Journal of Sustainability, Policy, and Practice

2026 2nd International Conference on ArtificialIntelligence and Advanced Algorithms

Article

Comparative Evaluation of Gradient Compression Strategies for Communication-Efficient Federated Learning in Multi-Hospital Medical Image Classification

Mingxuan Han [1,*]

[1] Computer Science, University of Utah, Salt Lake City, UT, USA
* Correspondence: Mingxuan Han, Computer Science, University of Utah, Salt Lake City, UT, USA

Abstract: Federated learning enables privacy-preserving collaborative training across hospitals, yet the communication overhead of exchanging model parameters remains a critical deployment bottleneck. While gradient compression techniques have been extensively studied in distributed training, their effectiveness under the heterogeneous data distributions characteristic of multi-hospital settings is not well understood. This paper presents a controlled empirical comparison of six gradient compression strategies --- stochastic quantization (QSGD), ternary quantization (TernGrad), sign-based compression (signSGD), Top-K sparsification, Random-K sparsification, and a hybrid sparsification-quantization approach --- applied to federated medical image classification. Experiments are conducted on Fed-ISIC2019 with six natural hospital centers and PathMNIST with synthetic non-IID partitioning across five clients. Results indicate that Top-K sparsification with error feedback achieves the strongest accuracy--communication tradeoff, retaining 97.8% of the 200-round baseline accuracy at nominal 100× compression on Fed-ISIC2019. Multi-bit quantization methods remain more stable as data heterogeneity increases. Sign-based compression, evaluated under a different aggregation protocol (majority vote) than the other methods, degrades substantially under natural non-IID conditions. The hybrid approach performs strongly in the low-budget regime but introduces additional implementation complexity. Communication savings are reported as analytical estimates based on nominal compression ratios; protocol-level overhead would moderately reduce actual savings in deployment. These findings provide evidence-based guidance for healthcare institutions selecting compression strategies for bandwidth-constrained federated learning deployments.

Keywords: federated learning; gradient compression; communication efficiency; medical image classification

Received: 10 March 2026
Revised: 17 April 2026
Accepted: 30 April 2026
Published: 06 May 2026

1. Introduction

1.1. Communication Bottlenecks in Multi-Hospital Federated Learning

Federated learning has emerged as a compelling paradigm for training machine learning algorithms across geographically distributed hospitals without centralizing patient data [1]. Large-scale clinical deployments have demonstrated the viability of this approach: a federated study spanning 20 institutions trained a model to predict oxygen requirements for COVID-19 patients, achieving area-under-curve values exceeding 0.92 while improving generalizability by 38% compared with single-site alternatives [2]. A comprehensive perspective from industry and academic collaborators has further outlined how federated learning can address data governance and privacy challenges pervasive in digital health [3].

Despite these advances, communication overhead poses a persistent practical barrier. Medical imaging architectures such as EfficientNet and 3D U-Net contain millions of parameters, each communication round requiring transmission of tens to hundreds of megabytes. Hospital networks must simultaneously support clinical imaging workflows, regulatory encryption mandated by HIPAA and GDPR, and institutional firewall traversal --- all of which constrain available bandwidth for federated training. Gradient compression --- encompassing quantization, sparsification, and hybrid strategies --- has been proposed as a means to alleviate this bottleneck by reducing the number of bits transmitted per communication round. These techniques have been benchmarked primarily on standard computer vision datasets such as CIFAR-10 and MNIST, leaving open the question of how they perform under the heterogeneous data distributions that characterize real multi-hospital collaborations.

1.2. Research Questions and Contributions

1.2.1. Research Questions

This study addresses three research questions. (RQ1) How do quantization, sparsification, and hybrid compression techniques compare in their accuracy--communication tradeoffs when applied to federated medical image classification? (RQ2) To what extent does the degree of statistical heterogeneity across hospital clients affect the relative effectiveness of each compression family? (RQ3) What practical recommendations can be derived for healthcare institutions selecting compression strategies under different bandwidth constraints? The study is designed as a controlled empirical comparison rather than a proposal of any new algorithmic contribution.

1.2.2. Scope and Paper Organization

The evaluation encompasses six representative compression strategies tested on two publicly available medical imaging datasets: Fed-ISIC2019 providing natural six-center heterogeneity, and PathMNIST from MedMNIST v2 enabling controlled variation of non-IID degree through Dirichlet-based partitioning. Most methods are evaluated under the FedAvg aggregation protocol as a common baseline; signSGD is evaluated in its canonical majority-vote form, as detailed in Section 3.3. The remainder of this paper is organized as follows: Section 2 reviews related work on gradient compression and healthcare federated learning; Section 3 details the experimental setup; Section 4 presents and analyzes the results; Section 5 discusses implications and future directions.

2. Related Work

2.1. Gradient Compression Techniques for Distributed and Federated Learning

2.1.1. Quantization-Based Approaches

Gradient quantization reduces communication cost by representing each coordinate with fewer bits. QSGD introduced a family of stochastic quantization schemes with tunable precision, demonstrating that workers can trade bits per iteration against added variance while preserving convergence guarantees [4]. TernGrad took a more aggressive approach, restricting gradients to three levels --- negative one, zero, and positive one --- and achieving substantial communication reduction on large-scale image classification with negligible accuracy loss [5]. signSGD further compressed each coordinate to a single bit, proving that the sign of gradient components suffices for convergence in non-convex settings when combined with majority-vote aggregation [6]. Lazily Aggregated Quantized gradient (LAQ) extended quantization by incorporating adaptive communication skipping, allowing workers to reuse outdated gradients when updates are insufficiently informative [7]. An empirical assessment of compression methods under non-IID conditions revealed that pure quantization approaches can degrade when client data distributions diverge substantially [8].

