

CMST-FL: Cross-Missingness Stable Trust with Evidence-Guided Preprocessing for Federated Healthcare Signal Imputation

A Patient-Disjoint Two-Stage Model for Imputation Selection and Reliability-Weighted Aggregation

Benachir Rigalma, Soumia Ziti, Meryam Belhiah, Hamid Ouhnni
IPSS Team, Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco

Abstract—Missing physiological samples create preprocessing uncertainty and uneven client reliability in federated learning. This study proposes Cross-Missingness Stable Trust federated learning (CMST-FL), a two-stage framework that selects an imputer using development patients and then weights local updates from their marginal validation-quality gains across MCAR, MAR, and contiguous-block masks. Twelve BIDMC records screened five imputers, and 15 disjoint records evaluated five training strategies under three mechanisms, three rates, and three seeds. Iterative imputation minimized the predeclared multi-objective score (0.35) and achieved the lowest standardized imputation RMSE (0.7712) under the tested budget. Across 27 matched evaluation conditions, CMST-FL obtained mean test R^2 of 0.0130, RMSE of 0.3298, and MAE of 0.2624. Compared with FedAvg, its mean R^2 gain was 0.00658, and its RMSE reduction was 0.00110. CMST-FL was favored in all 27 conditions for both metrics, and the paired tests remained statistically significant after Holm correction. Client-level cluster resampling, however, produced wide intervals because only five clients were available, and the worst-client gain was not significant. Trust logs showed meaningful but moderate departures from equal sample weighting. The evidence therefore supports iterative preprocessing followed by CMST-FL as a promising benchmark-specific configuration, while the low absolute R^2 and small error reductions preclude claims of clinical or universal superiority.

Keywords—Federated learning; missing data; trust aggregation; physiological signals; multiple imputation; healthcare analytics

I. INTRODUCTION

Continuous physiological monitoring produces multivariate signals that are useful for prediction, but practical recordings contain absent or corrupted samples. The BIDMC PPG and Respiration Dataset offers synchronized physiological signals and manually derived respiration targets, making it suitable for controlled studies of missing-data robustness. The present work uses this public resource as a methodological benchmark rather than claiming clinical deployment [1], [2].

Federated learning (FL) enables clients to train a shared model while retaining raw records locally. FedAvg introduced sample-count aggregation for decentralized neural-network training [4]. FedProx later added a proximal local objective for heterogeneous networks [5]. A recent healthcare survey

identifies non-identically distributed data, missingness, and uncertain client quality as continuing obstacles [3]. q-FFL addresses fairness across clients [6], whereas FLTrust targets Byzantine updates [7]; neither directly evaluates whether a benign update remains useful across several missingness mechanisms.

This distinction motivates Cross-Missingness Stable Trust federated learning (CMST-FL). Instead of trusting a client because it performs well under only the current mask, CMST-FL asks a counterfactual question: does the local update improve the received global model across a fixed bank of plausible missingness mechanisms? It rewards mean improvement, penalizes cross-mechanism variability and worst-case harm, then smooths trust through rounds before combining it with sample count. The proposed research object is therefore a unified decision model—not only a new server weight—because preprocessing is selected on disjoint development patients before trust aggregation is evaluated.

The study addresses three questions. RQ1 asks which tested imputer provides the best predeclared balance of imputation fidelity, downstream validation performance, and preprocessing time. RQ2 asks whether CMST-FL consistently improves predictive performance over FedAvg and FedProx on patient-disjoint evaluation records. RQ3 asks whether the improvement persists across MCAR, MAR, and contiguous-block missingness at 10%, 30%, and 50% rates.

The main contributions are:

- A new two-stage CMST-FL model that connects leakage-controlled preprocessing selection to trust-weighted aggregation while separating development and evaluation patients;
- A mathematically explicit counterfactual cross-missingness utility with stability, worst-case, and temporal-memory terms requiring only client-side validation summaries and model parameters;
- A matched 27-condition comparison with centralized learning, FedAvg, FedProx and a guarded trust baseline, together with bootstrap intervals and paired Wilcoxon tests; and

- An evidence-bounded interpretation that reports both statistical consistency and the modest absolute magnitude of the observed gains.

The remainder of the study is organized as follows. Section II reviews missing-data and adaptive-aggregation research. Section III defines CMST-FL, Section IV presents the experimental protocol, and Section V reports preprocessing, predictive, cluster-aware, and trust-trajectory results. Section VI interprets the findings, Section VII records the evidence-supported configuration, Section VIII presents limitations, and Section IX concludes the study.

II. RELATED WORK

A. Missing-Data Preprocessing

Simple mean or median replacement is fast but can distort covariance and temporal structure [11]. K-nearest-neighbor imputation estimates absent values from similar observations [8], while iterative imputation models each feature conditionally on the others [9]. Denoising autoencoders learn nonlinear reconstructions and may be attractive for complex physiological data, although their ranking depends strongly on architecture, training budget, and sample size [10]. Missing-data methodology further emphasizes that missingness assumptions and uncertainty should be considered explicitly [11].

Recent federated imputation studies demonstrate growing interest in heterogeneous intensive-care time series and personalized imputation under heterogeneous missingness. CAFE, for example, learns personalized imputation weights from cross-client missing-data heterogeneity. These studies optimize imputation itself, whereas the present work first compares conventional and learned imputers under a fixed budget and then uses the selected imputer inside a reliability-weighted predictive FL pipeline. A method that is best in one dataset or budget cannot be assumed universally best; preprocessing is therefore treated as a patient-disjoint development-stage decision [12], [21].

B. Federated Aggregation and Trust

FedAvg is the sample-weighted reference [4], whereas FedProx changes local optimization through a proximal term [5]. q-FFL prioritizes clients according to loss to improve fairness [6], and FedFomo uses client-specific validation utility

[13]. FedLAW learns aggregation weights from client coherence [19]. FedAW introduces dataset-quality-aware weighting [20], while FedLW assigns larger weights to lower-loss clients [22]. These methods show that aggregation need not depend only on sample count, but none evaluates a local update across alternative missingness mechanisms in the manner proposed here.

