Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37334

Communication-Efficient Federated Learning under Non-IID Data: Aggregation and Compression Strategies

Chang Ma

College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, China
machang@stu.cdut.edu.cn

Abstract. As data in scenarios such as medical, transportation and the Internet of Things continue to be dispersed to institutions and edge devices, how to balance the quality of model training and communication costs without concentrating raw data has become an important issue in the deployment of federated learning. Federated learning faces a coupling bottleneck between client drift and communication overhead under non-independent and non-identically distributed (Non-IID) data. This paper reviews aggregation, compression and hybrid strategies with representative studies, and verifies Federated Averaging (FedAvg), Federated Proximal (FedProx) and their error-feedback Top-k combinations (FedAvg+EF-Top-k and FedProx+EF-Top-k) under the unified setting of Fashion-MNIST. The results show that compression can significantly reduce the cumulative uplink communication required to reach the target accuracy and enable more training rounds under a fixed total communication budget. However, strong heterogeneity is associated with greater run-to-run variability, and communication rounds, transmitted bits, and end-to-end time cannot be substituted for one another. Based on literature comparison and verification results, this paper proposes a selection strategy according to data heterogeneity, bandwidth and system constraints, and recommends the unified reporting of accuracy, communication volume, communication rounds, time, and energy consumption.

Keywords: Federated learning, Non-IID data, Communication efficiency, Aggregation strategy, Gradient compression

1. Introduction

With the help of federated learning technology, multiple clients can cooperate to complete model training tasks without exchanging their own original data throughout the process, and can also reduce the phenomenon of data islands and privacy risks [1, 2]. However, in practice, the data of the client usually does not meet the assumption standard of independent and identically distributed (IID), and will show non-IID characteristics. The direction of local updates on the client is not uniform, which will lead to client drift, and frequent transmission of model parameters or gradients will also consume bandwidth and energy. Data heterogeneity will also lengthen the training time required for the model to achieve the target accuracy, so that the statistical heterogeneity and communication overhead form a coupling effect [3, 4].

DOI: 10.54254/2755-2721/2026.37334

Such problems have been dealt with from many different directions by previous studies. System stability is prone to fluctuate in a heterogeneous data environment. Federated Proximal (FedProx) and Stochastic Controlled Averaging for Federated Learning (SCAFFOLD) have improved this situation by restricting or modifying local updates [3, 4]. Sattler et al. proposed sparse ternary compression (STC) to reduce the amount of data transmitted by parameters by means of sparsification. Xu et al. proposed the ternary federated averaging protocol (T-FedAvg) to achieve the same reduction effect of transmission through ternary quantisation [5, 6]. Chen et al. jointly optimised device selection, quantisation, and wireless resource allocation. Another study used layerwise asynchronous model update to reduce the synchronisation frequency of some parameters [7, 8]. The reinforcement-learning-based client-selection framework FAVOR proposed by Wang et al. selects more representative participating clients [9]. The above studies have optimised the update quality, single-round traffic, number of communication rounds or system time consumption. Due to the differences in data sets, non-IID division methods, model architecture and measurement specifications, the results of each study cannot be directly compared horizontally.

Existing reviews often organise relevant methods around specific dimensions, particularly data heterogeneity [3, 4]. Various empirical studies basically adopt self-set data division rules and cost definitions. The reported optimal results may not correspond to the same kind of problems. Under controlled conditions, there is still little evidence about the interaction between aggregation and compression. To fill this gap, this paper first reviews the development and current status of communication-efficient federated learning, which relies on the framework of aggregation, compression and hybrid strategies. It also compares the experimental conditions, results and application scope of various representative studies, and completes the pairing verification of Federated Averaging (FedAvg), FedProx, and their error-feedback Top-k variants under the unified experimental settings. The analysis of this paper focuses on why different evaluation indicators draw different conclusions, and how to select strategies under the constraints of data heterogeneity, bandwidth and equipment. It will not be separated from the corresponding experimental conditions, nor will it rank the cross-research results separately.

2. Development of federated learning

This paper divides the development of federated learning into four interconnected stages. The core of the first stage is the transformation of the training mode. The original centralised machine learning and data centre distributed training are switched to client-side collaborative training. FedAvg has established the basic paradigm of client-side local multi-step training and server-side aggregation [1, 2]. In the second stage, researchers focused on solving the problem of update offset and convergence under non-IID data. FedProx and SCAFFOLD are typical algorithms for heterogeneity-aware aggregation [3, 4]. In the third stage, the research focus has shifted to the transmission compression of model updating. The sparse direction is represented by STC, and the low-bit quantisation direction is represented by T-FedAvg [5, 6]. In the fourth stage, the research scope was further expanded, covering the fields of synchronisation-frequency optimisation, client scheduling, system resource collaboration, and layerwise asynchronous model update. Joint optimisation of device selection, quantisation, and wireless resource allocation, layerwise asynchronous model updating, and FAVOR are representative of these directions [7-9].

