Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Privacy-Preserving Federated Learning for Wearable Arrhythmia Detection: A Comprehensive Trade-off Analysis

Dhruv Shah

dhruvshah7779@gmail.com

University Institute of Computing, Chandigarh University

Research Article

Keywords: Federated learning, Differential privacy, Internet of Medical Things, Arrhythmia detection, Non-IID data, Wearable health

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11238805/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Wearable devices make continuous electrocardiogram (ECG) monitoring routine, but training diagnostic models on pooled patient data sits uneasily with how sensitive that data is. Federated learning avoids pooling, yet on battery-powered wearables it forces a compromise among diagnostic accuracy, formal privacy guarantees and communication cost. This study measures that compromise for a lightweight arrhythmia classifier with 4709 parameters, trained by federated averaging across ten simulated clients on the Massachusetts Institute of Technology–Beth Israel Hospital (MIT-BIH) Arrhythmia Database. Three factors were varied: communication volume relative to centralized training, the severity of non-independent and identically distributed (non-IID) label skew across clients, and the differential privacy (DP) budget ε.

Results

The centralized baseline reached 98.3% accuracy. Federated training without privacy noise reached 95.0% ± 1.1% (mean ± standard deviation over five seeds) while sending 44.9 times fewer uplink bytes than a single transfer of the raw training data at 64-bit precision. Accuracy fell from 95.8% to 89.9% as the Dirichlet concentration α dropped from 1.0 to 0.1, a loss of 5.9 percentage points under extreme label skew. Differential privacy behaved very differently: mean accuracy ranged from 58.4% to 79.2% across ε values from 0.1 to 10, with standard deviations of up to 27.5 percentage points. Accuracy peaked at ε = 2 and did not rise at ε = 5 or ε = 10, and no setting came close to the 95.0% obtained without noise.

Conclusions

For a model of this size, communication cost and client heterogeneity are manageable, whereas the privacy mechanism is the limiting factor. The non-monotonic, high-variance privacy results point to the per-round, unclipped noise injection used here and not to a fixed ceiling on private federated arrhythmia detection. A clipped, budget-accounted mechanism should be tested before any clinical deployment.

Background

Continuous electrocardiogram (ECG) monitoring through wearable Internet of Medical Things (IoMT) devices has changed how early arrhythmia detection can work in practice. The deep learning models behind that detection, though, have historically depended on pooling patient data in one place to train. That requirement sits badly with how sensitive the data is, and with how exposed it becomes during transmission and aggregation (Choudhury et al., 2020; Guerra-Manzanares et al., 2023).

Federated learning (FL) sidesteps the pooling requirement: clients train locally and only share model updates, not raw records (Gutierrez et al., 2024; Nguyen et al., 2021). That does not make the deployment problem disappear, though. A wearable running FL still has to balance three things that pull against each other: accuracy close to what a centralized model would achieve, a privacy guarantee strong enough to matter (typically differential privacy, DP), and a communication footprint small enough for a battery-powered device to sustain (Naeem, 2026). Figure 1 shows the pipeline evaluated here: clients train a lightweight model locally, add noise before transmission, and a server combines the parameters by federated averaging (FedAvg). Figure 2 lays out the trade-off the pipeline is built around.

The sections below work through this trade-off with measured numbers rather than treating it as a given. The Methods section describes the model and the three experimental axes. The Results section reports what happened when all three were run, including a result that runs somewhat against the framing I started with, and the Discussion section interprets it.

Lightweight ECG classification and embedded feasibility

Arrhythmia classifiers aimed at wearables have to satisfy two constraints at once: clinical sensitivity and a parameter budget small enough to run on-device. Recent work has pushed on embedded feasibility criteria specifically, including patient-independent evaluation and adherence to standard diagnostic protocols (Silva et al., 2026). Kim et al. (2022) ran a lightweight convolutional network on a low-power ARM Cortex-M4 microcontroller, and Thota et al. (2025) reported a model with 0.945 million parameters suited to wearable monitoring. What is less settled is what happens when that kind of small model moves from centralized to federated training, since local updates have to keep a shared model converging without the computational headroom larger architectures rely on.

Communication efficiency and privacy integration

Communication overhead is the more heavily studied of the three constraints. Gradient sparsification (Naeem, 2026), adaptive optimization (Chen et al., 2022) and quantization (Ikram et al., 2026) all target it directly. Combining these compression techniques with privacy mechanisms is where things get harder to reason about. Alwedaei and Moosa (2026) combined client-level differential privacy with Cheon–Kim–Kim–Song (CKKS) homomorphic encryption in a hierarchical federated setting. On MIT-BIH Arrhythmia they reported that encryption added no measurable accuracy loss, while DP noise cost between 6.15% and 12.23% accuracy. The question pursued here is narrower: whether a model small enough to be called lightweight can sidestep some of that cryptographic overhead by relying on communication-efficient optimization instead, while still reckoning with the noise-utility bottleneck that surveys such as Dritsas and Trigka (2025) describe at a general level.

