FedIIC: Towards Robust Federated Learning for Class-Imbalanced Medical Image Classification

Nannan Wu[1], Li Yu[1], Xin Yang[1], Kwang-Ting Cheng[2], and Zengqiang Yan[1](✉)

[1] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China
{wnn2000,hustlyu,xinyang2014,z_yan}@hust.edu.cn

[2] School of Engineering, Hong Kong University of Science and Technology, Hong Kong, China
timcheng@ust.hk

Abstract. Federated learning (FL), training deep models from decentralized data without privacy leakage, has shown great potential in medical image computing recently. However, considering the ubiquitous class imbalance in medical data, FL can exhibit performance degradation, especially for minority classes (*e.g.* rare diseases). Existing methods towards this problem mainly focus on training a balanced classifier to eliminate class prior bias among classes, but neglect to explore better representation to facilitate classification performance. In this paper, we present a privacy-preserving FL method named FedIIC to combat class imbalance from two perspectives: feature learning and classifier learning. In feature learning, two levels of contrastive learning are designed to extract better class-specific features with imbalanced data in FL. In classifier learning, per-class margins are dynamically set according to real-time difficulty and class priors, which helps the model learn classes equally. Experimental results on publicly-available datasets demonstrate the superior performance of FedIIC in dealing with both real-world and simulated multi-source medical imaging data under class imbalance. Code is available at https://github.com/wnn2000/FedIIC.

Keywords: Federated learning · Class imbalance · Contrastive learning · Classification

1 Introduction

Federated learning (FL), allowing decentralized data sources to train a unified deep learning model collaboratively without data sharing, has drawn great attention in medical imaging due to its privacy-preserving properties [13,22,25,40]. Existing studies of FL mainly focus on data heterogeneity across clients [19,20,31], while ignoring the widely-existed class imbalance problem in

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_65.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 692–702, 2023.
https://doi.org/10.1007/978-3-031-43895-0_65

medical scenarios. In clinical practice, the number of samples for different diseases may vary greatly due to varying incidence rates in the population. When conducting FL on cooperative medical institutions with global class-imbalanced data, the global model may suffer from significant performance degradation, which typically manifests as the recognition accuracy of minority classes (*e.g.* rare diseases) being lower than that of majority classes (*e.g.* common diseases) [34]. Deploying such a biased global/federated model is fatal, especially for misdiagnosing a rare disease [15,42]. Therefore, addressing class imbalance in federated learning is of great value.

Several FL frameworks have been proposed to tackle imbalanced data [9,41]. Following re-weighting [7], Wang *et al.* [39] presented a weighted form of cross entropy loss named ratio loss depending on a balanced auxiliary dataset for the server to calculate weights. Sarkar *et al.* [33] introduced focal loss [24] to up-weight hard samples. CLIMB [35] assigned larger weights to clients more likely to own minority classes via a meta-algorithm. Inspired by decoupling [17], CReFF [34] retrained a new classifier with balanced synthetic features in the server. All these methods aim to balance classes from the classifier perspective without exploring better representations with class-imbalanced data for performance improvement.

In this paper, we formulate the effect of class imbalance in FL into the attribute bias and the class bias [37]. The attribute bias means minority classes have more imbalanced background attributes in their class-specific attributes compared to majority classes, making them less distinguishable. The class bias represents the difference in prior probabilities across classes, resulting in biased predictions toward majority classes. To handle the two biases, we present a new class-balancing FL method named **FedIIC** from two perspectives: feature learning and classifier learning. The key idea of FedIIC is to alleviate the two biases through the calibration of the feature extractor and the classifier. Specifically, two-level supervised contrastive learning [18], *i.e.* intra- and inter-client contrastive learning, is built to calibrate the feature extractor for better feature learning. For classifier learning, difficulty-aware logit adjustment is adopted to calibrate the classifier dynamically for better decision boundaries. Extensive comparison experiments on both real-world and simulated multi-source data validate FedIIC's effectiveness.

