Communication-Efficient Federated Skin Lesion Classification with Generalizable Dataset Distillation

Yuchen Tian[1⋆], Jiacheng Wang[1⋆], Yueming Jin[2], Liansheng Wang[1(✉)]

[1] Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China, 361005
{tianyuchen,Jiachengw}@stu.xmu.edu.cn,lswang@xmu.edu.cn
[2] Department of Biomedical Engineering and Department of Electrical and Computer Engineering, National University of Singapore, Singapore, 119276
ymjin@nus.edu.sg

Abstract. Federated learning (FL) has recently been applied to skin lesion analysis, but the challenges of huge communication requirements and non-independent and identical distributions have not been fully addressed. The former problem arises from model parameter transfer between the server and clients, and the latter problem is due to differences in imaging protocols and operational customs. To reduce communication costs, dataset distillation methods have been adopted to distill thousands of real images into a few synthetic images (1 image per class) in each local client, which are then used to train a global model in the server. However, these methods often overlook the possible inter-client distribution drifts, limiting the performance of the global model. In this paper, we propose a generalizable dataset distillation-based federated learning (GDD-FL) framework to achieve communication-efficient federated skin lesion classification. Our framework includes the generalization dataset distillation (GDD) method, which explicitly models image features of the dataset into an uncertain Gaussian distribution and learns to produce synthetic images with features close to this distribution. The uncertainty in the mean and variance of the distribution enables the synthetic images to obtain diverse semantics and mitigate distribution drifts. Based on the GDD method, we further develop a communication-efficient FL framework that only needs to transmit a few synthesized images once for training a global model. We evaluate our approach on a large skin lesion classification dataset and compare it with existing dataset distillation methods and several powerful baselines. Our results show that our model consistently outperforms them, particularly in comparison to the classical FL method. All resources can be found at https://github.com/jcwang123/GDD-FL.

Keywords: Skin lesion classification · Dataset Distillation · Domain Generalization · Federated learning

⋆ Contributed Equally.

1 Introduction

Federated learning is an innovative approach to training deep learning models that allows for collaboration and sharing of knowledge without the need to centralize data. It involves transferring model parameters between different clients to improve model performance. Federated learning is particularly useful in clinical settings where privacy is of utmost importance, as it allows multiple healthcare providers to train models using their own data while keeping patient information secure. Recent studies have shown the potential of federated learning in predicting clinical outcomes [19, 2, 1, 8, 10].

However, federated learning methods require transmitting model parameters between the server and clients at each learning round [15], and the entire learning process typically involves hundreds of epochs. The resultant increase in communication costs has become one of the most significant challenges in federated learning. Moreover, some hospitals with strict privacy regulations do *not permit internet access*, rendering the communication-reliant federated learning methods infeasible. To address these challenges, previous studies have attempted to limit the number of communications to accelerate convergence and improve communication efficiency [4, 6, 14, 17, 21, 29]. However, such methods still require tens of communications, and parameter transmission remains time-consuming and laborious in the era of large models. Synthesis-based methods are proposed to transfer the local images into synthetic images using GANs [18] and centralize them into the server for task learning, but GANs are hard to train and the generated synthetic images cost a lot of transmission loads. Recently, data distillation has been introduced in the federated learning domain [24]. This technique distills local datasets into a few synthetic images, typically fewer than ten, and sends these synthetic data to a global server for global training. As the transmission requires only one round of communication, and the synthetic data contains no original information, this method inherits the advantages of low communication costs and excellent privacy protection.

Nevertheless, the previous studies mainly discuss the usefulness of small datasets, i.e., handwritten digits. Whether the distilled image retains the abundant semantics for lesion classification is unknown and needed to be investigated. More importantly, this method adopts the oldest distillation algorithm and has not taken into account the distribution drifts among different clients. The drifts will lead to differing distributions of each synthetic dataset, and consequently, the global model trained on such distributed data may exhibit limited performance, which can impact the accuracy and robustness of the model in real-world settings. Solving distribution drifts is a significant challenge in the federated learning community and has been widely studied [3, 12, 13, 26, 25, 16]. However, these strategies are primarily designed for parameter-communication methods, where the clients send their local model updates to the central server for aggregation. In contrast, data distillation aims to minimize the amount of communication by sending synthetic data instead of parameter updates. Therefore, these strategies may not be directly applicable to the data distillation approach.

