MH-pFLGB: Model Heterogeneous personalized Federated Learning via Global Bypass for Medical Image Analysis

Luyuan Xie[1,2⋆], Manqing Lin[1⋆], ChenMing Xu[1,2], Tianyu Luan[3], Zhipeng Zeng[1,2], Wenjun Qian[1,2], Cong Li[1,2⋆⋆], Yuejian Fang[1,2], Qingni Shen[1,2], and Zhonghai Wu[1,2⋆⋆]

[1]School of Software and Microelectronics, Peking University, Beijing, China
[2]National Engineering Research Center for Software Engineering, Peking University, Beijing 100871, China
[3]State University of New York at Buffalo

Abstract. In the evolving application of medical artificial intelligence, federated learning is notable for its ability to protect training data privacy. Federated learning facilitates collaborative model development without the need to share local data from healthcare institutions. Yet, the statistical and system heterogeneity among these institutions poses substantial challenges, which affects the effectiveness of federated learning and hampers the exchange of information between clients. To address these issues, we introduce a novel approach, MH-pFLGB, which employs a global bypass strategy to mitigate the reliance on public datasets and navigate the complexities of non-IID data distributions. Our method enhances traditional federated learning by integrating a global bypass model, which would share the information among the clients, but also serves as part of the network to enhance the performance on each client. Additionally, MH-pFLGB provides a feature fusion module to better combine the local and global features. We validate MH-pFLGB's effectiveness and adaptability through extensive testing on different medical tasks, demonstrating superior performance compared to existing state-of-the-art methods.

Keywords: Model heterogeneous · Personalized federated learning · Global bypass model.

1 Introduction

In the field of medical images, federated learning [20] has emerged as a key technique for its ability to protect the privacy of training datasets. This approach allows for the collaborative development of a unified global model, eliminating the need to directly share local data from individual healthcare facilities. However, the application of federated learning in healthcare faces challenges such as statistical heterogeneity [**?**], due to the diverse and non-uniformly distributed (non-IID) data across different institutions,

⋆ Equal Contribution.
⋆⋆ Corresponding author: Cong Li, li.cong@pku.edu.cn; Zhonghai Wu, wuzh@pku.edu.cn

and system heterogeneity [**?**], due to the unique architecture of local models by each institution. These challenges compromise the efficiency of federated learning and hinder the seamless exchange of information between client models. Addressing the issues of statistical and system heterogeneity presents a critical and impactful challenge in the application of federated learning within healthcare facilities.

Previous works only focused on statistical heterogeneity and proposed personalized federated learning methods [5,24,2,27,14,1,15]. Compared to traditional single model settings [20,10,13], personalized federated learning allows each client to learn their own model, effectively alleviating the problem of statistic heterogeneity. However, these methods still require models with the same structure for each client. Recent works including FedMD [12], FedDF [16], DS-pFL [9] and KT-pFL [30] tackle statistic and system heterogeneity in federated learning by sharing soft predictions among clients. These approaches have advanced the field by addressing heterogeneity issue, but depend heavily on public datasets for generating these soft predictions. However, collecting the medical dataset for public usage would involve a certain level of privacy requirements and complex censoring processes. Besides, the extensive size of public datasets would largely increase the computational cost, thus limiting the application of these techniques. All of these problems would significantly raise the cost of deploying those methods.

To eliminate the reliance on public datasets, we propose a global bypass strategy to address the challenges of heterogeneous models under the distribution of non-IID data. Unlike traditional approaches that rely on soft prediction generated from public datasets, our method adds a global bypass model to the local clients to share the information among the clients and help the local clients. In each client, the global bypass would not only learn the information from local data, but also help the previous local network to make its prediction. In the server, we aggregate the global bypass to share the information among each client. Additionally, we design the global bypass to be small so the computational cost is less than what would be required for local training on a public dataset.

