

AI-Based Preprocessing and R-Squared Weighted Federated Aggregation for Robust Healthcare Signal Imputation

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

Abstract—Federated learning is an emerging paradigm for collaborative model training in healthcare applications where sensitive patient data cannot be centralized due to privacy concerns; however, missing values in physiological signals remain a major challenge, often leading to degraded model accuracy and unstable convergence. To address this issue, this work presents an AI-based preprocessing framework that combines advanced imputation methods, including K-Nearest Neighbors (KNN), Autoencoder reconstruction, and Generative Adversarial Networks (GAN), with classical baselines such as mean and median imputation. Building on this preprocessing stage, this study proposes a novel R-squared weighted aggregation strategy that dynamically scales clients' contributions based on their prediction quality. The experiments were conducted on BIDMC physiological datasets with missingness rates ranging from 0.1 to 0.5 across multiple federated clients. Results show that weighted aggregation converges to lower RMSE and MAE than classical FedAvg, especially in the early rounds where high client heterogeneity exists. The proposed method adapts to client quality, thereby providing robust convergence and stability, resulting in balanced performance. These results demonstrate that combining AI-based preprocessing with R-squared weighted aggregation effectively improves the reliability of federated learning in healthcare IoT systems. The proposed approach presents a scalable and feasible solution for real-world scenarios where data incompleteness and client heterogeneity are inevitable.

Keywords—Missing data imputation; healthcare signals; weighted aggregation; AI-based preprocessing; federated learning

I. INTRODUCTION

Federated learning has emerged as a significant paradigm in contemporary artificial intelligence, particularly in sensitive fields such as healthcare, finance, and education [1]. Its main advantage is that models can be trained collaboratively across multiple institutions without requiring direct access to raw data, thereby preserving privacy and security [2]. This distributed approach allows knowledge sharing while respecting data governance rules, which is especially critical in the medical domain where patient confidentiality is paramount [3]. Federated learning supports distributed training and provides an opportunity to exploit heterogeneous datasets while addressing ethical and legal concerns [4]. Despite its promise, federated learning faces persistent challenges related to heterogeneity and incompleteness of client data [5]. In practice, datasets often contain missing values, imbalanced distributions, and varying quality levels [6]. Classical

aggregation methods such as FedAvg treat all clients' updates equally, which can lead to unstable convergence and reduced accuracy when some clients provide noisy or unreliable updates [7]. This gap between theoretical promise and practical reliability underscores the need for adaptive strategies that account for differences in client data quality [8].

The central research question driving this work is: How can federated learning be improved to guarantee reliable performance in the presence of heterogeneous and incomplete client data, particularly in healthcare applications where accuracy and stability are paramount?

A. Novelty and Contributions

This study proposes a new federated learning framework in two aspects. First, it combines a variety of AI-based imputation strategies (K-Nearest Neighbors, Autoencoder reconstruction, and Generative Adversarial Networks) with classical baselines (mean and median) to handle missing data in healthcare signals. Second, it proposes an R-squared weighted aggregation mechanism that dynamically weights the clients' contributions based on the quality of their predictions to reduce the effect of client heterogeneity. The scientific contributions of this work can be summarized as follows: Firstly, it systematically integrates advanced imputation strategies with a client-quality-aware aggregation scheme in federated learning; Secondly, it provides the formal derivation of both classical FedAvg and the proposed R-squared weighted aggregation, emphasizing the differences and benefits of the proposed approach; Thirdly, it carries out extensive experiments on BIDMC physiological datasets with different missingness rates, including direct comparisons among imputation methods and statistical validation of aggregation performance. Finally, it shows practical impact by improving reliability and stability in federated healthcare IoT systems, providing a scalable approach for real-world scenarios where data incompleteness and client diversity are unavoidable. This study proposes a framework to solve this problem that combines several AI-based preprocessing techniques with a new R-squared weighted aggregation mechanism. The main contributions are three. This study proposes classical baselines (mean, median) and extends them with state-of-the-art imputation methods (KNN, Autoencoder, GAN) for better preprocessing of incomplete healthcare signals. This study proposes an R-squared weighted aggregation mechanism to dynamically scale the contribution of clients according to their prediction quality. The study provides a comprehensive experimental evaluation

on the BIDMC physiological datasets with different levels of missingness, including statistical validation and comparison of imputation methods. The rest of the study is organized as follows. Section II presents the related work on imputation techniques and federated aggregation strategies. Section III elaborates on the methodology, including dataset preparation, preprocessing approaches, and evaluation metrics. Section IV reports the comparative results between classical FedAvg and the proposed R^2 -weighted aggregation. Section V discusses the strengths, limitations, and practical implications of the method. Section VI concludes the study and outlines directions for future research.