The main contributions are summarized as follows. (1) A new viewpoint of realistic medical FL scenarios where global training data is class-imbalanced. (2) A novel privacy-preserving framework FedIIC for balanced federated learning. (3) Superior performance in dealing with class imbalance under both real-world and simulated multi-source decentralized settings.

2 Methodology

2.1 Preliminaries and Overview

Considering a typical FL scenario for multi-class image classification with K participants, each participant is assumed to own a private dataset $D_k = \{(x_i, y_i)\}_{i=1}^{N_k}$, $k \in [K]$, where N_k is the data amount of D_k, and denote each image-label pair as $(x_i \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \mathcal{Y} = [L])$. The goal of FL is to

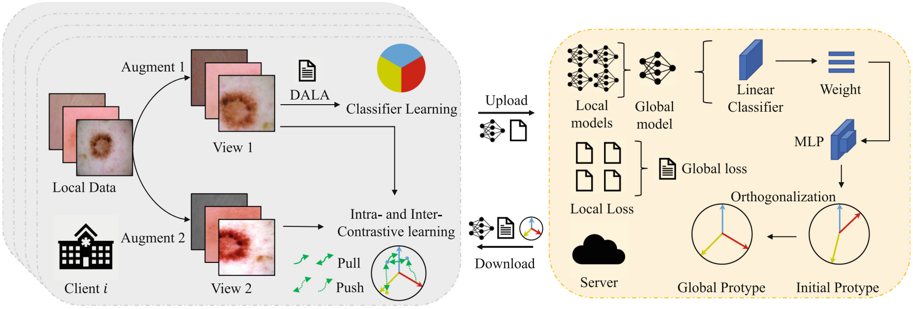

Fig. 1. Overview of the proposed FedIIC.

train a global model $f(g(\cdot))$ with the union of all cooperative data sources $D := \cup_{k\in[K]}D_k$ without privacy leakage, where $f(\cdot)$ and $g(\cdot)$ represent the linear classifier and the feature extractor respectively. Note that D is set as class-imbalanced in this paper.

Assuming each image has two kinds of latent attributes, *i.e.* $\mathcal{Z}_c$ and $\mathcal{Z}_a$, representing the class-specific attributes (determining the category of the image, *e.g.* texture, color, *etc.*) and the variant background attributes (*e.g.* brightness, contrast, *etc.*) respectively [37], based on the Bayes theorem, the posterior probability of classification can be formulated as

$$P(y \mid x) = P(y \mid \mathcal{Z}_c, \mathcal{Z}_a) = \frac{p(\mathcal{Z}_c \mid y)}{p(\mathcal{Z}_c)} \cdot \frac{p(\mathcal{Z}_a \mid y, \mathcal{Z}_c)}{p(\mathcal{Z}_a \mid \mathcal{Z}_c)} \cdot p(y), \tag{1}$$

where the last two items represent the attribute bias and the class bias respectively, which widely exist in class-imbalanced data and affect the posterior probability. For robust FL with class-imbalanced data, the key idea is to alleviate the two biases simultaneously, instead of focusing on the latter as [34]. Hence, we propose FedIIC to address class imbalance from the two perspectives as illustrated in Fig. 1. Details are presented in the following.

2.2 Intra-Client Contrastive Learning

Limited local data affects data diversity (*i.e.*, limited $(\mathcal{Z}_c, \mathcal{Z}_a)$ combinations), especially for minority classes, making $\mathcal{Z}_c$ less distinguishable. To emphasize more on the learning of $\mathcal{Z}_c$, supervised contrastive learning (SCL), proven to be effective for representation learning [16,21,27,45], is introduced in local training. The basic loss function of SCL can be formulated as

$$\mathcal{L}_{SCL} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{j\in P(i)} \log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{a\in A(i)} \exp(z_i \cdot z_a/\tau)}, \tag{2}$$