In this paper, we propose a novel and generalizable data distillation-based federated learning (GDD-FL) framework to address the challenges of communication costs and distribution drifts in skin lesion classification. We first propose a generalizable data distillation (GDD) method that distills each client's local dataset into a small number of synthetic images and makes synthetic data from different sites located in similar distributions. It is achieved by approximating the possible Gaussian distribution of mean and variance values in one client's synthetic data and randomly sampling a new distribution to produce synthetic images. Unlike current data distillation methods that align synthetic images to a fixed distribution, our GDD method produces synthetic images with uncertain distribution so they obtain better diversity. Based on the GDD method, we further build a communication-efficient federated learning framework for skin lesion classification. In this process, each client applies the GDD method to distill its local dataset into a small synthetic dataset and sends it to the global server. The global server then trains a brand-new model using the gathered data. By minimizing the communication between clients and the server, our method reduces communication costs and improves privacy protection. We evaluate the performance of our method on the ISIC-2020 dataset in IID and Non-IID federated settings and compare it with the classical federated learning method and other data distillation methods. The experimental results demonstrate that our GDD-FL framework consistently outperforms other methods in terms of classification accuracy while reducing communication costs and protecting privacy. Our proposed framework has great potential for applications in real-world scenarios where large datasets are distributed across different clients with limited communication resources.

2 Method

In summary, we introduce the approximation of the uncertain distribution of a real dataset in Sec. 2.1, how to optimize learnable synthetic images in Sec. 2.2. and the communication-efficient federated learning framework in Sec. 2.3.

2.1 Generalizable Dataset Distillation

The goal of dataset distillation is to condense the large-scale training set $\mathcal{T} = \{(x_1, y_1), ..., (x_{|\mathcal{T}|}, y_{|\mathcal{T}|})\}$ with $|\mathcal{T}|$ image and label pairs into a small synthetic set $\mathcal{S} = \{(s_1, y_1), ..., (s_{|\mathcal{S}|}, y_{|\mathcal{S}|})$ with $|\mathcal{S}|$ synthetic image and label pairs so that models trained on each $\mathcal{T}$ and $\mathcal{S}$ obtain comparable performance on unseen testing data: $\mathbb{E}_{x\sim P_\mathcal{D}}\mathcal{L}(\Phi_{\theta^\mathcal{T}}(x), y) \simeq \mathbb{E}_{x\sim P_\mathcal{D}}\mathcal{L}(\Phi_{\theta^\mathcal{S}}(x), y)$, where $P_\mathcal{D}$ is the real data distribution, $\mathcal{L}$ is the loss function (i.e. cross-entropy loss). Φ is a task-specific deep neural network, i.e. ResNet-18, parameterized by θ, and $\Phi_{\theta^\mathcal{T}}$ and $\Phi_{\theta^\mathcal{S}}$ are the networks that are trained on $\mathcal{T}$ and $\mathcal{S}$ respectively. Similar to techniques [28, 23], our goal is to synthesize data that approximates the distribution of the real training data, instead of selecting a representative subset of training samples as in [27, 32]. The process has been visualized in Fig. 1.

To obtain a small dataset with similar semantics to the real dataset, we approximate the possible Gaussian distribution of real data and align the learnable