Methods

The classifier is a compact multilayer perceptron trained federatedly across ten simulated wearable clients. It is evaluated on three axes: privacy-utility sensitivity, communication overhead and statistical heterogeneity.

Data and preprocessing

All experiments use the 48 records of the Massachusetts Institute of Technology–Beth Israel Hospital (MIT-BIH) Arrhythmia Database (Moody & Mark, 2001), obtained from PhysioNet. A window of 180 samples centered on each annotated beat was cut from the first recorded channel and z-scored per beat. Beat annotations were mapped to the five classes of the Association for the Advancement of Medical Instrumentation (AAMI): normal (N), supraventricular ectopic (S), ventricular ectopic (V), fusion (F) and unknown or paced (Q). This gave roughly 110 000 beats. Each run used a stratified 80/20 train-test split.

Model and federated training

The classifier is a multilayer perceptron with two hidden layers of 24 and 12 ReLU units, giving 4709 parameters in total. It was implemented with scikit-learn and trained with Adam at an initial learning rate of 0.01. Ten simulated clients each trained the shared model for one local epoch per round over 15 communication rounds, and the server combined client parameters by FedAvg, weighting each client by its number of training beats. Every configuration was repeated with five random seeds, which change the train-test split, the client partition, the model initialization and the noise draws. Results are reported as mean ± standard deviation (SD, divisor n) across seeds.

Privacy-utility sensitivity

Laplace noise with scale $1/\varepsilon$ was added to each client's parameters before aggregation, for ε values of 0.1, 0.5, 1.0, 2.0, 5.0 and 10.0. Sensitivity was fixed at 1 and parameters were not clipped. Noise was redrawn independently in every communication round instead of being drawn from a single privacy budget spent across the whole training run. This design choice turns out to matter for how the results should be read (see Discussion).

Communication overhead

Following the general framing of Alhonainy and Rao (2025), the cumulative bytes sent by clients across all rounds were compared with a centralized baseline that ships the full raw training set once. Each client upload was counted at 4 bytes per parameter (18 836 bytes per upload), so the federated total was 2 825 400 bytes for ten clients and 15 rounds. The downlink broadcast of the global model was not counted. The raw training arrays were counted at the 64-bit precision used in the simulation, which gave 126 807 152 bytes.

Statistical heterogeneity

Client data were partitioned by Dirichlet sampling over the AAMI class labels, with α values of 0.1, 0.5 and 1.0. Lower α produces more extreme label skew across clients (Reguieg et al., 2023; Reyes-Amezcua et al., 2026). Where sampling left a client with no beats of some class, one beat of that class was moved from the client holding the most, where possible, because the incremental training routine requires every class to be present. Figure 3 illustrates what the partitioning produces at the two extremes.

Use of artificial intelligence tools

Artificial intelligence (AI) tools were used in preparing this work. The large language model Claude (Anthropic) assisted with drafting and editing the manuscript text, with identifying candidate literature, with writing and debugging the Python code that runs the federated learning simulation, with writing the scripts that produce Fig. 1 to Fig. 5, and with typesetting the manuscript. I ran all experiments myself on the MIT-BIH data on my own computer, and the numbers reported here come directly from the output file of those runs. I reviewed and edited all AI-assisted material and take full responsibility for the content of the manuscript. No AI tool was used to generate, modify or select experimental data or results.

Results

The centralized baseline reached 98.3% accuracy. The identical architecture trained federatedly, with no privacy mechanism, reached 95.0% ± 1.1%, a modest 3.3-point gap that most deployments would probably accept given what it buys in return. The federated run sent 2 825 400 bytes across all rounds, against 126 807 152 bytes for shipping the centralized training set once, a 44.9-fold reduction (Table 1).

Table 1
Summary of the main federated learning results on MIT-BIH.

Metric	Value
Centralized accuracy (single run)	98.3%
Federated accuracy, no DP (mean ± SD, five seeds)	95.0% ± 1.1%
Communication reduction versus centralized (uplink only)	44.9×
Model parameters	4709
Model size per upload	18.4 KB

Note. SD, standard deviation; DP, differential privacy. The centralized baseline is a single run.

The non-IID sweep is the cleaner of the two remaining results. Accuracy moved from 95.8% ± 0.4% at α = 1.0 (near-uniform client data) to 95.0% ± 1.1% at α = 0.5 and 89.9% ± 4.4% at α = 0.1 (extreme heterogeneity), as shown in Fig. 4. A drop of 5.9 percentage points under the most adversarial

partitioning tested is a real cost, but a graceful one: the model degrades without collapsing, and the SD grows alongside the accuracy drop, as expected when client-to-client variance increases.