where I denotes the index set of the multi-view batch generated by different augmentations (*e.g.* the two views in Fig. 1), $|\cdot|$ measures the number of elements in a set, $A(i) = I \backslash \{i\}, P(i) = \{s \in A(i) | y_s = y_i\}$, τ represents the temperature, and z denotes the l_2-normalized embedding of a sample x. Note that in this paper, we use a 2-layer MLP $h(\cdot)$ to obtain z before it is normalized as [3], *i.e.* $z = \frac{h(g(x))}{\|h(g(x))\|_2}$. In the multi-view batch, $\mathcal{L}_{SCL}$ keeps the embeddings of the same class closer while pushing the embeddings of different classes further away, which helps the model learn better $\mathcal{Z}_c$ of each class due to richer $\mathcal{Z}_a$. However, SCL can not perfectly address class imbalance as the majority classes would benefit more from Eq. 2 following traditional training losses (*e.g.* the cross entropy loss). To overcome this problem, we propose to employ a dynamic temperature $\tau' := P\tau = (p^i p^j)^t \tau$ in Eq. 2 inspired by [16,45], where p^i is the prior probability of class i in the local dataset and t is a parameter set as 0.5 by default. Hence, the loss function is rewritten as

$$\mathcal{L}_{Intra} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(z_i \cdot z_j / \tau')}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau')}, \tag{3}$$

named intra-client contrastive learning. Through P, sample pairs of the minority classes are up-weighted compared to those of the majority classes, leading to better balance.

2.3 Inter-client Contrastive Learning

Given limited local data under FL, the effectiveness of intra-client contrastive learning may be bounded. How to better utilize cross-client data from the global perspective is crucial for further performance improvement. Inspired by learning from prototypes [4,12,31], we propose inter-client contrastive learning. Assuming a set of shared class-wise prototypes $V = \{v^1, v^2, ..., v^L\}$ across clients, the local model can be trained by

$$\mathcal{L}_{Inter} = \sum_{i \in I} \frac{-1}{|P(i)|} \log \frac{\exp(z_i \cdot v^{y_i} / \tau)}{\sum_{j=1}^{L} \exp(z_i \cdot v^j / \tau)}, \tag{4}$$

where y_i is the label of sample i. When minimizing $\mathcal{L}_{Inter}$, the embedding of each sample will get closer to the prototype of the same class while farther from the prototypes of different classes, encouraging local models to learn common attributes (*i.e.* class-specific attributes) for samples with the same classes.

To this end, how to produce high-quality prototypes is the key to inter-client contrastive learning. In previous studies, one common method to generate prototypes is uploading and aggregating local information. For example, Mu *et al.* [31] and Chen *et al.* [4] uploaded features to the server directly to generate prototypes. However, it may cause privacy leakage under well-designed attacks and will introduce extra communication costs. Different from these methods, in FedIIC, we propose a new method to generate global prototypes without uploading extra information. Considering that the essence of linear classification is similarity calculation based on vector inner product, the weights of a well-trained linear classifier are nearly co-linear with the feature vectors of different classes [11,32,45]. Therefore, the weights of a linear classifier denoted as $W = \{w^1, w^2, ..., w^L\}$, can represent the corresponding features of L classes learned by the feature extractor $g(\cdot)$ to some extent. Specifically, given a global model $[f_g(\cdot), g_g(\cdot), h_g(\cdot)]$ after model aggregation in the server, the weights of $g_g(\cdot)$ are fed to $h_g(\cdot)$ to calculate the initial prototypes $\widetilde{V} = \{\widetilde{v}^1, \widetilde{v}^2, ..., \widetilde{v}^L\}$ as shown in Fig. 1. Considering that features of different classes should have low inter-class similarity, we further fine-tune $\widetilde{V}$ via gradient descent by

$$\widetilde{V} \leftarrow \widetilde{V} - \nabla \sum_{i \in Y} \max_{j \in Y, j \neq i} \left(\frac{\widetilde{v}^i}{\|\widetilde{v}^i\|_2} \cdot \frac{\widetilde{v}^j}{\|\widetilde{v}^j\|_2} \right). \tag{5}$$

In this way, the cosine similarity of any $(\widetilde{v}^i, \widetilde{v}^j)$ pair in $\widetilde{V}$ is minimized to be equal, resulting in $\widetilde{V}$ with lower inter-class similarity. This operation is called orthogonalization. Finally, the class-wise prototypes V are defined as the element-wise l_2-normalization of $\widetilde{V}$ and are sent to clients for inter-client contrastive learning.

