Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36498

History-Aware Multi-Objective Client Selection for Federated Learning under Statistical and System Heterogeneity

Haoxuan Geng

School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, China
haoxuan.geng22@student.xjtlu.edu.cn

Abstract. Federated learning reduces raw-data movement, but partial participation under statistical and system heterogeneity makes the selected client cohort a major source of optimisation bias and delay. This study proposes a history-aware multi-objective selector that combines opt-in data-quality metadata, exponentially smoothed training utility, marginal label-balance gain, participation deficit, and an online latency penalty without querying every client's current private loss. A paired protocol compares Random FedAvg, Power-of-Choice, an Oort-inspired policy, a class-balance greedy policy, and the proposed method on Digits, MNIST, and CIFAR-10. The confirmatory benchmark uses 50 clients, three Dirichlet non-IID levels, ten fresh seeds, and Holm-adjusted Wilcoxon tests. Relative to Random FedAvg, the proposed method improves final-five-round accuracy in all six MNIST/CIFAR-10 conditions by 1.11–5.53 percentage points, with four adjusted-significant conditions. Component ablations show distinct roles. Label balance improves accuracy AUC in all six conditions, the adaptive penalty reduces simulated synchronous latency, and the fairness term improves participation uniformity without eliminating all dataset-dependent concentration. Accuracy gains remain positive when 6%, 10%, or 20% of clients participate per round. The evidence supports a practical accuracy–balance–latency trade-off rather than universal dominance, and the study defines clear boundaries for privacy and hardware claims.

Keywords: Federated learning, Client selection, Non-IID data, Edge–cloud collaboration, Multi-objective optimisation

1. Introduction

Federated learning (FL) trains a shared model through repeated local optimisation and server aggregation while raw observations remain at their data owners. Federated Averaging (FedAvg) established the communication-efficient form of this protocol [1]. In practical edge-cloud deployments, however, client data are non-independent and identically distributed (non-IID), only a fraction of devices participate in each round, and computation or network conditions differ substantially [2-4]. Client selection therefore determines not only which updates arrive quickly, but also which parts of the population influence the global trajectory.

Prior work addresses complementary dimensions of this decision. FedCS selects clients under resource constraints, Oort combines statistical utility with execution speed, and clustered sampling

DOI: 10.54254/2755-2721/2026.36498

seeks representative low-variance cohorts [5-7]. Power-of-Choice favours high-loss clients from a candidate pool, FedCor models loss correlations, and FedBalancer coordinates informative data with round deadlines [8-10]. More recent class-aware approaches reduce cohort imbalance through label statistics or update-derived heterogeneity [11, 12]. These methods reveal an unresolved tension. Utility selection may over-sample difficult or noisy clients, class balancing may repeatedly select a narrow cohort, latency filtering may suppress statistically useful devices, and fresh loss queries can create hidden selection overhead.

Reproducible platforms and sampling theory broaden the design space, but do not remove these operational choices [13-15]. A deployable policy needs to use information available before selection, avoid an oracle assumption that every client evaluates every global model, and expose rather than conceal trade-offs among learning progress, label representation, latency, and participation. Fixed scores are interpretable but cannot respond to changing latency pressure. Fully learned policies may require additional state exchange or extensive warm-up.

Partial participation also changes the empirical training distribution induced by the selection policy. If clients with particular labels, losses, or latencies receive higher inclusion probabilities, their local objectives contribute more frequently to the realised update trajectory even though FedAvg remains sample weighted within each round. Selection can therefore improve short-horizon optimisation while introducing longer-horizon representation bias. This distinction motivates reporting terminal accuracy together with trajectory-level, cohort-level, and system-level endpoints instead of treating test accuracy as the only criterion.

The investigation is organised around three research questions. RQ1 asks whether history-aware multi-objective selection improves learning trajectory and final-stage accuracy relative to Random FedAvg. RQ2 asks how any accuracy change co-varies with selected-label representation, simulated synchronous latency, participation uniformity, and pre-selection loss-query overhead. RQ3 asks which score components produce these effects and whether the direction persists across non-IID severity, participation rate, and model architecture. These questions separate efficacy, mechanism, and robustness before any claim of practical superiority is made.

This paper develops a lightweight history-aware multi-objective selector. Firstly, it combines one-time opt-in metadata with server-side histories observed only after participation. Secondly, greedy cohort construction evaluates marginal label complementarity rather than ranking clients independently. Finally, a projected online update adjusts the latency penalty while a participation-deficit term guards against starvation. The method retains standard FedAvg aggregation and introduces no current-loss query for unselected clients.

The empirical study makes three contributions. It uses paired partitions, initialisations, minibatch rules, and formal seeds across methods. It evaluates four complementary baselines on Digits, MNIST, and CIFAR-10 at multiple non-IID levels, with planned endpoints, confidence intervals, paired tests, and multiplicity correction. It then separates component roles through standard ablations, participation-rate sensitivity, and an exploratory full-resolution convolutional check. The objective is not to claim universal superiority, but to identify when the integrated policy helps, what it costs, and which claims require hardware or privacy-preserving validation.

2. Methodology

2.1. Research design and observability

The study uses a controlled paired repeated-measures design. Within each dataset, heterogeneity level, and seed, every policy receives the same train/test split, Dirichlet client partition, simulated

DOI: 10.54254/2755-2721/2026.36498

latency profile, model initialisation, local minibatch rule, number of selected clients, and communication budget. Only the selection decision changes. The primary endpoint is mean test accuracy over the final five rounds, which is less sensitive to non-IID oscillation than a single terminal value. Secondary endpoints are accuracy area under the communication-round curve (accuracy AUC), mean simulated synchronous latency, time to a predefined target, selected-label entropy, Jain's participation index, and pre-selection query count.

