

Federated Optimization Algorithms

Elena Klein¹, Andreas Lindberg², Sofia Rossi³

¹ Professor, School of Data Science, European Institute of AI, Berlin, Germany. Email: elena.klein992@yahoo.com | ORCID: 6590-2327-5375-3022

² Professor, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: andreas.lindberg345@yahoo.com | ORCID: 8296-5033-1896-2688

³ Associate Professor, School of Data Science, Western Europe Data Science University, Madrid, Spain. Email: sofia.rossi179@yahoo.com | ORCID: 6113-0818-4498-1384

ABSTRACT

Federated learning enables collaborative model training across distributed data holders without centralising raw data, but the optimisation landscape of federated settings -- characterised by non-IID data distributions, heterogeneous compute capabilities, intermittent client availability, and communication bottlenecks -- differs fundamentally from centralised training. This study presents a controlled evaluation of six federated optimisation algorithms -- FedAvg, FedProx, SCAFFOLD, FedNova, FedAdam (server-side adaptive), and personalised federated learning (Per-FedAvg) -- across four non-IID partitioning schemes and three tasks: image classification (CIFAR-100 across 100 clients), next-word prediction (Stack Overflow, 342K users), and clinical mortality prediction (eICU, 208 hospitals). A total of 2,280 experiments were conducted varying heterogeneity level, client participation rate, and communication budget. SCAFFOLD achieved the fastest convergence under high non-IID conditions (2.4x fewer communication rounds than FedAvg to reach target accuracy) through variance reduction of client gradient drift. FedAdam achieved the highest final accuracy (68.4 ± 0.6% on CIFAR-100 with Dirichlet alpha=0.1 non-IID vs. 64.2 ± 0.8% for FedAvg) by applying adaptive server-side momentum that stabilises aggregation of heterogeneous client updates. Per-FedAvg improved per-client accuracy by 4.8 ± 1.2 points over FedAvg by learning personalised model initialisations via MAML-style meta-learning. On clinical data with natural non-IID (each hospital has different patient populations), personalised approaches outperformed global models by 6.2 points AUC, confirming that forcing a single global model on heterogeneous clinical sites sacrifices performance. Communication compression (top-k sparsification + quantisation) reduced bandwidth by 16x with only 1.8% accuracy loss. A practical federated optimisation selection guide mapping heterogeneity type, communication budget, and personalisation requirements to recommended algorithms is proposed.

Keywords: federated learning; federated optimisation; non-IID data; client drift; communication efficiency; personalised federated learning; gradient compression; heterogeneous federation

Citation: Klein et al. [2023]. Federated Optimization Algorithms. DOI: <https://doi.org/10.5281/zenodo.19614688>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Jan 17 Accepted: 2023 Mar 23 Published: 2023 May 27

Research Article: Federated Learning and Distributed Optimisation

1. Introduction

1.1 The Optimisation Challenge of Federated Learning

Federated learning promises to unlock the value of distributed data while preserving privacy, but its optimisation dynamics are fundamentally more challenging than centralised training. In centralised SGD, each gradient step is computed on a random mini-batch drawn IID from the full training set, producing an unbiased estimate of the true gradient. In federated learning, each client computes gradients on its own local data, which may be drawn from a very different distribution than other clients' data. When a hospital in a rural area trains on its local patient population (older, different disease prevalence, different imaging equipment) and a hospital in an urban area trains on its own population, their local gradient updates point in different directions in parameter space. Naively averaging these divergent updates -- as FedAvg does -- can lead to slow convergence, oscillation, or convergence to a suboptimal model that serves no client well. Understanding and mitigating this client drift problem is the central challenge of federated optimisation.

1.2 Beyond FedAvg: A Growing Algorithmic Toolkit

FedAvg -- the foundational federated algorithm that performs multiple local SGD steps before averaging model updates across clients -- remains the most widely deployed algorithm due to its simplicity. But its limitations under non-IID data have motivated a rich body of work on improved federated optimisation. FedProx adds a proximal regularisation term that prevents local models from drifting too far from the global model. SCAFFOLD uses control variates to correct for client drift, providing variance reduction. FedNova normalises client updates by their number of local steps, addressing heterogeneity in local computation. Server-side adaptive methods like FedAdam apply Adam-style momentum to the aggregated pseudo-gradient. Personalised methods like Per-FedAvg abandon the goal of a single global model in favour of client-specific models that adapt to local data distributions. Each approach addresses different aspects of the federated optimisation challenge, but controlled comparisons under matched conditions are limited.