2.4 Difficulty-Aware Logit Adjustment

After calibrating the feature extractor $g(\cdot)$, one common method to calibrate the linear classifier $f(\cdot)$ is logit adjustment (LA) [2,30] to alleviate the impact of class imbalance in local training. Specifically, Zhang *et al.* [43] proposed to add per-class margins to logits and re-compute the cross entropy (CE) loss by

$$\mathcal{L}_{LA} = \sum_{i \in I} -\log \frac{\exp(f(g(x_i))_{y_i} - \delta_{y_i})}{\sum_{y' \in \mathcal{Y}} \exp(f(g(x_i))_{y'} - \delta_{y'})}, \tag{6}$$

where δ_y denotes the positive per-class margin and is inversely proportional to the local class frequency $p(y)$. In this way, during local training, the logits of minority classes will increase to compensate for the item, which in turn trains the model to emphasize more on minority classes. However, the frequency-dependent margin may not be appropriate for medical data. For instance, some disease types/classes may have large intra-class variations and are difficult to diagnose even with a large amount of data, which may result in even smaller per-class margins. To address this, in FedIIC, the per-class margin is calculated based on not only the class frequency but also difficulties inspired by [44]. Specifically, we define $\delta_y := \log([\bar{l}_{ce}(y)]^q / p(y))$, where $\bar{l}_{ce}(y)$ is the average CE loss of all samples belonging to class y in any round and q is a hyper-parameter set as 0.25 by default. $\bar{l}_{ce}(y)$ is calculated as follows. At any round r, the total sample number of class y, denoted as N_r^y, belonging to clients of communication is first calculated. After receiving the global model from the server and before local training, each client i uploads $l_{ce}^i(y)$, *i.e.* the total loss of class y, to the server. Finally, $\bar{l}_{ce}(y)$ is calculated as $\frac{1}{N_r^y} \sum_i l_{ce}^i(y)$. This process to calculate average loss value can be privacy-preserving under the existing secure multi-party computation framework based on homomorphic encryption [35]. Based on the newly defined δ_y, Eq. 6 is renamed as $\mathcal{L}_{DALA}$. Note that the calculation of $\mathcal{L}_{DALA}$ does not rely on the multi-view batch like $\mathcal{L}_{Intra}$ and $\mathcal{L}_{Inter}$. For a fair

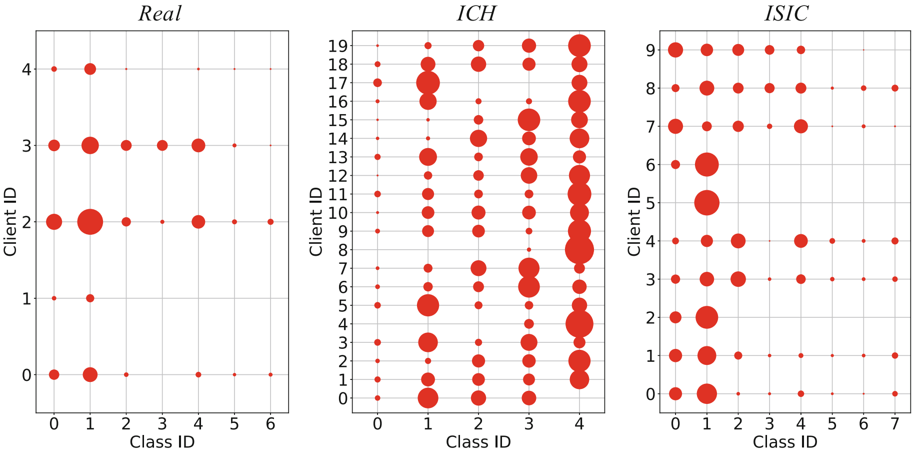

Fig. 2. Illustration of imbalanced data distributions. The radius of each solid circle represents each client's data amount of a specific class.

