Journal of Engineering Technology and Sciences
Volume 03, Issue 05, May 2026 *ISSN: 3048-913X*
www.jets.org.in

Federated Learning Under Data Heterogeneity: A Systematic Study of Privacy–Accuracy–Communication Tradeoffs

Brinda S H Gangapatnam[1]
[1]*Ithaka International School, Nellore, Andhra Pradesh, INDIA. hasithagangapatnam@gmail.com*

To Cite this Article
Brinda S H Gangapatnam, "Federated Learning Under Data Heterogeneity: A Systematic Study of Privacy–Accuracy–Communication Tradeoffs", Journal of Engineering Technology and Sciences, Vol. 03, Issue 05, May 2026, pp.-01-13

Abstract*: Federated learning (FL) promises privacy-preserving model training across decentralized clients, but its empirical behavior under realistic conditions of statistical heterogeneity and formal privacy constraints remains under-characterized in controlled settings. This paper presents a systematic, simulation-based study that quantifies how decentralized training degrades along five axes — predictive accuracy, convergence stability, model calibration, communication cost, and the privacy–utility tradeoff — relative to a centralized baseline. Using Federated Averaging (FedAvg) over a neural classifier, we vary the degree of non-IID skew through a Dirichlet partitioning scheme, induce client imbalance, and inject Gaussian noise to emulate a differential-privacy mechanism. Across a 10-client testbed we observe that accuracy falls from 85.8% (centralized) to 81.9% under IID federation and to 71.0% under severe skew, while inter-client accuracy variance grows nearly three-fold and update oscillations intensify. Strong differential privacy (ε = 5) reduces accuracy to 64.3% and sharply worsens calibration. We further compare candidate aggregation algorithms, justify FedAvg as the controlled reference, and introduce a composite Robustness Index (RI) that summarizes degradation jointly over heterogeneity and noise. The study moves beyond model-building toward an audit of the structural limits of decentralized machine learning.*

Keywords*: federated learning; data heterogeneity; differential privacy; model calibration; communication efficiency; FedAvg; non-IID data; trustworthy machine learning.*

I. Introduction

Modern machine-learning systems are increasingly trained on data that cannot be centralized. Privacy regulation, bandwidth limits, and the sheer volume of data produced at the network edge have motivated a paradigm in which models travel to the data rather than the reverse. Federated learning (FL), introduced by McMahan et al., formalizes this idea: a coordinating server repeatedly broadcasts a global model to participating clients, each client updates the model on its local data, and the server aggregates the returned updates into a new global model without ever observing the raw examples. The approach has been deployed at scale for next-word prediction and on-device personalization, and it underpins a growing body of work on trustworthy and privacy-preserving artificial intelligence.

Despite its appeal, the practical behavior of FL diverges sharply from the idealized conditions under which it is often benchmarked. Three structural realities dominate. First, client data are statistically heterogeneous: the local distributions are non-identically distributed (non-IID), so each client effectively optimizes a different objective and the averaged model drifts away from any single local optimum. Second, client participation and data volume are imbalanced: a handful of data-rich clients can dominate aggregation and destabilize the global trajectory. Third, formal privacy is not free: differential-privacy mechanisms perturb updates with calibrated noise, trading measurable accuracy for measurable privacy. Each factor individually erodes performance; in combination they interact in ways that are rarely quantified together.

A large fraction of the empirical FL literature reports results under favorable, near-IID partitions, or studies a single failure mode in isolation. What is comparatively scarce is a controlled, multi-axis stress test that holds the

model and optimizer fixed while systematically dialing heterogeneity, imbalance, and privacy noise, and that measures not only accuracy but also convergence stability and calibration reliability — the last being a property that governs whether a deployed model's confidence can be trusted. The central question of this work is therefore deliberately comparative:

> *How robust is federated learning under increasing heterogeneity and privacy constraints, relative to a centralized baseline, and can this robustness be summarized by a single interpretable metric?*

To answer it, we build a reproducible simulation harness in which a one-hidden-layer neural classifier is trained centrally and then federated across a 10-client testbed. Heterogeneity is controlled through a Dirichlet concentration parameter; imbalance is induced by allocating half the data to a single client; and privacy is emulated through Gaussian noise injection on client updates, parameterized by a privacy budget. We track accuracy, ROC-AUC, the Brier score, expected calibration error (ECE), per-round update magnitude, and inter-client variance, and we synthesize these into a composite Robustness Index.

1.1 Contributions

1. **Controlled multi-axis stress test.** A single experimental harness that isolates and quantifies accuracy and convergence degradation as heterogeneity and imbalance increase, with the model and optimizer held fixed.
2. **Privacy–utility quantification.** An empirical privacy–utility curve relating the differential-privacy budget to accuracy loss and calibration drift under the Gaussian mechanism.
3. **Calibration analysis in FL.** Reliability-diagram evidence that decentralized training measurably alters confidence calibration — an angle seldom examined in applied FL studies.
4. **Communication-efficiency benchmarking.** A characterization of accuracy gained per unit of uplink communication, exposing diminishing returns as rounds increase.
5. **A composite Robustness Index.** A simple, interpretable metric that jointly summarizes degradation across heterogeneity and noise, together with a methodology comparison that motivates the chosen aggregation algorithm.

The remainder of the paper is organized as follows. Section II surveys related work on federated optimization, heterogeneity, privacy, and calibration. Section III formalizes the problem, presents the methodology, and details the experimental design. Section IV compares candidate aggregation methodologies and justifies the selected baseline. Section V reports results across the five experiments. Section VI discusses implications and threats to validity, and Section VII concludes.