Robust aggregation and FLTrust principally target anomalous or adversarial updates [7], [14]. Dynamic trust management has also adapted trust thresholds and smoothing parameters according to client performance, behavior, and variability [23]. CMST-FL targets a different reliability question: rather than identifying malicious behavior or using absolute local metrics, it measures the local update's marginal predictive gain relative to the exact received global model, repeats that counterfactual evaluation across a missingness bank, and uses soft sample-aware weights instead of a fixed exclusion decision.

C. Research Gap and Scope

Two gaps are addressed. First, preprocessing and aggregation are often tuned on the same final records, which can overstate generalization. Second, a client weight based on a single validation condition may reward a brittle update. CMST-FL combines patient-disjoint selection with cross-mechanism validation. It does not provide a Byzantine-resilience proof, causal inference, privacy accounting, or clinical certification; these remain outside the present scope.

D. Novelty Positioning

The individual ingredients of CMST-FL—weighted averaging, validation-based quality signals, and temporal smoothing—have precedents. FedLAW studies learned aggregation weights [19], FedAW uses data-quality signals [20], CAFE exploits missing-data heterogeneity for imputation [21], FedLW uses client loss [22], and dynamic trust management adapts trust thresholds and smoothing parameters [23]. To the best of our knowledge, among the methods reviewed in Table I, none combines patient-disjoint imputer selection, local-versus-global gain across a missingness bank, stability and worst-case penalties, and temporally smoothed soft aggregation. The claim is therefore a new integration and formulation, not the invention of trust-weighted FL.

TABLE I. FEATURE-LEVEL POSITIONING OF CMST-FL AGAINST CLOSE METHOD FAMILIES

Method	Weight/decision signal	Missingness bank	Local-global gain	Stability + worst case	Temporal memory	Imputer selection
FedAvg	Sample count	No	No	No	No	No
FedProx	Sample count + proximal local loss	No	No	No	No	No
FedLAW / FedAW / FedLW	Coherence, data quality or loss	No	No	No	Method-specific	No
CAFE	Personalized imputation weights	Missing-data heterogeneity	No	No	No	Learns imputation
Dynamic trust	Performance and consistency indicators	No	No	No	Yes	No
CMST-FL	Relative gain across mask bank	Yes	Yes	Yes	Yes	Yes

III. PROPOSED CMST-FL METHOD

A. System Model

Let K clients hold disjoint patient records. At round t , the server broadcasts global parameters $w(t)$. Client i applies an imputer fitted only on its local training partition, performs one local epoch, and returns updated parameters $w_i(t+1)$. No raw waveform, target value, or missingness mask is transmitted. Each client evaluates the received and locally updated models on the same local validation data under a fixed missingness bank. Fig. 1 summarizes the patient-disjoint two-stage workflow.

B. Cross-Missingness Utility

For each bank condition b in B , Q is the bounded validation-quality score implemented in the experiment. It combines bounded R^2 with scale-normalized RMSE and MAE. For client i , the scale was $s_i = \max(P95(y_i) - P5(y_i), SD(y_i), \epsilon)$. The following expressions define Q and its normalized terms; the counterfactual gain in Eq. (1) then compares the local update with the exact global model received at that round.

$$Q = \text{clip}(0.50R_\beta^2 + 0.30 \exp(-\text{NRMSE})$$

$$+ 0.20 \exp(-\text{NMAE}), 0, 1)$$

$$R_\beta^2 = \{\text{clip}(R^2, -1, 1) + 1\}/2$$

$$\text{NRMSE} = \text{RMSE}/s_i, \text{ NMAE} = \text{MAE}/s_i$$

$$g_{i,b}(t) = Q(w_i(t+1); V_{i,b}) - Q(w(t); V_{i,b}). \quad (1)$$

The experiment used $B = \{\text{MCAR } 30\%, \text{ MAR } 30\%, \text{ block } 30\%\}$. A useful update should have positive average gain, low variation across mechanisms, and no severely negative condition. The penalized utility is

$$u_i(t) = \text{mean}_{b \in B} [g_{i,b}(t)] - \lambda \text{sd}_{b \in B} [g_{i,b}(t)]$$

$$- \mu \max\left(0, -\min_{b \in B} g_{i,b}(t)\right). \quad (2)$$

Here λ controls the stability penalty and μ penalizes worst-case harm. Unlike a weight based on one absolute local metric, the gain is relative to the exact global model received at that round. This construction reduces persistent preference for clients with intrinsically easier validation records.

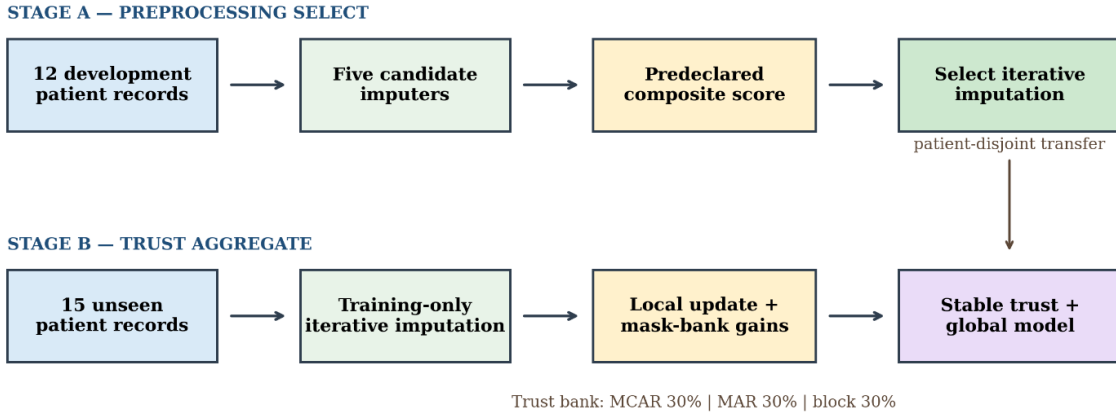


Fig. 1. Unified two-stage CMST-FL model. Stage A selects the imputer on development patients; Stage B transfers only the selected preprocessing decision to unseen patients and performs cross-missingness stable-trust aggregation.