comparison with other methods trained by the CE loss, only one view of the multi-view batch is used to calculate $\mathcal{L}_{DALA}$. The overall loss function in local training is written as

$$\mathcal{L} = \mathcal{L}_{DALA} + k_1\mathcal{L}_{Intra} + k_2\mathcal{L}_{Inter}, \tag{7}$$

where k_1 and k_2 are trade-off hyper-parameters. After minimizing $\mathcal{L}$ during the local training phase of each client, the global model is updated by FedAvg [28].

3 Experiments

Datasets. Three FL scenarios with class-imbalanced global data are used for evaluation, which are described as follows:

1. Real Multi-Source Dermoscopic Image Datasets (denoted as ***Real***) consisting of five data sources from three datasets, including PH^2 [29], Atlas [1], and HAM10000 [38] where each source is treated as an individual client. For evaluation, we construct a separate test set by randomly sampling from the training set of ISIC 2019 [5,38] and ensure that the test set has no overlap with the above five data sources.
2. Intracranial Hemorrhage Classification (denoted as ***ICH***). The RNSA ICH dataset [10], containing five ICH subtypes, is adopted for experiments. The same pre-processing strategies in [14,26] are adopted, and images with only one single hemorrhage type are selected. Following [14,26], data is split according to 7:1:2 for training, validation, and testing respectively. To simulate heterogeneous multi-source data, following [34], Dirichlet distribution, *i.e.* $Dir(\alpha = 1.0)$, is used to divide the training set to 20 clients.

Table 1. Quantitative comparison results under the *Real*, *ISIC*, and *ICH* settings. For *Real*, the average results (%) from the last five rounds are reported. For *ISIC* and *ICH*, the results (%) based on the best model (evaluated by the validation set) on the testing set are reported. The best results are marked in bold.

Methods	Year	Datasets								
		Real			*ISIC*			*ICH*		
		BACC	F1	ACC	BACC	F1	ACC	BACC	F1	ACC
FedAvg [28]	AISTATS'17	45.21	44.47	44.57	49.41	54.31	72.50	73.75	77.35	84.83
FedProx [20]	MLSys'20	45.61	44.90	44.89	69.00	69.46	80.50	79.62	82.45	86.78
MOON [19]	CVPR'21	44.40	43.28	43.68	66.31	71.27	81.38	77.05	78.81	84.87
FedProc [31]	FGCS'23	38.83	37.98	39.36	31.16	35.45	66.88	73.29	76.23	84.89
FedRS [23]	KDD'21	45.23	44.50	44.46	24.93	26.01	61.39	72.44	76.51	84.13
FedLC [43]	ICML'22	46.73	45.88	45.60	45.84	41.89	70.33	76.53	78.96	84.92
FedFocal [33]	IJCAI'20	44.00	43.31	42.96	47.68	38.29	56.99	63.04	54.80	52.30
PRR-Imb [4]	TMI'22	50.49	47.60	47.48	49.97	46.52	68.18	71.72	69.98	78.85
CLIMB [35]	ICLR'22	46.07	45.91	45.86	49.70	52.32	71.65	72.64	76.08	84.73
CReFF [34]	IJCAI'22	51.13	48.56	49.46	71.52	57.83	72.92	82.21	74.64	81.63
FedIIC (ours)	-	**55.12**	**51.57**	**51.67**	**78.84**	**78.05**	**85.71**	**84.22**	**84.73**	**87.77**

3. Skin Lesion Classification (denoted as ***ISIC***).The training data of ISIC 2019 [5,38], containing eight classes, is used for evaluation. Following [14,26], we split the dataset by 7:1:2 for training, validation, and testing respectively. Similarly, Dirichlet distribution, *i.e.* $Dir(\alpha = 1.0)$, is used to generate highly heterogeneous data partitions of 10 clients.