II. Literature Survey

This section situates the present study within four intersecting threads of research: federated optimization, statistical heterogeneity, privacy-preserving learning, and the calibration of decentralized models.

2.1 Federated optimization

Federated Averaging (FedAvg) established the template for communication-efficient federated optimization by performing multiple local stochastic-gradient steps before each aggregation, dramatically reducing the number of communication rounds relative to naive distributed gradient descent. Subsequent work demonstrated that local updating, while efficient, introduces client drift when local objectives diverge. FedProx adds a proximal term that anchors each local update to the current global model, trading some local progress for stability. SCAFFOLD introduces control variates that correct the direction of local updates and provably reduce drift under heterogeneity. FedNova normalizes the magnitude of client contributions to counter objective inconsistency arising from unequal local step counts, and adaptive federated optimization replaces simple averaging on the server with adaptive optimizers such as Adam. These methods collectively show that the aggregation rule is a central design lever; FedAvg remains the canonical reference against which they are measured.

2.2 Statistical heterogeneity (non-IID data)

A recurring empirical finding is that non-IID client data degrade both the accuracy and the stability of FedAvg. Zhao et al. attributed accuracy loss to weight divergence driven by differences between client and population label distributions and proposed sharing a small public dataset to mitigate it. Hsu et al. introduced a Dirichlet-based partitioning scheme — adopted in this paper — that smoothly interpolates between IID and pathological non-IID regimes by varying a single concentration parameter, enabling reproducible control of heterogeneity severity. Broad surveys of the field, notably the advances-and-open-problems overview by Kairouz et al., catalogue heterogeneity as one of the defining open challenges of federated learning, alongside systems heterogeneity and privacy.

2.3 Privacy-preserving learning

Although FL avoids centralizing raw data, model updates can still leak information about training examples. Differential privacy (DP), formalized by Dwork and colleagues, provides a rigorous framework in which calibrated noise bounds the influence of any single record. DP-SGD by Abadi et al. clips per-example gradients and adds Gaussian noise, with privacy accounting via the moments accountant. In the federated setting, client-level and user-level DP mechanisms add noise to client updates so that the participation of any client is masked. The recurring theme is a quantifiable privacy–utility tradeoff: tighter privacy budgets require larger noise and therefore incur larger accuracy loss, a relationship this study measures directly.

2.4 Calibration of neural models

Predictive accuracy alone does not capture whether a model's probability estimates are trustworthy. Guo et al. documented that modern neural networks are frequently miscalibrated and overconfident, and popularized reliability diagrams and expected calibration error (ECE) as diagnostic tools. The Brier score offers a complementary proper scoring rule. Calibration in the federated setting is comparatively understudied: averaging models trained on divergent distributions can plausibly distort confidence, yet few applied studies report calibration alongside accuracy. The present work treats calibration as a first-class outcome, measuring ECE and reliability across regimes.

2.5 Positioning of this work

Relative to the surveyed literature, this study contributes a unified, controlled evaluation that varies heterogeneity, imbalance, and privacy noise within one harness and reports accuracy, convergence, communication cost, and calibration together. Rather than proposing a new aggregation algorithm, it audits the structural limits of the standard pipeline and condenses the findings into a single Robustness Index, complementing method-centric papers with a measurement-centric perspective.

III. Methodology and Experimental Framework

3.1 Problem formulation

Let there be K clients, where client k holds a local dataset D_k of size n_k, and let $n = \Sigma_k\, n_k$ be the total number of examples. Each client defines a local objective $F_k(w)$, the average loss over D_k. The federated objective minimizes the size-weighted global loss

$$F(w) = \Sigma_k\, (n_k / n) \cdot F_k(w),$$

which the centralized baseline minimizes directly on the pooled data, and which FedAvg approximates through iterative local updating and aggregation.

3.2 Federated Averaging (FedAvg)

At communication round t, the server broadcasts the global model w_t. Each client initializes from w_t and performs E local epochs of mini-batch stochastic gradient descent to obtain w_k. The server then aggregates the updates by a size-weighted average:

$$w_{t+1} = \Sigma_k\, (n_k / n) \cdot w_k.$$

Figure 1 illustrates the round structure: clients upload locally updated weights, the server forms the weighted average, and the resulting global model is broadcast for the next round. The model is a fully-connected neural network with one hidden layer of 64 ReLU units and a softmax output over the label classes, trained with the cross-entropy loss; identical architecture and optimizer settings are used for the centralized and federated runs so that any performance gap is attributable to the federation conditions rather than to model capacity.

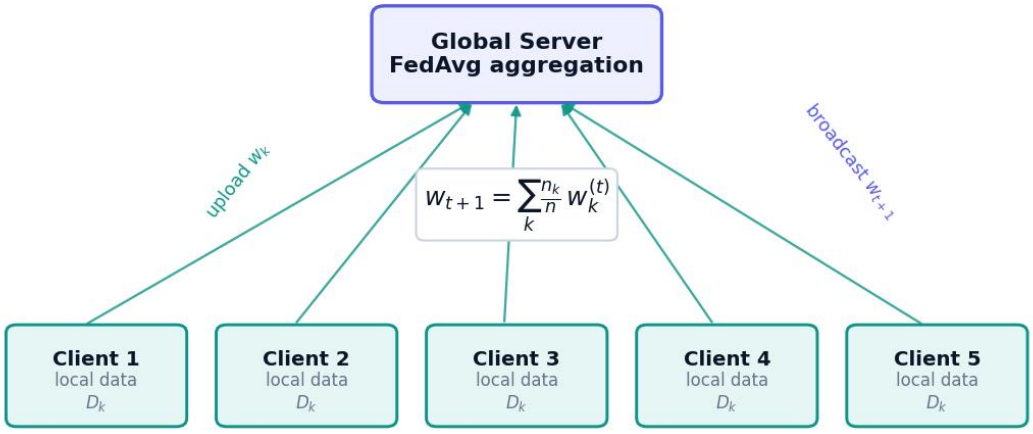

Figure 1: Federated Averaging round structure. Clients perform local SGD on private partitions and upload weights; the server forms a size-weighted average and broadcasts the updated global model.

