lead over 95 subjects and resample them to 250 Hz; 2) We divide the resampled recordings into episodes following the annotation on the time ticks. The episodes labeled with VT or ventricular fibrillation (VF) are defined as VA *episodes* while the episodes with other labels are defined as non-VA *episodes*; and 3) We segment each episode into 2-s *segments*.

LTAFDB: The second application is AF detection over the LTAFDB [27], which records the cardiac rhythm by ECG. The data preprocessing scheme is: 1) We utilize ECG recordings of the lead I over 84 subjects with the sampling rate at 128 Hz. We apply a band-pass FIR filter with a pass-band frequency of 0.5 Hz and a stop-band frequency of 50 Hz with an order of 5 to remove the noise; 2) We divide the recordings into episodes in the same manner as AAEL. There are 7358 AF and 46 347 non-AF episodes; 3) We segment each episode into 10-s segments.

HAR-UCI: The third application is human activities recognition (HAR) over the HAR-UCI dataset [31]. There are six activities (i.e., walking, upstairs, downstairs, sitting, standing, and laying) recorded by a smartphone over 30 subjects with a sampling rate of 50 Hz. Each sample contains 9-channel signals, including triaxial acceleration, body acceleration, and angular velocity. There are 30 episodes (signal episodes) of each type of action over 30 subjects (i.e., six episodes with six action for each subject). The signal has already been preprocessed by applying noise filters and then segmented into 2.56-s segments.

2) Evaluated Methods and Metrics: We compare PMFed against the methods falling under three categories: 1) FL algorithms that train a global model robust to non-IID local data; 2) existing SOTA meta FL algorithms that train a global model generalized to non-IID data; and 3) existing FL methods that utilize BN to overcome domain shifting. For (1), we implement FedAvg-FT [11] and FedProx [8]. For (2), we implement FedReptile [13], FedMeta [32], and Per-FedAvg [12]. For (3), we implement two SOTA methods, FedBN [14] and SiloBN [15]. For our method, we evaluate the performances of PMFed and conduct ablation studies on each component of the PMFed. We implement PMFed-Meta as an ablation study for local meta-learning mechanism where PMFed-Meta conducts local model training with one data set instead of setting a support set and query set in meta-learning. We further implement PMFed-MA as an ablation study for the proposed momentum-based model aggregation strategy where PMFed-MA aggregates the global model parameters by following the manner of FedAvg. We also implement fine-tuned PMFed (PMFed-FT) as an ablation study for the adaptive model personalization mechanism where PMFed-FT personalizes the global model with the simple fine-tuning strategy instead of the proposed adaptive model personalization.

We invoke metrics F1 score (F1) accuracy (ACC) to comprehensively evaluate methods. F1 is defined as $F_1 = 2 \times [(\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})]$ where $\text{Precision} = [\text{TP}/(\text{TP} + \text{FP})]$ and $\text{Recall} = [\text{TP}/(\text{TP} + \text{FN})]$.

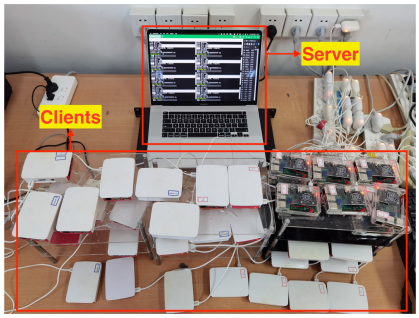


Fig. 6. IoT platform for performance evaluation.

In addition to the detection performances over segments, the detection performances over episodes are reported since the prediction over a single segment cannot sufficiently determine the health condition of the subject in real-world application scenarios. In VA and AF detection, a VA or AF episode would be determined if there are 4 consecutive VA or AF predictions on the input segments. Otherwise, the episode would be determined as non-VA or non-AF. In HAR, the episode would be determined by the greatest number of segments with certain labels. We further evaluate the effect of intersubject variability by comparing the performances achieved by PMFed and other baseline FL methods, and the intrasubject variability by comparing the performances achieved by PMFed, the simple fine-tuning approach, and the global model by PMFed.

3) Implementation Details: The performances of FL methods are evaluated on the platform with a MacBook Pro as server and Raspberry Pi 3Bs as clients with PyTorch (1.12.0) shown in Fig. 6. The server is a MacBook Pro 2020. The client devices are Raspberry Pi 3Bs. We further deploy the evaluated CNNs on STM32F469NI discovery kit [33] to evaluate practical performances of inference. The board is equipped with 2-MB flash and 324-KB SRAM.

The CNNs designed in [4], [34], and [35] for VA detection, AF detection, HAR are utilized for all evaluated methods. We invoke those networks with necessary modifications (e.g., change filter size and reduce the number of convolutional layers) to fit the input dimensions and limited hardware recourses. For each task, we randomly split subjects by 8:2 for training and testing. We perform 10-time Monte Carlo splitting on subjects of each dataset. The detection performances are reported based on the average performance of all 10 splits.