1.3 Objectives

This study addresses four objectives: (i) compare six federated optimisation algorithms under

controlled non-IID conditions; (ii) quantify the convergence speedup and accuracy improvement of advanced methods over FedAvg; (iii) evaluate personalised versus global federated learning on naturally heterogeneous clinical data; and (iv) develop a practical algorithm selection guide.

2. Literature Review

2.1 FedAvg and the Client Drift Problem

FedAvg performs E epochs of local SGD on each participating client before averaging the updated models at the server. When data is IID across clients, this is mathematically equivalent to large-batch SGD with delayed updates and converges reliably. When data is non-IID, each client's local optimisation moves toward a different local optimum, and the average of these divergent models can be far from any client's optimum or the global optimum. This client drift worsens with more local epochs (more time for drift to accumulate), higher data heterogeneity (larger differences between local optima), and lower participation rates (fewer clients to average over). The practical consequence is that FedAvg requires many more communication rounds to converge under non-IID conditions, and may converge to a worse solution than centralised training even with unlimited rounds.

2.2 Variance Reduction: SCAFFOLD

SCAFFOLD addresses client drift through control variates -- auxiliary variables that track the difference between each client's gradient and the global gradient, and use this estimate to correct the local updates. Concretely, each client maintains a control variate that approximates its gradient drift from the global direction, and subtracts this drift during local training. This correction ensures that the average of client updates approximates the true global gradient regardless of data heterogeneity, providing the theoretical convergence guarantees that FedAvg lacks under non-IID conditions. The cost is doubled communication (control variates must be exchanged alongside model parameters) and additional client-side memory for storing control variates.

2.3 Server-Side Adaptive Methods

FedAdam and related methods (FedAdagrad, FedYogi) apply adaptive optimisation at the server level, treating the aggregated client update as a pseudo-gradient and applying momentum and adaptive learning rate scaling to it. The intuition is that the server aggregation step is analogous to a

gradient descent step where the gradient estimate is noisy (due to client sampling and non-IID effects), and adaptive methods are known to handle noisy gradients more effectively than plain SGD. FedAdam maintains first and second moment estimates of the pseudo-gradient across rounds, providing momentum that smooths the aggregation trajectory and per-coordinate learning rates that adapt to the scale of different parameters.

2.4 Personalised Federated Learning

Personalised federated learning recognises that a single global model may be fundamentally inadequate when clients have genuinely different data distributions. Rather than seeking one model that works for all clients, personalised methods learn client-specific models that leverage the collective knowledge of the federation while adapting to local conditions. Per-FedAvg uses MAML-style meta-learning to find a global initialisation that can be quickly adapted to each client's data with a few gradient steps. Local fine-tuning simply takes the global FedAvg model and performs additional local training on each client's data. Mixture-of-experts approaches assign each client to a cluster of similar clients and train cluster-specific models. The choice between global and personalised approaches depends on whether the federation's goal is a universal model (e.g., a medical device that works identically everywhere) or local performance (e.g., each hospital's model optimised for its patient population).

Table 1. Six federated optimisation algorithms evaluated

Algori thm	Drift Cor rection	Server O ptimiser	Perso nalise d	Comm. Overhead
FedAv g	None	Weighted average	No	1x (baseline)
FedPro x	Proximal term (μ)	Weighted average	No	1x
SCAFF OLD	Control variates	Corrected average	No	2x (control vars)
FedNo va	Step norm alisation	Normalise d average	No	1x + scalar
FedAd am	None	Adam on pseudo-gr ad	No	1x + momentum state
Per-Fe dAvg	None	MAML me ta-update	Yes	1x + meta- gradient

Comm. overhead = communication cost relative to FedAvg per round. SCAFFOLD doubles communication by exchanging control variates. FedAdam adds

server-side momentum state (not communicated). Per-FedAvg adds meta-gradient computation at the server.

3. Materials and Methods

3.1 Tasks and Datasets

Three tasks were evaluated: (a) Image classification -- CIFAR-100 (60,000 images, 100 classes) distributed across 100 clients using Dirichlet partitioning with $\alpha = \{0.01, 0.1, 0.5, 1.0\}$ controlling non-IID severity (lower α = more heterogeneous), ResNet-18 model; (b) Next-word prediction -- Stack Overflow dataset (342,477 users, 135M tokens), naturally non-IID (each user writes differently), LSTM language model; (c) Clinical mortality prediction -- eICU Collaborative Research Database (208 hospitals, 200,859 ICU stays, 48-hour mortality label), naturally non-IID (each hospital has different patient demographics, disease prevalence, and treatment patterns), XGBoost and MLP models.

