Revisit the Stability of Vanilla Federated Learning Under Diverse Conditions

Youngjoon Lee[1], Jinu Gong[2], Sun Choi[3], and Joonhyuk Kang[1†]

[1] School of Electrical Engineering, KAIST, South Korea
[2] Department of Applied AI, Hansung University, South Korea
[3] Data Center and AI Group, Intel Corporation, United States
yjlee22@kaist.ac.kr, jkang@kaist.ac.kr

Abstract. Federated Learning (FL) is a distributed machine learning paradigm enabling collaborative model training across decentralized clients while preserving data privacy. In this paper, we revisit the stability of the vanilla FedAvg method under diverse conditions. Despite its conceptual simplicity, FedAvg exhibits remarkably stable performance compared to more advanced FL techniques. Our experiments assess the performance of various FL methods on blood cell and skin lesion classification tasks using Vision Transformer (ViT). Additionally, we evaluate the impact of different representative classification models and analyze sensitivity to hyperparameter variations. The results consistently demonstrate that, regardless of dataset, classification model employed, or hyperparameter settings, FedAvg maintains robust performance. Given its stability, robust performance without the need for extensive hyperparameter tuning, FedAvg is a safe and efficient choice for FL deployments in resource-constrained hospitals handling medical data. These findings highlight the value of the vanilla FedAvg as a reliable baseline for clinical practice.

Keywords: Federated Learning · Resource-Constrained · Stability.

1 Introduction

Deep learning [9, 24] has demonstrated remarkable success across various domains by leveraging large-scale datasets for training neural networks. While traditional deep learning approaches assume centralized data availability, many real-world scenarios, particularly in healthcare, face privacy constraints that prevent direct data sharing [1, 10, 20]. The growing concerns about data privacy and security have led to the emergence of federated learning (FL) [19], which enables collaborative model training while keeping sensitive data localized [4, 14, 23]. In particular, the deployment of FL in medical institutions has enabled collaborative learning across multiple hospitals without compromising patient confidentiality [11, 12, 21]. However, the effectiveness of FL heavily depends on the choice of the key hyperparameters across diverse scenarios.

† Corresponding author

Table 1. Comparison of optimization techniques used in different FL methods. ✗ indicates the use of vanilla optimization (local SGD for *ClientOpt* or mean aggregation for *ServerOpt*), while ✓ indicates the adoption of advanced optimization techniques.

Optimization	FedAvg	FedProx	FedDyn	FedCM	FedSAM	FedGAMMA	FedSpeed	FedSMOO
ClientOpt	✗	✓	✓	✓	✓	✓	✓	✓
ServerOpt	✗	✗	✓	✓	✗	✓	✗	✓

Recent FL techniques have developed various approaches to address client drift challenges caused by data heterogeneity across distributed clients [13, 17, 31]. FedProx [16] and FedDyn [2] tackle client drift through different regularization approaches - FedProx adds a proximal term to constrain local training, while FedDyn introduces dynamic correction terms to align local and global objectives. FedCM [30] takes a different approach by incorporating momentum in client training process to improve convergence stability. Another line of research focuses on sharpness-aware optimization [8], starting with FedSAM [22] which applies local perturbations to find flatter minima. Building upon this, FedGAMMA [5] extends the concept to global optimization, FedSpeed [26] combines it with proximal regularization for longer training intervals, and FedSMOO [25] integrates dynamic regularization with sharpness-awareness. As shown in Table 1, these methods introduce additional optimization mechanisms either at client-side, server-side, or both, all requiring careful hyperparameter tuning.

In this paper, we demonstrate that such complexity may not be necessary, as the simple averaging mechanism of FedAvg achieves comparable or better performance. Moreover, FedAvg remains a standard benchmark for recent FL studies, validating its enduring relevance despite newer sophisticated methods. Indeed, the straightforward nature of FedAvg makes it less susceptible to implementation errors and easier to debug compared to more complex alternatives. Furthermore, we numerically show that vanilla FedAvg's simple averaging mechanism effectively captures the essential aspects of distributed learning across various scenarios. The consistent performance of FedAvg raises important questions about the practical value of more complex FL variants. The main contributions of this paper are as follows:

- We validate that FedAvg achieves comparable convergence speed to recent FL methods across communication rounds.
- We evaluate the top-1 test accuracy of FL methods across various classification models, revealing FedAvg's consistent performance regardless of model architecture.
- We investigate the sensitivity of FL methods to hyperparameter variations, showing that FedAvg maintains stable performance without requiring extensive tuning.
- We show that advanced FL methods can potentially surpass FedAvg, but finding their optimal hyperparameters is challenging.