Specifically, we propose framework Model Heterogeneous personalized Federated Learning via Global Bypass (MH-pFLGB). Our global bypass consists of a body and a head module. The body is a light-weighted encoder for feature extraction and the head is a small module designed to fit the outputs of different tasks. To better fuse the information from the local model and global bypass model, we designed a fusion module named features weighted fusion to fusion the features from the body of the local and global model. The fusion is based on allowing models to learn how to better select weights for global and local features. This design would better utilize global knowledge and integrate it with local features, so that it can improve the performance of local models for each client.

Our contributions are summarized as follows:

- We introduce a novel personalized federated learning approach for dealing with heterogeneous models named MH-pFLGB. This approach leverages a global bypass mechanism that obviates the need for public medical datasets, thereby reducing the additional burdens associated with local training.

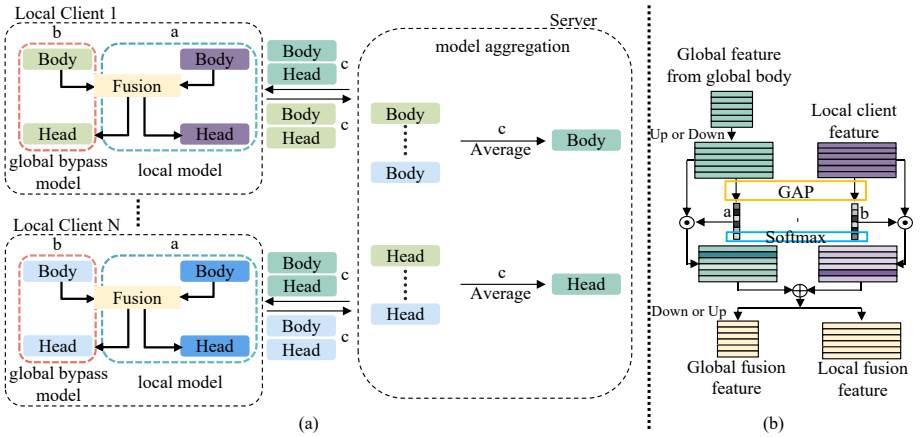

Fig. 1: (a) Overview of our proposed MH-pFLGB framework. Each training process consists of 3 steps. From a to c: **a.** Local model training. **b.** Global bypass model training. **c.** Upload, aggregation, and download. (b) Features weighted fusion. More details can be found in Section 2.1 and Section 2.2.

- We design a global bypass model to transfer information among different clients and enhance the result of each local client. Additionally, we integrate a feature fusion module to more effectively combine features from the local model and the global bypass.
- We demonstrate the efficacy and versatility of our MH-pFLGB through rigorous testing on a variety of medical tasks, such as image classification and image segmentation. Our method surpasses current state-of-the-art results in all these areas, underscoring its potential and adaptability across a broad range of medical applications.

2 Methods

2.1 Pipeline

The pipeline of MH-pFLGB is shown in Fig. 1 (a). Each local client consists of an architecture heterogeneous local model and a global bypass model that shares the same network architecture among other clients. The local body extracts personalized features from local clients, while the global body shares learned parameters among clients. Both local and global models are divided into a body model to extract features, and a head model to generate the network output using the features. Our training process consists of 3 steps: **a.** Local model training, **b.** Global bypass model training, and **c.** Global aggregation. We will explain the details of those steps in the following.

Local model training. In the local model training stage, the local model learns from both the local dataset and the global insights provided by the bypass model. At this stage, we freeze the global bypass model and only train the local model. For client

i, its local model training loss function $\mathcal{L}_{loc,i}$ is:

$$\mathcal{L}_{loc,i} = \lambda^l_{loc}\mathcal{L}^l_{loc}(\hat{y}^l_i, y_i) + \lambda^g_{loc}\mathcal{L}^g_{loc}(\hat{y}^g_i, y_i), \tag{1}$$

where $\hat{y}^l_i$ and $\hat{y}^g_i$ are the predictions from the local head and global head. $\mathcal{L}^l_{loc}$ and $\mathcal{L}^g_{loc}$ represent the loss functions for the local and global model output, respectively. λ^l_{loc} and λ^g_{loc} are their corresponding weights. y_i is the label of input data x_i. Note that, even though the global bypass is fixed, we still calculate the loss function on its output to maintain generalizability when training the local model.