The server observes a one-time profile for client i containing sample count n_i, a ten-bin label histogram, and simulated latency l_i. Exact histograms are an experimental assumption rather than a privacy guarantee. Dynamic utility is updated only after a selected client completes local training. Unselected clients retain their last history and receive a bounded staleness bonus. Consequently, the proposed method never requires every client to evaluate the current global model before a round. Power-of-Choice remains query-based by design, and its candidate queries are counted explicitly.

2.2. Information contract and temporal order

The selection information is separated by when and how it becomes available. Before training begins, each client provides an opt-in static profile containing sample count, class histogram, and a latency proxy. At the start of round t, the server combines that profile with only its previously observed state: exponentially smoothed local loss, staleness, and cumulative selections. It then constructs S_t without requesting a fresh evaluation from non-selected clients. Only members of S_t train and return the observations needed to update history after aggregation. This temporal ordering prevents a hidden all-client inference pass from being treated as cost-free selection.

The contract is operational rather than privacy-complete. Exact counts and histograms can reveal distributions; the implementation provides neither differential privacy nor secure metadata aggregation. Coarsened, noised, or update-derived descriptors would require a separate utility-privacy evaluation. Cold-start clients use static quality and staleness until a loss history is observed. Candidate-set normalisation and deterministic client-ID tie-breaking make fixed-seed selection exactly reproducible.

2.3. History-aware multi-objective score

Data quality combines normalised label entropy and square-root sample size,

$$q_i = 0.60H_i + 0.40\widehat{n}_i \tag{1}$$

After client i participates at round t, its historical utility is updated from observed local training loss $l_i(t)$,

$$u_i(t) = 0.70u_i(t-1) + 0.30l_i(t) \tag{2}$$

Before selection, normalised utility receives 0.15 times normalised square-root staleness. During greedy cohort construction, $c_i(t)$ is the normalised entropy obtained by adding candidate i to the clients already chosen in the round. The fairness term $f_i(t)$ is the normalised deficit between client i 's observed participation and the count expected under uniform selection. The candidate score is

$$s_i(t) = 0.15q_i + 0.35u_i(t) + 0.25c_i(t) + 0.15f_i(t) - \lambda_t\hat{l}_i \tag{3}$$

DOI: 10.54254/2755-2721/2026.36498

The highest-scoring remaining client is appended and the aggregate label histogram is updated until k clients have been selected. The latency coefficient begins at 0.10 and follows a clipped projected update after the selected cohort's maximum latency L_t is observed,

$$\lambda_{t+1} = clip\left[\lambda_t + 0.08\frac{L_t - B}{B},\ 0.05,\ 0.35\right] \tag{4}$$

where B is the 75th percentile of the condition-specific latency profile. The update increases latency pressure after a budget violation and relaxes it after a low-latency round. All weights and bounds are fixed before evaluation on formal seeds. Figure 1 shows how static profiles and server-side histories feed the marginal cohort score, while only post-participation observations update utility, participation counts, and the online latency weight.

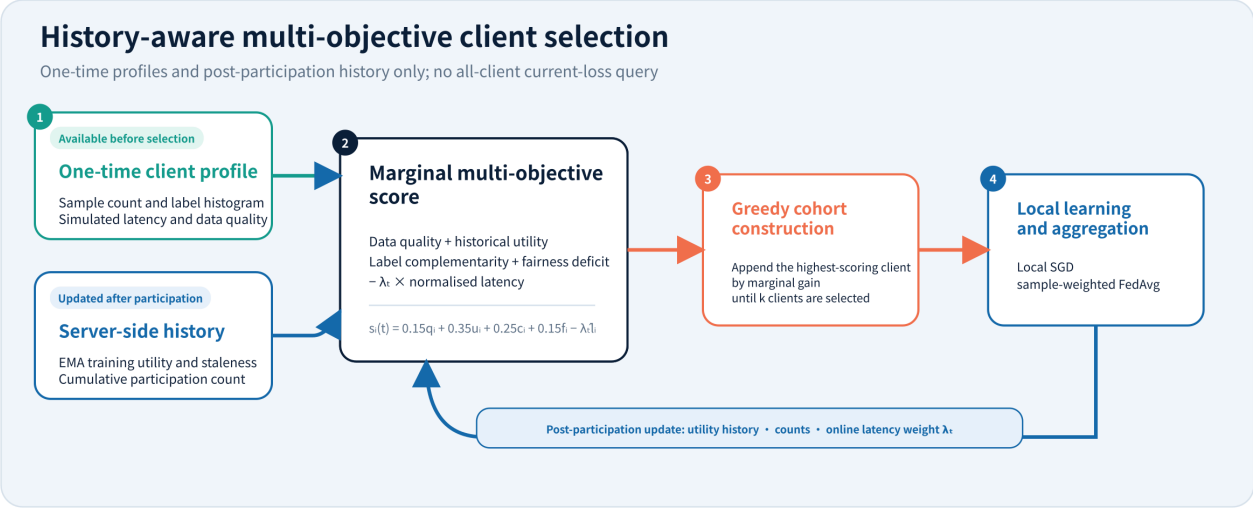

Figure 1. History-aware multi-objective client selection workflow (picture credit: original)

The complementarity term is recomputed after every greedy addition, so the method evaluates a candidate's contribution to the current cohort rather than assigning each client one independent label score. This distinction matters when two individually diverse clients carry redundant classes. The fairness deficit and staleness bonus act at the client level, whereas label entropy acts at the cohort level. The projected update for λ_t is not trained on test accuracy: it reacts only to the observed synchronous latency relative to B. Consequently, the controller can change system pressure online without leaking evaluation labels into selection.