With the gradual progress of research, the optimisation goal is no longer limited to the average of basic parameters, but expanded to the comprehensive balance of model performance, communication efficiency and system constraints. After this change, the communication cost cannot be judged only by the size of the model. If the heterogeneous environment increases the number of

DOI: 10.54254/2755-2721/2026.37334

convergence rounds, the cumulative traffic will also increase when the single-round transmission volume remains unchanged; if the compression disturbance makes the update quality decline, the nominal compression ratio may not bring the same overall benefit.

The training closed loop shown in Figure 1 includes four core links: global model distribution, local training at the client, update upload, and server aggregation. The model and update scale, the number of clients participating in each round, and the communication frequency jointly determine the uplink and downlink costs. Various optimisation strategies are also implemented around these dimensions.

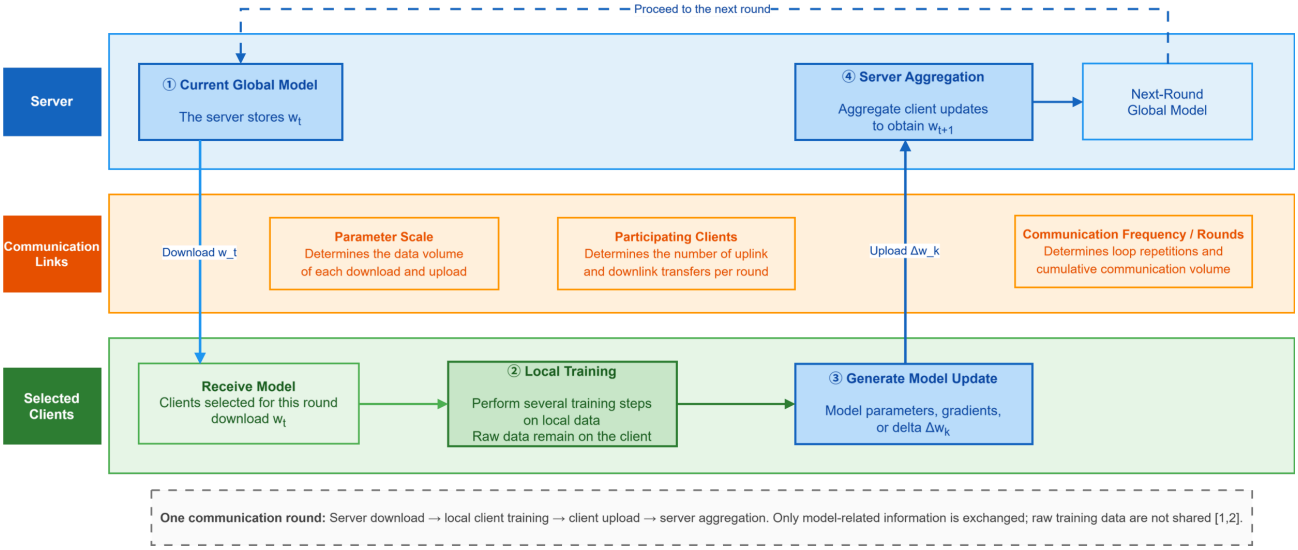

Figure 1. Federated learning training workflow and communication bottlenecks (picture credit: original)

3. Core techniques and taxonomy

FedAvg makes a weighted average of the local updates of each node, and the weight is determined according to the data volume of the client. However, the differences in labels, sample size or feature distribution between clients will make the local optimisation direction deviate from the global goal [3, 4]. For the improvement of update quality and the control of communication cost, the existing relevant methods can be divided into three types of strategies: aggregation, compression and hybrid.

3.1. Aggregation strategies

Various aggregation strategies adjust local training objectives, parameter update methods or server-side rules to alleviate client drift. The FedProx algorithm adds a proximal constraint in local training targets to adapt to scenes with a high degree of heterogeneity, but its proximal coefficient needs to be adjusted [10]. SCAFFOLD algorithm uses control variates to correct parameter update direction and reduce variance, but this requires additional state maintenance and transmission work [3, 4]. In the layerwise asynchronous model update, different modules of the model will be synchronised according to different frequencies, and the weighting operation will be completed according to the update freshness [8]. The FAVOR algorithm infers the data distribution from the model update information, dynamically selects the clients participating in the training, and reduces the selection deviation and communication rounds [9].