Global bypass model training. During the global model training phase, we freeze the local model and fine-tune the global bypass model. This enables the body of the global model to learn the information from each client. The loss function $\mathcal{L}_{glob,i}$ is represented as:

$$\mathcal{L}_{glob,i} = \lambda^g_{glob}\mathcal{L}^g_{glob}(\hat{y}^g_i, y_i) + \lambda^l_{glob}\mathcal{L}^l_{glob}(\hat{y}^l_i, y_i). \tag{2}$$

$\mathcal{L}^g_{glob}$ and $\mathcal{L}^l_{glob}$ represent the loss function for training the global and local model at the global training stage, respectively. λ^g_{glob} and λ^l_{glob} are their corresponding weights. Other variables are defined the same as in eq.(1). Similar to the local training stage, local loss function $\mathcal{L}^l_{glob}$ is designed to avoid client drift.

Global Aggregation. As the global model is uploaded to the server, the global aggregation process aggregates the model parameters, with distinct processes for both the body and head of the global model. This aggregation employs weight averaging, as outlined in [20]. Finally, the aggregated model is downloaded and distributed for the next round of training.

At the inference stage, we fuse global and local features, and the fused features output prediction results through the local head. The global head only participates in the training stage and does not participate in the inference stage. This is similar to adding a regularization term during local training, effectively preventing overfitting of the local model.

2.2 Features Weighted Fusion

In order to better fuse global and local features, we propose a new feature method named Feature Weighted Fusion. As shown in Fig. 1(b), the feature from global body x_g ensures that the dimension is the same as local client feature x_l through upsampling or downsampling as

$$\hat{x}_g = f_{up}(x_g) \; or \; f_{down}(x_g), \tag{3}$$

where $f_{up}(\cdot)$ and $f_{down}(\cdot)$ respectively represent upsampling or downsampling operations. Specifically, in the classification task, we use 1×1 convolution for both upsampling and downsampling. In the segmentation task, we adopt deconvolution for upsampling and convolution for downsampling, along with a global average pooling operation on the results. Having $\hat{x}_g$, a Softmax operator is applied on $\hat{x}_g$ and x_l 's channel-wise digits:

$$a_i = \frac{\exp(\hat{x}_{g,i})}{\exp(\hat{x}_{g,i})+\exp(x_{l,i})};\; b_i = \frac{\exp(x_{l,i})}{\exp(\hat{x}_{g,i})+\exp(x_{l,i})},\; 0 < i \leq C, \tag{4}$$

Table 1: The results of classification task in different resolutions. The x2↓, x4↓, and x8↓ are downsampling half, quarter, and eighth of high-resolution images. We evaluate ACC and MF1 results on BreaKHis dataset. The larger the better. **Bold** number means the best. Only local training, FedMD, FedDF, pFedDF, DS-PFL, KT-PFL, and MH-pFLGB use heterogeneous models in each client. The four client models are set to ResNet$\{17, 11, 8, 5\}$, respectively. Other methods use the unified model settings (ResNet17). MH-pFLGB achieves the best performance.