2.4. Local learning, baselines, and ablations

Selected clients perform local stochastic gradient descent and return updated parameters. The server applies sample-weighted FedAvg,

$$w_{t+1} = \sum_{i \in S_t} \frac{n_i}{\sum_{j \in S_t} n_j} w_i^{t+1} \tag{5}$$

Random FedAvg samples uniformly without replacement. Power-of-Choice samples ten candidates, queries their loss on at most 256 local observations, and selects the five highest-loss candidates [8]. The Oort-inspired policy ranks historical loss-sample-size utility, exploration,

DOI: 10.54254/2755-2721/2026.36498

staleness, and latency; the qualifier is retained because the full Oort systems stack is not reproduced [6]. Class-balance greedy selects clients that maximise the aggregate label entropy and serves as a strong component-oriented control motivated by Fed-CBS [11].

Five leave-one-component-out variants remove data quality, training utility, label balance, participation fairness, or latency penalty while preserving paired experimental controls. The study also varies the selected fraction to 3, 5, or 10 of 50 clients. The exploratory architecture check replaces compact features and a multilayer perceptron with full-resolution MNIST and CIFAR-10 inputs and a small two-convolutional-layer network.

All comparators share the same local optimiser, aggregation rule, model initialisation, data order, and round budget inside a paired condition. The class-balance control is intentionally strong because it isolates the mechanism most likely to benefit severe Dirichlet partitions. Power-of-Choice is allowed its ten-client loss-query candidate pool and the resulting queries are retained as an explicit endpoint. The Oort-inspired comparator uses historical utility, exploration, staleness, and latency but is not presented as a systems-level reproduction of Oort. This naming discipline prevents component comparisons from being mistaken for claims against complete production systems.

Constructing a cohort requires at most $O(kK)$ score evaluations plus an $O(C)$ histogram update per accepted client. This is small relative to local optimisation in the ten-class tasks, but it is not zero overhead. The method avoids an extra current-model loss evaluation on non-selected clients while still storing histories and scoring centrally.

2.5. Datasets, protocol, and statistical analysis

The controlled Digits study uses 1,437 training and 360 test observations from the scikit-learn Digits dataset, ten clients, Dirichlet concentration $\alpha = 0.30$, three selected clients, 40 rounds, and seeds 1001-1010. The standard benchmark uses the official MNIST and CIFAR-10 splits [16, 17]. MNIST is deterministically average-pooled to 14×14; CIFAR-10 is resized to 8×8 RGB for the main selection-focused benchmark. Both use 50 clients, five selected clients, 40 rounds, and $\alpha \in \{0.10, 0.30, 1.00\}$. Ten development-independent seeds 3001-3010 support confirmatory inference. Table 1 summarises the locked design.

Table 1. Experimental design matrix

Dataset	Clients / selected	Non-IID α	Rounds × confirmatory seeds	Main model
Digits	10 / 3	0.30	40 × 10	64-64-10 MLP
MNIST	50 / 5	0.10, 0.30, 1.00	40 × 10	64-unit MLP
CIFAR-10	50 / 5	0.10, 0.30, 1.00	40 × 10	64-unit MLP

The compact standard benchmark uses one local epoch, batch size 64, and learning rate 0.08. Digits uses two local epochs, batch size 32, and learning rate 0.05. No test observation enters selection, local training, stopping, or weight calibration. Synchronous latency is the maximum simulated latency among selected clients and is not hardware wall-clock time.

Means, sample standard deviations, and two-sided 95% t confidence intervals are reported across independent seeds. Proposed-versus-comparator differences are paired by seed and evaluated with a two-sided Wilcoxon signed-rank test. Holm correction is applied within each planned dataset-heterogeneity-metric family. Missing time-to-target outcomes are not imputed. Jain's index

$$J(x) = \frac{(\sum_i x_i)^2}{K \sum_i x_i^2} \tag{6}$$

DOI: 10.54254/2755-2721/2026.36498

measures uniformity of cumulative client selections and is not a demographic or predictive-fairness measure [18].

The evidence is generated as an auditable chain. Dataset archives are checksum verified, methods share paired experimental worlds, and round logs retain selections, accuracy, latency, entropy, queries, and adaptive weights. Reported summaries are recomputed from these records. Confirmatory and exploratory checks retain their stated roles rather than being pooled into one significance claim.

3. Results

3.1. Controlled Digits evidence

On Digits, the complete method attains 95.761% final-five-round accuracy compared with 95.500% for Random FedAvg. The paired gain of 0.261 percentage points is not significant after Holm correction ($p = 0.781$), so the primary endpoint does not support a terminal-accuracy claim. Accuracy AUC improves by 2.215 points ($p = 0.023$) and selected-label entropy improves by 0.043 ($p = 0.014$). Simulated latency to 90% accuracy decreases by 0.476 s, but the adjusted result remains marginal ($p = 0.059$). Jain participation uniformity decreases by 0.150 relative to random selection. The controlled evidence therefore supports faster, more label-balanced learning rather than universal terminal superiority.