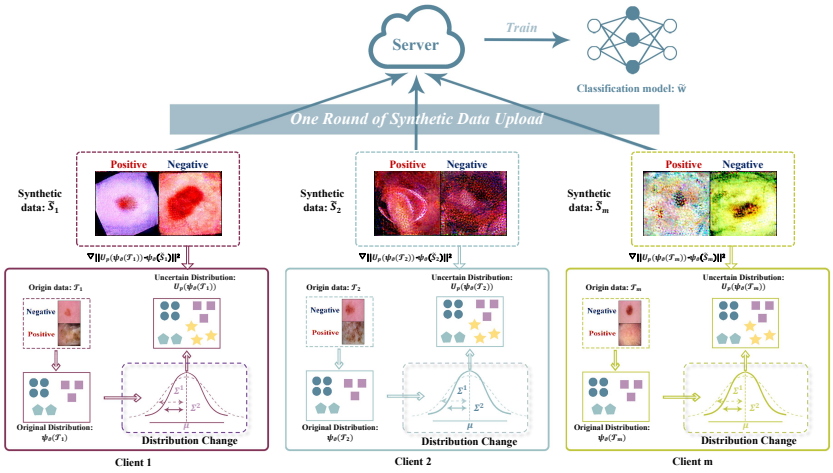

Fig. 1. Overall framework of the generalizable data distillation-based federated learning (GDD-FL). Unlike existing data distillation methods, GDD considers the possible distribution drifts inter-clients and proposes to change the target distribution with random deviations ($\sum$) so that the synthetic images' distribution can align the distributions of other clients.

synthetic data to the distribution. Typical data distillation methods adopt certain mean and variance values to determine the distribution. Instead, to simulate the possible client drift, we estimate the uncertainty of data distribution, and randomly sample new distributions. Specifically, the uncertainty of mean and variance values is estimated as:

$$\Sigma_\mu^2(x) = \frac{1}{|x|}\sum_{i=1}^{|x|}(\mu(x_i) - \mathbb{E}[\mu(x)])^2, \Sigma_\sigma^2(x) = \frac{1}{|x|}\sum_{i=1}^{|x|}(\sigma(x_i) - \mathbb{E}[\sigma(x)])^2, \quad (1)$$

where $\Sigma_\mu(x)$ and $\Sigma_\sigma(x)$ represent the uncertainty estimation of the feature mean μ and feature standard deviation σ, respectively.

After the estimation of possible client shifts, we randomly sample new feature statistics from the estimated distribution as $\hat{\mu}(x) \sim \mathcal{N}(\mu, \Sigma_\mu^2)$ and standard deviation $\hat{\sigma}(x) \sim \mathcal{N}(\sigma, \Sigma_\sigma^2)$ for the corresponding distribution:

$$\hat{\mu}(x) = \mu(x) + \epsilon_\mu \Sigma_\mu(x) \quad and \quad \hat{\sigma}(x) = \sigma(x) + \epsilon_\sigma \Sigma_\sigma, \quad (2)$$

where $\epsilon_\mu, \epsilon_\sigma \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In the end, the feature after the simulated client shift is formed as: $\hat{x} = \hat{\mu}(x) \times \frac{x-\mu(x)}{\sigma(x)} + \hat{\sigma}(x)$. After the distribution change, we optimize the learnable synthetic images to obtain the same distribution with $\hat{x}$, where the details are introduced next.

2.2 Distillation Process

The training details are presented in Algorithm 1. During each learning epoch, we randomly sample initial parameters ϑ for a typical ConvNet [5], denoted

as Ψ_ϑ, feeding the synthetic images and distribution-changed real data into the network to align the distribution. Before alignment, the read data is modified through the uncertain distribution change and Siamese augmentations. Specifically, the equations in Sec. 2.1 are used to approximate the uncertain distributions. We use U_p to denote the distribution change, where $p = 0.5$ is a controlling variable that represents the probability of performing the change to avoid introducing excessive noise. The differentiable Siamese augmentation [9] is denoted as $\mathcal{A}(\cdot)$, processing the real data and synthetic data respectively for better semantic alignment [30]. Finally, the optimization problem with uncertainty estimation is solved as: $\min_{\mathcal{S}} \mathbb{E}_{\substack{\vartheta \sim P_\vartheta \\ \omega \sim \Omega}} \| \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \psi_\vartheta(\mathcal{A}(U_p(x_i))) - \frac{1}{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} \psi_\vartheta \mathcal{A}(s_j) \|^2$.