In the federated training of PMFed, only the subjects from the training set would be utilized for the global training of all methods. We construct the support set τ^s and the query set τ^q for each training subject in each round. We randomly pick 5, 20, and 5 segments for τ^s and 5, 40, and 5 segments for τ^q of each targeting class in VA, AF, and HAR tasks, respectively. The momentum parameter γ is set at 0.3, 0.2, and 0.1, and the maximal training rounds R is set at 500, 1000, and 500 in VA, AF, and HAR tasks. In testing, the subjects from testing set would be considered as the unseen clients since all of these subjects' data would not be utilized during the training stage of FL. For each testing subject, we extract a small portion of segments to construct the *personalizing set*, which is used to fine-tune the model for personalization. We randomly pick 5, 20 and 5 segments for τ^s of each targeting class in VA, AF, and HAR tasks, respectively. The rest segments as *testing*

TABLE I
DETECTION PERFORMANCES OVER SEGMENTS OF THREE TASKS

	VA		AF		HAR	
	F1	ACC	F1	ACC	Macro-F1	ACC
FedAvg [9]	88.02	87.74	73.51	87.33	90.29	90.29
FedAvg-FT [11]	89.03	89.28	75.51	90.81	92.05	92.56
FedProx [8]	89.15	89.36	75.77	90.93	92.89	92.91
FedMeta [32]	74.43	81.01	52.07	62.20	54.60	58.71
FedReptile [13]	88.60	88.69	75.35	90.49	91.92	92.81
Per-FedAvg [12]	88.83	88.93	72.68	88.05	91.87	92.74
FedBN [14]	87.51	87.79	71.68	87.40	92.04	92.09
SiloBN [15]	87.53	87.81	73.52	88.39	91.99	92.02
PMFed-Meta	89.04	90.63	79.29	92.41	93.10	92.98
PMFed-MA	91.22	92.15	79.87	93.12	93.02	92.72
PMFed-FT	91.29	92.64	78.73	92.41	93.51	93.29
PMFed	**93.07**	**94.07**	**81.90**	**94.75**	**93.59**	**93.39**

TABLE II
DETECTION PERFORMANCES OVER EPISODES OF THREE TASKS

	VA		AF		HAR	
	F1	ACC	F1	ACC	Macro-F1	ACC
FedAvg [9]	89.37	87.46	72.65	86.02	89.44	90.23
FedAvg-FT [11]	91.47	88.25	74.59	88.82	92.01	93.61
FedProx [8]	91.47	88.25	74.52	88.81	92.29	93.89
FedMeta [32]	72.96	79.41	55.37	67.22	63.84	72.50
FedReptile [13]	91.84	88.37	74.34	88.07	92.59	94.17
Per-FedAvg [12]	91.88	88.37	70.95	84.90	92.31	93.89
FedBN [14]	92.62	88.38	73.09	85.94	92.87	94.44
SiloBN [15]	92.62	88.38	72.74	85.84	92.87	94.44
PMFed-Meta	85.56	87.86	80.19	92.74	91.90	93.33
PMFed-MA	90.51	90.37	80.30	92.13	93.15	93.28
PMFed-FT	93.20	92.31	78.72	92.24	93.10	94.44
PMFed	**95.52**	**93.89**	**80.34**	**93.64**	**93.47**	**94.72**

set of the testing subject would be utilized to evaluate the detection method. All methods except FedAvg would fine-tune the global model using the personalizing set of each testing subject with the step of 5.

B. Experimental Results

1) Detection Performances: VA Detection: As shown in Table I, FedAvg-FT improves its F1 score by 1.01% and accuracy by 1.54% when compared with the performances of FedAvg. It indicates that fine-tuning the global with local data could further improve the detection performance. As for SOTA meta-federated algorithms FedReptile and Per-FedAvg, they achieve relatively similar performances compared with FedAvg-FT. It indicates that the SOTA meta-FL methods cannot effectively adapt to each testing subject due to inter- and intra-subject variability. The performances of FedBN and SiloBN cannot exceed FedAvg-FT in terms of F1 and accuracy. It further indicates that the intrasubject variability would degrade the performance of the personalized model. As for our method PMFed, it achieves the best performance among all methods, with the highest F1 score of 93.07% and accuracy of 94.07%. The accuracy achieved by PMFed outperforms FedProx, Per-FedAvg, and FedBN by 4.7%, 5.1%, and 6.3%, respectively. Table II illustrates detection performances on VA episodes. Compared with FedAvg, FedAvg-FT achieves a 2.1% increase in F1 score from a baseline of 89.37% and a 0.79% increase in accuracy from a baseline of 87.46%. The meta-learning approaches, FedReptile and Per-FedAvg, achieve relatively similar performances in terms of two metrics when compared with FedAvg-FT. As for PMFed, it again achieves the best performance on two metrics. It has the highest detection accuracy (93.89%), and the highest F1 score (95.52%). It indicates that our method could alleviate the intra- and inter-subject variability problems by generating a well-generalized model initialization and adaptively fine-tuning the model. As for the ablation study, PMFed-FT achieves the second-best metrics among all methods for VA detection on both segments and episodes. It indicates that the local meta-learning and momentum-based model aggregation strategy could effectively improve the generalization of the global model by overcoming the intersubject variability even with simple fine-tuning. The performances of PMFed-Meta and PMFed-MA further demonstrate the importance of the local meta-learning and momentum-based model aggregation strategy to the global model quality.

AF Detection: As shown in Table I, the simple fine-tuning strategy helps FedAvg-FT to improve its accuracy by 2% and F1 score by 3.48% when compared with FedAvg. As for meta-FL algorithms, FedReptile and Per-FedAvg do not give out better detection performances than FedAvg-FT and FedProx. FedMeta achieves the worst detection performances among all evaluated methods since the method is sensitive to inter- and intra-subject variable biosignal data and therefore cannot generate a well-generalized model. The performances show that our proposed PMFed could effectively improve the model generalization. Our proposed method PMFed achieves the highest accuracy (94.75%) and F1 score (81.90%), which improve by 6.70% and 9.22% when compared with SOTA meta-FL method Per-FedAvg. As shown in Table II, PMFed improves by 9.39% in F1 score and 8.74% in accuracy over episodes when compared with Per-FedAvg. It shows that the proposed adaptive model personalization could further improve detection performances by overcoming the intrasubject variability.