3.2. Confirmatory MNIST and CIFAR-10 results

Table 2 reports the primary endpoint for all six standard conditions. Relative to Random FedAvg, the proposed effect is positive throughout, but its magnitude contracts as the Dirichlet concentration increases and client distributions become less concentrated. On MNIST, the gain falls from +3.80 percentage points at alpha = 0.10 to +1.99 at alpha = 0.30 and +1.11 at alpha = 1.00; only the most heterogeneous condition remains significant after Holm correction. The Random baseline simultaneously rises from 80.41% to 90.44%, leaving less missing label coverage for deliberate cohort construction to recover. CIFAR-10 shows the same contraction, from +5.53 to +3.63 and +2.51 points, but all three comparisons remain adjusted-significant. The larger CIFAR-10 effects indicate that targeted complementarity remains valuable on the more difficult task even after average accuracy increases from 23.64% to 37.81% under Random FedAvg. The proposed method has the highest mean accuracy in MNIST at alpha = 0.10 and 1.00 and in CIFAR-10 at alpha = 1.00. Class-balance greedy leads by only 0.18, 0.75, and 0.13 points in the other three conditions, while Oort-inspired differs from the proposed method by just 0.07 points in severe CIFAR-10. Thus, the table supports a consistent advantage over random selection, especially under severe heterogeneity, but not universal dominance over specialised selectors. Figure 2 visualises the proposed-minus-Random paired effects and their 95% confidence intervals, making both the common positive direction and the wider uncertainty in the two non-significant MNIST conditions explicit.

Table 2. Final-five-round mean test accuracy on confirmatory seeds, including proposed-minus-Random effects and Holm-adjusted p-values

Dataset	α	Random	Power-of-Choice	Oort-inspired	Class balance	Proposed	Difference (pp)	Holm p
MNIST	0.10	80.41	81.78	80.85	82.33	84.22	+3.80	0.016
MNIST	0.30	87.35	88.36	88.13	89.52	89.34	+1.99	0.055

DOI: 10.54254/2755-2721/2026.36498

Table 2. (continued)

MNIST	1.00	90.44	91.49	90.94	91.27	91.55	+1.11	0.078
CIFAR-10	0.10	23.64	23.45	29.10	29.92	29.17	+5.53	0.012
CIFAR-10	0.30	31.69	31.68	33.65	35.45	35.32	+3.63	0.008
CIFAR-10	1.00	37.81	37.48	37.36	39.92	40.32	+2.51	0.008

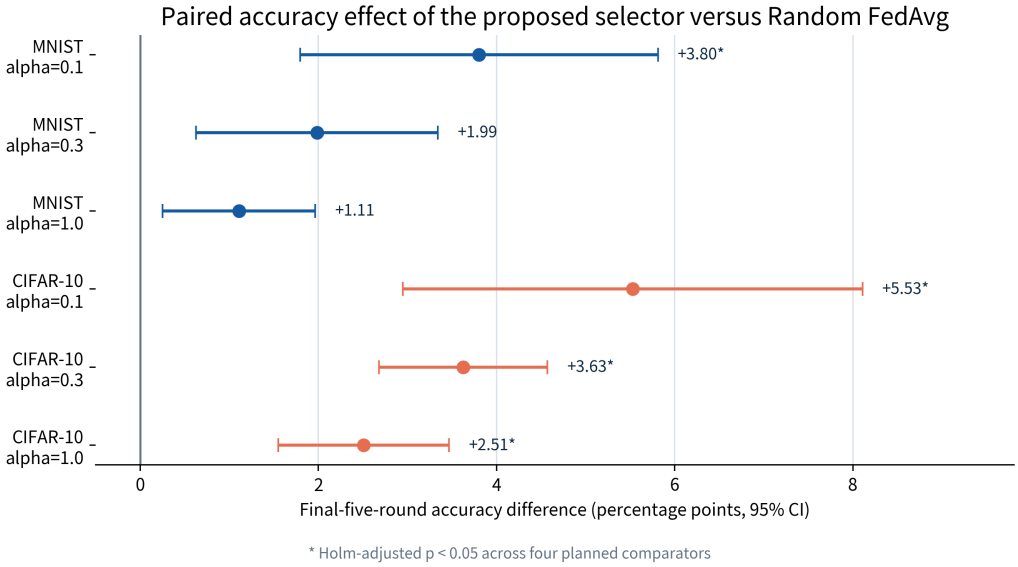

Figure 2. Paired final-five-round accuracy effect relative to Random FedAvg (picture credit: original)

Class-balance greedy is the strongest specialised baseline. Proposed-minus-class-balance accuracy ranges from -0.75 to $+1.88$ points, and none of the six final-five-round comparisons is adjusted-significant. This equivalence in accuracy conceals different system behaviour. On CIFAR-10, the proposed method reduces mean synchronous latency relative to class balancing by 34.3-118.6 ms, with all three adjusted comparisons significant. It also improves Jain uniformity over class balancing in five of six conditions. Conversely, class balancing achieves higher selected-label entropy in five adjusted-significant conditions. The integrated selector therefore trades a small amount of label purity for lower latency and broader participation while retaining similar accuracy.

Figure 3 makes this comparison visible without collapsing the six heterogeneity conditions into one grand mean. The proposed and class-balance points are close on MNIST and on moderate or mild CIFAR-10, while both separate from Random FedAvg most clearly in the severely heterogeneous CIFAR-10 condition. This pattern supports the interpretation that cohort label coverage is the dominant source of the accuracy gain. It also explains why the complete method should be evaluated on latency and participation alongside accuracy: a specialised class-aware selector already captures much of the statistical benefit.