Data distributions of the three training settings are illustrated in Fig. 2, and imbalance ratios are 35.43, 19.59 and 57.60, respectively.

Implementation Details. EfficientNet-B0 [36], pre-trained by ImageNet [8], is adopted as the backbone trained by an Adam optimizer with betas as 0.9 and 0.999, a weight decay as 5e-4, constant learning rates of 1e-4 for *Real* and 3e-4 for both *ICH* and *ISIC*, and a batch size of 32. For *ICH*, the multi-view batch for contrastive learning is generated by following [14,26]. For both *Real* and *ISIC*, the multi-view batch is generated by 1) RandAug [6] and 2) SimAugment [3]. The hyper-parameters k_1 and k_2 in Eq. 7 are set as 2.0. For federated training, the local training epoch is set as 1 and the global training round is set as 200 for *ICH* and *ISIC* and 30 for *Real*. At each round, all clients (*i.e.*, 100%) are included for model aggregation.

3.1 Comparison with State-of-the-Art Methods

Ten related approaches are included for comprehensive comparison, including FedAvg [28], FedProx [20] addressing data heterogeneity, MOON [19] and FedProc [31] utilizing contrastive learning in FL, FedFocal [33] utilizing focal loss

Table 2. Component-wise study.

FedAvg	DALA	Intra	Inter	*ISIC*		*ICH*	
				BACC	F1	BACC	F1
✓				49.41	54.31	73.75	77.35
✓	✓			50.65	42.12	81.26	76.47
✓	✓	✓		51.30	43.84	81.96	82.68
✓	✓		✓	75.78	76.55	83.81	82.39
✓	✓	✓	✓	**78.84**	**78.05**	**84.22**	**84.73**

Table 3. Parameter-wise study.

Param		BACC	F1
t	0.0	76.33	75.43
	0.5	**78.84**	**78.05**
orth.	w/o	72.64	75.34
	w	**78.84**	**78.05**
d	0.0	77.32	**78.41**
	0.25	**78.84**	78.05

[24] for balancing, FedRS [23] addressing the class-missing problem, FedLC [43] applying frequency-dependent logits adjustment in FL, PRR-Imb [4] training personalized models with heterogeneous and imbalanced data, and CLIMB [35] and CReFF [34] addressing class-imbalance global data in FL. All the methods share the same experimental details described above for a fair comparison. *More implementation details and visualization results can be found in supplemental materials.*

Following the ISIC 2019 competition, balanced accuracy (BACC) is used as the primary metric for class-imbalanced testing sets. Two key metrics in classification, *i.e.* F1 score (F1) and accuracy (ACC) are also employed for evaluation. Comparison results are summarized in Table 1. As can see, FedIIC achieves the best performance against all previous methods across the three metrics, outperforming the second-best approach (CReFF) by 3.99%, 7.32%, and 2.01% in BACC on *Real*, *ISIC*, and *ICH* respectively.

3.2 Ablation Study

To validate the effectiveness of each component in FedIIC, a series of ablation studies are conducted on *ISIC* and *ICH* following the same experimental details described in Sect. 3. Quantitative results are summarized in Table 2. Under severe global imbalance, FedAvg is struggling. With the introduction of DALA, the performance is improved in BACC but degraded in F1. It is consistent with the quantitative results between CReFF and FedAvg on *ICH* in Table 1, indicating the limitation of only eliminating class bias through classifier calibration while ignoring attribute bias. The above results validate the necessity of addressing the imbalance in feature learning for performance improvement. Therefore, introducing either intra- or inter-client contrastive learning for better representation learning under class imbalance is beneficial in both BACC and F1. By combining all the components, FedIIC achieves the best overall performance, outperforming FedAvg with large margins.