DOI: 10.54254/2755-2721/2026.37334

3.2. Compression strategy

Quantisation or sparsification is the main means to reduce the amount of data transmitted in a single round in the compression strategy [11]. Quantisation is to convert floating-point updates to low-bit format. T-FedAvg uses ternary coding to complete the compression of uplink and downlink data [6]. Magnitude-based sparsification transmits only updates with the largest absolute values [12]; the Top-k compressor used in this study retains the k largest-magnitude coordinates. The Top-k method belongs to this kind of method [12]. When the compression force is too large, precision loss or error accumulation will occur. At this time, error feedback is generally used to compensate for the unsent updated content [13]. Joint optimisation of quantisation, equipment selection and wireless resource allocation will have a corresponding impact on end-to-end time [7].

3.3. Hybrid strategies

When the hybrid strategy is adopted, either the aggregation and compression operations are combined, or multiple compression mechanisms are combined. STC integrates sparsification, ternary quantisation, and lossless coding, and compresses both uplink and downlink updates. Under the preset conditions of Non-IID, small batch and low participation rate, it achieves a good accuracy-communication compromise effect [5]. Such schemes are suitable for scenarios where data heterogeneity and bandwidth constraints coexist, but the parameter configuration and system implementation are relatively complex.

The three dimensions of update quality, single-round transmission volume and multi-mechanism synergy effect are respectively optimised by three types of strategies. The overall efficiency cannot be judged only by the change of a single index. Figure 2 integrates the three elements of learning efficiency, transmission efficiency and system conditions to build a unified evaluation framework.

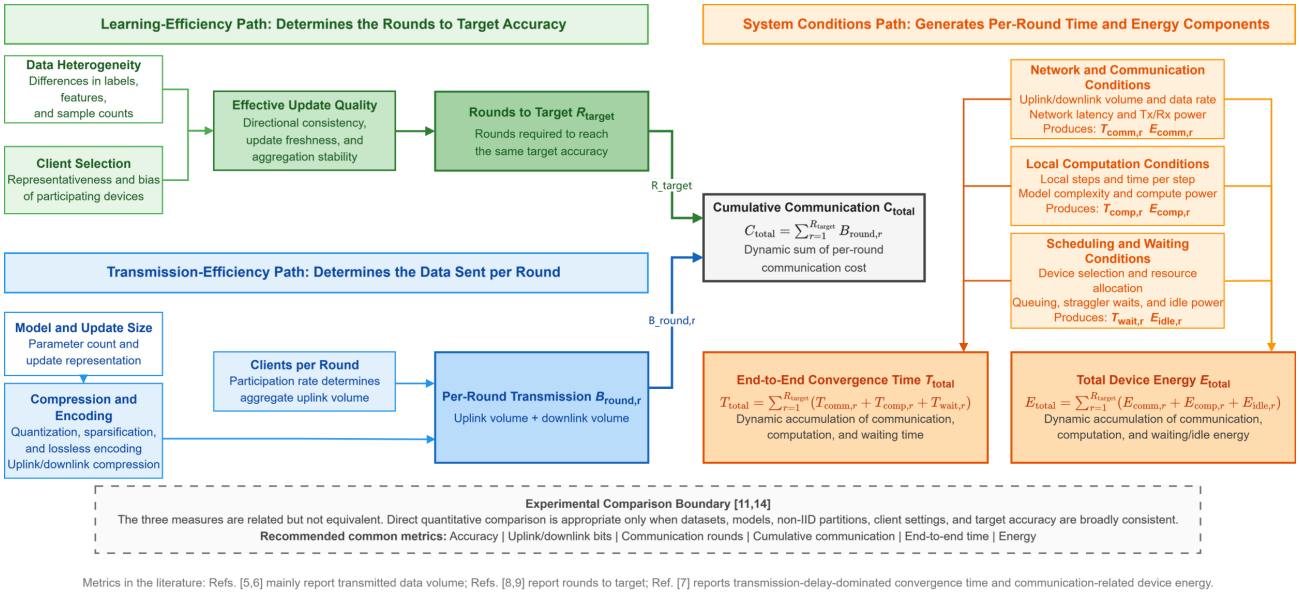

Figure 2. A unified evaluation framework for communication-efficiency experiments (picture credit: original)

In Figure 2, $B_{\text{round},r}$ denotes the total uplink and downlink transmission volume in round r, and R_{target} denotes the number of rounds required to reach the target accuracy; together, they determine

DOI: 10.54254/2755-2721/2026.37334

the cumulative communication volume C_{total}. The terms $T_{\text{comm},r}$, $T_{\text{comp},r}$, and $T_{\text{wait},r}$ denote communication, computation, and scheduling-and-waiting time, respectively. The terms $E_{\text{comm},r}$, $E_{\text{comp},r}$, and $E_{\text{idle},r}$ denote communication, computation, and idle energy consumption. These components accumulate over R_{target} rounds to form the end-to-end time T_{total} and total device energy consumption E_{total}.