The privacy sweep does not tell as tidy a story, and it would be dishonest to write it up as though it did. Table 2 and Fig. 5 show accuracy that stays roughly flat and noisy from ε = 0.1 through ε = 1.0 (58% to 61%, with SDs above 20 points in every case), reaches its best value at ε = 2.0 (79.2% ± 3.9%), and then falls back at ε = 5.0 and ε = 10.0 instead of continuing toward the no-DP baseline of 95.0%. That last point is the finding worth sitting with: weaker privacy protection (higher ε) should move accuracy closer to the no-DP result, not plateau 15 to 20 points below it with a spread nearly as large as at ε = 0.1.

Table 2
Accuracy under differential privacy noise for each privacy budget ε (five seeds).

Privacy budget ε	Mean accuracy (%)	SD (percentage points)
0.1	59.0	27.5
0.5	60.8	21.1
1.0	58.4	21.2
2.0	79.2	3.9
5.0	74.8	6.4
10.0	74.8	22.4
No noise (reference)	95.0	1.1
Note. SD, standard deviation across five seeds.		

Discussion

The communication and heterogeneity results are the easier ones to interpret. A 3.3-point gap to the centralized model, 44.9-fold fewer uplink bytes and a 5.9-point loss under the most skewed partition describe a small model that tolerates federation well. The privacy results need more care.

Two explanations are worth weighing for the privacy sweep, and they point to different next steps. The first is that the noise scheme itself is the limiting factor, not privacy in any fundamental sense. Noise is drawn independently for every client on every round, fifteen separate injections per run, instead of being budgeted once across the whole training process as formal DP accounting would require. The mechanism also assumes unit sensitivity without clipping client parameters, so the ε values are nominal noise levels and not formal guarantees. Independent per-round noise of this kind compounds in ways that do not track a single interpretable ε, which would explain both the large spread at every setting and the failure of larger ε to buy accuracy back reliably. If this is the dominant effect, the fix is

methodological: clip updates, adopt a privacy accountant and spend a fixed total budget across rounds. The curve in Fig. 5 should then straighten out considerably.

The second possibility is that a model with only 4709 parameters has little redundancy to absorb noise, because each weight does more work than in a large network, where noise on any single weight barely moves the output. If this is the real mechanism, a properly accounted scheme would still lose accuracy at tight budgets, but more smoothly and more monotonically than observed here.

The current experiment cannot separate the two explanations, and I prefer to say so rather than pick the more flattering one. One external reference point favors the first. Alwedaei and Moosa (2026), in a preprint, reported DP-induced accuracy losses of 6.15% to 12.23% on MIT-BIH Arrhythmia for privacy budgets between ε = 2 and ε = 8, with accuracy SDs below 1.31% across three seeds. The scheme tested here lost 15.8 to 20.2 points at ε = 2 to 10, and its seed-to-seed spread (Table 2) was larger by a factor of at least about three. The models, noise mechanisms and training schedules differ, so this is a contrast and not a like-for-like comparison. It does suggest that the instability reflects the noise scheme more than a hard limit on the task, and it is the reason I would not recommend this specific noise-injection scheme for a deployed clinical system without the accounting fix.

Two next steps follow. Adaptive allocation of the privacy budget, of the kind Mangala et al. (2026) proposed (their hybrid noise mechanism gave better privacy-utility trade-offs in a non-federated healthcare setting), is a more promising direction than tuning ε within the present per-round scheme. And because bandwidth is close to a solved problem for a model of this size (Table 1), the remaining design space lies in adaptive and communication-efficient optimization methods such as those of Chen et al. (2022) and Naeem (2026), which may help recover some of the accuracy lost to noise, rather than in compressing an already small transmission.

Several limitations apply. The train-test split is stratified at the beat level and not by patient, so absolute accuracies are probably optimistic compared with patient-independent evaluation (see Silva et al. (2026) on fair evaluation). The comparisons across settings share this limitation. The centralized baseline is a single run trained for up to 200 epochs, whereas the federated models received 15 rounds of one local epoch, so the 3.3-point gap mixes the effect of federation with a difference in training budget. The communication comparison counts uplink traffic only, and counts raw data at 64-bit precision but model parameters at 32 bits. With 32-bit raw data the reduction falls to about 22-fold, and counting the downlink broadcast of the global model as well would roughly halve it again to about 11-fold. The conclusion that federation sends about an order of magnitude less data survives either adjustment. Finally, all clients were simulated on one machine, so latency, energy use and device failures were not measured.

Conclusions

Federated learning for wearable arrhythmia detection, evaluated on real MIT-BIH data with a small on-device model, holds up well on two of three axes. It gets within 3.3 points of centralized accuracy while

sending 44.9 times fewer uplink bytes, and it degrades gracefully rather than catastrophically under extreme non-IID client heterogeneity. The third axis, differential privacy, is where the current picture is incomplete. Accuracy under noise injection is highly variable and does not improve monotonically with a weaker privacy budget, which points more toward a limitation of the per-round, unclipped noise injection used here than toward a fixed ceiling on what privacy-preserving federated arrhythmia detection can achieve. Future work will replace this scheme with a clipped, budget-accounted DP mechanism, extend the evaluation to PTB-XL for cross-dataset validation, and test the architecture against real wearable hardware constraints.