HAR Detection: Note that the data quantity of the HAR-UCI dataset is evenly spanned over each subject. Therefore, the main purpose of the experiment is to evaluate the generalization of the global model. Tables I and II show the detection performances in terms of macro-F1 (i.e., the averaged F1 on each action) and accuracy over segments and episodes. As shown in Table II, FedBN and SiloBN outperform SOTA meta-FL methods with only the feature distribution skew problem. It again indicates that the generalization of model initialization is critical in subject-specific detection. PMFed achieves the best performances with 93.59% in macro-F1 and 93.39% in accuracy on segments, and 93.47% macro-F1 and 94.72% accuracy on episodes. The performances show that PMFed could generate the model initialization with better generalization when compared with SOTA methods. PMFed slightly outperforms SOTA methods in HAR in terms of detection performance metrics. This is because data from the HAR dataset is evenly distributed among subjects in terms of quantity and morphological characteristics such that the inter- and intra-subject variability is relatively lower than the other two tasks. The purpose of setting the experiment is to validate that our method is capable of maintaining comparable

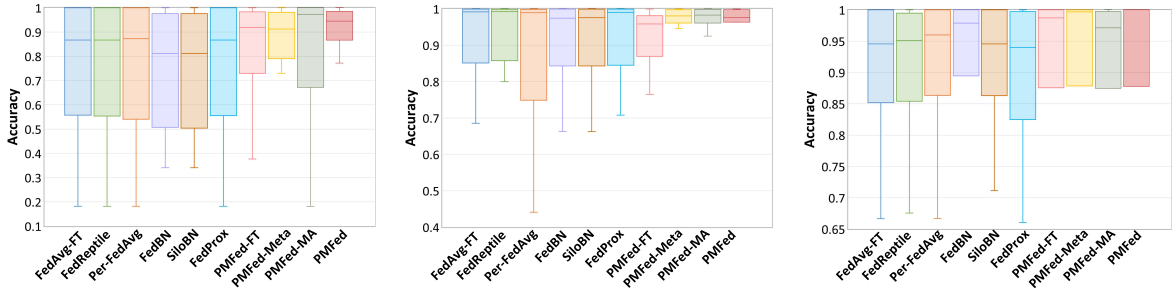


Fig. 7. Box plots of detection accuracy of individual testing subject's personalized model of FL methods over three health monitoring tasks. (a) Box plot (VA). (b) Box plot (AF). (c) Box plot (HAR).

performance even under conditions characterized by low-variability biosignal data.

2) Individual Performances: To demonstrate the performance improvement on individuals, Fig. 7 illustrates the distribution of detection accuracy of individual testing subjects from the same split across all evaluated methods over three tasks. As shown in the figure, SOTA methods yield a substantial variance in detection accuracy across testing subjects for all three tasks. In contrast, PMFed ensures a markedly smaller range of detection accuracy variance among test subjects. The distribution of individual detection accuracy achieved by PMFed has a narrower interquartile range and a narrower whisker range when compared with the other SOTA methods. To be more specific, in Fig. 7(a) which presents the VA detection, the detection accuracy for testing subjects fluctuates between 18% to 100% with FedProx, 19% to 100% with Per-FedAvg, and 34% to 100% with FedBN. On the other hand, PMFed consistently achieves an accuracy range of 77% to 100%. Likewise, in the AF detection shown in Fig. 7(b), the detection accuracy for testing subjects varies from 70% to 100% in FedProx, 44% to 100% in Per-FedAvg, and 66% to 100% in FedBN, while PMFed consistently attains an accuracy range from 96% to 100%. As for the HAR depicted in Fig. 7(c), the median accuracy achieved by PMFed is higher than the other SOTA methods and the interquartile range of PMFed is narrower when compared with other methods. Although PMFed achieves an accuracy range from 88% to 100% as FedBN represented by whiskers, it still outperforms FedProx which scores between 66% and 100% and Per-FedAvg which scores between 67% to 100%. It reveals that our proposed PMFed can effectively personalize the deep model for all testing subjects by overcoming the domain shifting problem during model personalization. Our method not only maintains consistent performance but also enables minority subjects to achieve superior detection accuracy.

As shown in Fig. 7(b), the box plot generated by PMFed-FT is markedly distinct, exhibiting a significantly longer interquartile range and a lower minimum, compared to the three boxes produced by PMFed, PMFed-Meta, and PMFed-MA. Given that PMFed-FT represents an ablation study focused on Subject-Specific Model Personalization, this disparity suggests that the personalized approach can markedly enhance detection accuracy for individual subjects. It also underscores the importance of the full PMFed system which effectively personalizes the global model to address intrasubject variability, which is crucial for boosting individual detection performance.

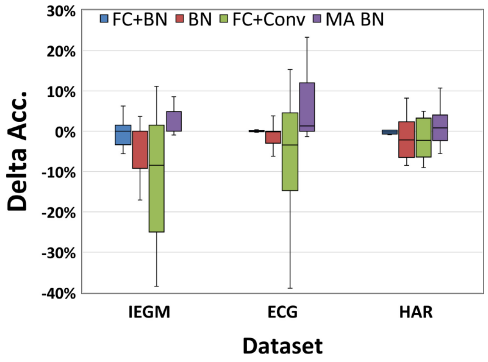


Fig. 8. Distribution of the delta accuracy of individual subjects under four strategies when compared to the general aggregation strategy.