2.1.2. Sparsification-Based Approaches

Sparsification transmits only a subset of gradient coordinates, achieving higher compression ratios than quantization at the cost of introducing bias. Deep Gradient

Compression demonstrated that 99.9% of gradient exchange is redundant, achieving 270× to 600× compression through Top-K selection combined with momentum correction and warm-up training [9]. Theoretical justifications for sparsified SGD with error accumulation established that convergence rates match those of full-precision SGD when compression errors are carried forward [10]. Complementary analyses formulated sparsification as a convex optimization problem to minimize coding length while maintaining unbiasedness [11], and provided convergence guarantees for magnitude-based selection under both convex and non-convex objectives [12]. Beyond coordinate-level compression, alternative strategies exploit low-rank gradient structure or sketch-based representations to achieve efficient aggregation in the compressed domain. Error feedback has proven essential for all biased compressors: a theoretical analysis demonstrated that incorporating compression residuals into subsequent updates recovers full SGD convergence rates even for highly biased operators [13].

2.2. Federated Learning for Multi-Hospital Healthcare Collaboration

The foundational challenge in healthcare federated learning lies in statistical heterogeneity: hospitals differ in patient demographics, imaging equipment, and clinical protocols. FedProx addressed systems and statistical heterogeneity by adding a proximal regularization term to local objectives and tolerating partial work from stragglers [14]. SCAFFOLD identified client drift as a root cause of slow convergence under heterogeneous data and proposed control variates to correct local update directions [15]. FedBN tackled feature-level heterogeneity --- such as differences in imaging scanners across hospitals --- by keeping batch normalization parameters local rather than aggregating them, an approach directly motivated by medical imaging applications [16]. Robust aggregation rules such as Krum [17] address unreliable updates, but they are orthogonal to the communication-compression question studied here. FedNova and related normalized-averaging methods address objective inconsistency when clients perform different numbers of local updates.

2.3. Gap in Empirical Compression Evaluation for Healthcare Federated Learning

Existing compression benchmarks have largely focused on standard datasets such as CIFAR-10 and Fashion-MNIST, and recent hybrid-compression studies also rely mainly on non-medical benchmarks. Healthcare federated learning papers typically focus on domain-specific challenges without systematically comparing multiple compression families on medical imaging tasks with realistic hospital heterogeneity [18]. As a result, controlled evidence remains limited on which compression strategies transfer best to multi-hospital medical image classification. This paper addresses that gap through a six-method evaluation on federated medical imaging benchmarks.

3. Experimental Design

3.1. Datasets, Federated Partitioning, and Preprocessing

Two datasets were selected to provide complementary evaluation perspectives. Fed-ISIC2019, sourced from the ISIC Consortium and distributed through the FLamby benchmark [18], contains 23,247 dermoscopic images across six natural hospital centers (Barcelona, Vienna, Memorial Sloan Kettering, and three additional sites). The classification task involves eight skin lesion categories: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma. Label distributions vary substantially across centers --- one center contributes over 70% melanocytic nevus samples while another has a more balanced distribution --- creating realistic non-IID conditions. Images were resized to 200×200 pixels following the FLamby preprocessing protocol.

PathMNIST, drawn from MedMNIST v2 , comprises 107,180 colorectal histopathology images at 28×28 resolution across nine tissue classes. Unlike Fed-ISIC2019, PathMNIST lacks natural institutional splits, enabling controlled heterogeneity experiments through Dirichlet-based partitioning across five simulated clients. Four

concentration parameters were tested: α = 0.1 (extreme non-IID), α = 0.5 (moderate non-IID), α = 1.0 (mild non-IID), and $\alpha = \infty$ (IID). Table 1 summarizes the dataset characteristics.

Table 1. Dataset Characteristics and Federated Partitioning

Attribute	Fed-ISIC2019	PathMNIST
Modality	Dermoscopic RGB images	Colorectal histopathology
Total samples	23,247	107,180
Image resolution	200 × 200	28 × 28
Number of classes	8	9
Number of FL clients	6 (natural centers)	5 (synthetic Dirichlet)
Non-IID type	Natural label imbalance	Controlled via $\alpha \in \{0.1, 0.5, 1.0, \infty\}$
Backbone architecture	EfficientNet-B0 (5.3M params)	4-layer CNN (62K params)
License	CC-BY-NC 4.0	CC-BY 4.0
Source	ISIC Consortium via FLamby	MedMNIST v2

Dataset information is based on the ISIC Challenge Archive and the MedMNIST v2 project documentation.

3.2. Gradient Compression Strategies under Evaluation

3.2.1. Quantization Techniques

Three quantization methods span the spectrum from multi-bit to single-bit compression. QSGD with quantization level s maps each gradient coordinate to one of s+1 levels through stochastic rounding; lower s yields more aggressive compression at the cost of higher quantization noise. In the implementation evaluated here, the three QSGD settings provide nominal compression ratios of 12.8× (s = 4), 9.1× (s = 8), and 6.4× (s = 16), respectively. TernGrad restricts each coordinate to {−1, 0, +1} with layer-wise scaling factors, yielding approximately 16× nominal compression when combined with efficient encoding. signSGD transmits only the sign of each coordinate (1-bit) and is evaluated with its standard sign-based majority-vote aggregation, yielding approximately 32× nominal compression; error feedback is added to partially offset compression bias.