Abbreviations

AAMI
Association for the Advancement of Medical Instrumentation
AI
Artificial intelligence
CKKS
Cheon–Kim–Kim–Song (homomorphic encryption scheme)
DP
Differential privacy
ECG
Electrocardiogram
FedAvg
Federated averaging
FL
Federated learning
IoMT
Internet of Medical Things
MIT-BIH
Massachusetts Institute of Technology–Beth Israel Hospital
non-IID
Not independent and identically distributed
SD
Standard deviation

Declarations

Ethics approval and consent to participate

Not applicable. This study used only the publicly available, de-identified MIT-BIH Arrhythmia Database (PhysioNet). No new human or animal data were collected, and no ethics approval or participant consent was required.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author Contribution

Dhruv Shah was responsible for conceptualization, methodology, investigation, data analysis, and manuscript preparation. The author reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

Availability of data and materials

The MIT-BIH Arrhythmia Database analyzed in this study is publicly available from PhysioNet (https://doi.org/10.13026/C2F305). The source code of the federated learning simulation, the file with the reported results (real_results.json) and the scripts that generate the figures are deposited in a public repository (https://github.com/Dhruv-79/FedCardio-Lite).

References

1. Alhonainy A, Rao P (2025) *Caching techniques for reducing the communication cost of federated learning in IoT environments* [Preprint]. https://doi.org/10.48550/arXiv.2507.17772. arXiv:2507.17772
2. Alwedaei SM, Moosa J (2026) *Privacy-utility tradeoffs in hierarchical federated learning under combined differential privacy and CKKS homomorphic encryption for distributed healthcare networks* [Preprint]. https://doi.org/10.21203/rs.3.rs-10182043/v1. Research Square
3. Chen Z, Cui H, Wu E, Xi Y (2022) *Efficient adaptive federated optimization of federated learning for IoT* [Preprint]. https://doi.org/10.48550/arXiv.2206.11448. arXiv:2206.11448

4. Choudhury O, Gkoulalas-Divanis A, Salonidis T, Sylla I, Park Y, Hsu G, Das A (2020) *Differential privacy-enabled federated learning for sensitive health data* [Preprint]. https://doi.org/10.48550/arXiv.1910.02578. arXiv:1910.02578
5. Dritsas E, Trigka M (2025) Federated learning for IoT: A survey of techniques, challenges, and applications. J Sens Actuator Networks 14(1). Article 9. https://doi.org/10.3390/jsan14010009
6. Guerra-Manzanares A, Lechuga Lopez LJ, Maniatakos M, Shamout FE (2023) Privacy-preserving machine learning for healthcare: Open challenges and future perspectives. In *Trustworthy machine learning for healthcare: First international workshop, TML4H 2023, proceedings* (Lecture Notes in Computer Science, Vol. 13932, pp. 25–40). Springer. https://doi.org/10.1007/978-3-031-39539-0_3
7. Gutierrez DMJ, Hassan HM, Landi L, Vitaletti A, Chatzigiannakis I (2024) *Application of federated learning techniques for arrhythmia classification using 12-lead ECG signals* [Preprint]. arXiv:2208.10993. https://doi.org/10.48550/arXiv.2208.10993
8. Ikram F, Thakur D, Guzzo A, Fortino G (2026) *Quantization in federated learning: Methods, challenges and future directions* [Preprint]. arXiv:2606.26822. https://doi.org/10.48550/arXiv.2606.26822
9. Kim S, Chon S, Kim J-K, Kim J, Gil Y, Jung S (2022) Lightweight convolutional neural network for real-time arrhythmia classification on low-power wearable electrocardiograph. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1915–1918). IEEE. https://doi.org/10.1109/EMBC48229.2022.9871156
10. Mangala N, Rangwala M, Aishwarya S, Reddy BE, Buyya R, Venugopal KR, Iyengar SS, Patnaik LM (2026) Differential privacy for secure machine learning in healthcare IoT-cloud systems. Future Generation Comput Syst 183:108548. https://doi.org/10.1016/j.future.2026.108548
11. Moody GB, Mark RG (2001) The impact of the MIT-BIH Arrhythmia Database. IEEE Eng Med Biol Mag 20(3):45–50. https://doi.org/10.1109/51.932724
12. Naeem M (2026) Ultra-low communication overhead federated learning strategy for distributed wearable ECG networks. J King Saud Univ – Comput Inform Sci 38(7). Article 745. https://doi.org/10.1007/s44443-026-01144-0
13. Nguyen DC, Pham Q-V, Pathirana PN, Ding M, Seneviratne A, Lin Z, Dobre OA, Hwang W-J (2021) *Federated learning for smart healthcare: A survey* [Preprint]. arXiv:2111.08834. https://doi.org/10.48550/arXiv.2111.08834
14. Reguieg H, El Hanjri M, Kamili E, M., Kobbane A (2023) *A comparative evaluation of FedAvg and Per-FedAvg algorithms for Dirichlet distributed heterogeneous data* [Preprint]. https://doi.org/10.48550/arXiv.2309.01275. arXiv:2309.01275
15. Reyes-Amezcua I, Lopez-Tiro F, Larose C, Daul C, Mendez-Vazquez A, Ochoa-Ruiz G (2026) *FedAgain: A trust-based and robust federated learning strategy for an automated kidney stone identification in ureteroscopy* [Preprint]. https://doi.org/10.48550/arXiv.2603.19512. arXiv:2603.19512
16. Silva G, Silva P, Moreira G, Freitas V, Gertrudes J, Luz E (2026) A systematic review of ECG arrhythmia classification: Adherence to standards, fair evaluation, and embedded feasibility. Res Biomedical Eng 42(3). Article 42. https://doi.org/10.1007/s42600-026-00485-z