4. Experimental designs and results across studies

4.1. Comparison principles

To evaluate the communication efficiency, at least three types of indicators should be considered, namely, the amount of single-round transmission, the number of rounds required to achieve the target accuracy, and the end-to-end convergence time. When using high-intensity compression, the number of bits transmitted in a single round will be reduced, and the compression error will increase the number of rounds required to achieve the target accuracy. When scheduling on the client side, the model size will remain unchanged. By improving the representativeness of participating nodes, the number of rounds required to achieve the target accuracy can be reduced. The end-to-end convergence time is affected by many factors, such as uplink and downlink bandwidth, local computing capacity, straggler waiting time, scheduling overhead and so on. It cannot be directly calculated by the cumulative number of transmission bits. The numerical results of different papers can be directly compared, which must meet the conditions that the data set, model, client partition mode, target accuracy, and communication-counting rules of the two papers are generally consistent [11, 14].

4.2. Representative studies and their results

Sattler et al. selected CIFAR-10, Fashion-MNIST and other tasks to carry out comparative tests on STC, FedAvg, and signSGD. In the IID experiment of CIFAR-10, STC with a sparsification rate of 1/400 requires only 183.9 MB of uplink traffic, while the uncompressed benchmark model requires 36,696 MB of uplink traffic, and FedAvg with a local interval of 100 steps requires 1,606.3 MB. In the IID experiment scenario with only 5/400 clients, the final accuracy of STC is 68.2%, and the final accuracy of FedAvg is 42.3% [5]. Xu et al. completed the performance evaluation of the T-FedAvg algorithm for MNIST and CIFAR-10 data sets. In 100 rounds of IID experiments on MNIST, the uplink and downlink data volume of T-FedAvg was 2.36 Mbit, and the corresponding data volume of FedAvg was 19.53 Mbit. In the CIFAR-10 task with five classes per client, based on the improved ResNet model, the accuracy of T-FedAvg is 76.69%, while the accuracy of FedAvg is 71.47% [6]. In another set of CNN experimental scenarios, according to the statistics of communication rounds, FedAvg converges slightly faster, and the low transmission volume per round does not mean that the total number of rounds required to achieve the target accuracy is less.

The efficiency of training rounds will be affected by both the client participation mode and the synchronisation scheme. FAVOR selects 10 clients from 100 candidate clients to participate in training in each round. On MNIST, Fashion-MNIST and CIFAR-10 data sets, the rounds required to achieve the target accuracy are reduced by 49%, 23% and 42%, respectively, compared with FedAvg [9], but its reinforcement-learning scheduler needs to complete additional training and state maintenance work. For the MNIST dataset and human activity recognition (HAR) task, Chen et al. completed the effect evaluation of layerwise asynchronous model update and temporally weighted aggregation. On the two data partitions of MNIST, TEFL can achieve 95% accuracy in about 30

DOI: 10.54254/2755-2721/2026.37334

rounds, while FedAvg needs about 75 rounds to achieve the same goal. In five successful HAR tests, TEFL needs 119–181 rounds to achieve 90% accuracy, and FedAvg only successfully achieves this accuracy in 4 tests, with 451–856 rounds [8]. The asynchronous update defined in this article refers to synchronising different layers of the model with different frequencies, rather than aggregating immediately after the client completes local training.

Chen et al. carried out the collaborative optimisation of device selection, 4-bit quantisation, and wireless resource allocation. In the simulation test environment, 10 of the 15 devices in each round completed data upload. Compared with the standard federated learning scheme, the recognition accuracy of the framework was improved by up to 3.6% and the convergence time was shortened by up to 87% [7]. The test results confirm the application potential of collaborative system optimisation, but the overall performance improvement is achieved by the joint action of multiple components and depends on specific wireless simulation conditions, which makes it impossible to determine the specific contribution of each component alone, nor can this scheme be extended to all network scenarios.

4.3. Overall analysis

This study concludes that the core of communication optimisation covers three aspects, including the amount of data transmitted in each round, the client objects participating in training, and the timing of model synchronisation. STC and T-FedAvg algorithms can achieve compression gains [5, 6]. The FAVOR algorithm focuses on improving the representativeness of participants [9]. The layerwise asynchronous model method promotes round efficiency optimisation by reducing the synchronisation frequency of some parameters [8]. The joint optimisation framework also incorporates the wireless resource dimension [7]. All the above methods can play a role in the corresponding scenarios, but they will also incur additional costs, specifically involving compression error, scheduling, state maintenance or system modelling.