Method	high-resolution		x2↓		x4↓		x8↓		Average	
	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑
Only Local Training	0.7891	0.7319	0.8027	0.7461	0.7538	0.6852	0.6956	0.5867	0.7603	0.6875
FedAvg [20]	0.7406	0.6425	0.7908	0.7405	0.6892	0.6031	0.5774	0.4681	0.6995	0.6136
FedAvg+FT	0.7749	0.7218	0.8124	0.7511	0.7327	0.6628	0.6234	0.5073	0.7359	0.6608
SCAFFOLD [10]	0.7442	0.6512	0.8097	0.7533	0.6725	0.5963	0.5866	0.4732	0.7033	0.6185
SCAFFOLD+FT	0.7761	0.7229	0.8237	0.7709	0.7523	0.6872	0.6142	0.5005	0.7416	0.6704
FedProx [13]	0.7354	0.6386	0.7873	0.7421	0.6944	0.6107	0.5821	0.4687	0.6998	0.6150
FedProx+FT	0.7827	0.732	0.8055	0.7549	0.7548	0.6811	0.6071	0.4829	0.7375	0.6627
Ditto [14]	0.7304	0.6221	0.7661	0.6482	0.6065	0.5022	0.5931	0.4741	0.6740	0.5617
APFL [2]	0.7444	0.6568	0.7992	0.7355	0.6227	0.5229	0.6133	0.4986	0.6949	0.6035
FedRep [1]	0.7991	0.7618	0.8229	0.7697	0.7762	0.7182	0.6328	0.5091	0.7578	0.6897
LG-FedAvg [15]	0.7972	0.7523	0.5655	0.4397	0.6131	0.5080	0.6080	0.4902	0.6460	0.5476
FedMD [12]	0.7599	0.7083	0.8321	0.7829	0.7721	0.6893	0.6495	0.5439	0.7534	0.6811
FedDF [16]	0.7661	0.7253	0.8132	0.7629	0.7826	0.7342	0.6627	0.5627	0.7562	0.6963
pFedDF [16]	0.8233	0.7941	0.8369	0.7965	0.8121	0.7534	0.6843	0.6022	0.7892	0.7366
DS-pFL [9]	0.7842	0.7609	0.8334	0.7967	0.7782	0.7258	0.6327	0.5229	0.7571	0.7016
KT-pFL [30]	0.8424	0.8133	0.8441	0.8011	0.7801	0.7325	0.7032	0.6219	0.7925	0.7422
MH-pFLGB (Ours)	**0.8952**	**0.8697**	**0.8963**	**0.8745**	**0.8681**	**0.8333**	**0.7793**	**0.7214**	**0.8597**	**0.8247**

where C is the the dimension of $\hat{x}_g$ and x_l. a and b are the calculated weights. The local fusion feature x_{lf} is obtained by multiplying and adding the corresponding weights and features.

$$x_{lf,i} = a_i \hat{x}_{g,i} + b_i x_{l,i},\ 0 < i \leq C, \tag{5}$$

where $x_{lf,i}$ is the i-th element of x_{lf}. x_{lf} obtains the global fusion feature x_{gf} for global head through downsampling or upsampling as:

$$x_{gf} = f_{down}(x_{lf})\ or\ f_{up}(x_{lf}). \tag{6}$$

3 Experiments Setup

Task and Dataset. We verify the effectiveness of MH-pFLGB on 3 non-IID tasks. For **medical image classification (different resolution)** task, our experiments are conducted on the *Breast Cancer Histopathological Image Database (BreaKHis)* [26]. We perform x2↓, x4↓, and x8↓ downsampling on the high-resolution images [36]. Each resolution of medical images is treated as a client. In this task, we employed ResNet $\{17, 11, 8, 5\}$ as the local model of each client, respectively. For **medical image classification (different label distributions)** task, we employ 2 datasets, including a breast cancer classification dataset *BreaKHis* (color images) and an Optical Coherence Tomography (OCT) disease classification dataset *OCT2017* (grayscale images) [11]. We

design eight clients, each corresponding to a distinct heterogeneous model. These models include ResNet [4], ShuffleNetV2 [18], ResNeXt [29], SqueezeNet [8], SENet [6], MobileNetV2 [23], DenseNet [7], and VGG [25]. For **medical image segmentation** task, we use four datasets for polyp segmentation. They are ColonDB [33], ETIS [32], ClinicDB [34] and Kvasir-SEG [35]. Each center's dataset treated as a separate client. Each client utilized a specific model, including Unet++ [31], FCN [17], Unet [22], and Res-Unet [3].

Table 2: The results of Image Classification Task with Different Label Distributions. This task includes breast cancer classification and OCT disease classification. We evaluate ACC and MF1 result in this task. The larger the better. **Bold** number means the best. MH-pFLID has the best performance.