As shown in Fig.7(c), a feature skew issue persists due to intersubject variability, as evidenced by the differing median lines in the box plots of each method. PMFed-MA exhibits the lowest median line, demonstrating that the Momentum-based Model Aggregation strategy effectively counteracts the adverse effects of aggregating client models trained on skewed features. Through the application of the strategy, PMFed achieves the highest median (approaching 100%), signifying a substantial proportion of subjects reaching near-perfect detection accuracy.

In summary, the key benefit of using the full PMFed is that it could optimize the FL process to obtain a well-generalized global model to overcome the feature distribution skew problem caused by intersubject variability, and optimize the fine-tuning process to overcome the overfitting problem caused by intrasubject variability, respectively.

3) Impact of Different Components: In preliminary experiments detailed in Section IV-B2, we assessed the influence of various neural network components on model aggregation by implementing four distinct strategies: 1) averaging both fully connected and BN layers (FC+BN); 2) averaging solely BN layers (BN); 3) averaging FC and convolutional layers (FC+Conv); and 4) applying momentum-based aggregation to BN layers (MA BN). Fig. 8 shows the resulting distributions of delta accuracy of individual subjects across all three datasets over the general aggregation strategy. Fig. 8 underscores the pivotal role of BN layers in determining the quality of the global models. It also indicates that simple average aggregation of BN layer parameters may not suffice for achieving a well-generalized global model.

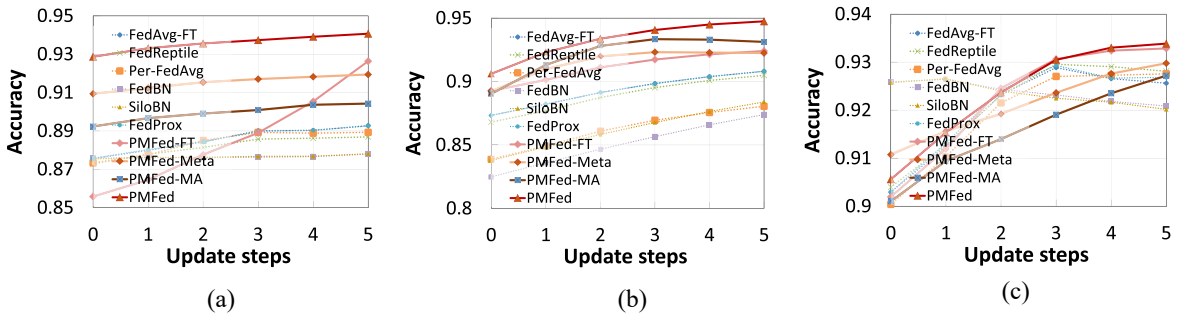


Fig. 9. Accuracy trends of FL methods during personalization over three health monitoring tasks. (a) Personalization accuracy (VA). (b) Personalization accuracy (AF). (c) Personalization accuracy (HAR).

4) Evaluation of Generalization and Personalization: To further demonstrate the effectiveness of the proposed meta-federated training and adaptive model personalization, we present the accuracy curves versus update steps of the global models personalized on subject-specific data for all FL methods in Fig. 9. The 5-step gradient descent is applied to personalize each global model using the personalizing set of the testing subjects. The curves of PMFed shown in Fig. 9(a) and (b) indicate that the proposed adaptive personalization mechanism could enable the global model to be with a better starting point by adaptively replacing the BN statistics. Based on a better starting point, the model personalized by PMFed could achieve the best detection performance in terms of accuracy when compared with SOTA methods.

As for the ablation study, as shown in Fig. 9, the global model trained with PMFed-FT could always gain a higher improvement than the SOTA methods. It indicates that the local meta-learning and momentum-based model aggregation strategy could improve the generalization of the global model by overcoming the intersubject variability. As for PMFed-MA and PMFed-Meta, the global model trained with these methods could also gain a higher or similar improvement than the SOTA methods. It indicates that the proposed adaptive model personalization mechanism could alleviate the intrasubject variability issue in model personalization.

5) Performance Comparison for Inter- and Intra-Subject Variability: For intersubject variability, as introduced in Section III-A, biosignals with the same label may exhibit varied morphological characteristics among different subjects. Therefore, we demonstrate the detection sensitivity achieved by PMFed and other SOTA methods on the same type of event or disease (VA, AF, Walking) across different subjects for all three datasets. Owing to this variability, SOTA methods perform inconsistently across different subjects. For example, as shown in Fig. 10(a), FedProx achieves nearly 100% detection sensitivity for VA on Subjects 184, 185, and 216. However, for Subject 329, FedProx's detection sensitivity can drop to below 25%. Similar disparities in detection performance are observed for all SOTA methods in detecting AF and Walking events across the ECG and HAR datasets. These outcomes underscore the significant impact of intersubject variability on the detection models. Conversely, as illustrated in Fig. 10, PMFed enables the deep model to consistently perform well across different subjects by overcoming intersubject variability. This suggests that the proposed optimizations can