17. Thota V, Prajapati H, Joshi Y, Rathi S (2025) *A lightweight CNN-attention-BiLSTM architecture for multi-class arrhythmia classification on standard and wearable ECGs* [Preprint]. https://doi.org/10.48550/arXiv.2511.08650. arXiv:2511.08650

Figures

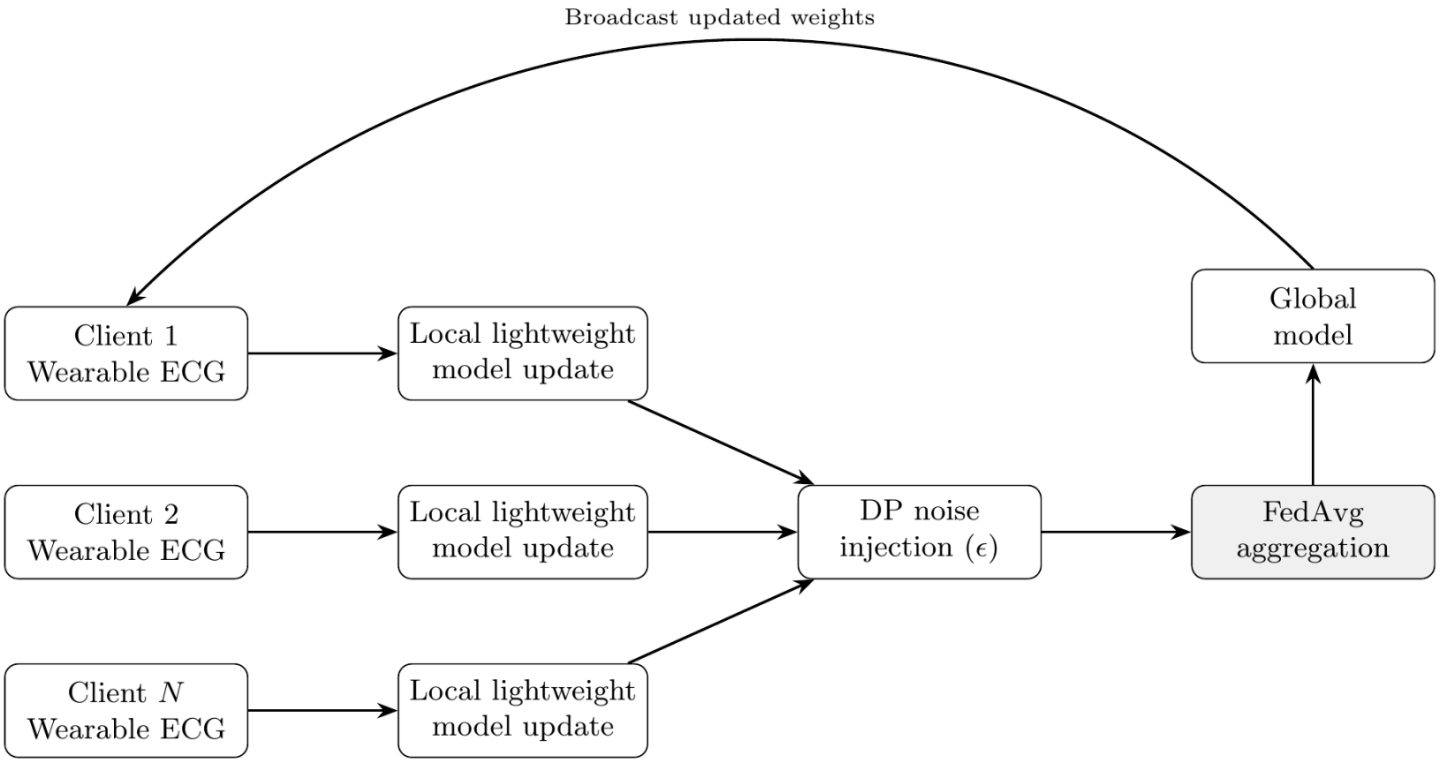


Figure 1

Federated learning pipeline evaluated in this study. Wearable clients train a lightweight model locally and add differential privacy (DP) noise before transmission. A federated averaging (FedAvg) step aggregates the updates into a global model that is broadcast back each round.