2.3 Communication-Efficient Federated Learning

Consider a federated learning task with m clients, the client k-th owns local dataset $\mathcal{T}_k$. We can obtain a set of synthetic datasets through our proposed GDD: $\widetilde{\mathcal{S}} = \{\widetilde{\mathcal{S}}_k | k = 1, 2, ..., m\}$. The server then collects all synthetic datasets from the local sites and uses the merged data $\widetilde{\mathcal{S}}$ to train a brand-new model from scratch. We consider a non-convex neural network objective in the server and train a machine learning model on $\widetilde{\mathcal{S}}$. For each iteration, we sample a mini-batch from the synthetic dataset, denoted as $(x, y) \in \widetilde{\mathcal{S}}$, and calculate the objective function $\mathcal{L}(x, y; w)$, where $\mathcal{L}$ represents the typical entropy loss. Note that the sampled mini-batch may contain synthetic images from multiple clients, which enhances feature diversity in each mini-batch. After optimizing for a total of E epochs, the parameter $\widetilde{w}$ is well-trained.

3 Experiment

3.1 Datasets and Evaluation Metrics

Datasets: For our experiments, we used the public skin lesion classification ISIC2020 [20] dataset provided by the International Skin Imaging Collaboration archive. The dataset contains a total of 33,126 samples in the public training set. Since the public test set is not available, we divided the training set into the train, validation, and test sets with 26,500, 3,312, and 3,314 samples.
Client Split: To simulate the federation, we used two types of splits, IID and Non-IID, as the prior work [11]. For IID federation, we randomly divided the train and validation sets into ten parts ($m = 10$) with equal numbers of positive and negative samples. For Non-IID federation, we used Dirichlet with $\alpha = 1$ to distribute local data. We evaluated the global model using the test set.
Evaluation Metrics: We used four widely adopted metrics, namely, Precision (P), Recall (R), F1 score, and AUC, to comprehensively evaluate the classification performance. Higher values indicate better classification performance.

3.2 Implementation Details

We use ResNet-18 [7] as the base classification model and a classical ConvNet [5] as the image feature extractor for data distillation training. To improve memory

Algorithm 1 Process of Generalizable Data Distillation

Input: Training set $\mathcal{T}$
Output: Synthetic samples $\mathcal{S}$ for C classes
function ClientDatasetDistillation($\mathcal{T}$)
Initialize $\mathcal{S}$ by sampling from random noise
for *each iteration* **do**
Sample $\vartheta \sim P_\vartheta$
Sample mini-batch $B_c^{\mathcal{T}} \sim \mathcal{T}, B_c^{\mathcal{S}} \sim \mathcal{S}$ and augmentation $\mathcal{A}_c$ for every class c
Compute $O_c^{\mathcal{T}} = \frac{1}{|B_c^{\mathcal{T}}|} \sum_{(x,y)\in B_c^{\mathcal{T}}} \Psi_\vartheta(\mathcal{A}_c(x))$ for every class c
Compute $O_c^{\mathcal{S}} = \frac{1}{|B_c^{\mathcal{S}}|} \sum_{(s,y)\in B_c^{\mathcal{S}}} \Psi_\vartheta(\mathcal{A}_c(s))$ for every class c
Compute $O_c^{U} = \frac{1}{|B_c^{\mathcal{T}}|} \sum_{(x,y)\in B_c^{\mathcal{T}}} \Psi_\vartheta(\mathcal{A}_c(U_p(x)))$ for every class c
Compute $\mathcal{L}_{\mathcal{S},\mathcal{T}} = \sum_{c=0}^{C-1} \|O_c^{\mathcal{T}} - O_c^{\mathcal{S}}\|^2$
Compute $\mathcal{L}_U = \sum_{c=0}^{C-1} \|O_c^{U} - O_c^{\mathcal{S}}\|^2$
Update $\mathcal{S} \leftarrow \mathcal{S} - \eta \nabla_{\mathcal{S}} (\mathcal{L}_U + \mathcal{L}_{\mathcal{S},\mathcal{T}})$
end for
return $\mathcal{S}$ for C classes
end function

usage and computational efficiency, all images are resized to (224×224). During the distillation training, we use the SGD optimizer [22] with an initial learning rate of 1 for 300 epochs and set the batch size to 64. For training the classification model, we use the SGD optimizer with an initial learning rate of 0.01. The model is trained for 50 epochs using a batch size of 64.