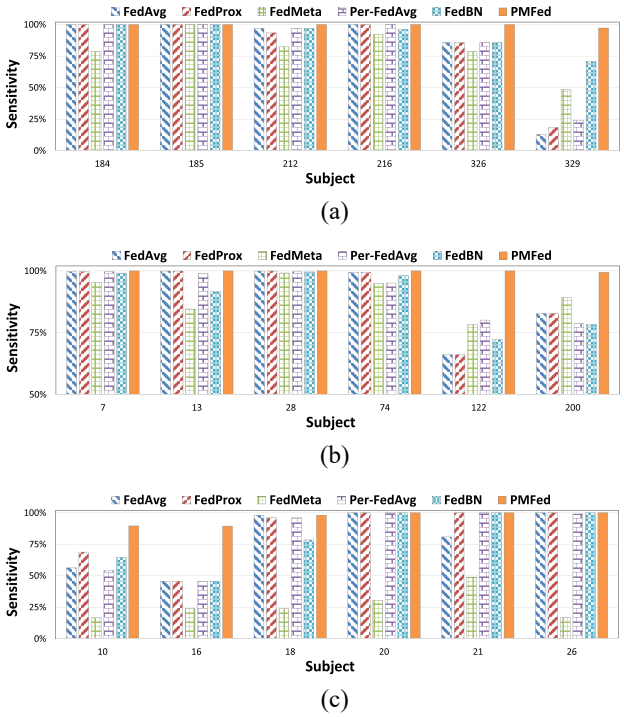


Fig. 10. Performance comparison of intersubject variability by individual detection sensitivity with PMFed and SOTA methods. (a) Individual sensitivity for VA detection. (b) Individual sensitivity for AF detection. (c) Individual sensitivity for HAR detection.

effectively enhance the generalization of the global model and mitigate the adverse effects of intersubject variability.

For intrasubject variability, as introduced in Section III-B, biosignals may exhibit varied temporal patterns within the same subject. This type of variability can significantly impede performance enhancement through fine-tuning. To elucidate this variability, we demonstrate the detection accuracy achieved by PMFed, PMFed-FT, and PMFed-Global (the global model trained using PMFed) on individual subjects from all three datasets. Due to this variability, not all global models benefit from detection improvement through fine-tuning with subject-specific data. For instance, as evidenced by the performance of PMFed-FT and PMFed-Global in Fig. 11, global models fine-tuned on data from Subject 74 (for AF detection), Subject 326 (for VA detection), and Subject 16 (for Walking detection) exhibit accuracy degradation. These results highlight the significant impact that varied temporal patterns

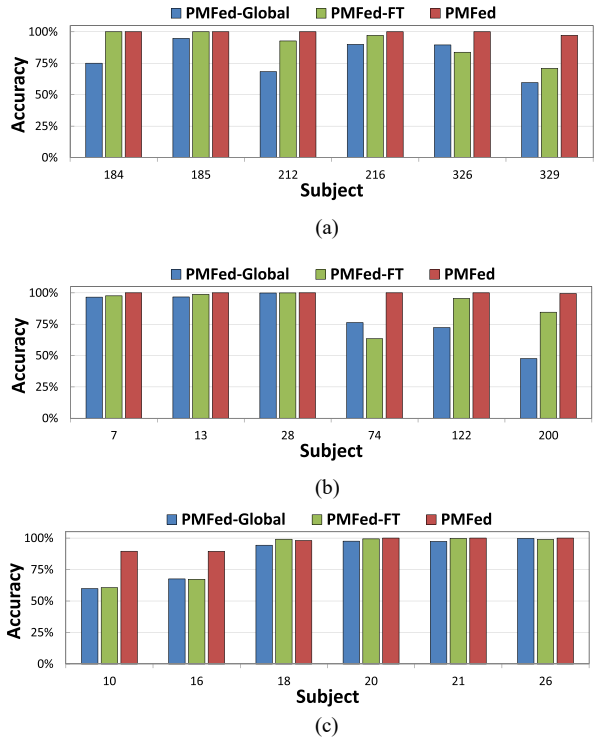


Fig. 11. Performance comparison of intrasubject variability by individual detection accuracy with PMFed, PMFed-Global, and PMFed-FT. (a) Individual accuracy for VA detection. (b) Individual accuracy for AF detection. (c) Individual accuracy for HAR detection.

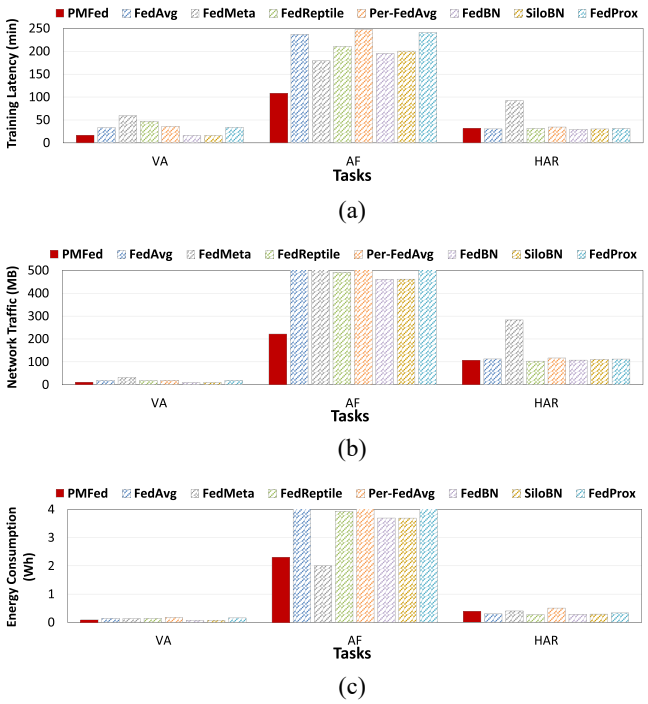


Fig. 12. Practical performances of FL methods in terms of federated training latency, network traffic, and energy consumption over three tasks. (a) Training latency of federated training. (b) Network traffic of federated training. (c) Energy consumption of federated training.

of biosignals can have on the personalization (fine-tuning) process. Conversely, as depicted in Fig. 11, PMFed allows the global model to avoid drifting too far from the learned representations during personalization, achieving an accuracy improvement for nearly all subjects. This suggests that the proposed optimizations can effectively address the overfitting issue caused by intrasubject variability.