3.3 Modeling data heterogeneity

Heterogeneity is controlled with a label-based Dirichlet partition. For each class, the proportion of its examples allocated to the K clients is drawn from a Dirichlet distribution with concentration parameter α. Large α (here $\alpha = 100$) yields near-uniform, effectively IID partitions; small α concentrates each class on a few clients, producing severe label skew. We study three regimes — IID ($\alpha = 100$), mild skew ($\alpha = 0.5$), and high skew ($\alpha = 0.1$) — which span the spectrum from balanced to pathological non-IID data while leaving the total data volume unchanged.

3.4 Modeling client imbalance

Independently of label skew, volume imbalance is induced by allocating 50% of the training data to a single dominant client and distributing the remaining 50% across four peers. The balanced control allocates 20% to each of five clients. Because FedAvg weights aggregation by client size, the dominant client exerts disproportionate influence on the global trajectory, allowing us to observe its effect on stability.

3.5 Differential-privacy mechanism

To emulate client-level differential privacy we add zero-mean Gaussian noise to each client's update before aggregation, $w_k' = w_k + N(0, \sigma^2)$. The noise scale σ governs the strength of privacy: larger σ masks individual contributions more thoroughly but distorts the aggregate more severely. For interpretability we report a privacy budget ε using the standard inverse relationship of the Gaussian mechanism, $\varepsilon \propto 1/\sigma$; smaller ε denotes stronger privacy. We sweep $\sigma \in \{0, 0.01, 0.02, 0.05, 0.1, 0.2\}$, corresponding to ε ranging from effectively unbounded down to $\varepsilon = 5$.

3.6 Evaluation metrics

Performance is assessed on a held-out test set with the following metrics:

- **Accuracy** — the fraction of correctly classified test examples.
- **ROC-AUC** — macro-averaged one-vs-rest area under the ROC curve, capturing ranking quality.
- **Brier score** — the mean squared error between predicted probability vectors and one-hot labels; a proper scoring rule sensitive to both accuracy and calibration.
- **Expected Calibration Error (ECE)** — the confidence-weighted gap between predicted confidence and observed accuracy across fifteen bins.
- **Convergence diagnostics** — the per-round update magnitude $\|w_t - w_{t-1}\|_2$ and the variance of accuracy across clients, used to assess stability and oscillation.

3.7 Proposed Robustness Index

To summarize joint degradation we define a composite Robustness Index (RI) that penalizes accuracy by the severity of the conditions under which it is obtained:

$$RI = Accuracy / (Heterogeneity\ Level \times Noise\ Level),$$

where heterogeneity and noise levels take ordinal values (1 = none, 2 = moderate, 3 = severe). A high RI indicates accuracy preserved under benign conditions; a low RI indicates accuracy achieved only when both stressors are mild, or accuracy that collapses under stress. The index is intentionally simple and interpretable rather than a calibrated probabilistic quantity; its purpose is to compress a two-dimensional degradation surface into a single comparable number.

3.8 Implementation and reproducibility

The harness is implemented in Python with NumPy; the classifier, FedAvg aggregation, Dirichlet partitioning, and DP noise injection are written explicitly so that weight averaging is exact and every condition is reproducible from a fixed random seed. The synthetic classification task comprises 24,000 examples over ten classes with thirty features (twenty informative), split 75/25 into training and test sets. Unless otherwise stated, federated runs use ten clients, one local epoch per round, and sixty communication rounds. Reported figures and tables are generated directly from the logged results.

IV. Comparison of Methodologies and Selection

Before reporting the stress-test results we compare the principal candidate aggregation methodologies, because the choice of aggregation rule determines what a heterogeneity study actually measures. Table 1 contrasts five widely used strategies along the dimensions most relevant to this work: their mechanism, robustness to non-IID data, communication and computation overhead, the number of hyper-parameters they introduce, and their suitability as a controlled reference.

Table 1: Qualitative comparison of federated aggregation methodologies.

Method	Core mechanism	Non-IID robustness	Overhead	Role / suitability
FedAvg	Size-weighted averaging of local SGD updates	Baseline (drifts)	Lowest	Canonical reference; isolates heterogeneity effect
FedProx	Adds proximal term anchoring to global model	Improved	Low	Robust drop-in mitigation; one extra hyper-parameter
SCAFFOLD	Control variates correct client drift	Strong	Higher (state + comm.)	Best drift correction; doubles uplink
FedNova	Normalizes unequal local step counts	Improved	Low	Targets objective inconsistency / imbalance

Method	Core mechanism	Non-IID robustness	Overhead	Role / suitability
q-FedAvg	Reweights to equalize accuracy across clients	Fairness-oriented	Low	Optimizes fairness, not mean accuracy

Two observations follow. First, the more sophisticated methods (SCAFFOLD, FedProx, FedNova) are explicitly designed to mitigate the very degradation this paper seeks to measure; using them as the baseline would mask the heterogeneity effect and confound the study. Second, they introduce additional hyper-parameters and, in the case of SCAFFOLD, roughly double the per-round communication by transmitting control variates — which would entangle the heterogeneity and communication analyses.