II. RELATED WORK

A. Missing Data Imputation

Missing data in physiological signals has been addressed using a variety of preprocessing techniques. Classical statistical baselines such as mean and median imputation offer computational efficiency but fail to capture nonlinear dependencies [18]. More advanced approaches include K-Nearest Neighbors (KNN), which leverages local similarity structures to estimate missing values [19], and Autoencoder-based reconstruction, which learns latent representations to infer missing entries from underlying nonlinear patterns [20]. Generative Adversarial Networks (GANs) have also been applied to model complex physiological signal distributions and generate realistic imputations [21]. Additional preprocessing strategies include outlier detection, class imbalance correction using SMOTE, normalization, and natural language processing for unstructured medical text [10–12]. Additional preprocessing strategies for healthcare data, including KNN-based imputation, outlier detection, and SMOTE-based balancing, have also been explored in prior work [9]. These methods highlight the importance of robust preprocessing in improving downstream learning performance.

B. Federated Aggregation Methods

Federated learning has introduced several aggregation strategies beyond the classical FedAvg algorithm. Weighted FedAvg incorporates dataset size into the aggregation process, while robust statistical methods such as Median, Trimmed Mean, and Krum aim to mitigate the influence of unreliable or adversarial clients [7, 13–15]. Byzantine-resilient approaches such as GeoMed and Clipped Clustering further enhance robustness under extreme heterogeneity. Adaptive weighting schemes based on gradient norms have also been proposed to improve convergence stability in non-IID environments [16, 17].

More recent methods explicitly address client heterogeneity. FedProx introduces proximal regularization to stabilize updates from diverse clients, FedNova normalizes local training to reduce objective inconsistency, SCAFFOLD employs control variates to correct client drift, and FedAdam applies adaptive optimization at the server level. These approaches modify gradient dynamics or optimization behavior to improve global model stability under heterogeneous conditions.

C. Distinction of the Present Work

The present study differs from prior aggregation strategies by incorporating a reliability-aware weighting mechanism based on the coefficient of determination (R^2). Unlike gradient-based or dataset-size-based weighting schemes, R^2 weighting directly reflects the explanatory performance of each client model. This provides an interpretable and transparent measure of client reliability, particularly relevant in healthcare scenarios where data quality varies significantly across institutions. The proposed framework integrates advanced imputation methods with R^2 -weighted aggregation to enhance robustness under heterogeneous and incomplete data conditions.

III. METHOD

A. Dataset

This study utilized several physiological signal records from the BIDMC dataset available on PhysioNet. These include, but are not limited to, `bidmc_20_Signals.csv`, `bidmc_28_Signals.csv`, `bidmc_31_Signals.csv`, `bidmc_35_Signals.csv`, and `bidmc_43_Signals.csv`. The multi-channel signals in each record are RESP, PLETH, ECG leads (II, AVR, V), and ABP for one patient. Signals were sampled at 125 Hz and lasted from 6 to 10 minutes. There are 53 patients in total in the dataset. The study used a train-validation-test split of 70%-15%-15%. Synthetic missingness was introduced at the rates 0.1, 0.3, and 0.5, both by MCAR (Missing Completely at Random) and MAR (Missing at Random) procedures to mimic real-world incompleteness. To prevent data leakage, the train-validation-test split was performed strictly at the patient level. Each BIDMC record corresponds to a unique patient, and no patient appears in more than one subset. This ensures that the model is never evaluated on physiological signals originating from a patient whose data contributed to training, thereby preserving the independence of evaluation and avoiding optimistic bias. Patient-level splitting is essential in physiological time-series tasks because intra-patient temporal patterns are highly correlated; mixing segments from the same patient across different subsets would artificially inflate performance.