C. Temporal Trust and Aggregation

The instantaneous trust is a temperature-scaled logistic transformation. Exponential smoothing reduces round-to-round weight volatility:

$$z_i(t) = \frac{1}{1 + \exp\left(-\frac{u_i(t)}{\tau}\right)}, \quad (3)$$

$$T_i(t) = \rho T_i(t-1) + (1 - \rho) z_i(t). \quad (4)$$

CMST-FL combines trust with the number n_i of local training examples. With trust floor epsilon and concentration exponent gamma, the normalized aggregation coefficient is

$$a_i(t) = \frac{n_i[\epsilon + T_i(t)]^\gamma}{\sum_{j=1}^K n_j[\epsilon + T_j(t)]^\gamma}, \quad (5)$$

$$w(t+1) = \sum_{i=1}^K a_i(t) w_i(t+1). \quad (6)$$

The floor prevents a client from receiving zero weight, whereas γ controls how strongly reliable clients are emphasized. FedAvg is recovered conceptually when the trust

multiplier is constant across clients. Parameters were fixed before Stage B: $\lambda=0.5$, $\mu=1.0$, $\tau=0.05$, $\rho=0.8$, $\epsilon=0.05$, $\gamma=2$, and $T_i(0)=0.5$.

D. Integrated Two-Stage Decision Model

The complete contribution is represented as an ordered model rather than an isolated aggregation equation. Let I^* be the development-selected imputer, N_{train} the training-only normalization operator, B the validation-mask bank, $T(T)$ the final client-trust vector, and $w(T)$ the final global parameters. The model returned by the procedure is

$$\mathcal{M}_{\text{CMST}} = \{I^*, N_{\text{train}}, B, T(T), w(T)\},$$

$$I^* = \underset{j}{\text{argmin}} S_j. \quad (7)$$

This representation makes the deployment decision explicit: the imputer and normalization are fixed before evaluation patients are exposed, whereas trust and model parameters evolve only inside Stage B. Table II converts the mathematical definition into five operational steps and their outputs.

TABLE II. OPERATIONAL DEFINITION OF THE TWO-STAGE CMST-FL MODEL

Stage	Input	Decision rule	Output
A—Screen	Development patients; five imputers	Minimize predeclared score S_j	Selected imputer I^*
B—Prepare	Unseen patient partitions	Fit I^* and normalization on training only	Leakage-controlled client data
B—Evaluate	Global and local models; bank B	Compute relative gains and penalized utility	Instantaneous trust
B—Stabilize	Current and previous trust	Logistic mapping and temporal smoothing	$T_i(t)$
B—Aggregate	Local models, samples and trust	Sample-aware soft weighting	Final global model $w(T)$

IV. EXPERIMENTAL DESIGN

A. Dataset, Features and Patient Split

The experiment used BIDMC CSV records derived from the PhysioNet dataset [1], [2]. RESP was the regression target, and PLETH, lead II, lead V, and aVR were predictors. Twenty-seven patient recordings were selected deterministically: 12 development records for preprocessing screening and 15 non-overlapping evaluation records for the aggregation study. Each record was uniformly subsampled to at most 200 rows to respect the two-day compute budget. Within every record, the first 60% of samples formed the training partition, the next 20% validation, and the final 20% test, preserving temporal order.

Five clients received patient-record partitions created with a fixed seed. Training-only pooled sufficient statistics were used for feature and target standardization. This design prevents a test statistic, future temporal segment, or evaluation patient from influencing imputer fitting, normalization, preprocessing selection, or trust calculation.

B. Missingness and Preprocessing Screen

Synthetic masks were generated per patient record and seed. MCAR independently masked entries. MAR selected missing rows using a logistic propensity derived from another

feature. Block missingness created contiguous missing segments. A safeguard retained at least one observed feature per row. Native missing values were preserved and were excluded from the synthetic-mask imputation metric.

Stage A compared mean, median, distance-weighted five-neighbor KNN, iterative chained regression (10 iterations, median initialization), and a denoising autoencoder. The autoencoder used two ReLU hidden layers, Adam, and three epochs; therefore, its result represents only the tested low-budget configuration. Screening comprised three mechanisms at 30% missingness and seeds 11 and 22. Each method was evaluated with centralized learning and five-round FedAvg.

The predeclared score S_j combined percentile ranks, with lower values better. R^2 ranks were reversed so that higher R^2 received a lower rank:

$$S_j = 0.35r_{\text{imp},j} + 0.20r_{\text{F, RMSE},j} + 0.10r_{\text{F, } R^2,j} + 0.15r_{\text{C, RMSE},j} + 0.05r_{\text{C, } R^2,j} + 0.15r_{\text{time},j}. \quad (8)$$

Thus, direct imputation carried 35% weight, downstream federated validation 30%, centralized validation 20%, and log-transformed preprocessing time 15%. Only the winning method was advanced to Stage B. Table III consolidates the patient split, model budget, missingness conditions, and fixed aggregation settings.

TABLE III. REPRODUCIBLE EXPERIMENTAL PROTOCOL

Item	Specification
Development records	12 patient records; no Stage B patients
Evaluation records	15 disjoint patient records
Clients	5
Rows per record	At most 200, uniformly spaced
Split	60% train / 20% validation / 20% test, chronological
Stage A screen	5 imputers $\times$ 3 mechanisms $\times$ 1 rate $\times$ 2 seeds
Stage B evaluation	5 strategies $\times$ 3 mechanisms $\times$ 3 rates $\times$ 3 seeds
Central / FL training	10 epochs / 10 rounds; 1 local epoch per round
Network	4–128–64–1 multilayer perceptron; ReLU; Adam 0.001
FedProx	Proximal coefficient 0.01
CMST bank	MCAR, MAR, and block, each at 30%
Seeds	Stage A: 11, 22; Stage B: 11, 22, 33
Execution platform	64-bit Windows 10 Pro laptop; Intel Core i3-5005U CPU at 2.00 GHz; 2 physical cores and 4 logical processors; 7.9 GB RAM; Intel HD Graphics 5500; TensorFlow detected no GPU accelerator, so execution was CPU-only.
Software constraints	Flower $\geq 1.12, < 1.13$; TensorFlow $\geq 2.16, < 2.19$; NumPy $\geq 1.26, < 2.0$; scikit-learn $\geq 1.4, < 2.0$; SciPy $\geq 1.11, < 2.0$

C. Aggregation Baselines

Stage B compared: 1) centralized learning with the same architecture and total epoch count; 2) FedAvg; 3) FedProx with proximal coefficient 0.01; 4) CMST-FL; and 5) FedGuard, an experimental server-side selector that interpolated sample weights and trust/alignment weights using candidate coefficients 0, 0.25, and 0.50 subject to validation gates. FedGuard is treated as a secondary guarded baseline, while CMST-FL is the proposed method.