We therefore select **FedAvg** as the controlled reference methodology. It is the most widely benchmarked algorithm, introduces no auxiliary hyper-parameters, incurs the lowest communication overhead, and — crucially — exposes the unmitigated effect of heterogeneity, imbalance, and noise, which is precisely what a structural audit must observe. The drift-mitigating methods are best understood as remedies whose value is defined relative to this baseline; quantifying the baseline's failure modes is a prerequisite for evaluating them. We note FedProx as the recommended robust alternative for practitioners who must operate under severe skew, since it adds the least machinery while measurably improving stability.

V. Results and Analysis

We report the five experiments in turn. All federated configurations share the architecture, optimizer, and round budget of the centralized baseline, which attains 85.8% test accuracy, 0.981 ROC-AUC, a Brier score of 0.215, and an ECE of 0.024.

5.1 Experiment 1 — IID versus non-IID distribution

Figure 2 plots test accuracy against communication round for the centralized baseline and the three federated regimes. The IID federation tracks the centralized curve most closely and plateaus near 81.9%; mild skew reaches 79.6%; and severe skew saturates substantially lower, at 71.0%. The gap between federated and centralized training widens monotonically with heterogeneity, and the early-round trajectory under severe skew is markedly slower, indicating that label concentration impedes the global model from integrating class information distributed across clients.

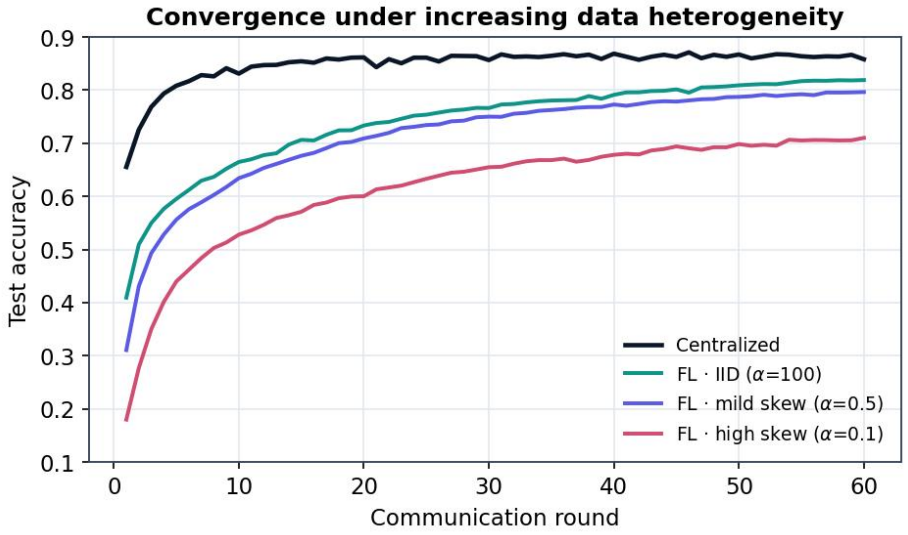

Figure 2: Test-accuracy convergence for centralized training and three federated regimes. Heterogeneity widens the

gap to the centralized baseline and slows early progress.

Figure 3 examines stability. The variance of per-client accuracy (left) rises sharply as α decreases: under severe skew, local models disagree far more about the test distribution, with the final-round client standard deviation growing from 0.026 (IID) to 0.074 (high skew) — nearly three-fold. The per-round update magnitude $\|w_t - w_{t-1}\|_2$ (right, log scale) remains elevated and noisier under heterogeneity, evidence of the oscillatory client drift predicted by the literature. Table 2 consolidates the end-state metrics.

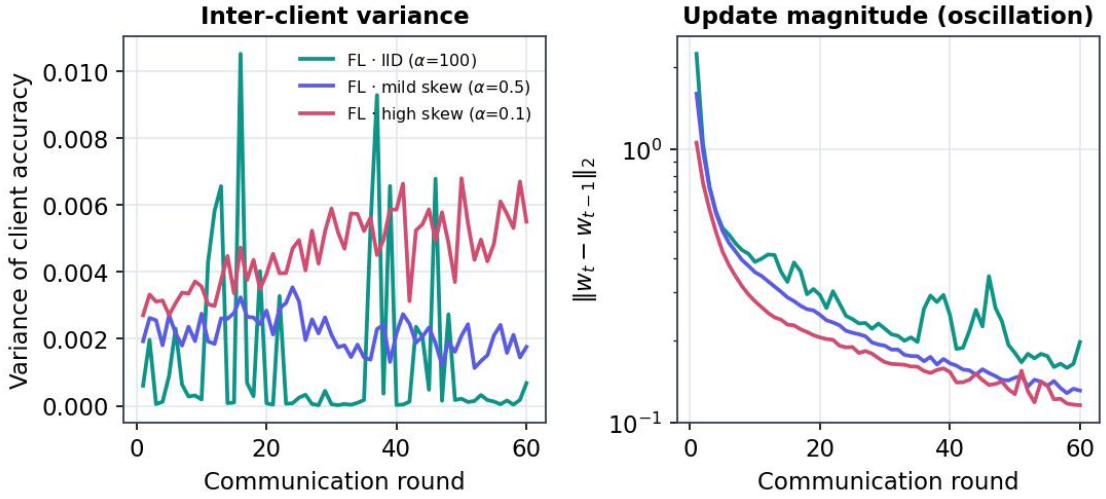

Figure 3: Stability diagnostics across heterogeneity regimes. Left: inter-client accuracy variance. Right: per-round update magnitude (log scale), a proxy for oscillation.