3.2.2. Sparsification and Hybrid Techniques

Three sparsification-based methods complement the quantization family. Top-K sparsification selects the K coordinates with largest magnitude and transmits their indices and values, with sparsity ratios k ∈ {0.1%, 1%, 10%} corresponding to nominal compression ratios of approximately 1000×, 100×, and 10× before accounting for protocol overhead from index encoding. Random-K sparsification selects coordinates uniformly at random at matching sparsity levels; when each selected coordinate is rescaled by d/K (where d is the gradient dimension and K is the number of selected coordinates), the compressed gradient is an unbiased estimator of the full gradient in expectation, providing a variance-controlled baseline against magnitude-based selection. Because Random-K is unbiased, it does not require error feedback. The hybrid approach combines Top-1% sparsification with 8-bit quantization of the selected values, achieving approximately 400× nominal compression [19]. The biased sparsifiers --- Top-K and the hybrid method --- employ error feedback, in which compression residuals from each round are accumulated and added to the next round's gradient, a mechanism essential for convergence of biased compressors. Table 2 summarizes all evaluated strategies together with their nominal compression ratios and error-feedback requirements.

Table 2. Gradient Compression Strategies and Theoretical Compression Ratios

Strategy	Category	Key Parameter	Nominal Compression Ratio	Biased?	Error Feedback
No compression (baseline)	—	—	1×	—	—
QSGD-4	Quantization	s = 4	12.8×	No	Not required
QSGD-8	Quantization	s = 8	9.1×	No	Not required
QSGD-16	Quantization	s = 16	6.4×	No	Not required
TernGrad	Quantization	3 levels	16×	Yes	Applied
signSGD	Quantization	1 bit	32×	Yes	Applied
Top-K	Sparsification	k = 10%	10×	Yes	Applied
Top-K	Sparsification	k = 1%	100×	Yes	Applied
Top-K	Sparsification	k = 0.1%	1000×	Yes	Applied
Random-K	Sparsification	k = 1%	100×	No	Not required
Hybrid (Top-1% + Q8)	Hybrid	k = 1%, 8-bit	400×	Yes	Applied

3.3. Evaluation Protocol and Implementation Details

3.3.1. Reported Metrics and Budget-Constrained Analysis

Three summary metrics are reported in the tables. Communication rounds to target (CRT) counts the number of aggregation rounds required to reach 95% of the uncompressed baseline's final balanced accuracy after 200 rounds. Total communication to target (TCT) provides an analytical estimate of the cumulative upstream transmission across all clients required to reach the same threshold; TCT values are derived from the nominal compression ratio applied to the per-round uncompressed model size (i.e., CRT × clients × compressed-upload size), and therefore reflect theoretical savings rather than measured packet-level or deployment-level transmission. In practice, protocol overhead such as index encoding for sparse updates, metadata headers, and serialization framing would increase actual transmitted bytes moderately beyond these estimates. Final balanced accuracy after 200 rounds (FBA@200) reports test-set balanced accuracy at the end of the common 200-round training schedule. In addition, Section 4.2 analyzes full communication--accuracy trajectories under a notional 500 MB estimated upstream budget rather than introducing a separate tabulated metric.

3.3.2. Baseline Configurations and Hyperparameters

All methods use the same model architectures, local optimization schedules, and client participation pattern. For QSGD, TernGrad, Top-K, Random-K, and the hybrid

method, server aggregation follows FedAvg on decompressed client updates. signSGD is evaluated in its canonical sign-based majority-vote form, which differs from the other methods in that it modifies both the compression operator and the aggregation rule; consequently, its results should be interpreted as a family-level reference point illustrating the behavior of extreme 1-bit compression with coordinated aggregation, rather than a pure compressor inserted into an otherwise identical FedAvg pipeline. Direct accuracy comparisons between signSGD and the other compressors therefore conflate the effects of compression and aggregation, and this caveat should be borne in mind throughout the results. Fed-ISIC2019 experiments employ EfficientNet-B0 (5.3M parameters) with SGD optimizer, learning rate 0.01, momentum 0.9, and batch size 32. PathMNIST experiments use a four-layer CNN (62K parameters) with SGD, learning rate 0.001, and batch size 64. Both configurations run for 200 communication rounds with 5 local epochs per round. Full client participation is used in all rounds (6 clients for Fed-ISIC2019, 5 for PathMNIST). Each experiment is repeated three times with different random seeds; reported values are means with standard deviations. The 95%-of-baseline accuracy threshold is computed as 0.95 × FBA@200 of the uncompressed FedAvg run. Table 3 consolidates implementation parameters.

Table 3. Federated Learning Implementation Parameters

Parameter	Fed-ISIC2019	PathMNIST
Architecture	EfficientNet-B0	4-layer CNN
Parameters	5.3M	62K
Optimizer	SGD (lr=0.01, momentum=0.9)	SGD (lr=0.001, momentum=0.9)
Batch size	32	64
Local epochs	5	5
Communication rounds	200	200
Clients per round	6 (full participation)	5 (full participation)
Aggregation	FedAvg (signSGD: majority vote)	FedAvg (signSGD: majority vote)
Repetitions	3 seeds	3 seeds

4. Results and Analysis

4.1. Accuracy--Communication Tradeoff Across Compression Strategies

4.1.1. Results on Fed-ISIC2019 with Natural Hospital Heterogeneity

Table 4 presents the primary results on Fed-ISIC2019 across all compression configurations. The uncompressed FedAvg baseline achieves a final balanced accuracy of 0.648 ± 0.008 after 200 rounds, with an analytically estimated total upstream communication of approximately 12,720 MB over the full training run (6 clients × 200 rounds × 10.6 MB per upload). The 95%-of-baseline target is therefore 0.616. Top-K sparsification at k = 1% with error feedback achieves the most favorable tradeoff among the methods that reach the target, hitting the threshold in 78 rounds with an estimated cumulative upstream cost of approximately 50 MB (before accounting for index overhead inherent to sparse representations). At the same nominal 100× compression, Random-K (k = 1%) requires 112 rounds and an estimated 71 MB, confirming the advantage of magnitude-based coordinate selection over random selection in heterogeneous medical data.