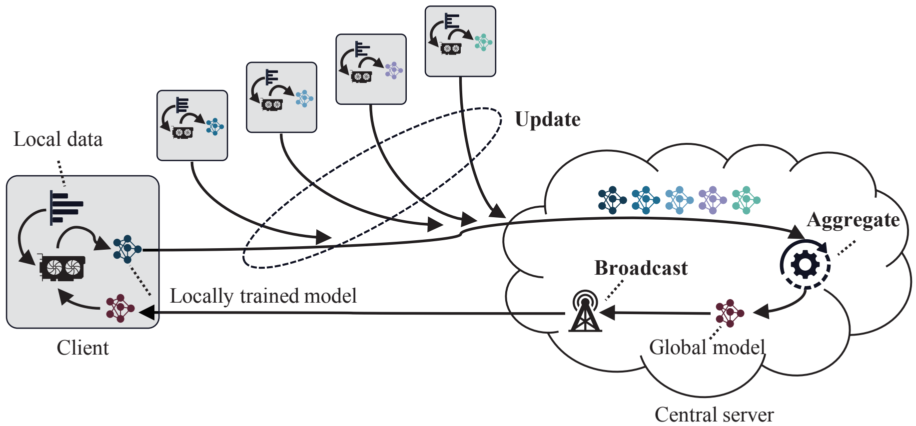

Fig. 1. Overview of the general FL framework. The process consists of three main steps - (1) broadcast of global model from central server to all clients, (2) local training at randomly selected clients using their private data, and (3) aggregation of locally trained models at the central server to improve the global model.

2 Problem and Method

2.1 Federated Setting

We consider a federated network consisting of central server and N clients. Each client $n = 1, \ldots, N$ holds different local dataset $\mathcal{D}^n$ with potentially varying number of data points. The goal of FL is to train a globally shared model $\theta \in \Theta$ by leveraging the clients' datasets $\{\mathcal{D}^n\}_{n=1}^N$ without direct data sharing. Mathematically, the training objective can be formulated as minimizing

$$F(\theta) \triangleq \frac{1}{N} \sum_{n=1}^{N} F_n(\theta), \tag{1}$$

where $F(\theta)$ is the global objective function and $F_n(\theta)$ denotes the local objective function for the nth client defined as

$$F_n(\theta) = \frac{1}{|\mathcal{D}^n|} \sum_{x \in \mathcal{D}^n} f_n(\theta; x), \tag{2}$$

with $f_n(\theta; x)$ being the loss function for data sample x.

2.2 General FL Framework

In FL, the central server communicates with the clients over R communication rounds to minimize the objective function $F(\theta)$, as illustrated in Fig. 1. The standard FL framework consists of the following steps:

Algorithm 1: General FL Framework.

Server Execution:
for $r = 0, \ldots, R-1$ **do**
  Broadcast:
    Transmit global model θ_r to all clients
    Randomly sample M clients, $\mathcal{S}_r$
  Update:
    For each client $m \in \mathcal{S}_r$ **in parallel**:
      $\theta_r^{m,L^m} \leftarrow ClientOpt(\theta_r, \mathcal{D}^m, L^m, \Phi_{client})$
  Aggregate:
    $\theta_{r+1} \leftarrow ServerOpt\Big(\{\theta_r^m\}_{m\in\mathcal{S}_r}, \Phi_{server}\Big)$
end

Broadcast At each communication round r, the central server broadcasts its aggregated model θ_r to all clients. Note that the global model is randomly initialized at the very first round ($r = 0$). Then the central server randomly samples a subset of $M \ll N$ clients from the total N clients, denoted as $\mathcal{S}_r$, to participate in the update process. Only the selected clients participate in the model training while the remaining clients remain idle until the next round.

Update After receiving θ_r from the central server, each client n initializes its local model as $\theta_r^n \leftarrow \theta_r$, where θ_r^n is the local model of the n^{th} client. Then each randomly selected client $m \in \mathcal{S}_r$ trains its local model with $\mathcal{D}^m$ up to the maximum number of local epochs L^m as follows:

$$\theta_r^{m,L^m} \leftarrow ClientOpt(\theta_r, \mathcal{D}^m, L^m, \Phi_{client}). \tag{3}$$

The local training process varies depending on the FL method - vanilla FedAvg performs standard local SGD, while advanced methods incorporate additional optimization techniques with client-side hyperparameters Φ_{client}.