B. Preprocessing Approaches

1) *Mean and median imputation*: These classical statistical baselines replace missing entries with the average or median of the observed values in each feature. While computationally efficient, they fail to capture nonlinear dependencies and often underestimate variance [18].

2) *K-Nearest Neighbors (KNN) imputation*: This method estimates missing values by identifying the k most similar samples based on Euclidean distance and averaging their corresponding feature values. KNN leverages local similarity structures, making it more adaptive than mean or median imputation [19]. KNN imputation was configured with $k = 5$ using the Euclidean distance metric. These implementation details ensure reproducibility and allow replication of the experiments.

3) *Autoencoder reconstruction*: A deep neural network was trained to compress and reconstruct input signals. By

learning latent representations, the autoencoder can infer missing values from underlying nonlinear patterns. This approach is particularly effective in capturing hidden correlations among physiological signals [20]. The autoencoder consisted of three encoder layers (128–64–32 units), a latent dimension of 32, a symmetric decoder, ReLU activations, the Adam optimizer (learning rate 0.001), a batch size of 64, and 100 training epochs.

4) *Generative Adversarial Network (GAN) imputation*: A generative adversarial framework was employed, where a generator network produces plausible imputations and a discriminator network evaluates their realism. GANs are capable of modelling complex distributions, yielding imputations that closely resemble true physiological dynamics [21]. The generator was implemented with three dense layers (128–64–32 units) using ReLU and a Sigmoid output layer, while the discriminator consisted of three dense layers (128–64–32 units) with LeakyReLU activations. The optimizer was Adam (learning rate 0.0002), with batch size 64 and 200 epochs.

All hyperparameters were kept consistent across experiments to ensure reproducibility and allow exact replication of the results.

C. Relationship Between Imputation and Federated Training

The federated learning process in this study operates on imputed data, making the relationship between imputation and model training explicit. Let x_i denote the input features of client i , where missing values have been replaced using a chosen imputation method, and let y_i denote the corresponding target variable. The local learning objective for each client is to minimize a supervised loss function $L(f_i(x_i), y_i)$ where f_i is the client's local model. Because imputation modifies the input distribution, the quality of the imputed values directly affects the gradients computed during local training and, consequently, the updates sent to the server.

Formally, the input to each client model is the imputed dataset x_i^{imp} , the target is the original ground-truth signal y_i and the objective is to minimize the prediction error $L(f_i(x_i^{\text{imp}}), y_i)$. Imputation therefore acts as a preprocessing transformation that shapes the feature space on which federated optimization is performed. High-quality imputation reduces noise and improves the stability of local gradients, whereas poor imputation can introduce bias or distort client updates, amplifying heterogeneity across clients. This interaction highlights that imputation is not an isolated preprocessing step but an integral component influencing the behavior and convergence of the federated learning process.

Imputation performance was evaluated exclusively on the artificially masked values, ensuring that the error metrics reflect the true reconstruction ability of each method rather than the model's behavior on already-observed data. For each client, missingness was generated under a Missing At Random (MAR) mechanism: the probability of masking a value depended on observable features such as signal amplitude and local temporal variability. This procedure ensures that

missingness is not uniform but correlated with clinically meaningful characteristics, reflecting realistic sensor degradation patterns.

To guarantee reproducibility, the MAR masking process was implemented using a fixed random seed and a deterministic masking function applied consistently across all clients. Each client received a distinct but reproducible MAR pattern, allowing fair comparison of imputation methods and federated aggregation strategies. All experiments were executed using the same masking configuration, and the original unmasked dataset was preserved to compute ground-truth reconstruction errors.