All FL methods ran for 10 synchronous rounds with one local epoch, all five clients participating and failures disallowed. Strategy comparisons reused the same patient partitions, masks, and seeds. The implementation used Flower for federated communication [18] and TensorFlow/Keras for the multilayer perceptron [24].

D. Metrics and Statistical Analysis

Primary predictive metrics were test R^2 , RMSE, and MAE; higher R^2 and lower errors are better [15], [16]. Imputation RMSE and MAE were computed only at synthetically masked standardized feature entries. We defined worst-client R^2 as the minimum local test R^2 among the five clients for each federated method. Centralized worst-client entries were not used for fairness conclusions because the centralized evaluator produced a pooled metric rather than five local metrics.

Means were calculated across conditions, and 95% percentile bootstrap confidence intervals were estimated with 5,000 resamples [25]. Paired analyses used the 27 matched mechanism–rate–seed conditions and two-sided Wilcoxon signed-rank tests [26]. Holm correction controlled familywise error across the reported strategy–metric comparisons [17]. Because those 27 conditions reuse the same five clients, a supplementary client-cluster bootstrap resampled clients with replacement (50,000 resamples; seed 20260913). The cluster analysis is reported separately and is interpreted cautiously because five clusters cannot support precise population-level uncertainty.

V. RESULTS

A. Preprocessing Selection

Table IV and Fig. 2 summarize Stage A. Iterative imputation achieved the lowest selection score (0.35) and the best direct imputation RMSE (0.7712). KNN had slightly better FedAvg validation RMSE and R^2 , but weaker direct imputation error and centralized validation performance. Mean and median replacement were fastest. The DAE required 3.42 s and had weaker FedAvg validation performance under its three-epoch budget. Because no larger DAE budget was run, this outcome excludes only the tested configuration and does not establish general DAE inferiority.

TABLE IV. STAGE-A PREPROCESSING SCREEN (SIX DEVELOPMENT CONDITIONS)

Method	Imp. RMSE	Central R^2	FedAvg R^2	Time (s)	Score
Iterative	0.7712	0.1641	0.0846	0.273	0.35
KNN	0.9400	0.1341	0.0920	0.226	0.56
Mean	0.9515	0.1614	0.0744	0.078	0.57
DAE	0.8605	0.1539	0.0320	3.416	0.75
Median	0.9808	0.1603	0.0637	0.079	0.77

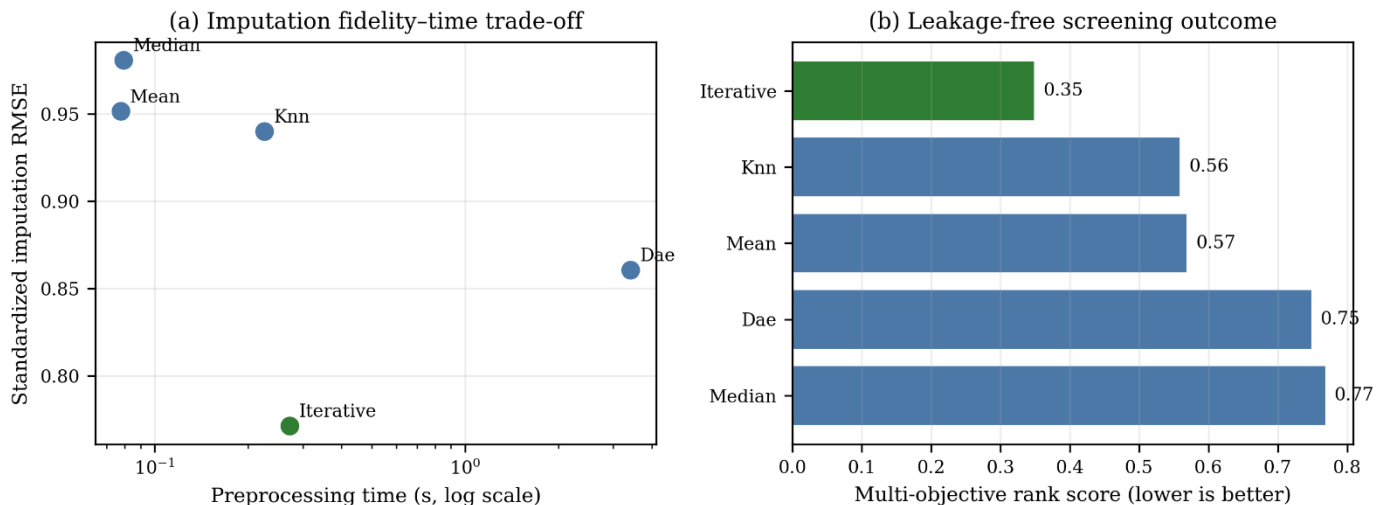


Fig. 2. Preprocessing selection. Iterative imputation provided the best predeclared multi-objective score under the tested data and compute budget; the result is not a universal ranking of imputers.

B. Overall Aggregation Performance

Table V summarizes the 27 Stage B runs per strategy, and Fig. 3 shows the corresponding means and bootstrap intervals. CMST-FL obtained the highest mean R^2 (0.0130) and the

lowest RMSE (0.3298) and MAE (0.2624). Relative to FedAvg, its average R^2 gain was 0.00658, while RMSE and MAE decreased by 0.00110 (0.33%) and 0.00091 (0.35%), respectively. The directions were consistent, but the absolute differences were small.