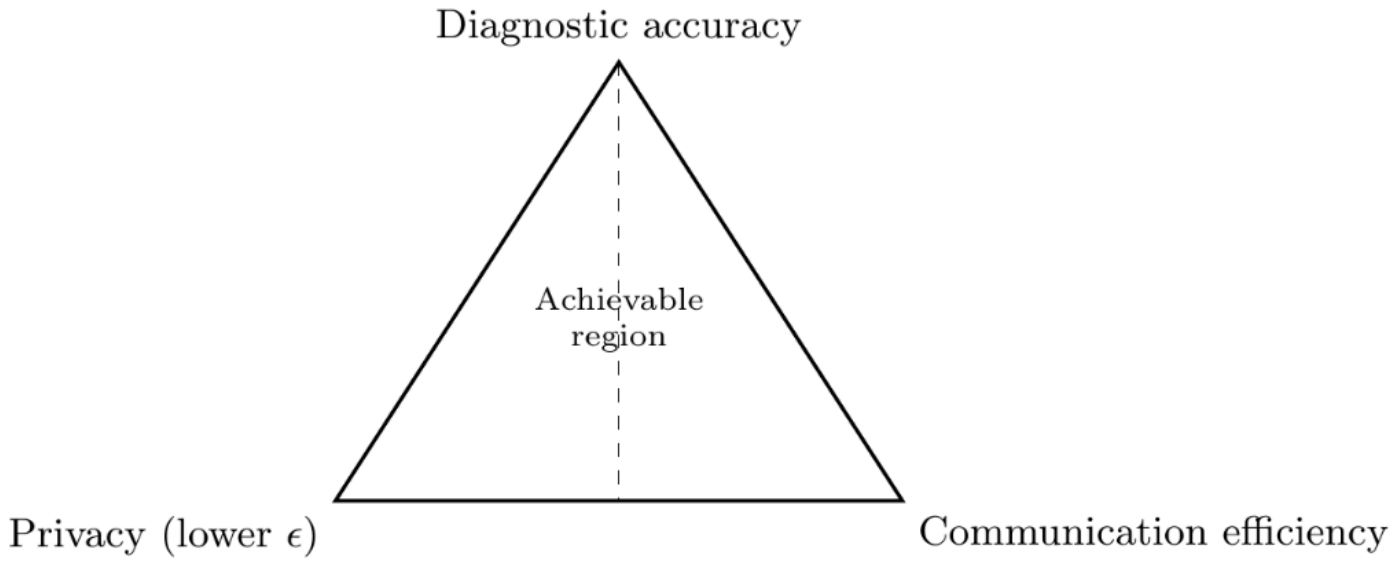


Figure 2

Privacy, communication efficiency and accuracy trade off against one another. Stronger privacy (lower ε) or more compression typically costs diagnostic accuracy. The Results show that the communication and heterogeneity axes hold up well for a lightweight model, whereas the privacy axis does not.