Table 2: End-state metrics by training regime (10 clients, 60 rounds).

Regime	Accuracy	ROC-AUC	Brier	ECE	Conv. round (95%)
Centralized	0.858	0.981	0.215	0.024	—
FL · IID (α=100)	0.819	0.974	0.273	0.039	34
FL · mild (α=0.5)	0.796	0.972	0.301	0.045	33
FL · high (α=0.1)	0.710	0.955	0.412	0.046	39

These results confirm the first hypothesis: non-IID data significantly degrade both final accuracy and convergence stability. The Brier score nearly doubles from the centralized baseline to the high-skew regime, and ECE rises in step, foreshadowing the calibration findings of Section 5.5.

5.2 Experiment 2 — Client imbalance

Holding the label distribution comparable, we contrast a balanced five-client split against an imbalanced split in which one client holds half the data. Figure 4 shows that final accuracy is barely affected (85.2% balanced versus 85.0% imbalanced), but the per-round update magnitude is consistently larger under imbalance (final-round $\|\Delta w\|$ of 0.337 versus 0.245), reflecting the dominant client's outsized pull on each aggregation step. In other words, volume imbalance alone — absent strong label skew — primarily manifests as reduced stability rather than reduced asymptotic accuracy.

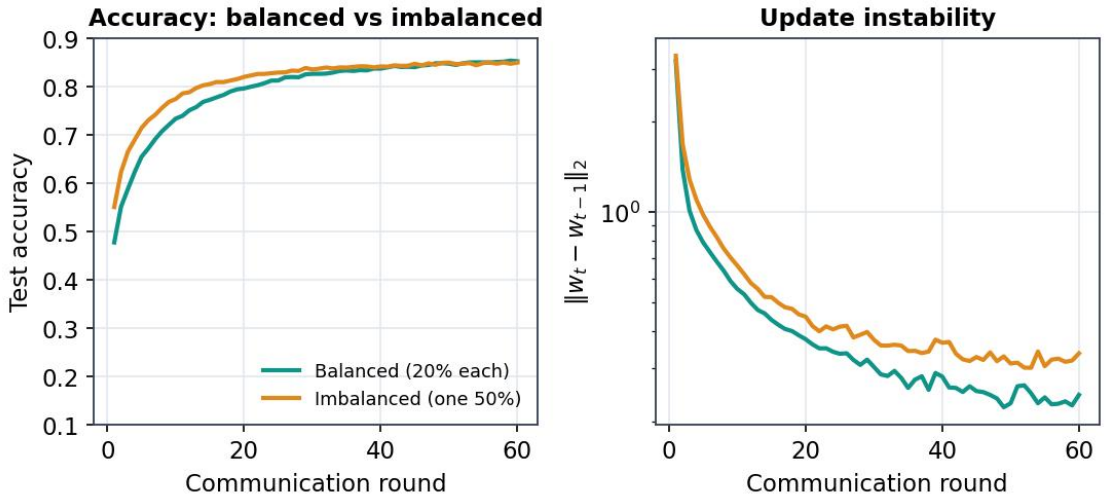

Figure 4: Effect of client imbalance. Accuracy is largely preserved (left), but update instability increases when one client holds 50% of the data (right, log scale).

Table 3: Balanced versus imbalanced client volume (5 clients).

Configuration	Accuracy	ECE	Final $\|\Delta w\|_2$
Balanced (20% each)	0.852	0.031	0.245
Imbalanced (one client 50%)	0.850	0.022	0.337

5.3 Experiment 3 — Differential-privacy noise injection

Figure 5 traces the privacy–utility tradeoff as the budget ε tightens (axis reversed so that stronger privacy lies to the right). Accuracy is essentially flat for large budgets ($\varepsilon \geq 50$, accuracy ≈ 80.7%) but degrades steeply once noise becomes substantial: at $\varepsilon = 10$ accuracy falls to 74.7%, and at $\varepsilon = 5$ it collapses to 64.3%. Calibration follows a non-monotonic path — moderate noise can incidentally reduce overconfidence, but strong noise ($\varepsilon = 5$) sharply inflates ECE to 0.125 as the aggregate model becomes erratic. Table 4 reports the full sweep.