3.3 Comparison of State-of-the-Arts

We mainly compare our method with the latest data distillation methods, namely DC [32], DSA [30], and DM [31]. Since these techniques have not been used in federated learning, we re-implement them in our settings. In addition, we compare the performance of the classical federated learning framework, FedAVG [15]. Furthermore, we demonstrate several centralized training results, where the "Upper Bound" refers to centralizing all data to train a classification model, and "R.S.@10" and "R.S.@100" denote randomly selecting 10/100 images per lesion class. Since the distillation method used in the prior work [24] is too old without novel designs, we focus on the latest distillation methods.

The quantitative results are shown in Tab. 1. It is seen that GDD outperforms other distillation techniques consistently across all settings. Notably, the improvement is more significant when distilling the dataset into 10 images per class, as the diversity of synthetic images is progressively enhanced. Compared with FedAVG, the results in the IID setting show that data distillation-based methods still have room for improvement. However, data distillation-based methods have a significant advantage over FedAVG in terms of low communication costs. Moreover, GDD-FL shows a substantial improvement in the Non-IID setting for AUC scores, i.e., 5.88% and 6.37% for distilling 1/10 images per class.

Table 1. Comparison with latest data distillation methods on the ISIC-2020 dataset. "*" denotes the implementation on our federated setting.

	Method	AUC	P	R	F1
	Upper Bound	80.97 ± 2.12	97.23 ± 0.03	91.11 ± 2.36	94.07 ± 1.54
	R.S.@10	56.86 ± 1.16	87.29 ± 0.19	59.63 ± 5.08	70.86 ± 3.99
	R.S.@100	60.22 ± 2.70	87.67 ± 0.08	65.56 ± 3.53	75.02 ± 2.54
***Transmit parameters:* 12640 MB**					
	FedAVG (IID) [15]	75.33 ± 3.21	96.55 ± 0.34	80.12 ± 4.97	87.57 ± 3.13
	FedAVG (Non-IID)	65.97 ± 4.17	83.22 ± 1.34	70.68 ± 6.26	76.44 ± 4.23
***Transmit 1 image per class:* 2.88 MB**					
IID	DC* [32]	66.72 ± 2.98	96.71 ± 0.09	75.26 ± 5.23	84.56 ± 3.25
IID	DSA* [30]	64.43 ± 2.67	96.76 ± 0.09	74.35 ± 5.78	83.99 ± 3.56
IID	DM* [31]	68.82 ± 0.22	96.76 ± 0.09	74.35 ± 5.77	83.99 ± 3.56
IID	GDD-FL (Ours)	**71.76 ± 0.04**	**96.93 ± 0.11**	**78.55 ± 2.74**	**86.78 ± 0.04**
Non-IID	DC* [32]	65.18 ± 8.09	96.73 ± 0.04	72.50 ± 12.22	82.42 ± 8.58
Non-IID	DSA* [30]	68.79 ± 0.88	96.70 ± 0.04	78.66 ± 2.93	86.73 ± 1.77
Non-IID	DM* [31]	68.41 ± 0.17	96.60 ± 0.12	72.88 ± 0.56	83.08 ± 0.32
Non-IID	GDD-FL (Ours)	**71.85 ± 1.75**	**97.17 ± 0.55**	**81.14 ± 3.28**	**88.41 ± 2.01**
***Transmit 10 images per class:* 28.81 MB**					
IID	DC* [32]	66.79 ± 1.41	96.47 ± 0.03	78.21 ± 4.37	86.39 ± 2.35
IID	DSA* [30]	65.15 ± 3.03	96.69 ± 0.06	76.38 ± 4.67	85.34 ± 2.13
IID	DM* [31]	69.29 ± 2.35	96.58 ± 0.79	78.21 ± 5.73	86.43 ± 2.24
IID	GDD-FL (Ours)	**73.38 ± 1.50**	**96.79 ± 0.28**	**79.80 ± 10.68**	**87.48 ± 7.09**
Non-IID	DC* [32]	65.70 ± 2.47	96.83 ± 0.03	76.42 ± 4.13	85.42 ± 2.48
Non-IID	DSA* [30]	69.99 ± 0.32	96.80 ± 0.02	79.40 ± 1.89	87.24 ± 3.75
Non-IID	DM* [31]	69.07 ± 7.25	96.85 ± 0.73	76.75 ± 4.51	85.64 ± 3.28
Non-IID	GDD-FL (Ours)	**73.34 ± 1.28**	**96.86 ± 0.34**	**81.37 ± 7.34**	**87.72 ± 10.73**