TABLE V. STAGE-B TEST PERFORMANCE; MEAN AND 95% BOOTSTRAP CI FOR R^2

Strategy	R^2 [95% CI]	RMSE	MAE	Worst-client R^2
Centralized	0.0045 [-0.0007, 0.0098]	0.3312	0.2649	—
FedAvg	0.0064 [0.0037, 0.0094]	0.3309	0.2633	-0.0738
FedProx	0.0061 [0.0033, 0.0091]	0.3310	0.2627	-0.0725
FedGuard	0.0077 [0.0052, 0.0104]	0.3307	0.2631	-0.0673
CMST-FL	0.0130 [0.0104, 0.0157]	0.3298	0.2624	-0.0735

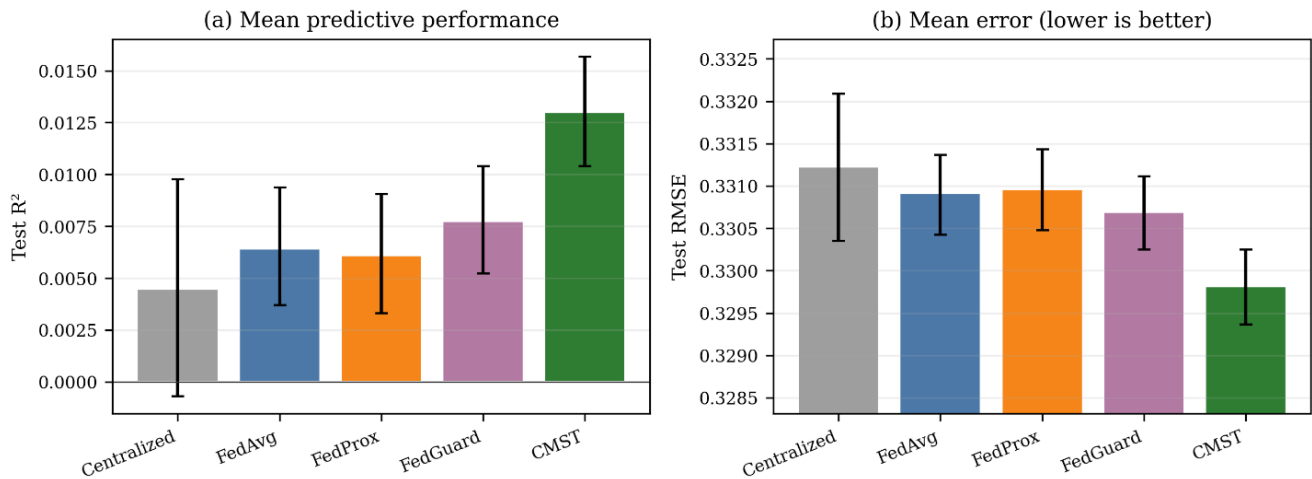


Fig. 3. Mean Stage B performance across 27 matched conditions. Error bars are 95% bootstrap confidence intervals. The truncated RMSE axis is used to make small between-method differences visible.

C. Paired Comparison with FedAvg

Table VI reports paired results. CMST-FL improved R^2 and RMSE in all 27 matched conditions and improved MAE in 25. Holm-adjusted p-values were below 0.000001 for those three metrics, and condition-level intervals excluded zero. In contrast, the mean worst-client R^2 gain was 0.00027, its interval crossed zero, and the adjusted p-value was 1.0. Thus, the experiment supports a small pooled predictive improvement, not a reliable fairness improvement.

D. Dependence-Aware and Trust-Trajectory Analysis

The client-cluster bootstrap retained the 135 matched client-condition differences but resampled the five client

clusters. The mean client-level R^2 improvement was 0.00594, with a 95% cluster interval of [-0.01229, 0.03037]. The mean RMSE reduction was 0.00102, with an interval of [-0.00207, 0.00518]. Both intervals crossed zero. This result does not contradict the condition-level paired test; it shows that uncertainty over a broader client population remains large when only five clients are observed.

Table VII and Fig. 4 connect CMST-FL's utility calculation to the aggregation trajectory. Across 1,350 client-round rows, temporal trust ranged from 0.361 to 0.745, and normalized weights ranged from 0.087 to 0.306. The mean absolute departure from the equal 0.20 weight was 0.034; penalized utility and weight were strongly associated ($r=0.918$).

TABLE VI. PAIRED CMST-FL IMPROVEMENT OVER FEDAVG (27 MATCHED CONDITIONS)

Metric	Mean improvement	95% bootstrap CI	Win rate	Holm p
R^2 gain	0.006581	[0.005287, 0.007907]	100.0%	<0.000001
RMSE reduction	0.001098	[0.000882, 0.001318]	100.0%	<0.000001
MAE reduction	0.000910	[0.000672, 0.001146]	92.6%	<0.000001
Worst-client R^2 gain	0.000267	[-0.006561, 0.007737]	40.7%	1.000

TABLE VII. CLIENT UTILITY, TEMPORAL TRUST AND NORMALIZED CMST-FL WEIGHT

Client	Raw utility	Penalized utility	Mean trust	Mean weight	Weight range
C1	0.0274	0.0256	0.5806	0.2147	[0.1796, 0.2449]
C2	0.0227	0.0210	0.5702	0.2078	[0.1872, 0.2310]
C3	-0.0062	-0.0178	0.4461	0.1343	[0.0867, 0.1960]
C4	0.0499	0.0485	0.6452	0.2609	[0.2039, 0.3055]
C5	0.0107	0.0087	0.5307	0.1823	[0.1631, 0.2066]