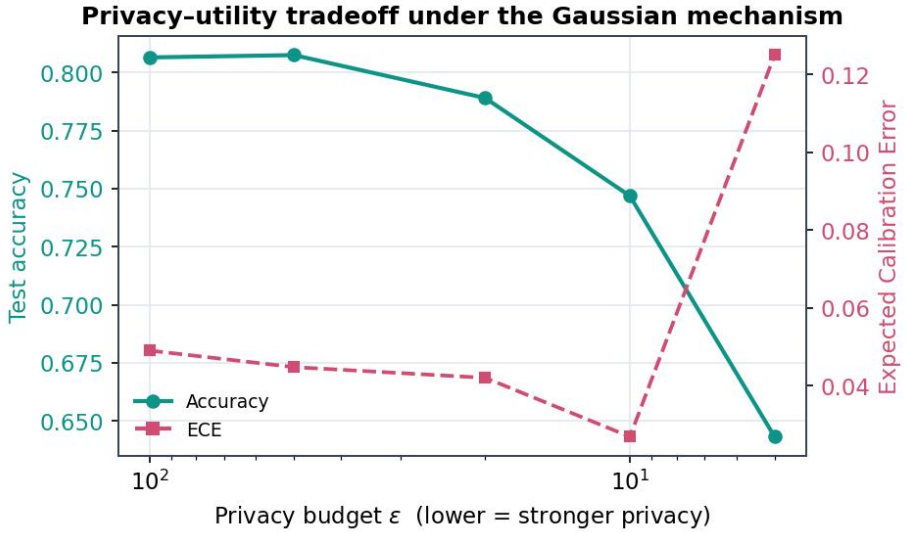

Figure 5: Privacy–utility tradeoff under the Gaussian mechanism. Accuracy (teal) is stable for loose budgets but collapses under strong privacy; ECE (rose) worsens sharply at ε = 5.

Table 4: Accuracy and calibration across privacy budgets (mild non-IID, 10 clients).

Noise σ	Budget ε	Accuracy	ECE	Regime
0.00	∞	0.811	0.052	No privacy
0.01	100	0.807	0.049	Negligible
0.02	50	0.808	0.045	Weak
0.05	20	0.789	0.042	Moderate
0.10	10	0.747	0.027	Strong
0.20	5	0.643	0.125	Severe

The curve quantifies a key operating insight: there exists a broad regime of effectively free privacy (ε ≥ 20) in which accuracy loss is under three points, beyond which the cost of additional privacy escalates rapidly. Practitioners can use such a curve to select a budget on the favorable side of the knee.

5.4 Experiment 4 — Communication efficiency

We next vary the number of communication rounds and measure accuracy against cumulative uplink cost, computed as rounds × clients × model size (the model has 2,634 parameters). Figure 6 (left) shows the familiar concave accuracy-versus-cost curve: accuracy climbs from 54.9% at five rounds to 80.7% at eighty rounds, but with steeply diminishing returns. The efficiency bar chart (right) makes the tradeoff explicit — accuracy gained per million transmitted parameters falls from 4.17 at five rounds to 0.38 at eighty rounds. Table 5 lists the operating points.

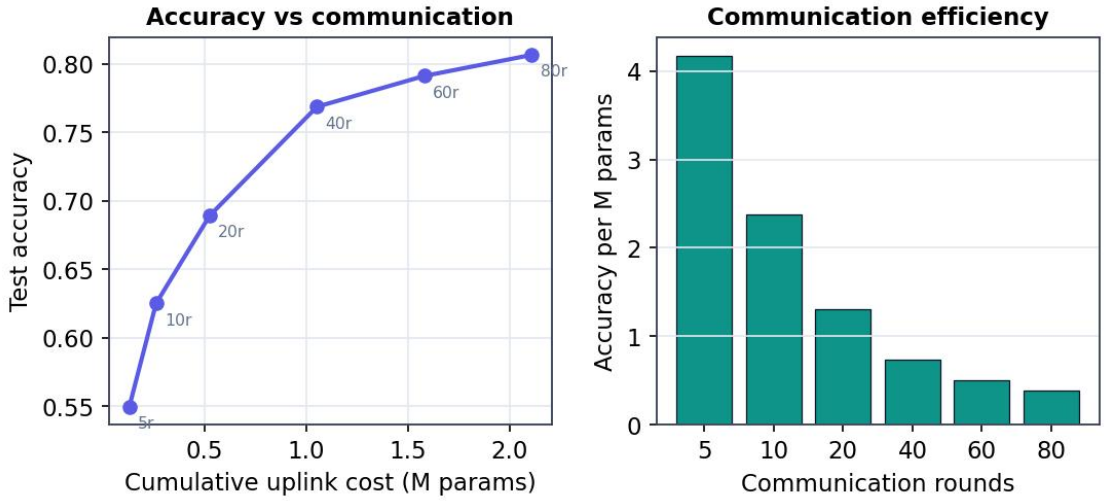

Figure 6: Communication efficiency. Left: accuracy versus cumulative uplink cost (round counts annotated). Right: accuracy gained per million transmitted parameters.

Table 5: Accuracy versus communication cost (mild non-IID, 10 clients).

Rounds	Accuracy	Uplink cost (M params)	Accuracy / M params
5	0.549	0.13	4.17
10	0.625	0.26	2.37
20	0.689	0.53	1.31

Rounds	Accuracy	Uplink cost (M params)	Accuracy / M params
40	0.769	1.05	0.73
60	0.791	1.58	0.50
80	0.807	2.11	0.38

The implication for deployment is that early rounds purchase the overwhelming majority of attainable accuracy at the lowest communication cost; extending training to chase the last few points is communication-inefficient and, in bandwidth-constrained edge settings, may not be justified.

5.5 Experiment 5 — Calibration stability

Finally, we ask whether federated training under heterogeneity damages calibration. Figure 7 presents reliability diagrams for the centralized model and for FedAvg under severe skew. The centralized model is well calibrated, with bars tracking the diagonal closely (ECE 0.024). Under severe skew the calibration gap (shaded) widens — the model is systematically overconfident, with its stated confidence exceeding observed accuracy across most bins, consistent with the elevated ECE reported in Table 2. This supports treating calibration as a distinct failure mode of decentralized training rather than a byproduct of accuracy loss alone, since overconfidence carries direct consequences for risk-sensitive deployment.