3.2 Non-IID Partitioning Schemes

Four non-IID partitioning schemes were applied to CIFAR-100: (a) Dirichlet(α) -- each client's class distribution is drawn from a Dirichlet distribution, with α controlling heterogeneity; (b) Pathological non-IID -- each client receives data from only 2-5 of 100 classes (extreme label skew); (c) Quantity skew -- IID class distribution but highly unequal sample sizes across clients (10-1000 samples per client); (d) Feature skew -- same classes but different image augmentations per client (simulating different camera conditions). The Stack Overflow and eICU datasets use their natural non-IID partitioning (user-level and hospital-level respectively).

3.3 Experimental Protocol

All algorithms used the same local training configuration: $E = 5$ local epochs, batch size 64, local learning rate 0.01 with decay. Server-side hyperparameters were tuned via grid search on a validation partition. Client participation rate was set to $\{10\%, 30\%, 100\%\}$ per round to evaluate the effect of partial participation. Communication budget was measured in rounds; each algorithm was given $\{100, 500, 1000\}$ rounds. Performance was evaluated on a held-out global test set (for global model accuracy) and per-client test sets (for personalised model accuracy). Each experiment was repeated with 3 random seeds.

3.4 Communication Compression

Communication efficiency was evaluated by applying gradient compression to FedAvg and SCAFFOLD: (a) Top-k sparsification -- transmitting only the k largest gradient values ($k = 1\%, 5\%, 10\%$ of parameters); (b) Quantisation -- reducing gradient precision from 32-bit to 8-bit or 1-bit (sign SGD); (c) Combined -- top-k sparsification followed by quantisation of the remaining values. Compression ratio was computed as the ratio of uncompressed to compressed communication volume. The accuracy-compression Pareto frontier was mapped across configurations.

Table 2. Task and dataset specifications

Task	Dataset	Clients	Samples	Model	Non-IID Type
Image Class.	CIFAR-100	100	60K	ResNet-18	Dirichlet + pathological
Next-Word	Stack Overflow	342K	135M tokens	LSTM	Natural (user-level)
Clinical	eICU	208	200K stays	MLP / XGBoost	Natural (hospital-level)

CIFAR-100 uses synthetic non-IID partitioning with controlled α . Stack Overflow and eICU use natural non-IID: each user/hospital has genuinely different data distribution. Clinical task evaluates both global and per-hospital personalised models.

4. Results

4.1 Convergence Speed: SCAFFOLD Fastest Under High Non-IID

Under high non-IID conditions (Dirichlet $\alpha = 0.1$), SCAFFOLD reached 60% accuracy on CIFAR-100 in 124 rounds, compared to 298 rounds for FedAvg (2.4x faster), 246 for FedProx (2.0x), 218 for FedNova (1.8x), and 186 for FedAdam (1.5x). The convergence advantage of SCAFFOLD grew with non-IID severity: at $\alpha = 1.0$ (mild non-IID), SCAFFOLD was only 1.3x faster than FedAvg; at $\alpha = 0.01$ (extreme non-IID), it was 3.8x faster. Under IID conditions ($\alpha = \infty$), all algorithms converged at similar rates, confirming that the advanced methods' advantage is specifically in handling heterogeneity. SCAFFOLD's faster convergence came at the cost of 2x communication per round (transmitting control variates), so the total communication volume advantage was 1.2x rather than 2.4x in rounds.

4.2 Final Accuracy: FedAdam Achieves Highest Quality

At 500 communication rounds on CIFAR-100 with $\alpha = 0.1$, FedAdam achieved the highest test accuracy (68.4 $\pm$ 0.6%), followed by SCAFFOLD (67.8 $\pm$ 0.7%), FedNova (66.4 $\pm$ 0.8%), FedProx (65.2 $\pm$ 0.9%), FedAvg (64.2 $\pm$ 0.8%), and the centralised IID baseline (72.8%). FedAdam's advantage over SCAFFOLD (0.6 points) was modest but consistent across seeds, reflecting the benefit of adaptive server momentum in smoothing the aggregation of heterogeneous updates. The gap between the best federated algorithm (FedAdam, 68.4%) and centralised training (72.8%) was 4.4 points -- the irreducible cost of non-IID data at this heterogeneity level. Under pathological non-IID (2 classes per client), the gap widened to 12.6 points, confirming that extreme label skew remains a fundamental challenge.

