pFLMoE: Decentralized Federated Learning via Mixture of Experts for Medical Data Analysis: Supplementary Materials

Luyuan Xie[1,2,3*] Tianyu Luan[4†] Wenyuan Cai[1] Guochen Yan[2,3] Zhaoyu Chen[1,2,3]
Nan Xi[4] Yuejian Fang[1,2,3] Qingni Shen[1,2,3] Zhonghai Wu[1,2,3] Junsong Yuan[1]

[1]School of Software and Microelectronics, Peking University [2]PKU-OCTA Laboratory for Blockchain and Privacy Computing
[3] National Engineering Research Center for Software Engineering, Peking University [4]State University of New York at Buffalo

1. Tasks and Datasets

We verify the effectiveness of pFLMoE on 5 non-IID tasks. Here, we provide additional details about the tasks and the datasets used.

A. Medical image classification (different resolution). We use the Breast Cancer Histopathological Image Database (BreaKHis) [17]. We treat the original image as a high-resolution image. Then, the Bicubic downsampling method is used to downsample the high-resolution image, obtaining images with resolutions of x2↓, x4↓, and x8↓, respectively. Each resolution of medical images was treated as a separate client, resulting in four clients in total. Each client has the same number of images with consistent label distribution, but the image resolution is different for each client. The dataset for each client was randomly divided into training and testing sets at a ratio of 7:3, following previous work. In this task, we employed a family of models such as ResNet$\{17, 11, 8, 5\}$.

B. Medical image super-resolution. We use BreaKHis dataset [17]. We perform x2↓, x4↓, and x8↓ Bicubic downsampling methods on the high-resolution images [19]. Each downsampled resolution of medical images is treated as a client, resulting in three clients in total. The dataset for each client was randomly divided into training and testing sets at a ratio of 7:3, following previous work. We used the RCNN [4] for the model heterogeneous framework. We used SRResNet$\{6, 12, 18\}$ [11] for the model heterogeneous framework.

C. Medical time-series classification. We used the Sleep-EDF dataset [5] for the classification task of time series under Non-IID distribution. We divided the Sleep-EDF dataset evenly among three clients. The ratio of the training set to the testing set for each client is 8:2. We designed three clients using the TCN [1], Transformer [23] and RNN [20].

D. Medical image classification (different label distributions). This task includes a breast cancer classification task and an OCT disease classification task. We designed eight clients, each corresponding to a distinct heterogeneous model. These models included ResNet [6], ShuffleNetV2 [13], ResNeXt [21], SqueezeNet [9], SENet [7], MobileNetV2 [15], DenseNet [8], and VGG [16]. Similar to the previous approach, we applied non-IID label distribution methods to the BreaKHis (breast cancer classification) [10] and ODIR-5K (ocular disease recognition) across the 8 clients.

For the breast cancer classification task, we have filled in the data quantity to 8000 and allocated 1000 pieces of data to each client. The ratio of training set to testing set for each client is 8:2.

For theocular disease recognition task, we randomly selected 6400 pieces of data, with 800 pieces per client. The ratio of training set to test set is also 8:2.

E. Medical image segmentation. Here, we focus on polyp segmentation [3]. The dataset for this task consisted of endoscopic images collected and annotated from four different centers, with each center's dataset treated as a separate client. Thus, there were four clients in total for this task. The number of each client are 1000, 380, 196 and 612. The ratio of the training set to the testing set for each client is 1:1. Each client utilized a specific model, including Unet++ [24], FCN [12], Unet [14], and Res-Unet [2].

2. Implementation Details

For different tasks, pFLMoE adopts different learning rates of two-stage and batch size. The specific settings are shown in Tab. 1. In experiments, all frameworks have a communication round of 100. For classification, $\mathcal{L}_{loc}$ and $\mathcal{L}_{MoE}$ are cross-entropy loss. For super-resolution tasks, $\mathcal{L}_{loc}$ and $\mathcal{L}_{MoE}$ are $L1$ loss. And for segmentation tasks, $\mathcal{L}_{loc}$ and $\mathcal{L}_{MoE}$ are Dice and cross-entropy loss. λ_{loc} and λ_{MoE} are set to

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

†Tianyu Luan is corresponding author: tianyulu@buffalo.edu

0.5. 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. For super-resolution, we adopted the Peak-Signal to Noise Ratio (PSNR) and structural similarity index (SSIM) to evaluate the performance. Additionally, Dice is used to evaluate the segmentation task performance across frameworks.

3. Baselines

In the medical image classification task (different resolution), we selected FedAvg, SCAFFOLD, FedProx, FedRep, LG-FedAvg, APFL, and Ditto with homogeneous models. We chose MH-pFLID, FedMD, FedDF, pFedDF, DS-pFL, and KT-pFL for heterogeneous model federated learning.

For medical image super-resolution, we compared various approaches, including local training of clients and a variety of personalized federated learning techniques, as well as methods for learning a single global model. Among the personalized methods, we also chose FedRep, LG-FedAvg, APFL, and Ditto. We also compare MH-pFLID with our method (pFLMoE) under the heterogeneous model setting.

The baseline used in the medical time-series classification task is the same as the medical image classification task (different label distrbutions).

For image segmentation tasks, we compared various approaches, including local training of clients and a variety of personalized federated learning techniques, as well as methods for learning a single global model. Among the personalized methods, we also chose FedRep, LG-FedAvg, APFL, and Ditto. We simultaneously added LC-Fed [18] and FedSM [22] which are effective improvements for FedRep and APFL in the federated segmentation domain. We also compare MH-pFLID with our method (pFLMoE) under the heterogeneous model setting.