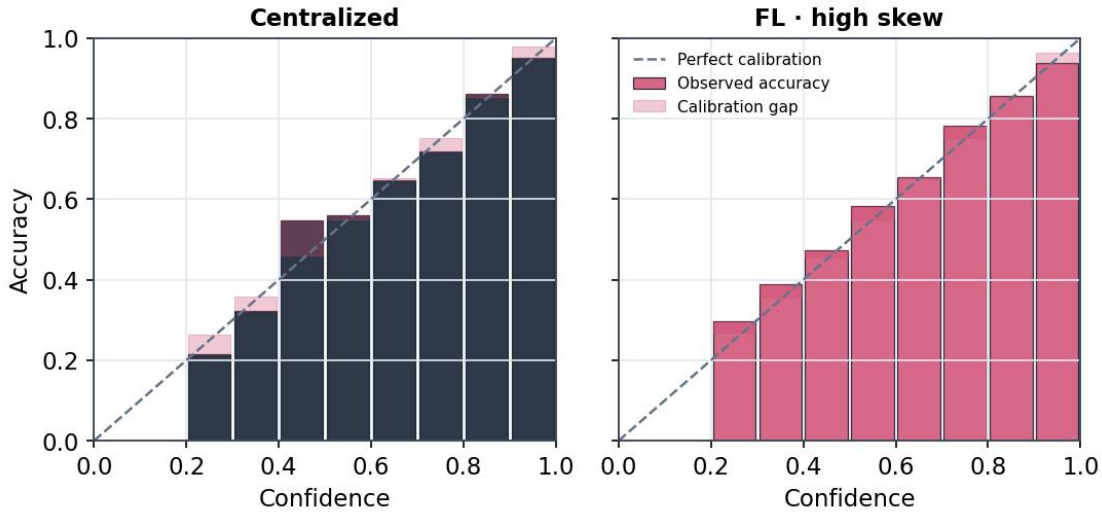

Figure 7: Reliability diagrams. The centralized model (left) is well calibrated; FedAvg under severe skew (right) is systematically overconfident, with a larger calibration gap.

5.6 Composite robustness surface

Figure 8 renders the proposed Robustness Index across the joint heterogeneity–noise grid. RI is highest (0.80) in the benign corner — IID data with no privacy noise — and decays along both axes, reaching its minimum (0.07) under severe skew combined with strong noise. The roughly multiplicative decay confirms that the two stressors compound: a system tolerable under either condition alone can become unacceptable under both. The single-number summary makes such joint regimes directly comparable, which is the metric's intended use.

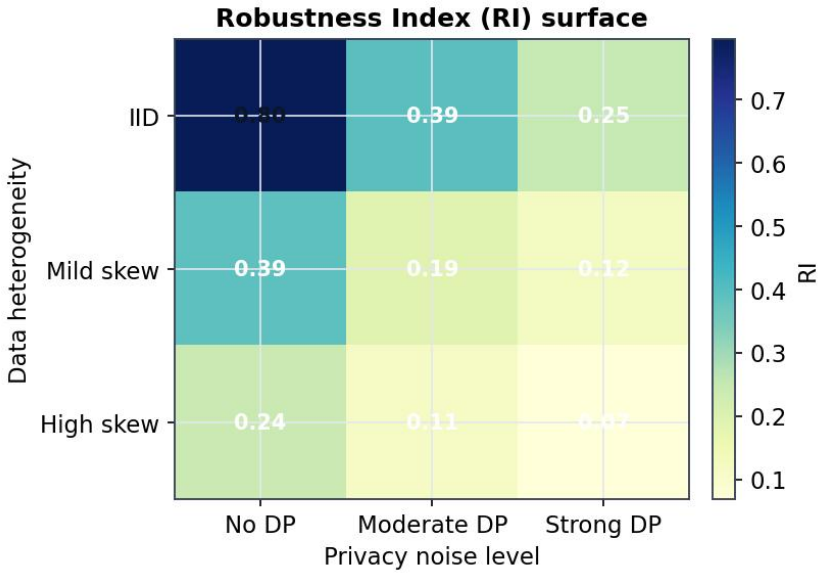

Figure 8: Robustness Index surface over heterogeneity and privacy noise. RI decays along both axes and is lowest where severe skew and strong noise coincide.

VI. Discussion

6.1 Principal findings

Three findings stand out. First, heterogeneity — not imbalance — is the dominant driver of accuracy loss: severe label skew costs roughly fifteen accuracy points relative to the centralized baseline, whereas volume imbalance alone costs almost none and instead manifests as reduced stability. Second, the privacy–utility relationship is strongly non-linear, with a wide low-cost regime followed by a sharp knee, implying that sensible budget selection can secure meaningful privacy at modest accuracy cost. Third, calibration degrades independently of and in addition to accuracy under heterogeneity, a result with direct bearing on whether federated models can be trusted in decision-critical applications.

6.2 Practical implications

For practitioners, the results argue for measuring heterogeneity before deployment and budgeting communication around the concave efficiency curve rather than a fixed round count. Where severe skew is unavoidable, a drift-mitigating aggregator such as FedProx is the lowest-cost remedy. Reporting calibration alongside accuracy should be standard practice for federated systems, and the Robustness Index offers a compact way to communicate joint operating risk to stakeholders.