Breast cancer classification																
Method	ResNet		shufflenetv2		ResNeXt		squeezeNet		SENet		MobileNet		DenseNet		VGG		Average	
	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1 ↑	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑
Only Local Training	0.59	0.455	0.845	0.8412	0.665	0.5519	0.84	0.7919	0.875	0.849	0.755	0.5752	0.855	0.6884	0.875	0.8515	0.7875	0.7005
FedMD [12]	0.692	0.5721	0.823	0.8027	0.704	0.6087	0.875	0.8544	0.907	0.8745	0.762	0.6627	0.835	0.6493	0.842	0.8001	0.8050	0.7281
FedDF [16]	0.721	0.5949	0.817	0.8094	0.723	0.6221	0.893	0.8735	0.935	0.9021	0.757	0.6609	0.847	0.6819	0.833	0.7826	0.8158	0.7409
pFedDF [16]	0.755	0.6536	0.853	0.8256	0.741	0.6237	0.894	0.8742	0.935	0.9021	0.796	0.7219	0.879	0.7095	0.874	0.8521	0.8409	0.7703
DS-pFL [9]	0.715	0.6099	0.792	0.7734	0.765	0.6547	0.899	0.8792	0.935	0.9021	0.794	0.7331	0.853	0.6691	0.851	0.8266	0.8255	0.7560
KT-pFL [30]	0.765	0.6733	0.87	0.8331	0.755	0.6432	0.885	0.8621	0.935	0.9021	0.78	0.6931	0.865	0.6819	0.905	0.9023	0.8450	0.7739
MH-pFLGB (Ours)	**0.830**	**0.7076**	**0.945**	**0.9399**	**0.820**	**0.7796**	**0.975**	**0.9608**	**0.962**	**0.9452**	**0.821**	**0.7080**	**0.893**	**0.7015**	**0.995**	**0.9928**	**0.9038**	**0.8419**
OCT disease classification																
Method	ResNet		shufflenetv2		ResNeXt		squeezeNet		SENet		MobileNet		DenseNet		VGG		Average	
	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑	ACC ↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑	ACC↑	MF1↑
Only Local Training	0.9162	0.9099	0.8922	0.8918	0.8694	0.8253	0.8472	0.8361	0.9388	0.9311	0.914	0.7236	0.9054	0.9	0.9262	0.9077	0.9012	0.8657
FedMD [12]	0.8828	0.8349	0.8856	0.8531	0.8246	0.7822	0.8254	0.8021	0.8552	0.8321	0.9254	0.7542	0.9254	0.9119	0.9552	0.9293	0.8850	0.8375
FedDF [16]	0.854	0.8229	0.913	0.8936	0.865	0.8241	0.8054	0.7749	0.8926	0.8733	0.9178	0.7361	0.8958	0.8831	0.963	0.9308	0.8883	0.8424
pFedDF [16]	0.9364	0.9152	0.92	0.913	0.881	0.8327	0.863	0.8239	0.941	0.8952	0.931	0.7249	0.897	0.8829	0.961	0.9234	0.9163	0.8639
DS-pFL [9]	0.8432	0.8079	0.864	0.8604	0.874	0.8356	0.835	0.7449	0.8874	0.8821	0.8998	0.7532	0.8592	0.8264	0.8814	0.8731	0.8680	0.8230
KT-pFL [30]	0.9532	0.9392	0.965	0.963	0.8594	0.8466	0.9136	0.9067	0.955	0.943	0.9622	0.8099	0.9038	0.8794	0.927	0.9022	0.9299	0.8988
MH-pFLGB (Ours)	**0.9634**	**0.9507**	**0.9950**	**0.9864**	**0.8824**	**0.8539**	**0.9672**	**0.9573**	**0.9714**	**0.9639**	**0.9674**	**0.8055**	**0.9088**	**0.9020**	**0.9496**	**0.9288**	**0.9506**	**0.9186**

Implementation Details. MH-pFLGB adopts learning rates of 1×10^{-4} and 1×10^{-5} for the local model training and global model training stage, respectively. The batch size is set to 8. In experiments, all frameworks have a communication round of 100. Local training epochs are 5 (4 epochs in the first stage and 1 round in the second stage for MH-pFLGB). For classification, all of the loss functions are cross-entropy loss, and all of the loss functions are Dice loss for segmentation tasks. λ^{l}_{local} and λ^{g}_{glob} are set to 0.9. λ^{g}_{loc} and λ^{l}_{glob} are 0.1. The performance evaluation of the classification task is based on two metrics, accuracy (*ACC*) and macro-averaged F1-score (*MF1*), providing a comprehensive assessment of the model's robustness [28]. Additionally, *Dice* is used to evaluate the segmentation task performance across frameworks [5]. Moreover, we implement MH-pFLGB using PyTorch 1.10 [21] and train it on an NVIDIA GeForce RTX 3090 Ti GPU. We have included more baseline, datasets, training settings, and model structure details in the supplementary materials.