Ablation studies of hyper-parameters in FedIIC are conducted on *ISIC* as stated in Table 3. Setting $t = 0$ encounters noticeable performance degradation, indicating the necessity of dynamic temperatures based on class priors in intra-client contrastive learning. Meanwhile, the performance gap between the initial prototypes $\widetilde{V}$ with and without orthogonalization validates the effectiveness of reducing inter-class similarity in prototypes. When introducing difficulty to logit adjustment (*i.e.*, $d = 0.25$), we observe an increase in BACC and a decrease in F1, which is consistent with the above analysis in Table 1 (*i.e.*, CReFF vs. FedAvg).

4 Conclusion

This paper discusses a more realistic federated learning (FL) setting in medical scenarios where global data is class-imbalanced and presents a novel framework FedIIC. The key idea behind FedIIC is to calibrate both the feature extractor and the classification head to simultaneously eliminate attribute biases and class biases. Specifically, both intra- and inter-client contrastive learning are introduced for balanced feature learning, and difficulty-aware logit adjustment is deployed to balance decision boundaries across classes. Experimental results on both real-world and simulated medical FL scenarios demonstrate FedIIC's superiority against the state-of-the-art FL approaches. We believe that this study is helpful to build real-world FL systems for clinical applications.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grants 62202179 and 62271220, in part by the Natural Science Foundation of Hubei Province of China under Grant 2022CFB585, and in part by the Research Grants Council GRF Grant 16203319. The computation is supported by the HPC Platform of HUST.

References

1. Argenziano, G., et al.: Interactive atlas of dermoscopy (2000)
2. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: NeurIPS, vol. 32 (2019)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML, pp. 1597–1607 (2020)
4. Chen, Z., Yang, C., Zhu, M., Peng, Z., Yuan, Y.: Personalized retrogress-resilient federated learning toward imbalanced medical data. IEEE Trans. Med. Imaging **41**(12), 3663–3674 (2022)
5. Combalia, M., et al.: BCN20000: dermoscopic lesions in the wild. arXiv:1908.02288 (2019)
6. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: RandAugment: practical automated data augmentation with a reduced search space. In: NeurIPS (2020)
7. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR, pp. 9268–9277 (2019)
8. Deng, J., et al.: ImageNet: a large-scale hierarchical image database. In: CVPR, pp. 248–255 (2009)
9. Duan, M., et al.: Self-balancing federated learning with global imbalanced data in mobile systems. IEEE Trans. Parallel Distrib. Syst. **32**(1), 59–71 (2020)
10. Flanders, A.E., et al.: Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol. Artif. Intel. **2**(3), e190211 (2020)
11. Graf, F., Hofer, C., Niethammer, M., Kwitt, R.: Dissecting supervised constrastive learning. In: ICML, pp. 3821–3830 (2021)