We also present the visualizations of our synthetic data in Fig. 2, where the first row shows negative samples and the second row shows positive samples. Each column represents the synthetic images distilled by a different client. We observed that the synthetic images underwent style changes based on the original dermoscopy images and contain more texture and style information useful for training a classification model. However, these semantics make the appearance abnormal from the human view, and therefore, it is hard to tell what these images exactly represent.

3.4 Detailed Analysis

Ablation Analysis: We also conduct an ablation study to evaluate the impact of our proposed distribution change. Results are shown in Tab. 1, where we compare the performance of GDD-FL to that of the baseline data distillation

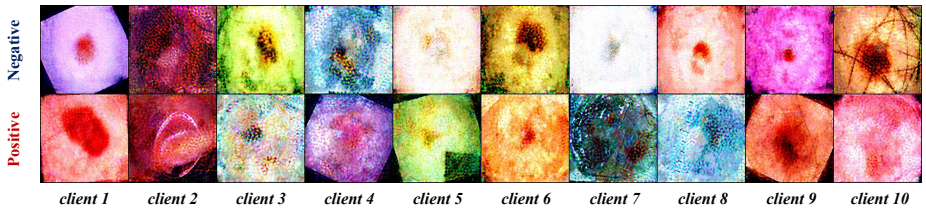

Fig. 2. Visualization of our synthetic images, including the positive and negative samples from ten distributed clients.

method, DM, and to the results of training a model with a randomly sampled subset of 10 or 100 images per class ("R.S.@10" and "R.S.@100"). The DM method is trained without distribution change. As seen from the table, when trained with the same number of samples, GDD-FL achieves significantly better AUC scores, with an improvement of nearly 6% over random sampling. Moreover, the comparison between DM and GDD-FL further confirms the effectiveness of our proposed distribution change. While the performance of DM drops slightly in the Non-IID setting, GDD-FL demonstrates stable performance across both IID and Non-IID settings. Notably, the significant improvement brought by GDD-FL suggests that the distribution change not only enhances generalization but also leads to more diverse semantics, thereby improving classification learning.

Computation Analysis: We further count the computational costs to make a comparison between GDD-FL and the traditional FedAVG. FedAVG trains the model in parallel for 1 hour, requiring 16848 MB GPU. Our method involves distillation training at local sites (1.2 hours, 9561 MB GPU) and classification training at the server (0.1 hours, 16848 MB GPU). It indicates that our method minimizes communication resources while using similar computational resources.

Privacy Protection: GDD-FL condenses numerous real images into a smaller set of synthetic images. By treating the synthetic images as learnable variables and inputting them along with real images into a fixed network, we minimize the discrepancy between their feature outputs. This training aligns the synthetic images with the overall distribution of the real dataset, rather than specific individual images. We also apply random perturbations to the real distribution, reducing privacy risks. Consequently, the synthetic data doesn't contain precise personal information and is not part of the original dataset.

4 Conclusion

In this paper, we introduce a communication-efficient federated skin lesions classification framework using generalizable data distillation, named GDD-FL. Unlike current data distillation methods that align synthetic images to a fixed distribution, our GDD simulates the possible inter-client distribution drifts and produces synthetic images with better diversity and distribution alignment. The experimental results on the ISIC-2020 dataset demonstrate that our GDD-FL framework consistently outperforms other methods in terms of classification accuracy while reducing communication costs and protecting privacy.