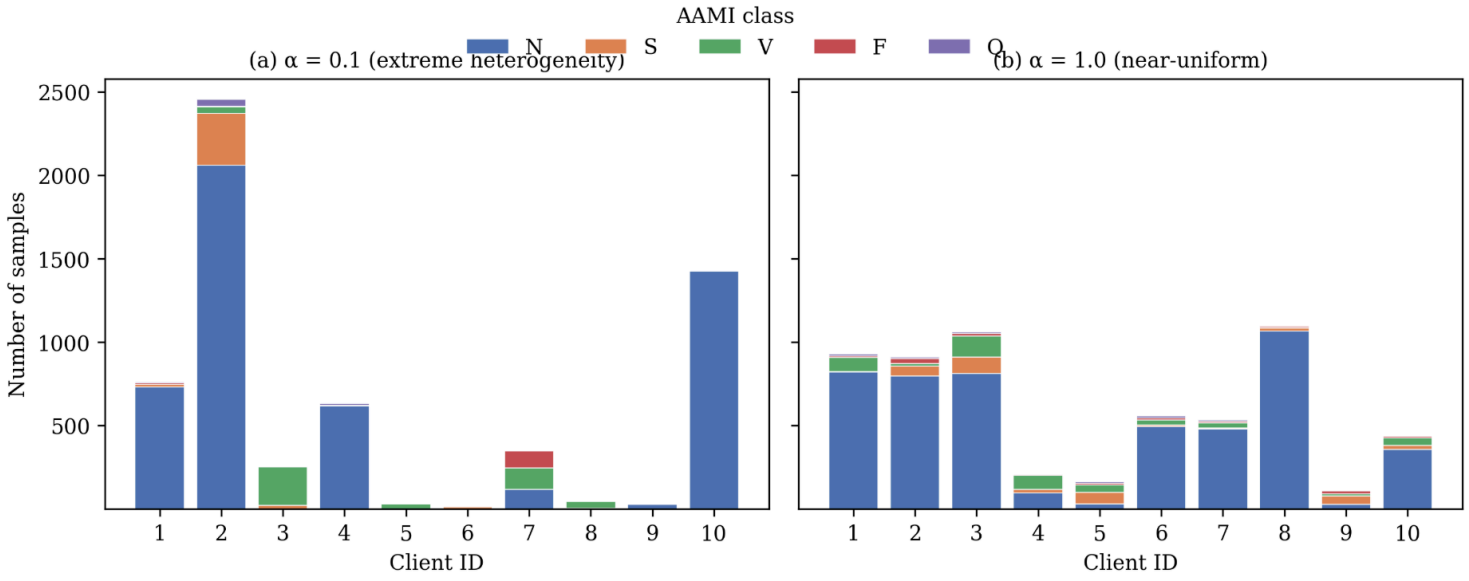


Figure 3

Illustration of Dirichlet label partitioning across ten clients. Panels show α = 0.1 (a) and α = 1.0 (b). Class imbalance follows a representative AAMI prior to illustrate the mechanism; these are not the per-client counts of the reported experiments.

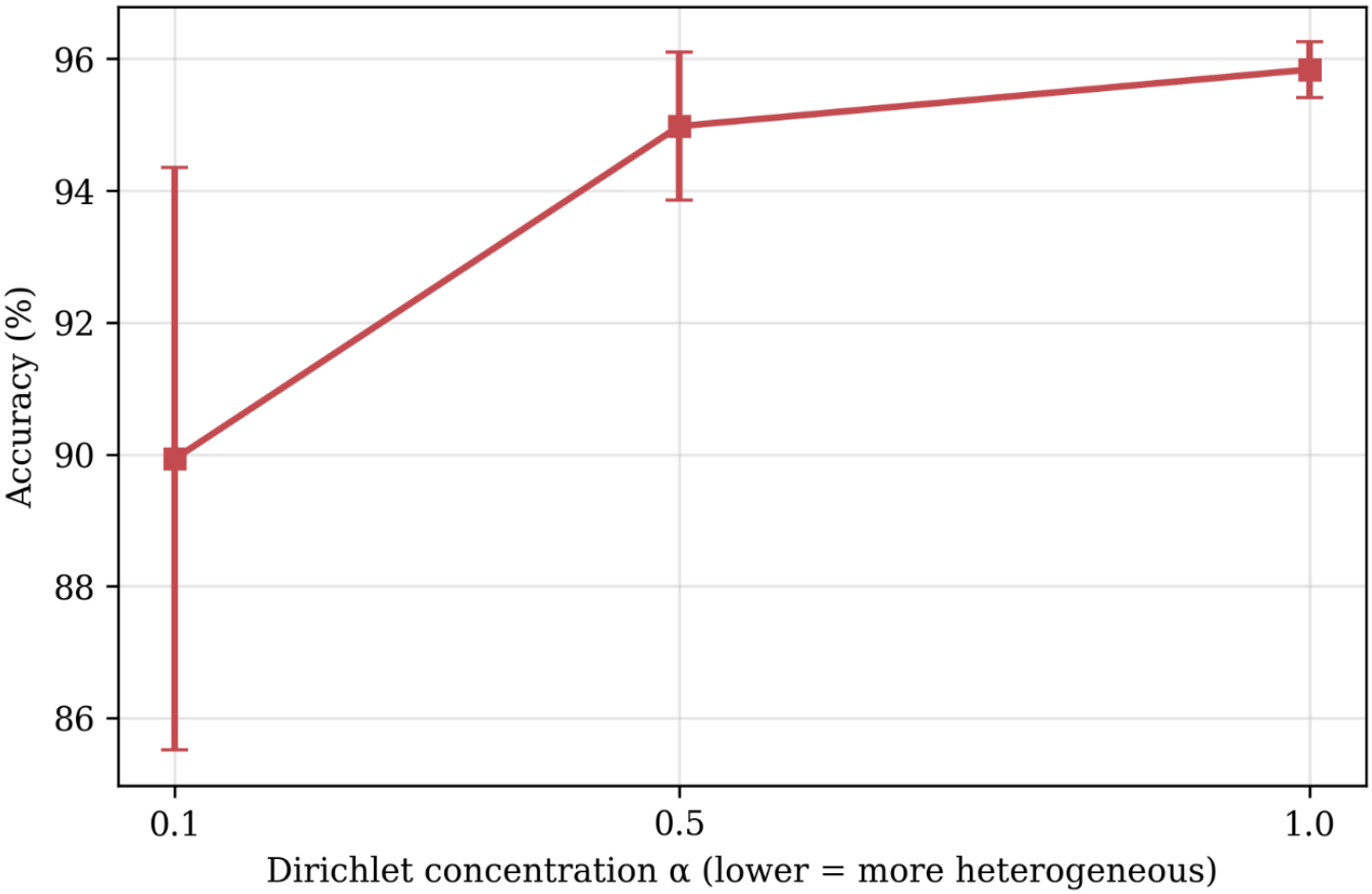


Figure 4

Accuracy falls gradually as non-IID label skew increases. Lower α means more extreme label skew across clients. Points show the mean over five seeds and bars show one SD.