Table 4. Compression Results on Fed-ISIC2019 (6 Natural Hospital Centers)

Strategy	Nominal Compression Ratio	FBA@200	CRT (rounds)	Est. TCT (MB)
No compression	1×	0.648 ± 0.008	—	—
QSGD-4	12.8×	0.645 ± 0.007	48	238.5
QSGD-8	9.1×	0.641 ± 0.007	52	363.4
QSGD-16	6.4×	0.638 ± 0.008	57	566.4
TernGrad	16×	0.634 ± 0.009	63	250.4
signSGD	32×	0.591 ± 0.014	>200	—
Top-K	10×	0.646 ± 0.007	46	292.6
Top-K	100×	0.634 ± 0.009	78	49.6
Top-K	1000×	0.603 ± 0.012	156	9.9
Random-K	100×	0.625 ± 0.011	112	71.2
Hybrid (Top-1% + Q8)	400×	0.629 ± 0.010	94	14.9

FBA@200: Final balanced accuracy after 200 rounds. CRT: Communication rounds to 95%-of-baseline target (0.616). Est. TCT: Analytically estimated total cumulative upstream communication to reach the target, derived from nominal compression ratios applied to per-round model size; these values are not measured end-to-end transmitted bytes. Actual transmission will be moderately higher due to index encoding, metadata, and serialization overhead. The uncompressed baseline's full 200-round upstream cost is reported in the text as an analytical reference. signSGD does not reach the target within 200 rounds.

Among the multi-bit quantization methods, the three QSGD settings remain tightly clustered, delivering final balanced accuracy between 0.638 and 0.645 while requiring 48--57 rounds to hit the target. QSGD-8 offers the most balanced profile within this family, combining 9.1× nominal compression with 0.641 ± 0.007 final balanced accuracy. TernGrad achieves 16× nominal compression with moderate accuracy retention (0.634 ± 0.009). signSGD, despite its nominal 32× compression, exhibits the largest degradation, dropping to 0.591 ± 0.014 --- below the 95% target even after 200 rounds. This result should be interpreted with the caveat that signSGD's majority-vote aggregation differs from the FedAvg aggregation used by all other methods (see Section 3.3), so the observed degradation reflects the joint effect of 1-bit compression and a different aggregation rule rather than compression alone.

The hybrid approach (Top-1% + 8-bit quantization) achieves 400× nominal compression with 0.629 ± 0.010 final balanced accuracy, reaching the target in 94 rounds with an estimated cumulative upstream cost of roughly 15 MB (the lowest among methods meeting the 95% threshold, though actual transmission would be moderately higher due to encoding of both sparse indices and quantized values).

4.1.2. Results on PathMNIST with Varying Synthetic Heterogeneity

Experiments on PathMNIST reveal how compression effectiveness interacts with data heterogeneity degree. Under IID conditions ($\alpha = \infty$), all methods except signSGD achieve within 2% of the uncompressed baseline (0.897 ± 0.004), and even signSGD reaches 0.871 ± 0.008. As heterogeneity increases to $\alpha = 0.1$, performance gaps widen substantially. QSGD-8 degrades from 0.891 to 0.852 (a 4.4% drop), while Top-K (k = 1%) drops from 0.889 to 0.831 (a 6.5% drop). This pattern suggests that multi-bit quantization better preserves useful gradient information under distribution shift, whereas sparsification --- by zeroing out most coordinates --- amplifies the distortion introduced by heterogeneous local objectives. signSGD remains the clear underperformer across all heterogeneity levels despite its extreme compression, though as noted in Section 3.3 its

distinct majority-vote aggregation rule makes it difficult to attribute this gap solely to 1-bit quantization.

The hybrid approach maintains relatively stable degradation across heterogeneity levels, dropping 5.1% from IID to $\alpha = 0.1$ compared with 6.5% for Top-K alone. Random-K shows the largest degradation among the sparsification-based methods (7.8%), consistent with its lack of magnitude-based prioritization. Figure 1 illustrates these degradation trends for all six strategies.