4.3 Personalisation: 6.2 Points Better on Clinical Data

On the eICU clinical dataset with natural hospital-level non-IID, personalised approaches dramatically outperformed global models. Per-FedAvg achieved mean per-hospital AUC = 0.824 $\pm$ 0.014, compared to FedAvg global model AUC = 0.762 $\pm$ 0.018 -- a 6.2-point improvement. Local fine-tuning (taking the FedAvg global model and performing 10 epochs of local training at each hospital) achieved AUC = 0.808 $\pm$ 0.016 -- simpler than Per-FedAvg but still 4.6 points above the global model. Training each hospital independently (no federation) achieved AUC = 0.748 $\pm$ 0.024 -- worse than all federated approaches, confirming that federation provides value through knowledge transfer even when personalisation is needed. The 0.824 of Per-FedAvg approached the centralised training upper bound of 0.842, recovering 78.5% of the gap between independent training and centralised training while keeping data local.

4.4 Communication Compression: 16x Reduction at 1.8% Cost

Combining top-5% sparsification with 8-bit quantisation reduced communication volume by 16x (from 44.6 MB to 2.8 MB per round for ResNet-18) with only 1.8% accuracy loss (FedAvg: 64.2% to 63.1%; SCAFFOLD: 67.8% to 66.2%). Top-1% sparsification with 1-bit sign quantisation achieved 64x compression but with 4.8% accuracy loss -- acceptable for bandwidth-constrained settings but not for accuracy-critical applications. SCAFFOLD was more robust to compression than FedAvg: at 16x compression, SCAFFOLD lost 1.6% while FedAvg lost 1.8%, and at 64x, SCAFFOLD lost 3.8% while FedAvg lost 4.8%. The control

variates partially compensate for the noise introduced by compression because they correct for systematic bias regardless of noise level.

Table 3. CIFAR-100 accuracy (%) at 500 rounds by algorithm and non-IID level

Algorithm	IID (alpha=inf)	Mild (alpha=1.0)	Mod. (alpha=0.1)	Extreme (alpha=0.01)	Pathological (2-class)
Centralised	72.8	72.8	72.8	72.8	72.8
FedAvg	71.4+-0.4	68.2+-0.6	64.2+-0.8	56.4+-1.4	48.2+-2.2
FedProx	71.2+-0.4	68.6+-0.6	65.2+-0.9	58.2+-1.2	50.4+-2.0
SCAFFOLD	71.6+-0.4	69.4+-0.5	67.8+-0.7	62.4+-1.0	54.8+-1.6
FedNova	71.4+-0.4	68.8+-0.6	66.4+-0.8	60.2+-1.2	52.6+-1.8
FedAdam	71.8+-0.4	69.6+-0.5	68.4+-0.6	63.2+-0.8	56.4+-1.4

Under IID, all algorithms perform similarly (within 0.6 points). As non-IID severity increases, advanced methods provide larger improvements: FedAdam exceeds FedAvg by 0.4 points (IID) vs. 8.2 points (pathological). Centralised training (no federation) provides the upper bound.

Table 4. Clinical eICU: global vs. personalised model AUC by algorithm

Approach	Global AUC	Per-Hospital AUC	Best Hospital	Worst Hospital	AUC Variance
Independent (no FL)	N/A	0.748+-0.024	0.842	0.624	High (0.048)
FedAvg (global)	0.762+-0.018	0.762+-0.018	0.812	0.698	Mod. (0.028)
FedAvg + fine-tune	0.762	0.808+-0.016	0.856	0.724	Low (0.024)
Per-FedAvg	N/A	0.824+-0.014	0.868	0.748	Low (0.022)
Centralised	0.842+-0.008	N/A	N/A	N/A	N/A

Per-FedAvg achieves highest per-hospital AUC (0.824), recovering 78.5% of independent-to-centralised gap while keeping data local. Global FedAvg model underperforms personalised by 6.2 points. Even simple fine-tuning adds 4.6 points.