4 Results and Discussion

4.1 Medical Image Classification (Different Resolutions)

In this task, we employ the ResNet family model to train breast cancer medical images under different resolutions. The higher the image resolution of this client, the deeper

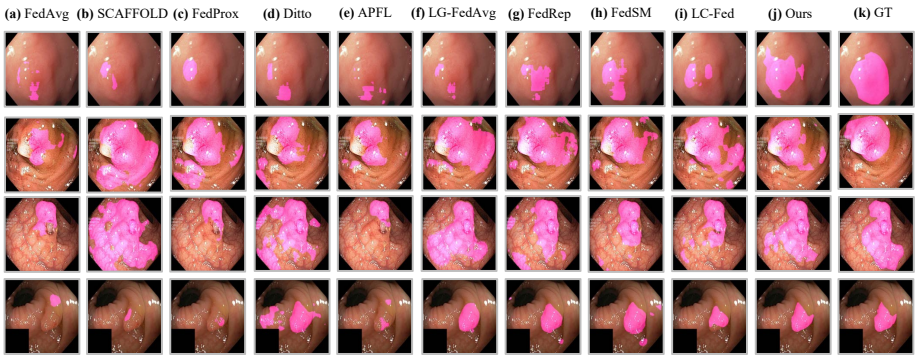

Fig. 2: Visualized comparison of Federated Learning in medical image segmentation. We randomly select four samples from different clients to form the visualization. (a-j) Segmentation results by a model trained with FedAVG, SCAFFOLD, FedProx, Ditto, APFL, LG-FedAvg, FedRep, FedSM, LC-Fed, and our method MH-pFLGB; (k) Ground truths (denoted as 'GT').

and more complex the model we adopt. In Table 1, compared to other federated learning frameworks, MH-pFLGB achieves the best performance. This indicates that MH-pFLGB, based on the global bypass model paradigm, effectively enables local heterogeneous models within the same family to fuse global knowledge, thereby enhancing the performance of local models. Furthermore, MH-pFLGB demonstrates a more significant advantage in terms of the MF1 metric, highlighting its ability to improve the robustness of local heterogeneous models.

4.2 Medical Image Classification (Different Label Distributions)

In Table 2, the experimental results for the medical image classification task with different label distributions, where each client uses heterogeneous models, show that MH-pFLGB achieves the optimal results. This demonstrates that, compared to heterogeneous federated learning methods based on soft predictions, the global bypass model approach of MH-pFLGB has advantages. It can more effectively utilize knowledge from other clients to guide local client learning. Compared to only local training, MH-pFLGB enhances the local performance of each heterogeneous model. This indicates that our proposed feature weighted fusion method fuses global and local features well, thereby improving the performance of local models.

4.3 Medical Image Segmentation

We validate the effectiveness of MH-pFLGB in medical image segmentation tasks. Table 3 presents the results of previous federated learning frameworks in the segmentation task, demonstrating that MH-pFLGB achieves the best outcomes. This indicates that our framework can effectively fuse global features and local heterogeneous model features from various clients, thus performing well in various downstream tasks. Meanwhile,

Table 3: For the medical image segmentation task, we evaluate the *Dice* result on Polyp dataset. The larger the better. Row "Only Local Training" and "MH-pFLGB" use heterogeneous models in each client. The four client models are set to Unet++, FCN, Unet, and ResUnet, respectively. For other methods, their clients use the unified model settings (Unet). MH-pFLGB achieves the best segmentation results.