At present, there are some limitations in the comparability and external validity of the completed experiments. Among the five studies involved, MNIST, Fashion-MNIST and CIFAR-10 are the main data sets in the core experiment, and a small number of HAR and EMG tasks are also used. When constructing Non-IID data, the number of categories that can be obtained by each client is generally limited. The experimental scale of these studies is usually maintained between 5 and 100 clients, and only a few special stress tests will increase the number of clients to 400. The network fluctuation, equipment offline, energy consumption, and privacy mechanism overhead in the real scene have not yet formed a unified evaluation standard [11, 14]. Cross-study comparison should clearly list the accuracy, transmission volume, communication rounds and end-to-end time, and clearly mark the specific experimental conditions. For the research that only publishes the compression ratio, it is necessary to clearly indicate whether the statistical scope covers the values, indices and coding metadata. For the research that only discloses the communication rounds, it is necessary to clearly mark the target accuracy and the disposal scheme for the failed runs. If the above requirements are not followed, the cost boundary corresponding to the same term may differ.

4.4. Validation experiment under a unified setting

In this paper, pairing verification is carried out on the Fashion-MNIST dataset to narrow the differences between different experimental settings. The experiment uses the FashionCNN model with 421,642 trainable parameters. 54,000 training samples are equally divided among 20 clients, and each client is assigned 10, 5, or 2 classes under the IID, moderate Non-IID, and strong Non-IID

DOI: 10.54254/2755-2721/2026.37334

conditions, respectively, so as to build three different conditions: IID, moderate Non-IID, and strong Non-IID. 6,000 additional samples are reserved for model development in the experiment. The official 10,000 test samples are only used for evaluation in the formal stage. The three data division methods ensure that the number of samples for each client is completely consistent, and the global category distribution is always balanced. The main difference is the label heterogeneity of the client. Each round of training will select five clients; each client will complete one local epoch. The batch size is set to 64, the learning rate of stochastic gradient descent (SGD) is set to 0.1, the maximum number of training rounds is 300, and 2026, 2027 and 2028 are set as three random seeds of the formal experiment.

There are four comparison methods included in this experiment, namely FedAvg (M1), FedProx (M2) [10], FedAvg (M3) combined with error-feedback Top-k, and FedProx (M4) combined with error-feedback Top-k [12, 13]. During the experiment, the proximal coefficients of FedProx were fixed at 0.01, 0.1 and 0.01 for three different levels of heterogeneity, respectively. The Top-k compressor system only processes the uplink update data and retains the 10% of coordinates with the largest absolute values. The residual part not transmitted by the client will be compensated when it participates in the task next time. When the heterogeneity level and random seed are consistent, the four methods use the same data partition method, initial model, data reading order and client scheduling table. The experimental results are presented with the mean and sample standard deviation after three runs with paired seeds.

The evaluation dimension adopted this time includes four items, namely, the accuracy rate after 300 rounds, the communication rounds required to reach an 80% or 85% accuracy rate for the first time, the cumulative uplink traffic when the target accuracy rate is reached, and the accuracy rate under the unified total communication budget. During statistics, uncompressed parameters and reserved Top-k values are calculated as 32 bits, and each reserved coordinate is additionally added with 19 bits of index data. According to such a statistical rule, the effective compression ratio corresponding to 10% of the coordinate retention ratio is not 10 times. This time, the first attainment is set as the core judgment basis, and the sensitivity of the model to single-round fluctuation is detected by using attainment for three consecutive rounds.

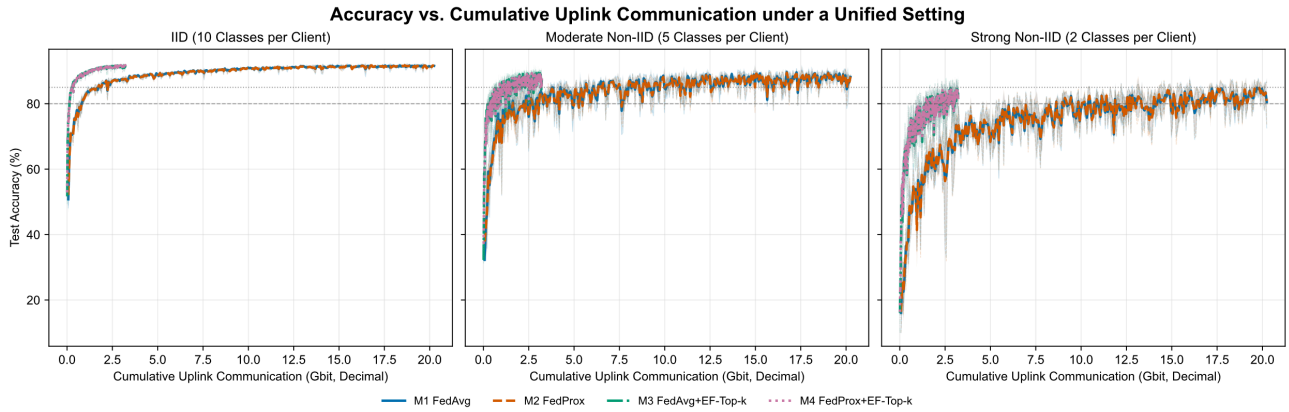

Figure 3. Accuracy versus cumulative uplink communication for different methods under a unified setting (picture credit: original)