DOI: 10.54254/2755-2721/2026.36498

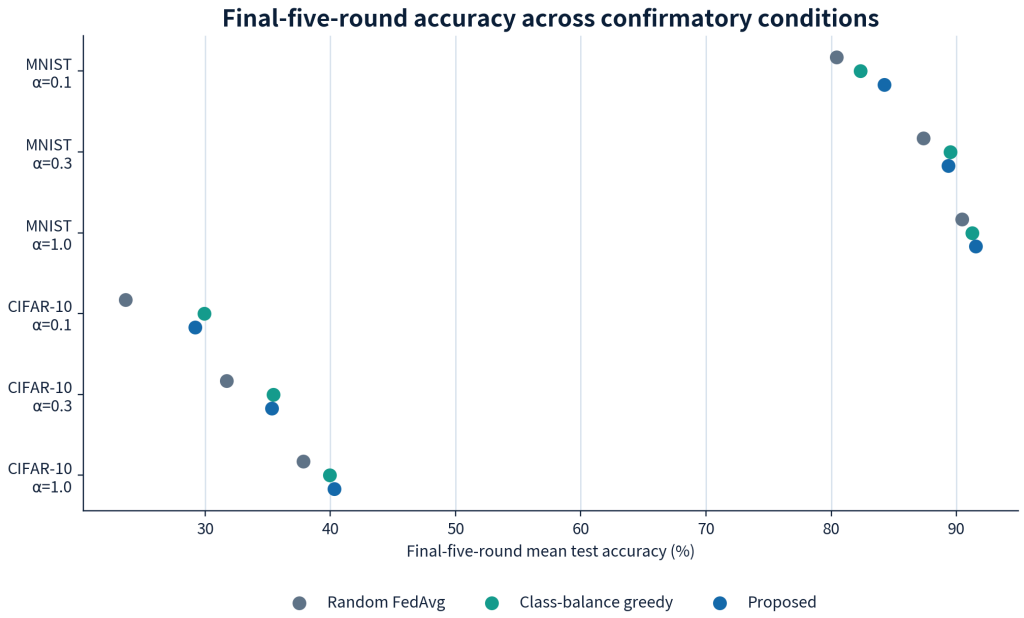

Figure 3. Final-five-round accuracy across the six confirmatory conditions (picture credit: original)

Relative to Random FedAvg, selected-label entropy is higher in all six conditions by 0.028-0.168, with every adjusted comparison significant. Mean synchronous latency is lower by 13.9-89.6 ms; four conditions are adjusted-significant and two are directionally consistent. Participation uniformity is dataset-dependent. The method matches or exceeds random selection on MNIST at moderate and mild heterogeneity, but concentrates participation strongly on CIFAR-10. This limitation is retained in the claim boundary rather than averaged away.

3.3. Component roles and participation sensitivity

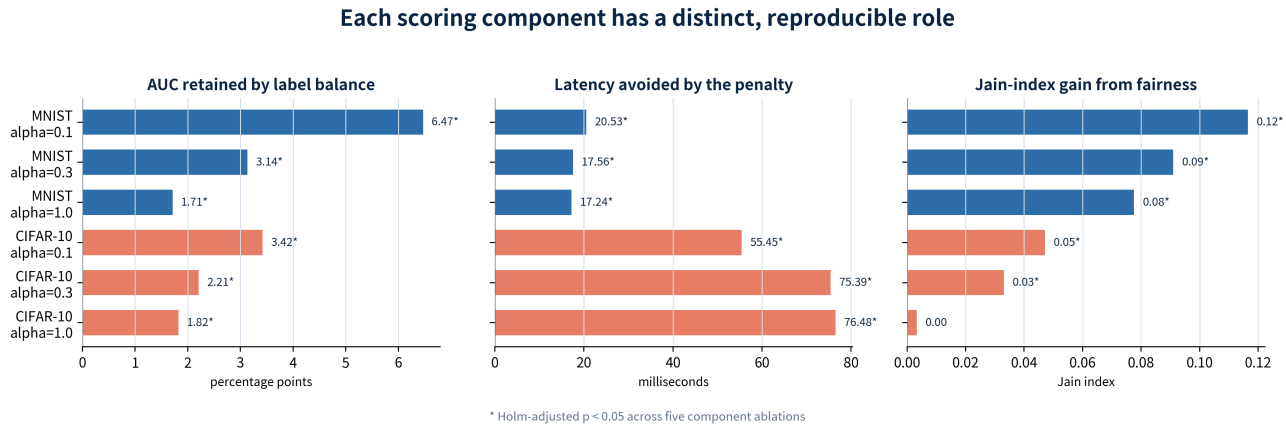

Figure 4. Leave-one-component-out effects across the six standard conditions (picture credit: original)

The standard ablations clarify why a single accuracy ranking is insufficient. Removing label balance reduces accuracy AUC in every standard condition by 1.71-6.47 points, and all six comparisons remain significant after Holm correction across five ablations. Removing the latency penalty increases mean synchronous latency by 17.2-76.5 ms in every condition, again with six

DOI: 10.54254/2755-2721/2026.36498

significant comparisons. Removing the fairness term lowers Jain's index in all six conditions and is significant in five. Data quality provides positive AUC effects in all six conditions, but only two are adjusted-significant. Training utility sometimes improves early learning yet can reduce final-five-round accuracy, illustrating the bias introduced by repeatedly favouring high-loss clients. Figure 4 reports complete-method minus leave-one-component-out effects, so a positive bar means that the removed component improved the displayed endpoint.