In the medical image classification task (different label distrbutions), we compared various methods, including local training of clients with heterogeneous models and existing heterogeneous model federated learning approaches (FedMD, FedDF, pFedDF, DS-pFL, and KT-pFL, MH-pFLID).

4. Training Settings

4.1. Evaluation Indicators

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. For the super-resolution task, we adopted the Peak Signal-to-Noise Ratio (PSNR) and structural similarity index (SSIM) to evaluate the performance. Additionally, Dice is used to evaluate the segmentation task performance across frameworks.

A. Accuracy. Accuracy is the ratio of the number of correct judgments to the total number of judgments.

B. Macro-averaged F1-score. First, calculate the F1-score for each recognition category, and then calculate the overall average value.

C. Peak Signal-to-Noise Ratio. The formula for Peak Signal-to-Noise Ratio ($PSNR$) is typically written as:

$$PSNR = 10 * log_{10}(\frac{R^2}{MSE}), \tag{1}$$

where R is the maximum possible pixel value in the image (for example, for an 8-bit image, R=255). MSE is the Mean Squared Error, calculated as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(I(i) - K(i))^2, \tag{2}$$

where $I(i)$ and $K(i)$ are the pixel values of the original image and the reconstructed image at position i, and N is the total number of pixels in the image.

D. Structural similarity index (SSIM). First, calculate the F1-score for each recognition category, and then calculate the overall average value.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{3}$$

where x and y are the two images being compared, μ_x and μ_y are the average luminance of x and y. σ_x^2 and σ_y^2 are the variances of x and y. σ_{xy} is the covariance between x and y. C_1 and C_2 are small constants to stabilize the division (typically $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$, where L is the dynamic range of the pixel values).

E. Dice. It is a set similarity metric commonly used to calculate the similarity between two samples, with a threshold of [0,1]. In medical images, it is often used for image segmentation, with the best segmentation result being 1 and the worst result being 0. The Dice coefficient calculation formula is as follows:

$$Dice = \frac{2 * (pred \cap true)}{pred \cup true} \tag{4}$$

Among them, $pred$ is the set of predicted values, $true$ is the set of groudtruth values. And the numerator is the intersection between $pred$ and $true$. Multiplying by 2 is due to the repeated calculation of common elements between $pred$ and $true$ in the denominator. The denominator is the union of $pred$ and $true$.

4.2. Loss Function

Many loss functions have been applied in this article, and here are some explanations for them.The cross entropy loss function is very common and will not be explained in detail here. We mainly explain Dice loss.

Dice Loss applied in the field of image segmentation. It is represented as:

Table 1. The two-stage learning rates and batch size of pFLMoE under 5 tasks.

	Medical image classification (different resolution)	**Medical image super-resolution**	**Medical time-series classification**	**Medical image classification (different label distributions)**	**Medical image segmentation**
Learning rate of Local Network Training	0.0005	0.0001	0.001	0.001	0.001
Learning rate of Mixture of Experts Decision	0.0001	0.00001	0.0001	0.0001	0.0001
Batch size	32	16	256	8	8

$$DiceLoss = 1 - \frac{2 * (pred \cap true)}{pred \cup true} \tag{5}$$

The Dice loss and Dice coefficient are the same thing, and their relationship is:

$$DiceLoss = 1 - Dice \tag{6}$$

In super-resolution task, we use L_1 loss to optimize the model.

4.3. Public Datasets for other Federated Learning of Heterogeneous Models

In this section, we mainly describe the setting of public datasets for methods such as FedMD, FedDF, DS-pFL and KT-pFL.

A. Medical image classification (different resolution). We select 100 pieces of data from each client and put them into the central server as public data, totaling 400 pieces of data as public data. In order to better obtain soft predictions for individual clients, the image resolution of the publicly available dataset will be resized to the corresponding resolution for each client.

B. Medical image classification (different label distributions). For the breast cancer classification task, we select 50 pieces of data for each client to upload, and the public dataset contains 400 images. For the Ocular Disease Recognition task, we also select 50 pieces of data for each client to upload, and the public dataset contains 400 images.

C. Medical time-series classification. We select 200 pieces of data for each client to upload, and the public dataset contains 600 images.

References

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 1

[2] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 1

[3] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021. 1

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 1

[5] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[9] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 1

[10] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018. 1

[11] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[13] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 1

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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, pages 234–241. Springer, 2015. 1

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[17] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016. 1

[18] Jiacheng Wang, Yueming Jin, and Liansheng Wang. Personalizing federated medical image segmentation via local calibration. In *ECCV*, pages 456–472. Springer, 2022. 2

[19] Luyuan Xie, Cong Li, Zirui Wang, Xin Zhang, Boyan Chen, Qingni Shen, and Zhonghai Wu. Shisrcnet: Super-resolution and classification network for low-resolution breast cancer histopathology image, 2023. 1

[20] Luyuan Xie, Cong Li, Xin Zhang, Shengfang Zhai, Yuejian Fang, Qingni Shen, and Zhonghai Wu. Trls: A time series representation learning framework via spectrogram for medical signal processing, 2024. 1

[21] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1

[22] An Xu, Wenqi Li, Pengfei Guo, Dong Yang, Holger R Roth, Ali Hatamizadeh, Can Zhao, Daguang Xu, Heng Huang, and Ziyue Xu. Closing the generalization gap of cross-silo federated medical image segmentation. In *CVPR*, pages 20866–20875, 2022. 2

[23] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff. A transformer-based framework for multivariate time series representation learning. 2021. 1

[24] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019. 1