References

1. Antunes, R.S., André da Costa, C., Küderle, A., Yari, I.A., Eskofier, B.: Federated learning for healthcare: Systematic review and architecture proposal. ACM Transactions on Intelligent Systems and Technology (TIST) **13**(4), 1–23 (2022)
2. Bdair, T., Navab, N., Albarqouni, S.: Fedperl: semi-supervised peer learning for skin lesion classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 336–346. Springer (2021)
3. Fallah, A., Mokhtari, A., Ozdaglar, A.: Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. Advances in Neural Information Processing Systems **33**, 3557–3568 (2020)
4. Gao, H., Xu, A., Huang, H.: On the convergence of communication-efficient local sgd for federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 7510–7518 (2021)
5. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4367–4375 (2018)
6. Hamer, J., Mohri, M., Suresh, A.T.: Fedboost: A communication-efficient algorithm for federated learning. In: International Conference on Machine Learning. pp. 3973–3983. PMLR (2020)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Hossen, M.N., Panneerselvam, V., Koundal, D., Ahmed, K., Bui, F.M., Ibrahim, S.M.: Federated machine learning for detection of skin diseases and enhancement of internet of medical things (iomt) security. IEEE journal of biomedical and health informatics **27**(2), 835–841 (2022)
9. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
10. Li, G., Togo, R., Ogawa, T., Haseyama, M.: Dataset distillation for medical dataset sharing. arXiv preprint arXiv:2209.14603 (2022)
11. Li, Q., Diao, Y., Chen, Q., He, B.: Federated learning on non-iid data silos: An experimental study. In: 2022 IEEE 38th International Conference on Data Engineering (ICDE). pp. 965–978. IEEE (2022)
12. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems **2**, 429–450 (2020)
13. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. arXiv preprint arXiv:2102.07623 (2021)
14. Malinovskiy, G., Kovalev, D., Gasanov, E., Condat, L., Richtarik, P.: From local sgd to local fixed-point methods for federated learning. In: International Conference on Machine Learning. pp. 6692–6701. PMLR (2020)
15. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
16. Mu, X., Shen, Y., Cheng, K., Geng, X., Fu, J., Zhang, T., Zhang, Z.: Fedproc: Prototypical contrastive federated learning on non-iid data. Future Generation Computer Systems **143**, 93–104 (2023)

17. Pathak, R., Wainwright, M.J.: Fedsplit: An algorithmic framework for fast federated optimization. Advances in neural information processing systems **33**, 7057–7066 (2020)
18. Pennisi, M., Proietto Salanitri, F., Palazzo, S., Pino, C., Rundo, F., Giordano, D., Spampinato, C.: Gan latent space manipulation and aggregation for federated learning in medical imaging. In: Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health: Third MICCAI Workshop, DeCaF 2022, and Second MICCAI Workshop, FAIR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18 and 22, 2022, Proceedings. pp. 68–78. Springer (2022)
19. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al.: The future of digital health with federated learning. NPJ digital medicine **3**(1), 119 (2020)
20. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific data **8**(1), 34 (2021)
21. Rothchild, D., Panda, A., Ullah, E., Ivkin, N., Stoica, I., Braverman, V., Gonzalez, J., Arora, R.: Fetchsgd: Communication-efficient federated learning with sketching. In: International Conference on Machine Learning. pp. 8253–8265. PMLR (2020)
22. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
23. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
24. Song, R., Liu, D., Chen, D.Z., Festag, A., Trinitis, C., Schulz, M., Knoll, A.: Federated learning via decentralized dataset distillation in resource-constrained edge environments. arXiv preprint arXiv:2208.11311 (2022)
25. Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., Zhang, C.: Fedproto: Federated prototype learning across heterogeneous clients. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8432–8440 (2022)
26. Wang, J., Jin, Y., Wang, L.: Personalizing federated medical image segmentation via local calibration. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI. pp. 456–472. Springer (2022)
27. Wang, T., Zhu, J.Y., Torralba, A., Efros, A.A.: Dataset distillation. arXiv preprint arXiv:1811.10959 (2018)
28. Welling, M.: Herding dynamical weights to learn. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 1121–1128 (2009)
29. Yuan, H., Ma, T.: Federated accelerated stochastic gradient descent. Advances in Neural Information Processing Systems **33**, 5332–5344 (2020)
30. Zhao, B., Bilen, H.: Dataset condensation with differentiable siamese augmentation. In: International Conference on Machine Learning. pp. 12674–12685. PMLR (2021)
31. Zhao, B., Bilen, H.: Dataset condensation with distribution matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6514–6523 (2023)
32. Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. arXiv preprint arXiv:2006.05929 (2020)