6.3 Threats to validity

Several caveats temper the conclusions. The study uses a synthetic classification task and a compact neural classifier to enable exact, reproducible control; absolute numbers will differ on larger models and real datasets, although the qualitative trends are consistent with the broader literature. The differential-privacy treatment uses a Gaussian-noise proxy with an order-of-magnitude budget mapping rather than a formally accounted mechanism with gradient clipping; the curve should be read as illustrative of the tradeoff shape, not as a certified privacy guarantee. The Robustness Index uses ordinal severity levels and is intended as an interpretable summary rather than a calibrated probabilistic measure. Finally, results reflect single-seed runs per configuration; multi-seed averaging would tighten the variance estimates.

6.4 Future work

Natural extensions include replacing the noise proxy with a formally accounted DP-FedAvg mechanism, repeating the stress test under FedProx and SCAFFOLD to quantify how much each mitigation recovers, scaling to standard federated benchmarks and larger architectures, and integrating uncertainty quantification so that calibration and epistemic uncertainty are reported jointly. A multi-seed study with confidence intervals would also strengthen the statistical claims.

VII. Conclusion

This paper presented a controlled, multi-axis audit of federated learning under realistic conditions of data heterogeneity, client imbalance, and differential-privacy noise. Holding the model and optimizer fixed, we quantified how FedAvg degrades relative to a centralized baseline across accuracy, convergence stability, communication cost, and calibration. Severe label skew reduced accuracy from 85.8% to 71.0% and tripled inter-client variance; volume imbalance chiefly harmed stability rather than accuracy; strong privacy ($\varepsilon = 5$) drove accuracy to 64.3% and sharply worsened calibration; and the accuracy-per-communication curve exhibited pronounced diminishing returns. We compared candidate aggregation methodologies and justified FedAvg as the controlled reference for such an audit, identifying FedProx as the lowest-cost mitigation. Finally, we proposed a composite Robustness Index that compresses the joint degradation surface into a single interpretable number. Beyond building a model, the work systematically maps the structural limits of decentralized machine learning — a measurement-centric complement to the method-centric federated-learning literature, and a template that practitioners can reuse to characterize their own deployments before trusting them.

Appendix A. Reproducibility and Hyperparameter Settings

All experiments were produced by a single NumPy harness with a fixed random seed, so every reported figure and table is reproducible end to end. Table 6 records the configuration shared across runs; deviations for individual experiments (number of clients, rounds, Dirichlet α, and noise σ) are stated in the relevant subsections of Section V.

Table 6: Shared experimental configuration.

Setting	Value	Notes
Dataset size	24,000 samples	75/25 train/test, stratified
Features / classes	30 / 10	20 informative features
Model	MLP, 1 hidden layer	64 ReLU units, softmax output
Parameters	2,634	used for communication cost
Loss / optimizer	Cross-entropy / SGD	learning rate 0.15
Batch size	64	mini-batch SGD
Local epochs (E)	1	per communication round
Default rounds	60	swept in Experiment 4
Default clients (K)	10	5 in Experiment 2
ECE bins	15	10 for reliability diagrams
Random seed	42	fixed for reproducibility

The differential-privacy treatment adds zero-mean Gaussian noise to client updates with scale σ and reports an order-of-magnitude budget $\varepsilon \propto 1/\sigma$; it is a tradeoff-shape proxy rather than a formally accounted mechanism, as discussed in Section 6.3.

Appendix B. Notation

Table 7 summarizes the symbols used throughout the paper.

Table 7: Summary of notation.

Symbol	Meaning
K	Number of participating clients
n_k , n	Local dataset size of client k; total number of examples
w_t	Global model weights at communication round t
w_k	Locally updated weights returned by client k
F_k , F	Local objective of client k; size-weighted global objective
E	Number of local epochs per round
α	Dirichlet concentration parameter (larger = more IID)
σ	Standard deviation of Gaussian privacy noise
ε	Differential-privacy budget (smaller = stronger privacy)
ECE	Expected Calibration Error
RI	Proposed composite Robustness Index

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. AISTATS, 2017, pp. 1273–1282.

[2] P. Kairouz et al., "Advances and Open Problems in Federated Learning," Foundations and Trends in Machine Learning, vol. 14, no. 1–2, pp. 1–210, 2021.

[3] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks (FedProx)," in Proc. MLSys, 2020.

[4] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning," in Proc. ICML, 2020, pp. 5132–5143.

[5] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization (FedNova)," in Proc. NeurIPS, 2020.

[6] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair Resource Allocation in Federated Learning (q-FedAvg)," in Proc. ICLR, 2020.

[7] S. J. Reddi et al., "Adaptive Federated Optimization," in Proc. ICLR, 2021.

[8] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated Learning with Non-IID Data," arXiv:1806.00582, 2018.

[9] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification," arXiv:1909.06335, 2019.

[10] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014.

[11] M. Abadi et al., "Deep Learning with Differential Privacy," in Proc. ACM CCS, 2016, pp. 308–318.

[12] R. C. Geyer, T. Klein, and M. Nabi, "Differentially Private Federated Learning: A Client-Level Perspective," arXiv:1712.07557, 2017.

[13] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning Differentially Private Recurrent Language Models," in Proc. ICLR, 2018.

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. ICML, 2017, pp. 1321–1330.

[15] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining Well-Calibrated Probabilities Using Bayesian Binning," in Proc. AAAI, 2015, pp. 2901–2907.

[16] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950.

[17] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," ACM Trans. Intelligent Systems and Technology, vol. 10, no. 2, pp. 1–19, 2019.

[18] D. J. Beutel et al., "Flower: A Friendly Federated Learning Research Framework," arXiv:2007.14390, 2020.