6) Practical Performances: We further evaluate the practical performances of FL methods on the developed framework.

Training Latency: Fig. 12(a) presents the latency (i.e., the total federated training time) of all evaluated methods. The latency is reported based on the mechanism that halts the federated training when the training accuracy is no longer increasing and the moving average (period of 16) keeps stable (standard variance less than 0.1) for 10 rounds.

In the VA detection task shown in Fig. 12(a), our method reduces training latency by 51.0% and 54.3% in comparison to FedProx from method category (1) and Per-FedAvg from method category (2) defined in Section V-A2, respectively. When compared with FedBN, our method demonstrates a training latency that is comparable. This is because SOTA methods in category (3) are primarily designed to expedite convergence over non-IID data through optimizations on BN. This underlines the critical role that BN layer parameters play in model convergence during training.

As for AF detection shown in Fig. 12(a), our method significantly reduces training latency by 55.0%, 56.2%, and 44.4% when compared to FedProx from category (1), Per-FedAvg from category (2), and FedBN from category (3), respectively.

Regarding the HAR task, our method achieves a training latency that is comparable with SOTA methods, as the convergence rate is similar across these methods. This is due to the even distribution of data from the HAR dataset among subjects, both in terms of quantity and morphological characteristics. As a result, the inter- and intra-subject variability is relatively lower than in the other two tasks. The purpose of setting the experiment is to validate that our method is capable of maintaining comparable performance even under conditions characterized by low-variability biosignal data.

Network Traffic: As shown in Fig. 12(b), PMFed achieves low-network traffic. PMFed again achieves the lowest network traffic in AF detection. PMFed reduces network traffic by 60.7%, 59.8%, and 51.9% when compared with FedProx in category (1), Per-FedAvg in category (2), and FedBN in category (3), respectively. As for VA detection, PMFed could reduce the network traffic by 44.5% and 46.3% when compared with FedProx and Per-FedAvg. When compared with FedBN, our method demonstrates comparable network traffic. This again indicates the critical role that BN layer parameters play in model convergence. As for the HAR task, all methods except FedMeta achieve comparable network traffic as the dataset is with low-intersubject variability.

Energy Consumption: As shown in Fig. 12(c), in AF detection, PMFed achieves the second lowest energy consumption. PMFed reduces energy consumption by 50.8%, 54.0%, and 37.5% when compared with FedProx in category (1), Per-FedAvg in category (2), and FedBN in category (3), respectively. As for VA detection, PMFed could reduce energy consumption by 42.1% and 47.2% when compared with FedProx and Per-FedAvg. When compared with FedBN, our method achieves a comparable energy consumption. This again indicates the critical role that BN layer parameters play in

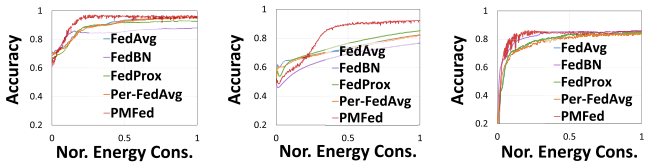


Fig. 13. Tradeoff between energy and accuracy of PMFed and SOTA methods. (a) VA detection. (b) AF detection. (c) HAR.

TABLE III
INFERENCE PERFORMANCES OF CNNS ON THREE TASKS

Task	Inference Latency	Flash Occupation	Work-in Memory
CNN-VA	9.50 ms	40.75 KB	5.24 KB
CNN-AF	64.53 ms	299.59 KB	24.87 KB
CNN-HAR	37.57 ms	284.70 KB	13.82 KB

model convergence. The energy consumption is reported based on the setting that the evaluated FL reaches full convergence. Specifically, in the context of HAR dataset, the uniform distribution of data episodes per class across subjects negates significant data quantity skew issues. This uniformity means that the advantages of the proposed PMFed, particularly its enhanced convergence speed observed in IEGM and ECG datasets with data quantity skewness, are not as pronounced in the HAR dataset. Consequently, the convergence rates of the methods evaluated on the HAR dataset are relatively uniform, leading to comparable energy consumption levels across all methods.

Energy and Accuracy Tradeoff: Fig. 13 illustrates the relationship between the normalized energy consumption of the edge device and the detection accuracy obtained by PMFed and SOTA methods across three datasets. This figure demonstrates that PMFed requires a much less amount of energy to achieve improved detection performance on datasets with high skewness, like the IEGM and ECG datasets, when compared to SOTA methods. Conversely, for the HAR dataset, achieving higher accuracy necessitates relatively similar energy consumption for all methods. This pattern underscores the impact of dataset characteristics on the efficiency of FL models, highlighting how inherent dataset variability influences both energy demands and accuracy outcomes.