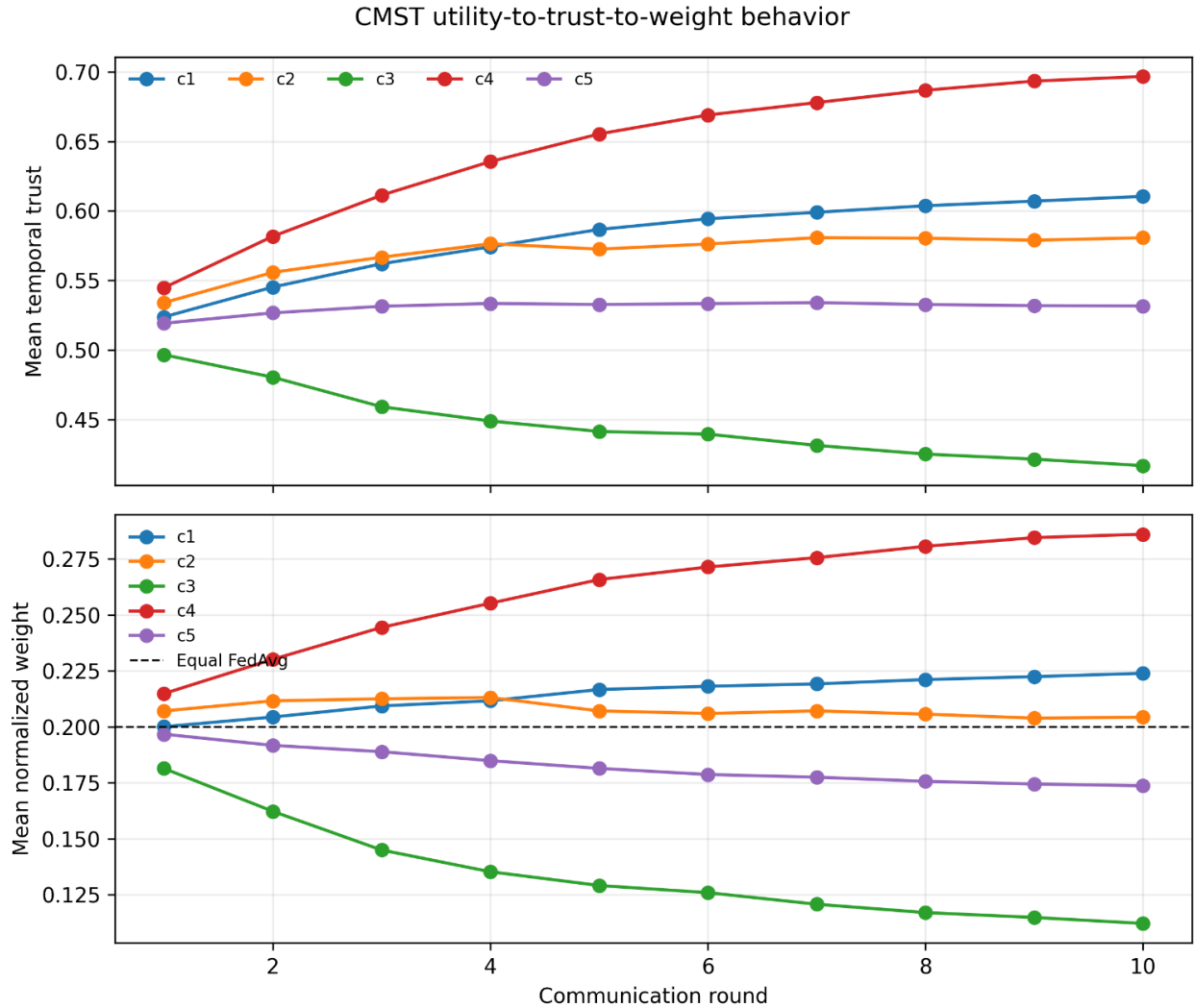


Fig. 4. CMST-FL temporal trust and normalized aggregation-weight trajectories, averaged over the 27 evaluation conditions. The dashed weight reference is the equal 0.20 allocation used because every client had 360 training examples.

Client C4 received the largest mean weight (0.261), whereas C3 received the smallest (0.134), confirming that CMST-FL materially changed the optimization weights. Nevertheless, all clients had 360 training examples, so this experiment isolates the trust multiplier and does not empirically validate behavior under unequal sample counts.

E. Robustness Across Missingness Conditions

Fig. 5 and Table VIII report performance across the held-out evaluation grid. The CMST-FL R^2 gain over FedAvg was

positive in every mechanism–rate cell after averaging three seeds. The largest mean gain occurred for MCAR at 30% (+0.01107), while the smallest occurred for block missingness at 50% (+0.00281). CMST-FL also had the highest R^2 among the four federated strategies in each of the 27 conditions. Because the validation bank used the same mechanism families, these results demonstrate robustness within MCAR, MAR, and block missingness—not generalization to genuinely unseen mechanisms.

CMST-FL gain across held-out evaluation conditions

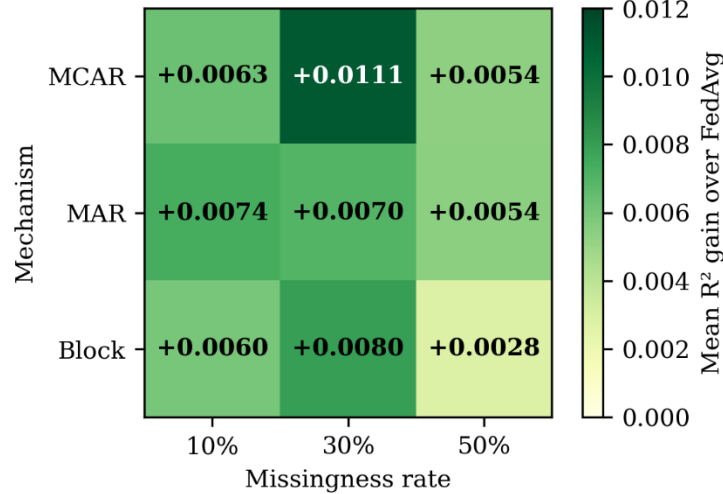


Fig. 5. Mean CMST-FL test R^2 gain over FedAvg by missingness mechanism and rate. Every cell is positive; each cell averages three seeds.

TABLE VIII. MEAN TEST R^2 BY MISSINGNESS CONDITION (THREE SEEDS)

Mechanism	Rate	Centralized	FedAvg	CMST-FL
MCAR	10%	0.0052	0.0097	0.0160
MCAR	30%	0.0065	0.0039	0.0149
MCAR	50%	0.0019	0.0047	0.0100
MAR	10%	0.0131	0.0065	0.0139
MAR	30%	0.0108	0.0074	0.0144
MAR	50%	0.0008	0.0064	0.0118
Block	10%	0.0166	0.0133	0.0193
Block	30%	0.0017	0.0049	0.0129
Block	50%	-0.0164	0.0009	0.0037

F. Centralized–Federated Gap

CMST-FL exceeded the matched centralized R^2 and reduced matched RMSE in 18 of 27 conditions. Averaged over all conditions, CMST-FL improved R^2 by 0.00851 and reduced RMSE by 0.00141 relative to the centralized reference. This result should not be read as a general advantage of FL: the centralized model has access to pooled data but was constrained to the same compact architecture, ten epochs, and a reduced dataset. The defensible conclusion is that CMST-FL reduced and, in this benchmark, slightly reversed the average centralized–federated gap under the fixed budget.