Participation-rate sensitivity supports the accuracy result beyond the locked 10% setting. When 3, 5, or 10 of 50 clients participate per round, proposed-minus-random final-five-round accuracy is respectively $+3.39$, $+1.99$, and $+2.06$ points on MNIST and $+5.63$, $+3.63$, and $+3.34$ points on CIFAR-10. All six effects remain significant after correcting across the three participation rates. Accuracy AUC and selected-label entropy are also positive in every condition, while synchronous latency is lower. Jain uniformity remains mixed, confirming that the fairness reward improves the score locally but is not a hard global constraint.

The largest effects occur when only 6% of clients participate, but the gain remains positive at 20% participation (Figure 5). Increasing k allows Random FedAvg to cover more labels by chance, which reduces the marginal advantage of deliberate complementarity. The non-monotonic MNIST pattern also cautions against fitting a simple linear effect of participation rate from three settings. The sensitivity analysis is therefore used to test persistence of direction, not to identify an optimal deployment rate. The stars in Figure 5 identify within-dataset Holm-adjusted p-values below 0.05.

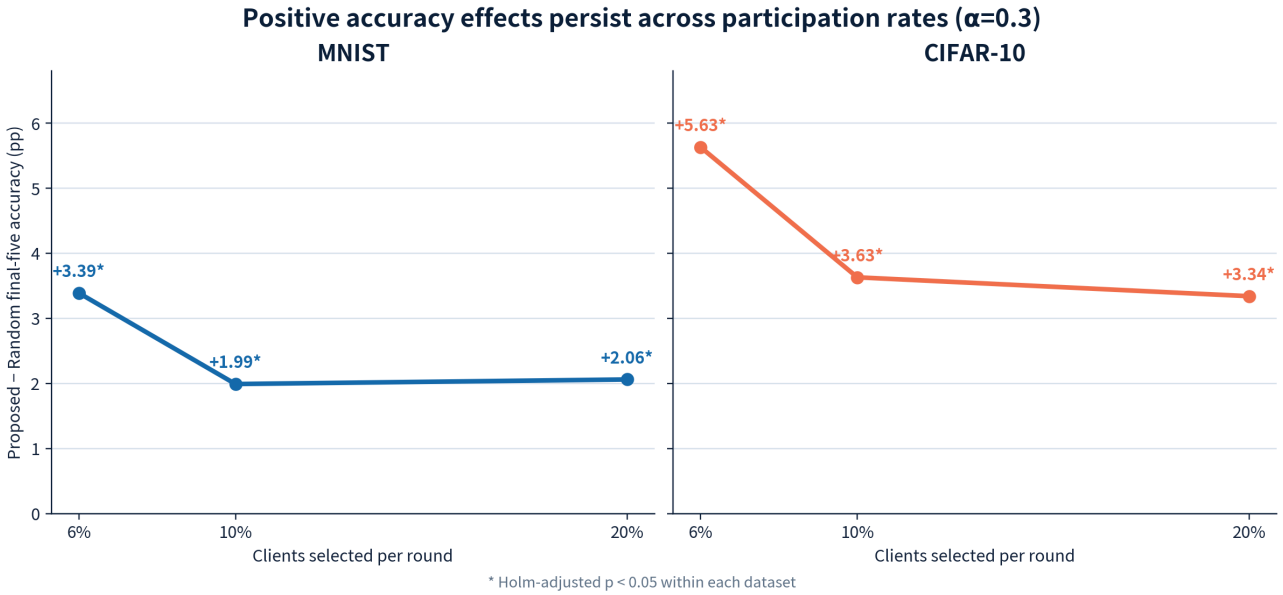

Figure 5. Participation-rate sensitivity at Dirichlet concentration 0.30 (picture credit: original)

3.4. Exploratory architecture robustness

An exploratory robustness check replaces the compact benchmark models with a full-resolution small CNN, uses $\alpha = 0.30$, and repeats five paired seeds for 30 rounds. Relative to Random FedAvg, the proposed selector raises final-five-round accuracy by 2.60 percentage points on MNIST and 8.69 points on CIFAR-10. Every paired seed is positive on both datasets. Accuracy AUC also increases by 7.16 and 5.12 points, while mean synchronous latency falls by 47.7 and 76.9 ms. The two-sided exact Wilcoxon $p = 0.0625$ and Holm-adjusted $p = 0.125$ reflect the minimum attainable two-sided value with five non-zero pairs; these runs are therefore directional architecture evidence,

DOI: 10.54254/2755-2721/2026.36498

not additional confirmatory significance claims. Relative to class-balance greedy, the mean final-five-round differences are $+4.31$ points on MNIST and $+1.66$ points on CIFAR-10, with wide intervals spanning zero. Figure 6 summarises these paired mean effects with 95% confidence intervals across the five exploratory seeds.