Inference Performances on MCU: Table III shows inference latency, flash occupation, and work-in memory overhead when executing inference on the STM32F469NI board. The practical performances on board indicate that the deep models can be properly deployed on the resource-constrained platform to conduct real-time and on-device health monitoring. Our FL approach is based on a computing framework where MCU handles data sensing and on-device inference, while the Raspberry Pi manages FL and model personalization. This configuration aligns with health monitoring contexts, where local devices like wearables and implants typically connect through an edge device (e.g., a smartphone or hub). In our framework, deep models are trained and fine-tuned on the edge device, such as a Raspberry Pi, to circumvent the limited computing power and memory of MCU-based monitors. This design efficiently supports the intensive computational demands of training and personalizing models, making it a practical solution for enhancing the capabilities of health monitoring systems.

In future work, we aim to delve into on-device training to enable FL directly on MCUs. While there are precedents in this area, existing solutions still fall short of system requirements, placing a heavy workload on the highly resource-constrained MCUs. Developing efficient TinyML-based training methods for MCU-level devices presents a compelling direction.

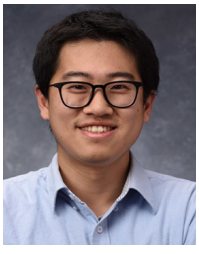
VI. CONCLUSION

In this article, we propose a PMFed framework for IoT-enabled health monitoring. PMFed aims to address inter- and intra-subject variability issues in health monitoring on biosignals by obtaining a well-generalized global model and properly personalizing the model. Experimental results show that PMFed outperforms SOTA FL methods in terms of various detection metrics while significantly reducing total training time and communication overhead in real-world health monitoring tasks.

REFERENCES

[1] P. Valsalan, T. A. B. Baomar, and A. H. O. Baabood, "IoT based health monitoring system," *J. Crit. Rev.*, vol. 7, no. 4, pp. 739–743, 2020.
[2] (Boston Sci., Marlborough, MA, USA). *LUX-dx^{TM} Insertable Cardiac Monitor (ICM) System*. (2017). [Online]. Available: https://www.bostonscientific.com/content/dam/bostonscientific/Rhyt%20Management/portfolio-group/lux-dx-icm/pdf/LUX-Dx-Clinic-Resource-Guide.pdf
[3] G. M. Balbim et al., "Using fitbit as an mHealth intervention tool to promote physical activity: Potential challenges and solutions," *JMIR mHealth uHealth*, vol. 9, no. 3, 2021, Art. no. e25289.
[4] Z. Jia, Z. Wang, F. Hong, L. Ping, Y. Shi, and J. Hu, "Learning to learn personalized neural network for ventricular arrhythmias detection on intracardiac EGMs," in *Proc. IJCAI*, 2021, pp. 2606–2613.
[5] S.-M. Lee, S. M. Yoon, and H. Cho, "Human activity recognition from accelerometer data using convolutional neural network," in *Proc. Int. Conf. Big Data Smart Comput. (BIGCOMP)*, 2017, pp. 131–134.
[6] P. Kairouz et al., "Advances and open problems in federated learning," 2019, *arXiv:1912.04977*.
[7] X. Lan, D. Ng, S. Hong, and M. Feng, "Intra-inter subject self-supervised learning for multivariate cardiac signals," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 4532–4540.
[8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, vol. 2, 2020, pp. 429–450.
[9] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Stat.*, 2017, pp. 1273–1282.
[10] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," 2020, *arXiv:2002.10619*.
[11] K. Wang, R. Mathews, C. Kiddon, H. Eichner, F. Beaufays, and D. Ramage, "Federated evaluation of on-device personalization," 2019, *arXiv:1910.10252*,
[12] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," 2020, *arXiv:2002.07948*.
[13] Y. Jiang, J. Konečnỳ, K. Rush, and S. Kannan, "Improving federated learning personalization via model agnostic meta learning," 2019, *arXiv:1909.12488*.
[14] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," 2021, *arXiv:2102.07623*.
[15] M. Andreux, J. O. D. Terrail, C. Beguier, and E. W. Tramel, "Siloed federated learning for multi-centric histopathology datasets," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Cham, Switzerland: Springer, 2020, pp. 129–139.