VI. DISCUSSION

A. Why CMST-FL Helped

Three design elements plausibly explain the consistency. First, trust is based on improvement relative to the received model, rather than absolute local performance. Second, the bank discourages an update that is beneficial under one mask but harmful under another. Third, temporal smoothing limits abrupt reweighting from a single noisy validation round. The sample-count factor retains FedAvg's preference for estimates supported by more local examples.

The results also show why significance alone is insufficient. All 27 paired R^2 and RMSE differences favored CMST-FL, producing very small p-values, yet the absolute error reduction was about one-thousandth. This is encouraging evidence of a stable directional effect, not evidence of a large clinical benefit. The low absolute R^2 values indicate that the compact four-input model explains little test variance and that the task remains difficult after aggressive subsampling.

B. Preprocessing Interpretation

Iterative imputation was preferred because it combined the lowest direct imputation error with strong centralized validation results and moderate runtime. It did not lead to the best individual downstream metric: KNN achieved slightly stronger FedAvg validation R^2 . The multi-objective score makes this trade-off explicit. Therefore, the article can answer which tested method was best under the defined score and budget, but it should not claim that iterative imputation is universally superior. Longer DAE training, different signal windows, or richer temporal models could change the ranking.

C. Comparison with FedProx and Guarded Weighting

FedProx produced nearly the same mean R^2 as FedAvg under the single fixed coefficient $\mu=0.01$ and one local epoch.

Because no coefficient sweep was performed, the result characterizes only this configuration and cannot establish a general limitation of proximal regularization. FedGuard improved mean R^2 over FedAvg but remained below CMST-FL. Its validation gate accepted a nonzero trust mixture in 38 of 270 server rounds (14.1%), so most rounds retained sample weighting.

D. Practical Implications

CMST-FL is most applicable when each participating institution can reserve a local validation segment and evaluate client updates under several plausible missingness conditions without sharing raw patient data. Compared with FedAvg, the additional client-side workload consists mainly of validation inference under the three conditions in the missingness bank. The server then calculates the utility, updates the temporal trust values, and normalizes the aggregation weights. These server-side operations are lightweight relative to local model training. However, institutions must maintain representative local validation data, and the choice of missingness conditions should reflect the types of signal loss expected in the intended application. The archived experiment controller recorded a total runtime of 5.081 hours for the complete two-stage pipeline. The experiment was executed on a 64-bit Windows 10 Pro laptop equipped with an Intel Core i3-5005U CPU

operating at 2.00 GHz, two physical cores, four logical processors, and 7.9 GB of RAM. Although the laptop included Intel HD Graphics 5500, TensorFlow detected no GPU accelerator; consequently, the experiment used CPU-only execution. The hardware specifications were verified on the same machine after the experiment was completed. The reported runtime should therefore be interpreted as an observed result for this specific hardware and experimental configuration rather than as a hardware-normalized performance benchmark. Nevertheless, completing the pipeline on modest hardware suggests that the proposed framework can be evaluated without specialized high-performance computing resources. Its practical value remains subject to confirmation using larger client populations, longer training schedules, and full-resolution physiological signals.

VII. EVIDENCE-SUPPORTED FINAL MODEL

The experiment supports one concrete model for the present benchmark: training-only iterative imputation followed by CMST-FL aggregation. The decision was not selected from the final test results. Iterative imputation was frozen after Stage A because it minimized the development score; the complete CMST configuration was then evaluated on disjoint patients. Table IX records the resulting model so that the contribution is reproducible and falsifiable.

TABLE IX. FINAL EMPIRICAL CONFIGURATION OF THE CMST-FL MODEL

Component	Selected specification	Evidence or role
Preprocessing	Iterative chained regression; 10 iterations; median initialization	Best composite score 0.35 and lowest imputation RMSE 0.7712
Normalization	Training-only pooled sufficient statistics	Prevents validation/test leakage
Predictor	PLETH, IL, V and aVR $\rightarrow$ RESP; dense 128–64–1 network	Same compact architecture across strategies
Trust evidence	Relative local-versus-global composite quality gain under MCAR, MAR and block at 30%	Measures cross-mechanism marginal utility
Stability	Mean gain – 0.5 SD – worst-case negative-gain penalty	Discourages brittle or harmful updates
Temporal memory	$\rho=0.8$; logistic temperature 0.05; initial trust 0.5	Reduces round-to-round volatility
Aggregation	$n_i(0.05 + T_i)^2$ normalized across clients	Equal $n_i=360$ here; unequal-sample behavior remains untested
Observed outcome	R^2 0.0130; RMSE 0.3298; MAE 0.2624	Best mean pooled metrics among tested strategies

The strongest result is consistency rather than magnitude: CMST-FL improved R^2 and RMSE over FedAvg in every matched condition, yet the absolute error reduction was small. Consequently, the final model should be presented as an evidence-supported prototype that narrows the tested federated performance gap, not as a guaranteed optimal or clinically validated model.

VIII. THREATS TO VALIDITY AND LIMITATIONS

The primary limitation is scale. The evaluation used 27 of the 53 BIDMC records, each restricted to 200 uniformly spaced rows, five clients, and one dataset. Training stopped after 10 synchronous rounds, so the ranking is a fixed-budget comparison rather than evidence of convergence. Larger client populations, longer training, full-resolution windows, and external datasets are required before population-level or clinical claims can be made.

Second, missingness was synthetically injected. MCAR, propensity-based MAR, and contiguous blocks represent

different patterns, but real sensor loss may be nonstationary, device-specific, or missing-not-at-random. The validation bank used the same three families as the evaluation grid, so genuinely unseen mechanisms were not tested. Third, the small feed-forward predictor did not model temporal windows, which helps explain the low R^2 values. Fourth, all clients had equal sample counts, leaving the sample-aware factor untested under quantity heterogeneity.

Fifth, Stage A used two seeds and Stage B three seeds. The client-cluster analysis exposes the resulting uncertainty but cannot overcome having only five clusters. The DAE received only three epochs, FedProx used only $\mu=0.01$, and CMST-FL hyperparameters were fixed. No component ablation removed the stability, worst-case, temporal, or nonlinear-weighting terms; consequently, the current experiment supports the complete formulation but cannot attribute the gain to an individual component. Hardware specifications were verified after execution on the same machine, although separate wall-clock times and communication overhead for the individual

strategies were not recorded. Future work will address these limitations with DAE and FedProx sensitivity analyses, incremental CMST-FL ablations, at least 30 rounds, more clients, unseen mask families, unequal sample sizes, and stronger temporal predictors.