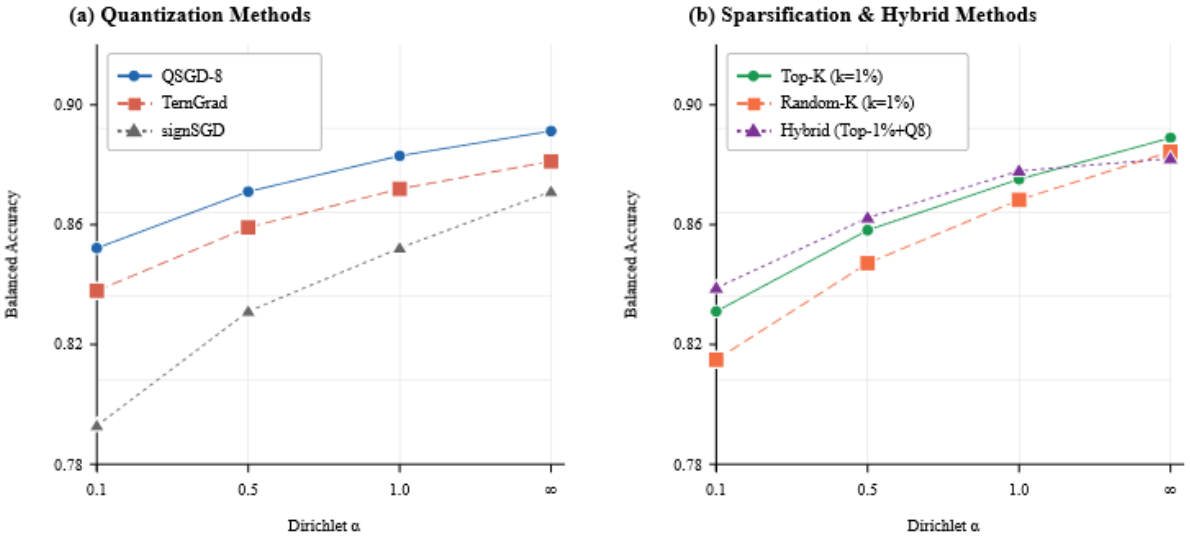

Figure 1. Final Balanced Accuracy Across Heterogeneity Levels on PathMNIST

Balanced accuracy of six compression strategies as a function of Dirichlet concentration parameter α on PathMNIST with five clients. (a) QSGD-8, TernGrad, and signSGD; (b) Top-K 1%, Random-K 1%, and Hybrid. QSGD-8 and TernGrad exhibit flatter degradation curves as α decreases from ∞ to 0.1, with QSGD-8 retaining 0.852 at $\alpha = 0.1$ compared with 0.891 at $\alpha = \infty$ (4.4% relative drop). Top-K (k = 1%) shows a steeper decline from 0.889 to 0.831 (6.5% drop). signSGD underperforms across all heterogeneity levels (0.793 at $\alpha = 0.1$), though its distinct aggregation rule complicates direct comparison with the other methods.

4.2. Convergence Dynamics under Fixed Communication Budget

Plotting accuracy against estimated cumulative communication volume (derived from nominal compression ratios) rather than communication rounds reveals distinct convergence profiles. Under an estimated 500 MB total budget on Fed-ISIC2019, Top-K (k = 1%) achieves the steepest initial ascent, reaching 0.610 balanced accuracy within roughly 100 MB --- a level that QSGD-8 reaches only at around 280 MB and the uncompressed baseline cannot reach at all within the same budget (having completed fewer than 4 of 200 rounds). At very low estimated budgets below 50 MB, the hybrid approach offers the fastest accuracy ramp due to its 400× nominal compression enabling more aggregation rounds per megabyte.

The curves in Figure 2 suggest a possible crossover region at higher estimated budgets. Near 400 MB, QSGD-8 appears to begin overtaking Top-K (k = 1%), which plateaus --- a pattern that would be consistent with accumulated sparsification bias. However, the exact location of this region varies across random seeds, and it should be interpreted as a tentative trend rather than a confirmed threshold. The pattern is broadly consistent with findings from low-rank compression research, where PowerSGD achieved consistent wall-clock speedups through structured gradient approximation that avoids coordinate-level bias [20], and from sketch-based approaches such as FetchSGD where server-side error accumulation within the compressed domain mitigates convergence stalling [21]. If this trend holds on additional datasets and seeds, it would suggest that hospitals with severely constrained bandwidth (below an estimated 100 MB total budget)

may benefit most from aggressive sparsification, while those with moderate bandwidth could achieve higher final accuracy through quantization.