[16] S. Selvaraj and S. Sundaravaradhan, "Challenges and opportunities in IoT healthcare systems: A systematic review," *SN Appl. Sci.*, vol. 2, no. 1, p. 139, 2020.
[17] N. Zanker, D. Schuster, J. Gilkerson, and K. Stein, "Tachycardia detection in ICDs by Boston scientific," *Herzschrittmachertherapie+ Elektrophysiologie*, vol. 27, no. 3, pp. 186–192, 2016.
[18] "iRhythm Zio patch." iRhythm Technologies. 2021. [Online]. Available: https://www.irhythmtech.com
[19] "Fitbit smartwatches." Fitbit.2022. [Online]. Available: https://www.fitbit.com/global/us/products/smartwatches
[20] J. M. Raja et al., "Apple watch, wearables, and heart rhythm: Where do we stand?" *Ann. Transl. Med.*, vol. 7, no. 17, p. 417, 2019.
[21] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat. Med.*, vol. 25, no. 1, p. 65, 2019.
[22] S. L. Oh et al., "A deep learning approach for parkinson's disease diagnosis from EEG signals," *Neural Comput. Appl.*, vol. 32, no. 15, pp. 10927–10933, 2020.
[23] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "Fedhealth: A federated transfer learning framework for wearable healthcare," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 83–93, Jul./Aug. 2020.
[24] S. Warnat-Herresthal et al., "Swarm learning for decentralized and confidential clinical machine learning," *Nature*, vol. 594, no. 7862, pp. 265–270, 2021.
[25] X. Tan, C.-C. H. Chang, and L. Tang, "A tree-based federated learning approach for personalized treatment effect estimation from heterogeneous data sources," 2021, *arXiv:2103.06261*.
[26] *Ann Arbor Electrogram Libraries*. Chicago, IL, USA: AnnArbor, 2003.
[27] S. Petrutiu, A. V. Sahakian, and S. Swiryn, "Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans," *Europace*, vol. 9, no. 7, pp. 466–470, 2007.
[28] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
[29] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. ICML*, 2017, pp. 1126–1135.
[30] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96, 1996, pp. 226–231.
[31] D. Micucci, M. Mobilio, and P. Napoletano, "Unimib Shar: A dataset for human activity recognition using acceleration data from smartphones," *Appl. Sci.*, vol. 7, no. 10, p. 1101, 2017.
[32] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, "Federated meta-learning with fast convergence and efficient communication," 201, *arXiv:1802.07876*.
[33] (STMicroelectron., Geneva, Switzerland). *Discovery Kit With STM32F469NI MCU*. (2020). [Online]. Available: https://www.st.com/en/evaluation-tools/32f469idiscovery.html
[34] C.-H. Hsieh, Y.-S. Li, B.-J. Hwang, and C.-H. Hsiao, "Detection of atrial fibrillation using 1D convolutional neural network," *Sensors*, vol. 20, no. 7, p. 2136, 2020.
[35] A. Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Appl. Soft Comput.*, vol. 62, pp. 915–922, Jan. 2018.



Zhenge Jia (Member, IEEE) received the B.S. degree in engineering and computer science from The Australian National University, Canberra, ACT, Australia, in 2017, and the Ph.D. degree in electrical and computer engineering from the University of Pittsburgh, Pittsburgh, PA, USA, in 2022.

He is currently a Postdoctoral Fellow with the Department of Computer Science and Engineering, the University of Notre Dame, Notre Dame, IN, USA. He published more than 15 papers in *Nature Machine Intelligence*, *The DAC Journal*, ICCAD, IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, and IJCAI. His research interests include personalized deep learning and on-device deep learning in healthcare.

Tianren Zhou (Member, IEEE) received the master's degree from the Department of Computer Science and Technology, Shandong University, Jinan, China, in 2023, where he is currently pursuing the Ph.D. degree.

His current research interests include tinyML and federated learning.

Zheyu Yan (Member, IEEE) received the B.S. degree from Zhejiang University, Hangzhou, China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA.

His research interests lie in software–hardware co-design of deep neural network accelerators using emerging technologies, especially nonvolatile memory-based compute-in-memory platforms.

Jingtong Hu (Senior Member, IEEE) received the B.E. degree in computer science and technology from Shandong University, Jinan, China, in 2007, and the Ph.D. degree in computer science from The University of Texas at Dallas, Richardson, TX, USA, in 2013.

He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA. Before that, he was an Assistant Professor with Oklahoma State University, Stillwater, OK, USA, from 2013 to 2017. His current research interests include hardware/software co-design for machine learning algorithms, on-device AI, and embedded systems.

Dr. Hu's works have received the Donald O. Pederson Best Paper Award from IEEE TCAD, Best Paper Award from ASP-DAC 2024, and five best paper nominations from DAC, ASP-DAC, and ESWEEK. He was a recipient of the University of Pittsburgh William Kepler Whiteford Faculty Fellowship, the Oklahoma State University Outstanding New Faculty Award, the Air Force Summer Faculty Fellowship, and the ACM SIGDA Meritorious Service Award. He has served on the Technical Program Committee of many international conferences, such as *The DAC Journal*, *The DATE Journal*, ASP-DAC, ESWEEK, and CPS-IoT Week. He served as a Guest Editor for *Sensors*, IEEE TRANSACTIONS ON COMPUTERS, and *ACM Transactions on Cyber–Physical Systems*. He is currently serving as an Executive Committee Member and the Education Chair of ACM SIGDA and an Associate Editor for IEEE EMBEDDED SYSTEMS LETTERS, the *Journal of Systems Architecture: Embedded Software Design*, and *ACM Transactions on Cyber–Physical Systems*.

Yiyu Shi (Senior Member, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2005, and the M.S and Ph.D. degrees in electrical engineering from The University of California at Los Angeles, Los Angeles, CA, USA, in 2007 and 2009, respectively.

He is currently a Professor with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA; the Site Director of National Science Foundation I/UCRC Alternative and Sustainable Intelligent Computing; and the Director of the Sustainable Computing Lab. His current research interests focus on hardware intelligence and biomedical applications.

Prof. Shi is the recipient of Facebook Research Award, the NSF CAREER Award, the IEEE Region 5 Outstanding Individual Achievement Award, and the IEEE Computer Society Mid-Career Research Achievement Award, among others. In recognition of his research, more than a dozen of his papers have been nominated for or awarded as the Best Paper in top journals and conferences, including the 2021 IEEE Transactions on Computer-Aided Design Donald O Pederson Best Paper Award and the 2023 IEEE/ACM International Conference on Computer-Aided Design William J. McCalla Best Paper Award. He has served on the Technical Program Committee of many international conferences. He is the Deputy Editor-in-Chief of IEEE VLSI CIRCUITS AND SYSTEMS NEWSLETTER and an Associate Editor of various IEEE and ACM journals. He is an IEEE CEDA Distinguished Lecturer and an ACM Distinguished Speaker.