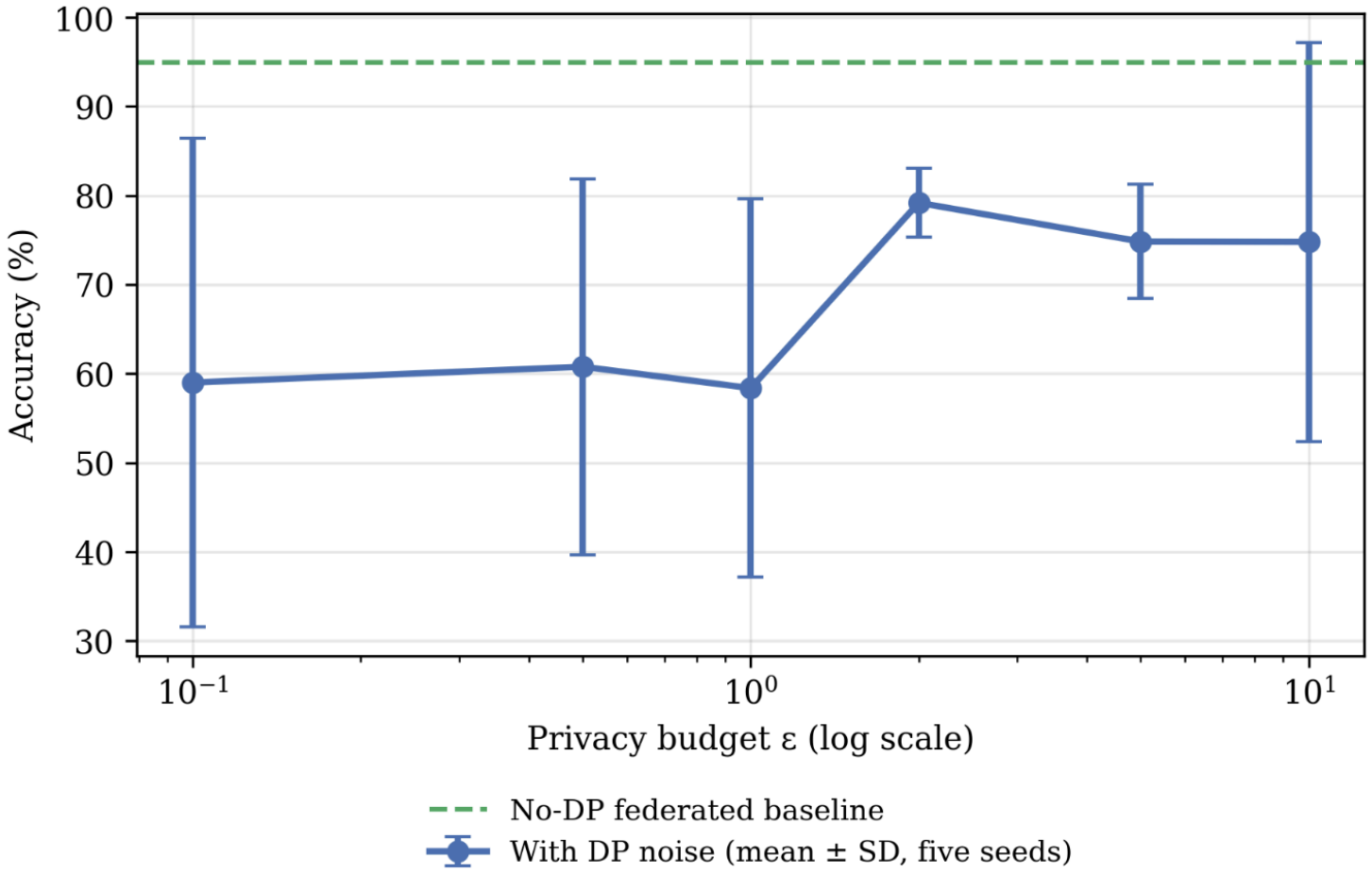


Figure 5

Accuracy under differential privacy is non-monotonic and highly variable across ε. Points show the mean over five seeds with one SD. The dashed line marks the no-DP federated baseline (95.0%), which no tested ε reaches. The ε axis is logarithmic.