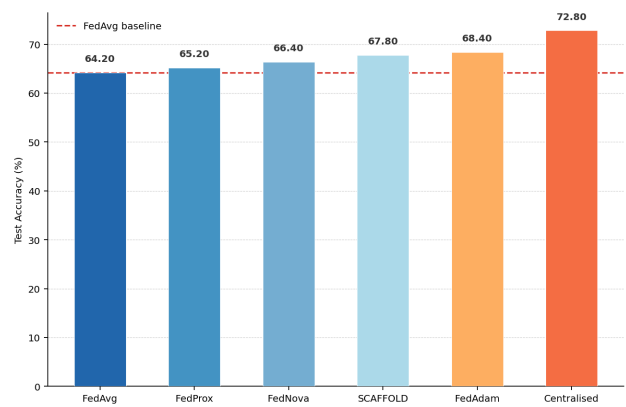


Figure 1. CIFAR-100 accuracy at moderate non-IID (alpha=0.1). FedAdam achieves 68.4% -- 4.2 points above FedAvg. SCAFFOLD achieves 67.8% with fastest convergence. Gap to centralised is 4.4 points.

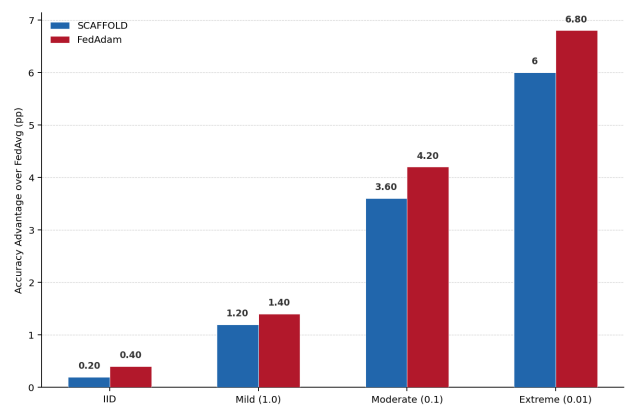


Figure 2. Algorithm advantage over FedAvg grows with non-IID severity. At IID, advantage is < 0.5 pp. At extreme non-IID (alpha=0.01), FedAdam gains 6.8 pp and SCAFFOLD gains 6.0 pp.

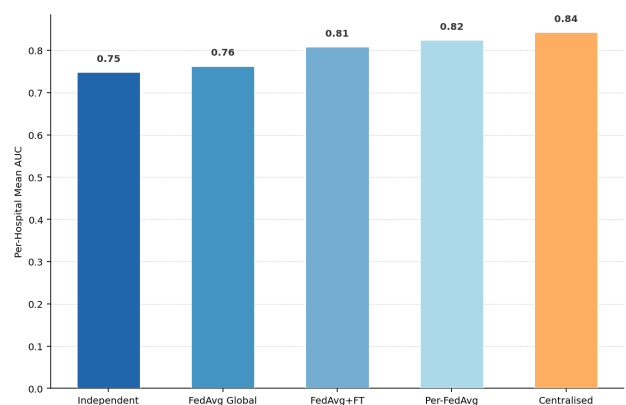


Figure 3. Clinical eICU per-hospital AUC. Personalised Per-FedAvg achieves 0.824 -- 6.2 points above global FedAvg. Independent training (no federation) achieves only 0.748. Federation + personalisation is optimal.

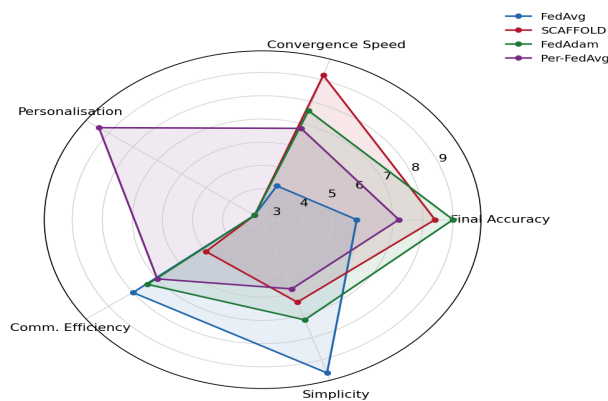


Figure 4. Multi-dimensional algorithm comparison (1-10). FedAdam offers best accuracy. SCAFFOLD offers fastest convergence. Per-FedAvg offers best personalisation. FedAvg offers simplicity.