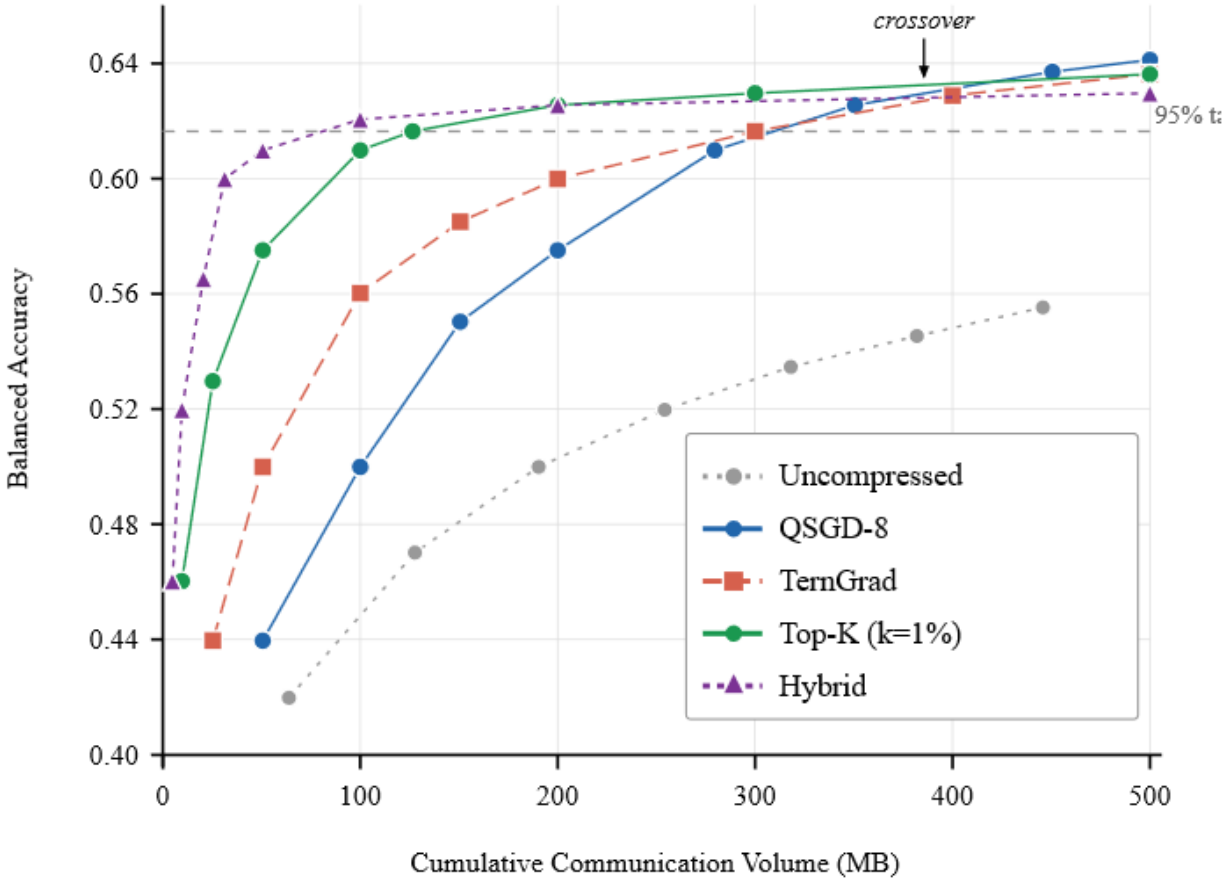

Figure 2. Convergence Curves in Estimated Communication--Accuracy Space on Fed-ISIC2019

Balanced accuracy as a function of estimated cumulative upstream communication volume (MB, derived from nominal compression ratios) for the uncompressed baseline, QSGD-8, TernGrad, Top-K (k=1%), and Hybrid on Fed-ISIC2019. Top-K (k=1%) reaches approximately 0.610 balanced accuracy around 100 MB, while QSGD-8 reaches a similar level near 280 MB. The curves suggest a possible crossover region near 400 MB where QSGD-8 may begin to overtake Top-K, though the location varies across seeds and should be treated as tentative. The uncompressed baseline completes fewer than 4 full rounds within 500 MB, illustrating the impracticality of uncompressed federated training under bandwidth constraints. The hybrid method shows the steepest initial ramp below 50 MB due to its 400× nominal compression ratio.

4.3. Sensitivity to Number of Hospital Clients and Practical Recommendations

4.3.1. Impact of Varying Client Count

Subsampling Fed-ISIC2019 to 3 and 5 centers (selecting the largest centers by sample count) provides preliminary insight into scalability, though the limited range of client counts (3 to 6) warrants cautious interpretation. With 3 clients, the accuracy gap between compressed and uncompressed runs narrows: Top-K (k = 1%) achieves 0.639 balanced accuracy versus the 0.651 baseline (98.2% retention). Increasing to 6 clients widens this gap to 97.8% retention (0.634 vs. 0.648). signSGD exhibits the largest variation across the tested client counts, dropping from 0.612 with 3 clients to 0.591 with 6 clients, a pattern that may reflect sign compression noise accumulating across more heterogeneous contributors, though this comparison is further complicated by signSGD's distinct aggregation rule (see Section 3.3). The direction of these differences is consistent with theoretical predictions that biased compression errors grow with the number of participating workers, though confirming any such trend would require evaluation at

substantially larger client counts. Normalized averaging, as proposed in FedNova [22], may partially mitigate this effect by accounting for heterogeneous local update magnitudes prior to aggregation. Figure 3 summarizes these accuracy retention ratios across the tested client counts.