Figure 3 shows that to achieve the same accuracy, the cumulative uplink traffic required by M3 and M4 is far lower than their corresponding uncompressed methods. Under IID data conditions, when the accuracy index of 85% is reached for the first time, the traffic demand of the uncompressed method is about 1.529 Gbit, and that of the compression method is about 0.272 Gbit.

DOI: 10.54254/2755-2721/2026.37334

In the moderate and strong Non-IID data environment, the uplink traffic required by the compression method to achieve the goal is generally reduced to 1/5.3 to 1/6.3 of the corresponding uncompressed method. In most scenarios, the number of rounds required by the compression method is either the same as that of the uncompressed method, or slightly more. The round efficiency is related to the bit efficiency, but they are not equal.

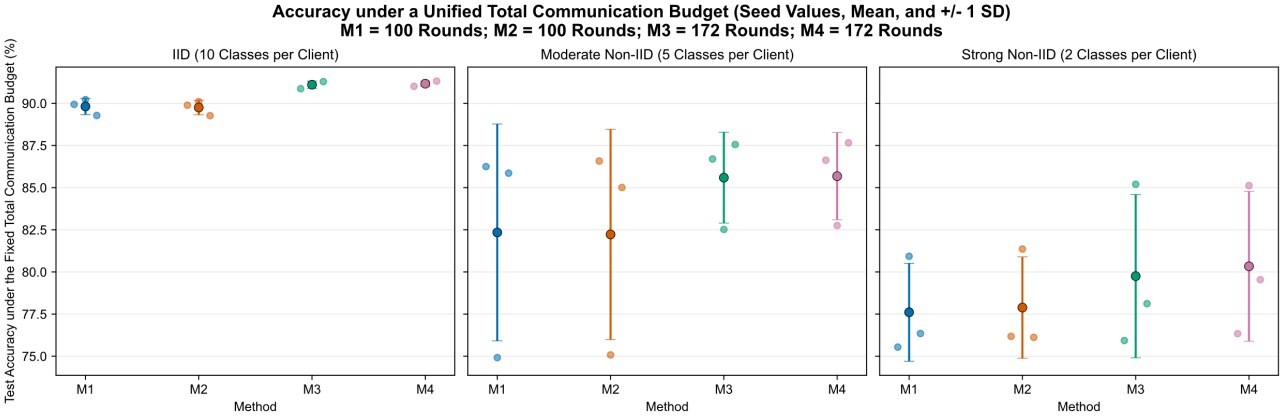

Figure 4. Test accuracy of the four methods under a unified total communication budget

Figure 4 sets the total traffic of 13.492544 Gbit consumed by uncompressed FedAvg to complete 100 rounds of training as the unified budget for all tests. M1 and M2 can complete 100 rounds of training under this budget, and M3 and M4, which compress only the uplink, can complete 172 rounds. In the IID data environment, the accuracy of the four methods within the budget limit is roughly in the range of 89.8% to 91.2%. When processing moderate and strong Non-IID data, the average accuracy of the compressed version method is about 3.3 and 2.3 percentage points higher than that of the corresponding uncompressed version, respectively. The communication saved through compression can be exchanged for more training rounds under a fixed budget. Only three test runs cannot support the statistical significance assertion with universal applicability.

When the data are strongly Non-IID, the standard deviation of the accuracy rate of the four methods after 300 rounds of training falls between 6.65 and 7.70 percentage points, and the range of this index under independent and identically distributed data is 0.13 to 0.53 percentage points. Set the robust criteria for maintaining 85% accuracy for three consecutive rounds, and the number of successful runs of M1 to M4 is 7 out of 9, 7 out of 9, 8 out of 9, and 7 out of 9, respectively. Single-round accuracy fluctuations may affect the first-attainment results. The average gain of FedProx is not large, and its direction will alter under different conditions. This experiment uses only one dataset, a small convolutional network, three random seeds and simulated clients, and does not directly measure the time or energy consumption. The experimental results can only verify the effectiveness of the evaluation framework, and cannot give a clear ranking of the methods.

5. Application scenarios and strategy selection

When selecting the appropriate method, the core bottleneck of the system should be identified first. Table 1 shows that when the degree of data heterogeneity is not high and the bandwidth is sufficient, the simple FedAvg method can be preferred. If the core problem is client drift, excessive single-round traffic or selection deviation, then the heterogeneity-aware aggregation, update compression or client scheduling schemes are adopted, respectively. When there are many constraints, hybrid or collaborative optimisation strategies can be selected.