Methods	Client1	Client2	Client3	Client4	Average
FedAvg [20]	0.5249	0.4205	0.5676	0.5500	0.5158
FedAvg+FT	0.6047	0.4762	0.7513	0.6681	0.6251
SCAFFOLD [10]	0.5244	0.3591	0.5935	0.5713	0.5121
SCAFFOLD+FT	0.5937	0.4312	0.8231	0.7208	0.6422
FedProx [13]	0.5529	0.4674	0.5403	0.6301	0.5477
FedProx+FT	0.7441	0.5701	0.7438	0.6402	0.6746
Ditto [14]	0.5720	0.4644	0.6648	0.6416	0.5857
APFL [2]	0.6120	0.5095	0.6333	0.5892	0.5860
LG-FedAvg [15]	0.6053	0.5062	0.7371	0.5596	0.6021
FedRep [1]	0.5809	0.3106	0.7088	0.7023	0.5757
FedSM [5]	0.6894	0.6278	0.8021	0.7391	0.7146
LC-Fed [27]	0.6233	0.4982	0.8217	0.7654	0.6772
Only Local Training	0.7049	0.4906	0.8079	0.7555	0.6897
MH-pFLGB (Ours)	**0.7525**	**0.7010**	**0.8469**	**0.7769**	**0.7693**

Table 4: Ablation studies of MH-pFLGB.

Methods	Breast Cancer		Segmentation
	ACC↑	MF1↑	Dice↑
MH-pFLGB	**0.9038**	**0.8419**	**0.7693**
w/o Global Head	0.8821	0.8192	0.7442
w/o Global Body	0.8295	0.7366	0.7198
w/o Features Weighted Fusion	0.8588	0.8009	0.7455
Only Local Training	0.7875	0.7005	0.6897

Table 5: GFLOPS and parameters of local heterogeneous models and messenger models in various tasks. The smaller the better. Among the four tasks, the GFLOPS and parameters of the messenger models are much smaller than those of the local models.

Tasks	Dataset	Model	GFLOPS	#Params
Medical Image Classification (Different Resolution)	BreaKHis (384x384x3 - 48x48x3)	ResNet17	3.495	4.231M
		ResNet11	0.667	2.104M
		ResNet8	0.140	1.558M
		ResNet5	0.044	1.359M
		Messeger	**0.01-0.07**	**0.035M**
Medical Image Classification (Different Label Distributions)	BreaKHis (384x384x3)	ResNet	10.020	11.111M
		Shufflenetv2	1.719	1.730M
		ResNeXt	41.245	7.930M
		squeezeNet	7.774	1.832M
		SENet	80.370	12.372M
		MobileNet	1.870	1.934M
		DenseNet	13.461	1.147M
		VGG	57.524	40.045M
		Messeger	**0.070**	**0.032M**
	OCT 2017 (256x256x1)	ResNet	4.351	11.090M
		Shufflenetv2	0.735	1.712M
		ResNeXt	18.256	7.910M
		squeezeNet	3.342	1.820M
		SENet	35.644	12.363M
		MobileNet	0.812	1.921M
		DenseNet	5.954	1.14M
		VGG	25.501	40.020M
		Messeger	**0.012**	**0.035M**
Medical Image Segmentation Task	ColonDB ETIS ClinicDB Kvasir-SEG (256x256x3)	Unet++	34.906	10.421M
		FCN	54.742	32.560M
		Unet	56.435	33.090M
		ResUnet	25.572	19.913M
		Messeger	**0.681**	**0.196M**

the visualization results in Fig. 2 show that the segmentation results of MH-pFLGB are closer to ground truth.

4.4 Ablation Experiments

To verify the effectiveness of MH-pFLGB's key components, we conduct a comparative analysis by removing each of the three elements (global head, global body, and feature-weighted fusion) during breast cancer classification tasks with different label distributions and segmentation tasks, as shown in Table 4. The experimental results indicate that more parameter sharing is beneficial for MH-pFLGB, and features weighted fusion effectively improves the performance of local heterogeneous models.