IX. CONCLUSION

This study introduced CMST-FL as a two-stage preprocessing-and-aggregation framework for incomplete physiological signals. Its distinguishing formulation evaluates each client update against the received global model with a composite quality score across several missingness mechanisms, penalizes unstable and worst-case harmful gains, and smooths trust across rounds. Stage A selected iterative imputation under the predeclared multi-objective score; Stage B found that CMST-FL improved R^2 and RMSE over FedAvg in all 27 matched conditions and achieved the best mean pooled metrics among the tested strategies.

The evidence therefore supports iterative training-only preprocessing followed by cross-missingness stable-trust aggregation for this benchmark. The novelty claim is specific to this integration and formulation. The low absolute R^2 , small error reductions, wide client-cluster intervals, fixed 10-round budget, and lack of ablations preclude claims of universal or clinical superiority.

REPRODUCIBILITY AND DATA AVAILABILITY

The BIDMC PPG and Respiration Dataset is publicly available from PhysioNet subject to its terms [1], [2]. The experimental archive preserves configuration files, deterministic seeds, patient partitions, missingness masks generated from stable identifiers, raw per-run JSON/CSV logs, console logs, and scripts used to create the publication tables. An anonymized code-and-results archive can be released with the manuscript. No new patient data were collected.

Evidence and figure provenance: Reported experimental values and Fig. 2, 3, and 5 were computed from the archived Stage A and Stage B CSV/JSON outputs. Fig. 4 was computed from the archived per-client utility, trust, and weight logs. Fig. 1 is a conceptual workflow. All figures were created programmatically; no generative-image model was used. Displayed values are rounded, and intervals and p-values are derived statistics rather than raw observations.

DECLARATIONS

Conflict of interest: The authors declare no conflict of interest. Ethical statement: The study used an openly available, de-identified dataset and involved no new human-participant data collection. Data availability: The source dataset is publicly available through PhysioNet [1], [2]; the analysis code and derived result tables can be supplied as supplementary material.

AI-assisted writing disclosure: A generative-AI assistant supported language editing, organization, equation typesetting, formatting, and evidence cross-checking; it did not generate experimental observations or numerical results. The authors remain responsible for independently verifying the code, logs, calculations, citations, claims, and final text.

REFERENCES

- [1] M. A. F. Pimentel, A. E. W. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton, "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1914–1923, 2017, doi: 10.1109/TBME.2016.2613124.
- [2] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000, doi: 10.1161/01.CIR.101.23.e215.
- [3] M. Ali, F. Naeem, M. Tariq, and G. Kaddoum, "Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 2, pp. 778–789, 2023, doi: 10.1109/JBHI.2022.3181823.
- [4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, PMLR, vol. 54, pp. 1273–1282, 2017.
- [5] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. MLSys*, 2020.
- [6] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," in *Proc. ICLR*, 2020.
- [7] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in *Proc. NDSS*, 2021, doi: 10.14722/ndss.2021.24434.
- [8] O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001, doi: 10.1093/bioinformatics/17.6.520.
- [9] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011, doi: 10.18637/jss.v045.i03.
- [10] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. ICML*, pp. 1096–1103, 2008, doi: 10.1145/1390156.1390294.
- [11] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Hoboken, NJ, USA: Wiley, 2019.
- [12] R. Vavekanand, A. A. Sathio, and M. Sultani, "A federated learning framework for deep imputation of missing data in heterogeneous ICU time series," *Sci. Rep.*, vol. 16, Art. no. 18097, 2026, doi: 10.1038/s41598-026-49311-7.
- [13] M. Zhang, K. Sapra, S. Fidler, S. Yeung, and J. M. Alvarez, "Personalized federated learning with first-order model optimization," in *Proc. ICLR*, 2021.
- [14] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 1142–1154, 2022, doi: 10.1109/TSP.2022.3153135.
- [15] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, Art. no. e623, 2021, doi: 10.7717/peerj-cs.623.
- [16] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not," *Geosci. Model Dev.*, vol. 15, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.
- [17] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.
- [18] D. J. Beutel et al., "Flower: A friendly federated learning research framework," *arXiv:2007.14390*, 2020.
- [19] Z. Li, T. Lin, X. Shang, and C. Wu, "Revisiting weighted aggregation in federated learning with neural networks," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 202, pp. 19767–19788, 2023.
- [20] Y. Tang, "Adapted weighted aggregation in federated learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 21, pp. 23763–23765, 2024, doi: 10.1609/aaai.v38i21.30557.
- [21] S. Min, H. Asif, X. Wang, and J. Vaidya, "CAFE: Improved federated data imputation by leveraging missing data heterogeneity," *IEEE Trans.*

- Knowl. Data Eng., vol. 37, no. 5, pp. 2266–2281, 2025, doi: 10.1109/TKDE.2025.3537403.
- [22] C. Yao, H. Wang, F. Yang, Y. Yang, and Z. Guo, “A loss-based weighted aggregation method for federated learning with heterogeneous computing resources,” in Proc. 9th Asia-Pacific Workshop on Netw. (APNet), pp. 307–309, 2025, doi: 10.1145/3735358.3737773.
- [23] P. Shepherd, T. Dagiuklas, B. Alkan, and J. Rodriguez, “Dynamic trust management for secure federated learning in critical industrial and IoT networks,” in Proc. IEEE ICC Workshops, pp. 371–376, 2025, doi: 10.1109/ICCWorkshops67674.2025.11162455.
- [24] M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” in Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI), pp. 265–283, 2016.
- [25] B. Efron, “Bootstrap methods: Another look at the jackknife,” Ann. Stat., vol. 7, no. 1, pp. 1–26, 1979, doi: 10.1214/aos/1176344552.
- [26] F. Wilcoxon, “Individual comparisons by ranking methods,” Biometrics Bull., vol. 1, no. 6, pp. 80–83, 1945, doi: 10.2307/3001968.