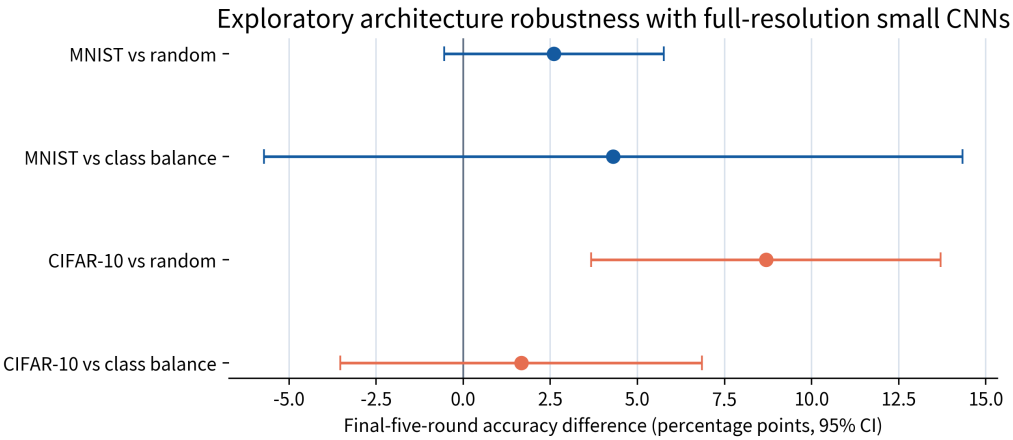

Figure 6. Exploratory full-resolution small-CNN accuracy effects (picture credit: original)

3.5. Reproducibility checks

The confirmatory tables comprise 12,000 per-round records, 3,000 client profiles, and 300 independent runs with unique composite keys. Selection-derived sample counts, maximum latency, label entropy, query counts, and Jain indices were reconstructed from the profiles. Recomputed run summaries, confidence intervals, Wilcoxon tests, and Holm adjustments match the locked outputs within floating-point tolerance. Independent MNIST and official-order CIFAR-10 reruns at $\alpha = 0.30$, seed 3001, reproduce all 200 per-dataset round records exactly, including selected-client sequences and every numeric field.

Validation is layered to expose silent corruption. Keys, factors, cohorts, sample conservation, selection-derived fields, summaries, and adjusted tests are independently checked; all 15 checks pass. Exact rerun equality is stronger than matching rounded tables because it also detects changes in preprocessing, random-number order, tie-breaking, or aggregation. This does not prove external generalisability, but it makes the reported evidence traceable to executable records.

4. Discussion

The results answer the three research questions with different levels of support. For RQ1, the proposed selector improves trajectory-level learning relative to Random FedAvg. Accuracy effects are positive across datasets, heterogeneity levels, and participation rates, but not every primary comparison is significant. The class-balance control remains statistically indistinguishable in final-five-round accuracy. The contribution is therefore not a universal accuracy win; it is an integrated policy that approaches the strongest class-aware control while reducing its latency and participation concentration in several conditions.

For RQ2, the trade-offs are measurable rather than implicit. The label-complementarity term is the most consistent driver of accuracy AUC. The projected latency penalty has a direct and reproducible system effect. The fairness term improves selection uniformity relative to its removal, but cannot always overcome the stronger utility and balance incentives. CIFAR-10 exposes this

DOI: 10.54254/2755-2721/2026.36498

limitation most clearly, where label-oriented selection concentrates on a subset of complementary clients despite the fairness reward. A deployment requiring minimum participation guarantees should replace the soft reward with quotas, constrained optimisation, or a primal-dual fairness constraint.

For RQ3, component effects depend on the endpoint. Utility history can accelerate learning, yet high-loss preference may favour difficult, atypical, or noisy clients and thereby hurt terminal accuracy. This observation agrees with the known convergence-bias tension in Power-of-Choice [8]. Label balance is consistently useful under the controlled Dirichlet partitions and connects to Fed-CBS [11], while the lower-latency outcome reflects the system-aware motivation of FedCS and Oort [5, 6]. Unlike a current-loss oracle, the proposed policy uses only previously observed utility. Its zero pre-selection loss-query count should nevertheless not be described as zero overhead because metadata, state maintenance, and score computation remain.

The mechanism is best understood as a controlled compromise rather than a single scalar notion of "best client." Label complementarity improves coverage when local distributions are concentrated; the latency term limits the synchronous cost of obtaining that coverage; and the participation deficit partially counteracts repeated selection of the same complementary clients. Their interaction explains both the gains and the failure modes. In CIFAR-10, a small subset can remain valuable for label coverage and utility often enough that a bounded fairness reward cannot restore random-like participation. Conversely, increasing the fairness coefficient without a constraint could sacrifice the accuracy benefit that motivates targeted selection. The results thus support exposing the three endpoint families separately rather than compressing them into an uncalibrated deployment utility.

The statistical interpretation is deliberately condition-level. The 40 communication rounds within one run are serially dependent and are not treated as 40 independent replicates. Instead, each seed contributes one predefined endpoint, and policy differences are paired within the same partition, initialisation, latency profile, and local randomisation. This choice avoids pseudo-replication and removes a large source of between-run variance, but ten seeds still provide limited resolution for nonparametric inference. Accordingly, confidence intervals and effect directions are reported alongside adjusted p values, and a non-significant condition is not interpreted as evidence of equivalence. Holm correction controls family-wise error for the planned comparisons without assuming independent tests, while the exploratory CNN results remain outside the confirmatory family. MNIST and CIFAR-10 are not pooled into one headline test because their difficulty, scale, and accuracy ranges differ. Future larger studies should pre-register a smallest effect of practical interest, increase seed counts through power or precision analysis, and report equivalence or non-inferiority tests when the goal is to show that latency or fairness improvements preserve accuracy.

For a research submission, the contribution is the observable-information contract, marginal cohort scoring, online latency response, and paired audit trail. The study neither establishes unrestricted state of the art nor proves convergence of the adaptive score. Production claims would require device traces, measured communication, dropout, secure metadata, and official systems comparisons; simulated milliseconds and exact histograms cannot support those claims alone.