Aggregate After receiving the set of models $\{\theta_r^m\}_{m\in\mathcal{S}_r}$, the central server aggregates these locally trained models using ServerOpt as:

$$\theta_{r+1} \leftarrow ServerOpt(\{\theta_r^m\}_{m\in\mathcal{S}_r}, \Phi_{server}), \tag{4}$$

where $ServerOpt(\cdot)$ can be any existing optimization-based FL method at the server-side. Note that this formulation supports not only simple averaging, but also advanced optimization techniques that incorporate the server-side hyperparameters Φ_{server} to refine the aggregated model. The improved global model θ_{r+1} is then used for the next round $r+1$, and this process repeats until r reaches the predefined number of rounds R for convergence. Overall procedure is described in Algorithm 1.

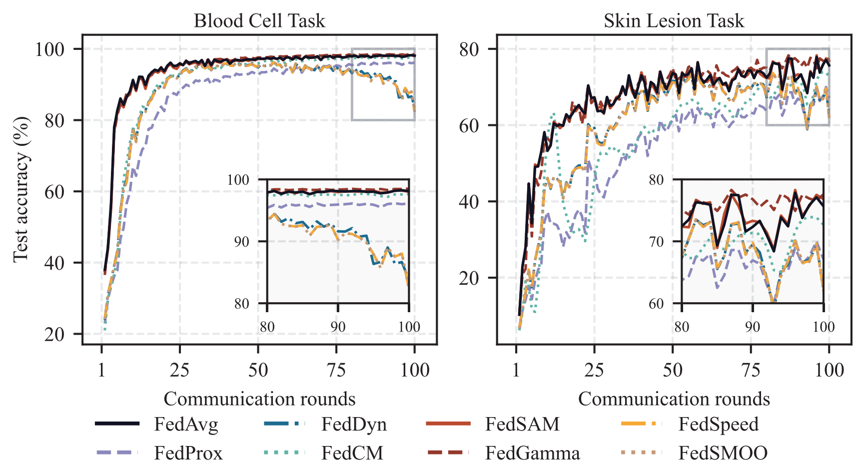

Fig. 2. Test accuracy vs. communication rounds for blood cell and skin lesion classification tasks. FedAvg shows comparable convergence speed and performance to state-of-the-art FL methods. Zoom-in plots highlight the comparable and stable performance of vanilla FL during final rounds.

3 Experiment and Results

3.1 Experiment Setting

To validate the stability of FedAvg, we evaluate the performance of state-of-the-art FL methods on two medical image classification tasks: blood cell [3] and skin lesion [29] classification task. For all experiments except the model comparison experiment, clients employ Vision Transformer (ViT) [7] as local model to ensure consistent evaluation of the FL methods. Moreover, we adopt label skew [15] to introduce data heterogeneity by distributing the non-IID data over $N = 100$ clients following a Dirichlet distribution ($\alpha = 1.0$). Additionally, $M = 10$ clients are randomly selected at each communication round to participate in the training process. All experiments are conducted on Intel Gaudi 2 AI accelerators.

3.2 Results

Impact of Stability To evaluate the stability of different FL methods, we analyze the test accuracy with respect to communication rounds on both blood cell and skin lesion classification tasks as shown in Fig. 2. The results demonstrate that FedAvg achieves comparable or even faster convergence compared to more sophisticated FL methods such as FedProx, FedCM, FedDyn, and FedSpeed. Specifically, FedAvg exhibits stable convergence patterns, reaching competitive test accuracies within 100 communication rounds for both datasets. The results challenge the common belief that advanced FL methods necessarily provide faster convergence.