DOI: 10.54254/2755-2721/2026.37334

Table 1. Strategy selection for different application scenarios

Application Scenario	Primary Bottleneck	Applicable Strategy	Selection Rationale	Main Limitation
Mild non-IID data and adequate bandwidth	Implementation cost	FedAvg	Simple structure suitable as a basic solution	Limited performance under strong heterogeneity
Strong non-IID data	Client drift	FedProx, SCAFFOLD	Constrain or correct local updates	Additional computation or state-maintenance cost
Low network bandwidth	Per-round communication volume	Quantisation, Top-k sparsification	Reduce the transmitted bits of parameters or gradients	Possible accuracy loss or error accumulation
Strong non-IID data and low bandwidth	Concurrent drift and communication cost	STC or other hybrid strategies	Jointly address update quality and compression	More complex parameter configuration and system implementation
Many candidate clients and few participation slots per round	Selection bias and communication rounds	Client scheduling	Improve the representativeness of participating clients	Additional training and state maintenance

In the application scenarios of edge computing and the Internet of Things, several key parameters, such as bandwidth, computing power, battery power and online availability, may change to varying degrees at the same time [15]. Before the formal deployment, the target accuracy and available communication budget should be determined first, and then the core bottleneck should be judged as client drift, limited transmission per round or insufficient participation opportunities. After entering the deployment phase, the system needs to monitor the straggler rate, the freshness of data updates and equipment energy consumption in real time, and then adjust the compression rate or the selection criteria of participants according to the real-time state of the system. The compression rate or round reduction observed in the simulation experiment will not be directly equivalent to the end-to-end benefits in the real environment.

In various actual deployment scenarios, the constraints and core bottlenecks faced by the system are different. Sheller et al. took the collaborative project of 10 medical institutions as an example and verified that multiple institutions can jointly complete the model training without exchanging original patient data. In this kind of inter-agency cooperation scenario, privacy protection requirements and differences in data distribution among institutions should also be taken into account [16]. When compression technology is used, or communication rounds are reduced, the impact of these operations on the quality of the inter-agency training model should be evaluated. In the federated Internet of Vehicles environment, vehicle mobility, changing channel conditions, and the number of participating vehicles per round will jointly affect system efficiency. The FedSSC algorithm developed by Liu et al. integrates vehicle selection, subchannel matching and power allocation to reduce communication overhead and delay [17]. Referring to these actual cases, the strategies in Table 1 should not be selected only according to the algorithm ranking on the standard data set, but should be determined in combination with the actual system limitations and core bottlenecks.

DOI: 10.54254/2755-2721/2026.37334

6. Current research topics and challenges

At present, the field of communication-efficient federated learning still faces a core problem, which is strongly Non-IID data. When each client contains only a small number of classes, or when sample sizes are severely imbalanced, client drift is aggravated [3, 4]. If more aggressive quantisation or sparse processing is adopted, the compression error may be amplified [11, 13]. There are differences in computing power, network quality, and online time between different devices, which will further increase the difficulty of scheduling and synchronisation [11, 15]. Algorithm convergence should be evaluated under the actual system conditions.

The design of communication optimisation cannot be carried out alone, but also needs to be promoted in coordination with the privacy mechanism. When using differential privacy technology, the added noise will reduce the accuracy of data processing, and secure aggregation will increase the computational or communication overhead [1, 2]. The data set standards adopted by different studies are not unified, the heterogeneous partition methods are different, the model framework is not unified, and the cost calculation methods are not consistent, so it is difficult to achieve a fair comparison across studies [11, 14]. Subsequent studies need to optimise existing algorithms and establish a unified evaluation method to include data, equipment, network and privacy factors into the evaluation scope.

7. Future research directions

Future research should move beyond optimising individual components alone and toward coordinating data, algorithms, and systems. Relevant personnel can dynamically adjust the aggregation weight, compression ratio, synchronisation frequency and client participation according to the differences in data quality and network conditions. When carrying out edge deployment, it is necessary to consider the factors of delay, energy consumption, reliability and privacy protection [15]. Such designs should be run many times in a unified benchmark test and report the results, while clearly defining the calculation method of uplink and downlink transmission and system time. Communication optimisation technology can also be used in distributed training of large-scale models. The training process needs to adopt a variety of parallel strategies. Its huge parameter scale makes cross-device communication a core bottleneck [18]. Multiple rounds of local training, lowering the synchronisation frequency and compressing the communication data can provide relevant ideas to solve the bottleneck. However, the data privacy, partial participation and statistical heterogeneity in the federated training scenario are different from those in data-centre training. Special verification work must be carried out for the conclusions drawn from these contents.