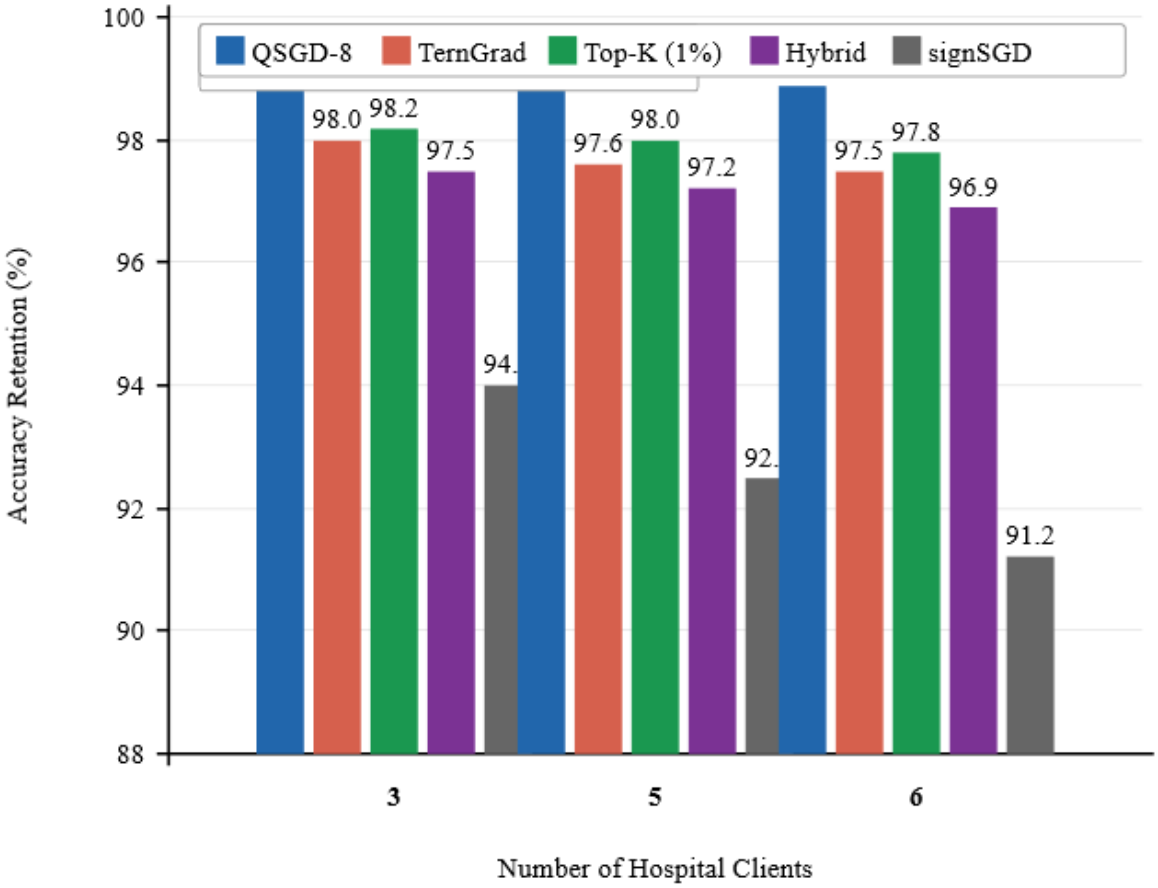

Figure 3. Accuracy Retention Ratio as a Function of Client Count on Fed-ISIC2019

Ratio of compressed method accuracy to uncompressed baseline accuracy at 3, 5, and 6 hospital clients for Top-K (k=1%), QSGD-8, TernGrad, signSGD, and Hybrid methods. Top-K (k=1%) decreases from 98.2% retention at 3 clients to 97.8% at 6 clients. QSGD-8 remains the most stable, declining from 99.1% to 98.9%. signSGD shows the largest difference across the tested client counts, from 94.0% at 3 clients to 91.2% at 6 clients, though this trend conflates the effects of its distinct majority-vote aggregation with compression noise. These results span a narrow client range (3 to 6) and should be treated as preliminary observations rather than evidence of a general scaling pattern.

4.3.2. Practical Deployment Recommendations

The experimental evidence supports budget-aware recommendations, where budget thresholds are expressed in terms of estimated communication volume derived from nominal compression ratios. When the estimated total upstream budget is below approximately 50 MB, the hybrid approach provides the fastest early accuracy growth and is the only method in this study that reaches the 95%-of-baseline target within roughly 15 MB of estimated cumulative upstream communication. Between approximately 50 and 150 MB, Top-K sparsification at k = 1% offers the best accuracy-per-byte tradeoff, reaching approximately 0.610 balanced accuracy by around 100 MB and surpassing the target by 78 rounds. For moderate estimated budgets of 150--500 MB, the convergence curves in Figure 2 suggest that QSGD-8 may offer an advantage, as its trajectory appears to continue improving after Top-K plateaus, ending at 0.641 after 200 rounds. When bandwidth is not a binding constraint, the uncompressed baseline remains the reference point.

A potentially complementary systems decision, not directly evaluated in the experiments above, is selective parameter communication. Keeping batch normalization layers local --- as motivated by cross-scanner heterogeneity in medical imaging --- reduces the parameter count subject to compression and aggregation. On Fed-ISIC2019 with EfficientNet-B0, excluding batch normalization parameters would reduce per-round transmission by approximately 2.1%, a modest systems-side saving that could compound over hundreds of rounds.

Table 5 consolidates the deployment recommendations.

Table 5. Recommended Compression Strategies by Bandwidth Scenario

Bandwidth Scenario	Total Budget	Recommended Strategy	Expected Accuracy	Nominal Compression Ratio
Ultra-constrained	< 50 MB	Hybrid (Top-1% + Q8)	~0.600 by 50 MB	400× nominal
Very constrained	50–150 MB	Top-K + error feedback (k = 1%)	~0.610 by 100 MB	100× nominal
Moderate	150–500 MB	QSGD-8	up to ~0.641	9.1× nominal
Relaxed	> 500 MB	No compression	~0.648 (200 rounds)	—

Recommendations are based on the Fed-ISIC2019 results with 6 natural hospital centers. Budget thresholds and accuracy values are analytically derived from nominal compression ratios and are not measured end-to-end transmission costs; actual bandwidth requirements will be moderately higher due to protocol overhead. The final column reports nominal compression ratios rather than measured communication savings. Accuracy values combine the budget-constrained communication curves in Figure 2 with the 200-round endpoint results in Table 4.