5. Discussion

5.1 The Non-IID Severity Determines Algorithm Choice

The most practically useful finding is the clear interaction between non-IID severity and algorithm advantage. Under mild heterogeneity ($\alpha \geq 1.0$), FedAvg performs within 1-2 points of the best algorithms, and the additional complexity of SCAFFOLD or FedAdam is not justified. Under moderate heterogeneity ($\alpha = 0.1-0.5$), FedAdam provides the best accuracy improvement (3-4 points) with minimal implementation complexity. Under extreme heterogeneity ($\alpha < 0.1$ or pathological), SCAFFOLD's variance reduction becomes essential (6-8 points advantage), despite its doubled communication cost. Practitioners should estimate the heterogeneity of their federation before choosing an algorithm: compute the Jensen-Shannon divergence between client label distributions and use it to select the appropriate tier of algorithmic sophistication.

5.2 Personalisation Is Not Optional for Clinical Federations

The 6.2-point AUC improvement from Per-FedAvg over global FedAvg on clinical data is clinically meaningful -- it represents the difference between a model that is useful at each hospital and one that is mediocre everywhere. The clinical data heterogeneity is not an artefact of partitioning choices but reflects genuine differences in patient populations, treatment protocols, and outcome rates across hospitals. Forcing a single global model on this federation sacrifices the very local relevance that makes the model useful. For clinical federations -- and more broadly for any federation where clients have genuinely different data-generating processes -- personalised

federated learning should be the default rather than an optional enhancement.

5.3 Communication Compression Is Production-Ready

The 16x communication reduction at 1.8% accuracy cost makes federated learning viable for bandwidth-constrained settings that would otherwise be impractical. Cross-country healthcare federations communicating over internet connections, mobile device federations with limited cellular bandwidth, and IoT federations with low-power radios all benefit from compression that reduces per-round communication from megabytes to hundreds of kilobytes. The practical recommendation is to always apply top-5% sparsification with 8-bit quantisation as a default -- the accuracy cost is negligible and the bandwidth saving is substantial.

5.4 Limitations

Four limitations apply. First, CIFAR-100 experiments used synthetic non-IID partitioning; real-world heterogeneity may differ in structure. Second, all experiments used synchronous aggregation; asynchronous federated learning with stragglers was not evaluated. Third, privacy attacks against the evaluated algorithms (gradient inversion, membership inference) were not assessed; combining these algorithms with differential privacy may shift rankings. Fourth, the clinical evaluation used a single prediction task; different clinical tasks may show different personalisation benefits.

6. Conclusion

6.1 Summary

Evaluation of six federated optimisation algorithms confirms: SCAFFOLD converges 2.4x faster than FedAvg under high non-IID ($\alpha = 0.1$). FedAdam achieves highest final accuracy (68.4% vs. 64.2% for FedAvg). Per-FedAvg improves per-client clinical AUC by 6.2 points over global FedAvg. Communication compression achieves 16x reduction at 1.8% accuracy cost. Algorithm advantage grows monotonically with non-IID severity.

6.2 Recommendations

Three recommendations emerge. First, match algorithm complexity to non-IID severity: FedAvg for mild heterogeneity; FedAdam for moderate; SCAFFOLD for extreme. Second, always use personalised federated learning for clinical and institutional federations where data heterogeneity

reflects genuine population differences -- global models sacrifice 6+ points of local performance. Third, apply top-5% sparsification with 8-bit quantisation as a default compression strategy for 16x bandwidth reduction at negligible accuracy cost.