8. Conclusion

When promoting the implementation of federated learning, Non-IID data and communication overhead will jointly form constraints. The four categories of approaches—aggregation, compression, scheduling, and collaborative optimisation—target update quality, per-round transmission, participant representativeness, and system resources, respectively, while differing in applicable conditions and additional costs. After literature comparison and unified verification, it is found that the number of communication rounds, transmission bits and end-to-end time measure efficiency from different dimensions and do not yield a reliable absolute ranking unless the data partition, model and system conditions are controlled. The premise of selecting an appropriate method is to analyse the main bottlenecks and take into account the accuracy, traffic volume,

DOI: 10.54254/2755-2721/2026.37334

number of communication rounds, time and energy consumption. The follow-up research should adopt more unified experimental settings, cover more real networks, different device population sizes and task types, and record the variations and failures during multiple operations, to improve the repeatability and practical deployment value of the research conclusions.

References

[1] Kairouz, P., McMahan, H. B., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning, 14*(1-2), 1-210. DOI: 10.1561/2200000083.

[2] Li, T., Sahu, A. K., Talwalkar, A., et al. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine, 37*(3), 50-60. DOI: 10.1109/MSP.2020.2975749.

[3] Zhu, H., Xu, J., Liu, S., et al. (2021). Federated learning on non-IID data: A survey. *Neurocomputing, 465*, 371-390. DOI: 10.1016/j.neucom.2021.07.098.

[4] Ye, M., Fang, X., Du, B., et al. (2024). Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys, 56*(3), 1-44. DOI: 10.1145/3625558.

[5] Sattler, F., Wiedemann, S., Müller, K. R., et al. (2020). Robust and communication-efficient federated learning from non-i.i.d. data. *IEEE Transactions on Neural Networks and Learning Systems, 31*(9), 3400-3413. DOI: 10.1109/TNNLS.2019.2944481.

[6] Xu, J., Du, W., Jin, Y., et al. (2022). Ternary compression for communication-efficient federated learning. *IEEE Transactions on Neural Networks and Learning Systems, 33*(3), 1162-1176. DOI: 10.1109/TNNLS.2020.3041185.

[7] Chen, M., Shlezinger, N., Poor, H. V., et al. (2021). Communication-efficient federated learning. *Proceedings of the National Academy of Sciences, 118*(17), e2024789118. DOI: 10.1073/pnas.2024789118.

[8] Chen, Y., Sun, X., & Jin, Y. (2020). Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation. *IEEE Transactions on Neural Networks and Learning Systems, 31*(10), 4229–4238. DOI: 10.1109/TNNLS.2019.2953131.

[9] Wang, H., Kaplan, Z., Niu, D., et al. (2020). Optimizing federated learning on non-IID data with reinforcement learning. *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 1698–1707. DOI: 10.1109/INFOCOM41043.2020.9155494.

[10] Li, T., Sahu, A. K., Zaheer, M., et al. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems, 2*, 429–450.

[11] Almanifi, O. R. A., Chow, C. O., Tham, M. L., et al. (2023). Communication and computation efficiency in federated learning: A survey. *Internet of Things, 22*, 100742. DOI: 10.1016/j.iot.2023.100742.

[12] Aji, A. F., & Heafield, K. (2017). Sparse communication for distributed gradient descent. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen: Association for Computational Linguistics, 440–445. DOI: 10.18653/v1/D17-1045.

[13] Stich, S. U., Cordonnier, J. B., & Jaggi, M. (2018). Sparsified SGD with memory. *Advances in Neural Information Processing Systems 31*, Curran Associates, Inc., 4447–4458.

[14] Asad, M., Moustafa, A., Ito, T., et al. (2021). Evaluating the communication efficiency in federated learning algorithms. *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 552–557. DOI: 10.1109/CSCWD49262.2021.9437738.

[15] Abreha, H. G., Hayajneh, M., & Serhani, M. A. (2022). Federated learning in edge computing: A systematic survey. *Sensors, 22*(2), 450. DOI: 10.3390/s22020450.

[16] Sheller, M. J., Edwards, B., Reina, G. A., et al. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports, 10*, 12598. DOI: 10.1038/s41598-020-69250-1.

[17] Liu, S., Guan, P., Yu, J., et al. (2023). FedSSC: Joint client selection and resource management for communication-efficient federated vehicular networks. *Computer Networks, 237*, 110100. DOI: 10.1016/j.comnet.2023.110100.

[18] Zeng, F., Gan, W., Wang, Y., et al. (2025). Distributed training of large language models: A survey. *Natural Language Processing Journal, 12*, 100174. DOI: 10.1016/j.nlp.2025.100174.