Several validity limits remain. The main benchmark intentionally uses compact models and downsampled inputs to isolate selection behaviour. Client partitions, availability, and latency are simulated; synchronous latency is not an end-to-end hardware measurement. Exact label histograms can reveal distributional information, and the method provides neither differential privacy nor secure aggregation. The class-balance and Oort-inspired baselines test mechanisms but are not complete systems reproductions. Ten confirmatory seeds permit paired nonparametric inference, yet broader

DOI: 10.54254/2755-2721/2026.36498

task and architecture coverage is needed for an international systems or machine-learning submission.

Future work should evaluate natural client partitions such as FEMNIST, trace-driven availability and device speed, dropout and stale-update stress, larger convolutional or transformer models, and official implementations of recent selectors. Privacy-preserving histogram estimates, update-derived heterogeneity, and secure aggregation should be evaluated jointly with utility. Finally, a constrained formulation should provide explicit latency and participation guarantees, accompanied by convergence or regret analysis for the online dual update and greedy complementarity term.

5. Conclusion

This paper introduced a history-aware multi-objective client selector for partially participating federated learning under statistical and system heterogeneity. The policy combines opt-in metadata, participation-conditioned utility history, marginal label complementarity, a fairness deficit, and an adaptive latency penalty without requiring an all-client current-loss query. Across six confirmatory MNIST and CIFAR-10 settings, its final-five-round accuracy advantage over Random FedAvg is positive in every case and significant in four. Sensitivity experiments retain positive significant effects at three participation rates, and a five-seed small-CNN study preserves the direction on both datasets without being treated as confirmatory inference. Standard ablations separate the mechanisms. Label balance consistently improves accuracy AUC, the latency term consistently reduces simulated synchronous delay, and the fairness term improves participation uniformity relative to its removal but does not impose a hard guarantee. Class-balance greedy remains a competitive accuracy baseline, and utility history has mixed terminal effects. The defensible contribution is therefore an interpretable accuracy-balance-latency compromise rather than universal dominance. The released round records, reconstructed metrics, recomputed statistical tests, and exact independent reruns support reproducibility. Trace-driven devices, natural client partitions, privacy-preserving descriptors, hard participation constraints, and larger models remain necessary before claiming production readiness or broad state-of-the-art superiority.

References

[1] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In A. Singh & J. Zhu (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54*, 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html

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

[3] Li, T., Sahu, A. K., Zaheer, M., et al. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems, 2*, 429–450. https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html

[4] Karimireddy, S. P., Kale, S., Mohri, M., et al. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning, 119*, 5132–5143. https://proceedings.mlr.press/v119/karimireddy20a.html

[5] Nishio, T., & Yonetani, R. (2019). Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019—2019 IEEE International Conference on Communications*, 1–7. https://doi.org/10.1109/ICC.2019.8761315

[6] Lai, F., Zhu, X., Madhyastha, H. V., & Chowdhury, M. (2021). Oort: Efficient federated learning via guided participant selection. In *15th USENIX Symposium on Operating Systems Design and Implementation*, 19–35. https://www.usenix.org/conference/osdi21/presentation/lai

DOI: 10.54254/2755-2721/2026.36498

[7] Fraboni, Y., Vidal, R., Kameni, L., et al. (2021). Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *Proceedings of the 38th International Conference on Machine Learning, 139*, 3407–3416. https://proceedings.mlr.press/v139/fraboni21a.html

[8] Cho, Y. J., Wang, J., & Joshi, G. (2022). Towards understanding biased client selection in federated learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, 151*, 10351–10375. https://proceedings.mlr.press/v151/jee-cho22a.html

[9] Tang, M., Ning, X., Wang, Y., et al. (2022). FedCor: Correlation-based active client selection strategy for heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10102–10111. https://openaccess.thecvf.com/content/CVPR2022/html/Tang_FedCor_Correlation-Based_Active_Client_Selection_Strategy_for_Heterogeneous_Federated_Learning_CVPR_2022_paper.html

[10] Shin, J., Li, Y., Liu, Y., et al. (2022). FedBalancer: Data and pace control for efficient federated learning on heterogeneous clients. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, 436–449. https://doi.org/10.1145/3498361.3538917

[11] Zhang, J., Li, A., Tang, M., et al. (2023). Fed-CBS: A heterogeneity-aware client sampling mechanism for federated learning via class-imbalance reduction. In *Proceedings of the 40th International Conference on Machine Learning, 202*, 41354–41381. https://proceedings.mlr.press/v202/zhang23y.html

[12] Chen, H., & Vikalo, H. (2024). Heterogeneity-guided client sampling: Towards fast and efficient non-IID federated learning. In *Advances in Neural Information Processing Systems 37*. https://doi.org/10.52202/079017-2093

[13] Lai, F., Dai, Y., Singapuram, S., et al. (2022). FedScale: Benchmarking model and system performance of federated learning at scale. In *Proceedings of the 39th International Conference on Machine Learning, 162*, 11814–11827. https://proceedings.mlr.press/v162/lai22a.html

[14] Chen, W., Horváth, S., & Richtárik, P. (2022). Optimal client sampling for federated learning. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=l0BP1lHpPW

[15] Reddi, S. J., Charles, Z., Zaheer, M., et al. (2021). Adaptive federated optimization. *International Conference on Learning Representations*. https://openreview.net/forum?id=LkFG3lB13U5

[16] LeCun, Y., Bottou, L., Bengio, Y., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791

[17] Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images*. University of Toronto.

[18] Jain, R., Chiu, D. M., & Hawe, W. R. (1984). *A quantitative measure of fairness and discrimination for resource allocation in shared computer systems*. Digital Equipment Corporation.