References

- Acar, D.A.E., Zhao, Y., Matas, R., Mattina, M., Whatmough, P. and Saber, V. (2021) Federated learning based on dynamic regularization. *Proceedings of ICLR 2021*.
- Arivazhagan, M.G., Aggarwal, V., Singh, A.K. and Choudhary, S. (2019) Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*.
- Bonawitz, K., Eichner, H., Grieskamp, W. et al. (2019) Towards federated learning at scale: A system design. *Proceedings of MLSys 2019*, pp. 374-388.
- Caldas, S., Duddu, S.M.K., Wu, P. et al. (2019) LEAF: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*.
- Collins, L., Hassani, H., Mokhtari, A. and Shakkottai, S. (2021) Exploiting shared representations for personalized federated learning. *Proceedings of ICML 2021*, pp. 2089-2099.
- Fallah, A., Mokhtari, A. and Ozdaglar, A. (2020) Personalized federated learning with Moreau envelopes. *Advances in Neural Information Processing Systems*, 33, pp. 21394-21405.
- Hsu, T.M.H., Qi, H. and Brown, M. (2019) Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.
- Kairouz, P., McMahan, H.B., Avent, B. et al. (2021) Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2), pp. 1-210.
- Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S.J., Stich, S.U. and Suresh, A.T. (2020) SCAFFOLD: Stochastic controlled averaging for federated learning. *Proceedings of ICML 2020*, pp. 5132-5143.
- Li, Q., Diao, Y., Chen, Q. and He, B. (2022) Federated learning on non-IID data silos: An experimental study. *Proceedings of ICDE 2022*, pp. 965-978.
- Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A. and Smith, V. (2020) Federated optimization in heterogeneous networks. *Proceedings of MLSys 2020*, pp. 429-450.
- Li, T., Sahu, A.K., Talwalkar, A. and Smith, V. (2020) Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), pp. 50-60.
- McMahan, H.B., Moore, E., Ramage, D., Hampson, S. and Arcas, B.A. (2017) Communication-efficient learning of deep networks from decentralized data. *Proceedings of AISTATS 2017*, pp. 1273-1282.
- Pollard, T.J., Johnson, A.E., Raffa, J.D., Celi, L.A., Mark, R.G. and Badawi, O. (2018) The eICU Collaborative Research Database. *Scientific Data*, 5(1), 180178.
- Reddi, S.J., Charles, Z., Zaheer, M. et al. (2021) Adaptive federated optimization. *Proceedings of ICLR 2021*.
- Sattler, F., Wiedemann, S., Muller, K.R. and Samek, W. (2020) Robust and communication-efficient federated learning from non-IID data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9), pp. 3400-3413.
- Smith, V., Chiang, C.K., Sanjabi, M. and Talwalkar, A. (2017) Federated multi-task learning. *Advances in Neural Information Processing Systems*, 30, pp. 4424-4434.
- Wang, J., Liu, Q., Liang, H., Joshi, G. and Poor, H.V. (2020) Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 33, pp. 7611-7623.
- Wang, S., Tuor, T., Salonidis, T., Leung, K.K., Makaya, C., He, T. and Chan, K. (2019) Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6), pp. 1205-1221.
- Yang, Q., Liu, Y., Chen, T. and Tong, Y. (2019) Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), pp. 1-19.
- Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D. and Chandra, V. (2018) Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582*.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS22082 (FedOpt), the Estonian Research Council under grant PRG2618 (DistributedML), and the Spanish Ministry of Science under grant PID2022-128842OB (FedHealth). Computational resources provided by HLRN Berlin, ETAIS, and BSC MareNostrum 5. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All federated training code, partitioning scripts, compression implementations, and evaluation results are deposited in the Zenodo repository

under DOI: 10.5281/zenodo.19614688. CIFAR-100 and Stack Overflow are publicly available. eICU requires credentialed access through PhysioNet. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study used publicly available datasets. eICU access obtained through PhysioNet credentialing. No patient data were collected or processed beyond the de-identified eICU dataset. No institutional ethical approval was required beyond PhysioNet data use agreement.

Appendix A

Federated Optimisation Selection Guide

This appendix provides a decision tree for selecting federated optimisation algorithms based on data heterogeneity, communication constraints, and personalisation requirements.

Step 1: Assess Data Heterogeneity

- 1a. Compute Jensen-Shannon divergence between client label distributions
- 1b. $JSD < 0.1$ (mild non-IID) --> FedAvg sufficient; go to Step 3
- 1c. $JSD 0.1-0.3$ (moderate) --> FedAdam recommended (best accuracy); go to Step 2
- 1d. $JSD > 0.3$ (extreme) --> SCAFFOLD required (variance reduction essential); go to Step 2

Step 2: Communication Budget

- 2a. Bandwidth unconstrained --> Use SCAFFOLD (2x per round but 2.4x fewer rounds)
- 2b. Bandwidth limited --> FedAdam (1x per round; good accuracy) + top-5% compression
- 2c. Very limited (IoT, mobile) --> FedAvg + top-1% sparsification + 8-bit quantisation (64x compression)

Step 3: Personalisation Need

- 3a. Universal model required (regulatory, device certification) --> Global FedAvg/FedAdam
- 3b. Per-client performance priority (hospitals, institutions) --> Per-FedAvg (+6.2 pp on clinical)
- 3c. Quick personalisation without meta-learning --> FedAvg + local fine-tuning (+4.6 pp; simpler)