5. Discussion

5.1. Key Findings and Their Significance for Healthcare Federated Learning

This evaluation reveals three principal findings regarding gradient compression in multi-hospital federated learning. First, Top-K sparsification with error feedback achieves the strongest accuracy--communication ratio on naturally heterogeneous medical data, retaining 97.8% of the 200-round baseline accuracy at nominal 100× compression on Fed-ISIC2019. Second, multi-bit quantization provides the most stable performance as heterogeneity increases: QSGD-8 shows only 4.4% relative accuracy degradation from IID to extreme non-IID on PathMNIST, compared with 6.5% for Top-K. signSGD degrades most sharply, though this outcome conflates 1-bit compression with its distinct majority-vote aggregation rule and therefore cannot be attributed to compression alone. Third, the hybrid approach occupies the strongest low-budget region of the communication--accuracy space, yet its implementation requires coordinating both sparsification masks and quantization codebooks, adding engineering complexity that may not be justified when simpler alternatives suffice.

These results have tangible implications for healthcare institutions participating in federated consortia. A six-hospital dermatology collaboration using Top-K (k = 1%) compression would nominally reduce per-round communication from approximately 10.6 MB to around 106 KB per client (based on the nominal compression ratio; actual savings depend on protocol-level overhead including index encoding), potentially enabling participation over standard clinical network connections without dedicated infrastructure. The poor performance of signSGD under natural heterogeneity --- acknowledging that its majority-vote aggregation rule differs from the FedAvg protocol

used by all other methods --- cautions against adopting the most aggressive compression schemes without empirical validation on representative data and matched aggregation protocols.

5.2. Limitations

Several limitations constrain the generalizability of these findings. All experiments operate in a simulated federated setting using a single machine with multiple processes; real cross-hospital deployments involve network latency, packet loss, and asynchronous participation that may interact with compression in ways not captured here. Communication volumes reported in this study are analytically estimated from nominal compression ratios rather than measured at the packet level; actual transmission costs would include index encoding overhead for sparse methods, metadata framing, and serialization, which would moderately increase the reported TCT values and narrow the gap between sparsification-based and quantization-based strategies in practice. The evaluation is restricted to image classification tasks on Fed-ISIC2019 and the MedMNIST v2 PathMNIST benchmark [23]; medical image segmentation and tabular clinical prediction involve different model architectures and gradient distributions that warrant separate investigation. The interaction between gradient compression and privacy mechanisms such as differential privacy or secure aggregation also remains unexplored. Additionally, the signSGD comparison is not a pure compressor-only ablation because that method is evaluated in its canonical majority-vote form, which couples a different aggregation rule with 1-bit compression; disentangling these two effects would require evaluating signSGD with FedAvg-compatible decompression, which we leave to future work.

Future work should extend this evaluation along several dimensions. Adaptive compression strategies that adjust sparsification ratios or quantization bit-widths based on detected heterogeneity across hospitals represent a promising direction. Validation on larger federated consortia with ten or more institutions would test whether the client-count differences observed here persist at scale. Real-world deployment studies using hospital-network federated learning platforms would provide wall-clock measurements and fault-tolerance evidence to complement the communication-volume analysis presented here.

References

1. H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 54, pp. 1273–1282, 2017.
2. I. Dayan, H. R. Roth, A. Zhong, A. Harouni, A. Gentili, et al., "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature Medicine*, vol. 27, no. 10, pp. 1735–1743, 2021.
3. N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. J. Cardoso, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, p. 119, 2020.
4. D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 1709–1720, 2017.
5. W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "TernGrad: Ternary gradients to reduce communication in distributed deep learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 1508–1518, 2017.
6. J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, PMLR 80, pp. 559–568, 2018.
7. J. Sun, T. Chen, G. B. Giannakis, and Z. Yang, "Communication-efficient distributed learning via lazily aggregated quantized gradients," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 3365–3375, 2019.
8. F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-i.i.d. data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3400–3413, 2020.
9. Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *International Conference on Learning Representations (ICLR)*, 2018.
10. S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 4447–4458, 2018.

11. J. Wangni, J. Wang, J. Liu, and T. Zhang, "Gradient sparsification for communication-efficient distributed optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.
12. D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 5973–5987, 2018.
13. S. P. Karimireddy, Q. Rebjock, S. U. Stich, and M. Jaggi, "Error feedback fixes SignSGD and other gradient compression schemes," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, PMLR 97, pp. 3252–3261, 2019.
14. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems (MLSys)*, 2020.
15. S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, PMLR 119, pp. 5088–5099, 2020.
16. X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," in *International Conference on Learning Representations (ICLR)*, 2021.
17. P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 119–129, 2017.
18. J. O. Du Terrail, S.-S. Ayed, E. Cyffers, F. Grimberg, C. He, R. Loeb, P. Mangold, T. Marchand, O. Marfoq, E. Mushtaq, B. Muzellec, C. Philippenko, S. Silva, M. Teléncz uk, S. Albarqouni, S. Avestimehr, A. Bellet, A. Dieuleveut, M. Jaggi, S. P. Karimireddy, M. Lorenzi, G. Neglia, M. Tommasi, and M. Andreux, "FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, pp. 5315–5334, 2022.
19. S. Hu, L. Jiang, and B. He, "Practical hybrid gradient compression for federated learning systems," in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4147–4155, 2024.
20. T. Vogels, S. P. Karimireddy, and M. Jaggi, "PowerSGD: Practical low-rank gradient compression for distributed optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 14269–14278, 2019.
21. D. Rothchild, A. Panda, E. Ullah, N. Ivkin, I. Stoica, V. Braverman, J. Gonzalez, and R. Arora, "FetchSGD: Communication-efficient federated learning with sketching," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, PMLR 119, 2020.
22. J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.
23. J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "MedMNIST v2 --- A large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Scientific Data*, vol. 10, p. 41, 2023.