4.5 GFLOPS and Parameters

We compare the GFLOPS and parameter of the global bypass model with local heterogeneous models in three tasks. The results of Table 5 show that the GFLOPS and parameters of the global bypass model are much smaller than those of the local heterogeneous model on all the tasks.

5 Conclusion

MH-pFLGB can effectively solve the problems of statistic heterogeneity and system heterogeneity faced in federated learning. MH-pFLGB, based on the global bypass model paradigm, offers a solution to these issues. MH-pFLGB introduces a lightweight global bypass model in each client and designs a feature weighted fusion to fuse local and global knowledge. These can enable local heterogeneous models to capture information from other clients well under statistic heterogeneity. Numerous experiments have demonstrated that our method outperforms existing federated learning frameworks with heterogeneous models in multiple tasks.

Acknowledgements. This work was supported by the National Key R&D Program of China under Grant No.2022YFB2703301.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. In *ICML*, pp. 2089–2099. PMLR, 2021.
2. Deng, Y., Kamani, M. M., and Mahdavi, M. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.
3. Diakogiannis, F. I., Waldner, F., Caccetta, P., and Wu, C. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.
4. He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.
5. Xu, A.; Li, W.; Guo, P.; Yang, D.; Roth, H. R.; Hatamizadeh, A.; Zhao, C.; Xu, D.; Huang, H.; and Xu, Z. 2022. Closing the generalization gap of cross-silo federated medical image segmentation. In *CVPR*, 20866–20875.
6. Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.
7. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
8. Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
9. Itahara, S., Nishio, T., Koda, Y., Morikura, M., and Yamamoto, K. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. *IEEE Transactions on Mobile Computing*, 22(1):191–205, 2023. https://doi.org/10.1109/TMC.2021.3070013.
10. Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, pp. 5132–5143. PMLR, 2020.
11. Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

12. Li, D. and Wang, J. Fedmd: Heterogenous federated learning via model distillation. *CoRR*, abs/1910.03581, 2019.
13. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.
14. Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *ICML*, pp. 6357–6368. PMLR, 2021.
15. Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., Salakhutdinov, R., and Morency, L.-P. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.
16. Lin, T., Kong, L., Stich, S. U., and Jaggi, M. Ensemble distillation for robust model fusion in federated learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2351–2363. Curran Associates, Inc., 2020.
17. Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
18. Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.
19. Xie, L., Lin, M., Luan, T., Li, C., Fang, Y., Shen, Q., and Wu, Z. Mh-pflid: Model heterogeneous personalized federated learning via injection and distillation for medical data analysis. *arXiv preprint arXiv:2405.06822*.
20. McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pp. 1273–1282. PMLR, 2017.
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
22. Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.
23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.
24. Sattler, F., Müller, K.-R., and Samek, W. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.
25. Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
26. Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7): 1455–1462, 2016. https://doi.org/10.1109/TBME.2015.2496264.
27. Wang, J., Jin, Y., and Wang, L. Personalizing federated medical image segmentation via local calibration. In *ECCV*, pp. 456–472. Springer, 2022.
28. Xie, L., Li, C., Zhang, X., Zhai, S., Fang, Y., Shen, Q., and Wu, Z. Trls: A time series representation learning framework via spectrogram for medical signal processing, 2024.
29. Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

30. Zhang, J., Guo, S., Ma, X., Wang, H., Xu, W., and Wu, F. Parameterized knowledge transfer for personalized federated learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10092–10104. Curran Associates, Inc., 2021.
31. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.
32. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *IJCARS*, volume 9, pp. 283–293. 2014.
33. J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE TMI*, volume 35, pp. 630–644. 2015.
34. J. Bernal, F. J. S´ anchez, G. Fern´andez-Esparrach, D. Gil, C. Rodrıguez, and F. Vilarino Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *CMIG*, volume 43, pp. 99–111, 2015.
35. D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen Kvasir-seg: A segmented polyp dataset. *MMM*, 2020.
36. Xie, L., Li, C., Wang, Z., Zhang, X., Chen, B., Shen, Q., and Wu, Z. Shisrcnet: Super-resolution and classification network for low-resolution breast cancer histopathology image, 2023.