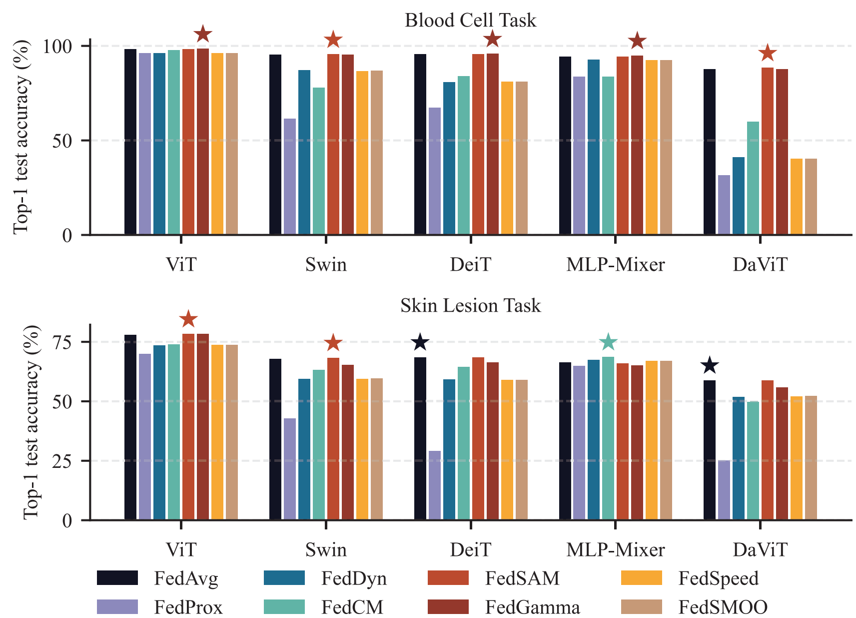

Fig. 3. Comparison of top-1 test accuracy across different AI models. Results show that FedAvg maintains stable performance regardless of the underlying model on both blood cell and skin lesion classification tasks. ★ denotes the best performance.

Furthermore, zoom-in plots of the final 20 rounds show that FedAvg performs competitively during convergence, suggesting that the added complexity of recent FL methods offers limited benefit without tuning. The consistent performance of FedAvg across both medical imaging tasks indicates its robust convergence properties despite its simplicity. This numerical evidence shows that FedAvg's simple averaging effectively captures the essential aspects of collaborative learning, making it a reliable choice for medical AI tasks.

Impact of Model Diversity To investigate how different AI models affect FL performance, we evaluate the top-1 test accuracy across representative vision models: ViT, DeiT [28], DaViT [6], Swin [18], and MLP-Mixer [27]. As shown in Fig. 3, for the blood cell classification task, FedGAMMA achieved the best performance on ViT, DeiT, and MLP-Mixer, while FedSAM led on Swin and DaViT. In the skin lesion classification task, FedAvg delivered the top results on DeiT and DaViT, FedSAM excelled on ViT and Swin, and FedCM attained the maximum performance on MLP-Mixer. Notably, even when FedAvg was not the best, its performance was nearly indistinguishable from the top method. Therefore, in resource-constrained environments, FedAvg is the cost-effective choice for achieving reasonable results without the need for model choice dilemma.

Table 2. Performance comparison of different FL methods on blood cell and skin lesion classification tasks. Results show that advanced FL methods exhibit significant performance variations across different hyperparameter settings, while FedAvg (marked with gray) maintains consistent performance without any tuning. The time costs for each FL method are measured when executed independently on a single Gaudi 2 chip.

Method	Key Φ	Blood Cell Task			Skin Lesion Task		
		IID	non-IID	Time/Round	IID	non-IID	Time/Round
FedAvg	-	98.45	98.25	31.04s	81.60	77.86	24.98s
FedProx	$\lambda = 0.1$	96.52	96.14	31.91s	74.86	69.98	25.39s
	$\lambda = 0.01$	98.39	98.13		81.40	77.96	
	$\lambda = 0.001$	98.42	98.25		81.65	77.91	
FedDyn	$\beta = 0.1$	96.58	96.23	33.28s	77.21	73.52	26.96s
	$\beta = 0.01$	96.58	96.23		77.21	73.52	
	$\beta = 0.001$	96.58	96.23		77.21	73.52	
FedCM	$\mu = 0.1$	97.84	97.72	32.54s	75.41	73.92	26.38s
	$\mu = 0.01$	87.99	87.46		66.93	50.07	
	$\mu = 0.001$	32.42	30.02		66.93	25.69	
FedSAM	$\rho = 0.1$	93.13	85.03	45.40s	70.17	68.43	32.11s
	$\rho = 0.01$	98.48	98.13		82.00	78.05	
	$\rho = 0.001$	98.42	98.22		81.70	**78.30**	
FedGAMMA	$\rho = 0.1$	92.17	80.06	45.75s	67.83	67.23	33.87s
	$\rho = 0.01$	98.60	**98.54**		82.89	76.61	
	$\rho = 0.001$	**98.68**	98.51		**83.74**	**78.30**	
FedSpeed	$\rho = 0.1$	37.27	42.09	45.13s	66.88	67.43	34.28s
	$\rho = 0.01$	96.49	95.97		75.06	72.82	
	$\rho = 0.001$	96.64	96.26		76.96	73.67	
FedSMOO	$\rho = 0.1$	35.78	40.13	53.68s	66.88	67.53	40.68s
	$\rho = 0.01$	96.49	95.97		75.11	72.77	
	$\rho = 0.001$	96.64	96.26		77.06	73.67	