12. Guo, Q., Qi, Y., Qi, S., Wu, D.: Dual class-aware contrastive federated semi-supervised learning. arXiv:2211.08914 (2022)
13. Jiang, M., Wang, Z., Dou, Q.: HarmoFL: harmonizing local and global drifts in federated learning on heterogeneous medical images. In: AAAI, pp. 1087–1095 (2022)
14. Jiang, M., et al.: Dynamic bank learning for semi-supervised federated image diagnosis with class imbalance. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III, pp. 196–206. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_19
15. Ju, L., et al.: Flexible sampling for long-tailed skin lesion classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III, pp. 462–471. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_44
16. Kang, B., Li, Y., Xie, S., Yuan, Z., Feng, J.: Exploring balanced feature spaces for representation learning. In: ICLR (2021)
17. Kang, B., et al.: Decoupling representation and classifier for long-tailed recognition. In: ICLR (2020)
18. Khosla, P., et al.: Supervised contrastive learning. In: NeurIPS, vol. 33, pp. 18661–18673 (2020)
19. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: CVPR, pp. 10713–10722 (2021)
20. Li, T., et al.: Federated optimization in heterogeneous networks. Proc. Mach. Learn. Syst. **2**, 429–450 (2020)
21. Li, T., et al.: Targeted supervised contrastive learning for long-tailed recognition. In: CVPR, pp. 6918–6928 (2022)
22. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: FedBN: federated learning on non-IID features via local batch normalization. In: ICLR (2021)
23. Li, X.C., Zhan, D.C.: FedRS: federated learning with restricted softmax for label distribution non-IID data. In: KDD, pp. 995–1005 (2021)
24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: CVPR, pp. 2980–2988 (2017)
25. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: FedDG: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: CVPR, pp. 1013–1023 (2021)
26. Liu, Q., Yang, H., Dou, Q., Heng, P.-A.: Federated semi-supervised medical image classification via inter-client relation matching. In: de Bruijne, M., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III, pp. 325–335. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_31
27. Marrakchi, Y., Makansi, O., Brox, T.: Fighting class imbalance with contrastive learning. In: de Bruijne, M., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III, pp. 466–476. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_44
28. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS, pp. 1273–1282 (2017)

29. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: PH2-A dermoscopic image database for research and benchmarking. In: EMBC, pp. 5437–5440 (2013)
30. Menon, A.K., et al.: Long-tail learning via logit adjustment. In: ICLR (2021)
31. Mu, X., et al.: FedProc: prototypical contrastive federated learning on non-IID data. Future Gener. Comput. Syst. **143**, 93–104 (2023). https://doi.org/10.1016/j.future.2023.01.019
32. Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proc. Natl. Acad. Sci. U.S.A. **117**(40), 24652–24663 (2020)
33. Sarkar, D., Narang, A., Rai, S.: Fed-Focal loss for imbalanced data classification in federated learning. In: IJCAI (2020)
34. Shang, X., Lu, Y., Huang, G., Wang, H.: Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features. In: IJCAI (2022)
35. Shen, Z., Cervino, J., Hassani, H., Ribeiro, A.: An agnostic approach to federated learning with class imbalance. In: ICLR (2022)
36. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: ICML, pp. 6105–6114 (2019)
37. Tang, K., Tao, M., Qi, J., Liu, Z., Zhang, H.: Invariant feature learning for generalized long-tailed classification. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV, pp. 709–726. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20053-3_41
38. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1), 1–9 (2018)
39. Wang, L., Xu, S., Wang, X., Zhu, Q.: Addressing class imbalance in federated learning. Proc. AAAI Conf. Artif. Intell. **35**(11), 10165–10173 (2021). https://doi.org/10.1609/aaai.v35i11.17219
40. Yan, Z., Wicaksana, J., Wang, Z., Yang, X., Cheng, K.T.: Variation-aware federated learning with multi-source decentralized medical image data. IEEE J. Biomed. Health Inform. **25**(7), 2615–2628 (2020)
41. Yang, M., Wang, X., Zhu, H., Wang, H., Qian, H.: Federated learning with class imbalance reduction. In: EUSIPCO, pp. 2174–2178 (2021)
42. Yang, Z., et al.: ProCo: prototype-aware contrastive learning for long-tailed medical image classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII, pp. 173–182. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_17
43. Zhang, J., et al.: Federated learning with label distribution skew via logits calibration. In: ICML, pp. 26311–26329 (2022)
44. Zhao, Y., Chen, W., Tan, X., Huang, K., Zhu, J.: Adaptive logit adjustment loss for long-tailed visual recognition. Proc. AAAI Conf. Artif. Intell. **36**(3), 3472–3480 (2022). https://doi.org/10.1609/aaai.v36i3.20258
45. Zhu, J., Wang, Z., Chen, J., Chen, Y.P.P., Jiang, Y.G.: Balanced contrastive learning for long-tailed visual recognition. In: CVPR, pp. 6908–6917 (2022)