Impact of Hyperparameter Selection We examine hyperparameter sensitivity under both IID and non-IID data by comparing the performance of vanilla FL and state-of-the-art methods. As presented in Table 2, FedProx is highly sensitive to the λ, showing substantially lower performance at $\lambda = 0.1$ than at $\lambda = 0.001$. Unlike FedProx, FedDyn demonstrates nearly uniform accuracy across different β values, implying minimal impact of its correction term within a certain range. Meanwhile, FedCM converges unstably at $\mu = 0.001$, recording only about 30% accuracy on the blood cell classification task. Nextly, FedSAM and FedGAMMA undergo considerable performance drops at $\rho = 0.1$, underscoring their sensitivity to aggressive sharpness exploration. Similarly, FedSpeed and FedSMOO exhibit reduced initial accuracy, stabilizing only at smaller ρ values.

Table 3. Comparison of convergence speed and test accuracy across FL methods using their sub optimal hyperparameters. Results show that even with optimal settings, advanced FL methods achieve marginal improvements over FedAvg (marked with gray) at the cost of hyperparameter tuning.

Method	Blood Cell Task		Skin Lesion Task	
	Top-1 Test Acc. (%)	Round	Top-1 Test Acc. (%)	Round
FedAvg	98.25	99	77.86	96
FedProx	98.25	99	77.96	96
FedDyn	96.23	55	73.52	82
FedCM	97.72	93	73.92	98
FedSAM	98.22	98	78.30	96
FedGAMMA	98.54	98	78.30	87
FedSpeed	96.26	55	73.67	56
FedSMOO	96.26	55	73.67	56

In contrast, FedAvg consistently surpasses 98% (blood cell) and 77% (skin lesion) without additional refinement across both IID and non-IID settings. Moreover, FedAvg requires less time per round than recent FL methods, resulting in lower communication overhead. The simple design of FedAvg eliminates the need for tuning complicated hyperparameters, thus saving both time and computational resources. Therefore, the reliability and efficiency of FedAvg make it a more practical choice for real-world medical AI applications in resource constrained settings, where achieving maximum impact at minimal cost is essential.

Ablation Study To evaluate state-of-the-art FL methods under their sub-optimal settings, we analyze their convergence behavior using the best-performing hyperparameters from our previous experiments. As shown in Table 3, FedGAMMA achieves 98.54% accuracy and converges in 98 rounds, which is 0.29% higher and 1 round faster than FedAvg. Methods like FedDyn, FedSpeed, and FedSMOO converge in 55 rounds but show lower accuracy around 96.2%, while FedProx (98.25%) and FedSAM (98.22%) perform similarly to FedAvg (98.25%). These results indicate that advanced FL methods may outperform FedAvg, but finding optimal hyperparameters remains a key challenge requiring further investigation.

4 Conclusion

In this work, we present a comprehensive study of FedAvg's stability compared to advanced FL methods across medical image classification tasks. Our numerical results show that vanilla FL maintains robust performance under diverse conditions without requiring hyperparameter tuning. Moreover, we have shown that advanced FL methods can slightly outperform FedAvg under optimal settings, the extensive tuning required undermines their practicality. This underscores the value of vanilla FL as a simple yet effective and reliable solution for medical FL in resource-constrained settings.

Acknowledgments. This research was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201787, contribution rate: 50%). Also, this research was supported in part by the NAVER-Intel Co-Lab. The work was conducted by KAIST and reviewed by both NAVER and Intel (contribution rate: 50%).

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Abouelmehdi, K., Beni-Hessane, A., Khaloufi, H.: Big healthcare data: preserving security and privacy. J. Big Data **5**(1), 1–18 (Jan 2018)
2. Acar, D.A.E., Zhao, Y., Matas, R., Mattina, M., Whatmough, P., Saligrama, V.: Federated learning based on dynamic regularization. In: Proc. ICLR. Vienna, Austria (May 2021)
3. Acevedo, A., Merino, A., Alférez, S., Molina, Á., Boldú, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data Br. **30**, 105474 (Jun 2020)
4. Antunes, R.S., André da Costa, C., Küderle, A., Yari, I.A., Eskofier, B.: Federated learning for healthcare: Systematic review and architecture proposal. ACM Trans. Intell. Syst. Technol. **13**(4), 1–23 (May 2022)
5. Dai, R., Yang, X., Sun, Y., Shen, L., Tian, X., Wang, M., Zhang, Y.: Fedgamma: Federated learning with global sharpness-aware minimization. IEEE Trans. Neural Netw. Learn. Syst. **35**(12), 17479–17492 (Dec 2024)
6. Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J., Yuan, L.: Davit: Dual attention vision transformers. In: Proc. ECCV. Tel Aviv, Israel (Oct 2022)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. ICLR. Vienna, Austria (May 2021)
8. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. In: Proc. ICLR. Vienna, Austria (May 2021)
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
10. Guan, H., Yap, P.T., Bozoki, A., Liu, M.: Federated learning for medical image analysis: A survey. Pattern Recognit. (Jul 2024)
11. Joshi, M., Pal, A., Sankarasubbu, M.: Federated learning for healthcare domain-pipeline, applications and challenges. ACM Trans. Comput. Healthc. **3**(4), 1–36 (Nov 2022)
12. Kairouz, P., McMahan, H.: Advances and Open Problems in Federated Learning, Found. Trends Mach. Learn., vol. 14. Now Publishers (Jun 2021)
13. Lee, Y., Park, S., Ahn, J.H., Kang, J.: Accelerated federated learning via greedy aggregation. IEEE Commun. Lett. **26**(12), 2919–2923 (Dec 2022)
14. Lee, Y., Park, S., Kang, J.: Fast-convergent federated learning via cyclic aggregation. In: Proc. IEEE ICIP. Kuala Lumpur, Malaysia (Oct 2023)
15. Li, Q., Diao, Y., Chen, Q., He, B.: Federated learning on non-iid data silos: An experimental study. In: Proc. IEEE ICDE. Kuala Lumpur, Malaysia (May 2022)

16. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proc. MLSys. Austin, United States (Mar 2020)
17. Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z.: On the convergence of fedavg on non-iid data. In: Proc. ICLR. Virtual Event, Ethiopia (May 2020)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE/CVF ICCV. Virtual Event (Oct 2021)
19. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Proc. AISTAT. Fort Lauderdale, United States (Apr 2017)
20. Nguyen, D.C., Pham, Q.V., Pathirana, P.N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O., Hwang, W.J.: Federated learning for smart healthcare: A survey. ACM Comput. Surv. **55**(3), 1–37 (Feb 2022)
21. Pfitzner, B., Steckhan, N., Arnrich, B.: Federated learning in a medical context: a systematic literature review. ACM Trans. Internet Technol. **21**(2), 1–31 (Jun 2021)
22. Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., Lu, Z.: Generalized federated learning via sharpness aware minimization. In: Proc. ICML. Baltimore, United States (Jul 2022)
23. Rauniyar, A., Hagos, D.H., Jha, D., Håkegård, J.E., Bagci, U., Rawat, D.B., Vlassov, V.: Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. IEEE Internet Things J. **11**(5), 7374–7398 (Mar 2024)
24. Simeone, O.: Machine learning for engineers. Cambridge University Press (2022)
25. Sun, Y., Shen, L., Chen, S., Ding, L., Tao, D.: Dynamic regularized sharpness aware minimization in federated learning: Approaching global consistency and smooth landscape. In: Proc. ICML. Hawaii, United States (Jun 2023)
26. Sun, Y., Shen, L., Huang, T., Ding, L., Tao, D.: Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. In: Proc. ICLR. Kigali, Rwanda (May 2023)
27. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. In: Proc. NeurIPS. Virtual Event (Dec 2021)
28. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Proc. ICML. Virtual Event (Jul 2021)
29. 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 (Aug 2018)
30. Xu, J., Wang, S., Wang, L., Yao, A.C.C.: Fedcm: Federated learning with client-level momentum. arXiv preprint arXiv:2106.10874 (Jun 2021)
31. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 (Jun 2018)