D. Federated Setup

The Federated Learning Experiments Were Implemented Using the Flower Framework (v1.0+), Which Provides a Flexible Architecture for Simulating Distributed Training on Multiple Clients. For This Study, We Set up Five Federated Clients, Each Representing a Different Hospital or Data Source, with Heterogeneous Preprocessing Strategies (mean, Median, KNN, Autoencoder, GAN) and Different Rates of Missingness (0.1-0.5). The Server Was Initialized with the FedAvg Strategy and Extended with the Proposed R-Squared-weighted Aggregation Function, Enabling Dual Logging of Classical and Weighted Metrics. The Training Was Performed over 100 Federated Rounds, Long Enough to Observe Early-round Heterogeneity Effects and Long-term Convergence Behavior. Each Client Was Launched Independently and connected to the Central Server Using gRPC to Ensure Synchronous Communication and Secure Parameter Exchange. At Each Round, the Server Aggregated the Model Updates, Producing Both Classical Averages and Weighted Results Based on Client-specific R^2 Scores. This Setup Is Similar to a Real-world Healthcare IoT Setting Where Several Institutions Collaboratively Train a Common Global Model While Preserving the Privacy of Local Data. The Federated Environment, by Simulating Diverse Preprocessing Methods and Missingness Rates Across Five Clients, Provided a Robust Testbed for Evaluating the Effectiveness of AI-based Preprocessing Combined with a Weighted Approach.

The experimental configuration assigns each client a different preprocessing method and a different missingness rate. This design introduces a confounding factor because the observed improvements cannot be attributed solely to the aggregation mechanism; they reflect the combined influence of preprocessing heterogeneity and missingness variability. This setup was intentionally chosen to simulate realistic non-IID healthcare environments, where institutions differ in data quality, preprocessing pipelines, and sensor reliability. To isolate the specific contribution of the aggregation strategy, future work will include controlled ablation studies in which all clients use identical preprocessing and identical missingness levels.

The heterogeneity used in this study is artificially simulated by assigning each client a distinct preprocessing method and a different missingness level. Although this configuration does not represent real multi-institution variability, it provides a controlled way to emulate the diversity typically observed across hospitals, sensors, and data-collection pipelines. This

synthetic heterogeneity allows the evaluation of aggregation robustness under challenging non-IID conditions. Future work will extend this analysis to genuine multi-institution datasets, enabling validation of the proposed approach in real federated healthcare environments.

E. Aggregation: Classical FedAvg Vs R²Weighted

The federated learning configuration consists of five clients, each simulating a distinct data-quality condition based on different missingness levels. All experiments were executed on a workstation equipped with an NVIDIA RTX 3080 GPU and an Intel i7 processor, while clients were simulated sequentially to ensure reproducibility. The training process was conducted for 1000 communication rounds, allowing both the classical FedAvg and the proposed R²-weighted aggregation to converge. Communication between clients and the server assumes a standard bandwidth of 100 Mbps, which is representative of typical hospital or clinical network environments. Although the current setup uses five simulated clients, scalability is limited primarily by memory constraints and communication overhead; increasing the number of clients would require distributed execution or asynchronous aggregation strategies, which the study identifies as a direction for future work.

Both strategies were executed in parallel during training, with performance metrics (RMSE, MAE, R²) logged at each round. This design allowed direct comparison between classical averaging and weighted aggregation, highlighting the conditions under which the proposed method provides significant improvements.

1) *Classical FedAvg*: In the standard approach, each client's model update is averaged uniformly, or proportionally to its dataset size, without considering the quality of the local model. This method is widely adopted due to its simplicity and stability, but it does not differentiate between strong and weak clients.

The aggregation rule is defined as follows:

$$\theta_{global}^{(t+1)} = \frac{1}{N} \sum_{i=1}^N \theta_i^{(t)}$$

This equation shows that the global model parameters are updated by taking the simple average of all client updates, treating each client equally regardless of data quality.

2) *R²Weighted approach*: To enhance robustness, the proposed method introduces a weighting scheme based on the coefficient of determination (R²) obtained from each client's evaluation. Clients with higher explanatory power (higher R²) are assigned greater weights, while weaker clients contribute less to the global update. This dynamic adjustment ensures that the global model prioritizes reliable local models, particularly in heterogeneous environments.

The aggregation rule is defined as follows:

$$\theta_{global}^{(t+1)} = \frac{\sum_{i=1}^N R_i^2 \cdot \theta_i^{(t)}}{\sum_{i=1}^N R_i^2}$$

Here, each client's update is weighted by its explanatory power (R²). Clients with higher R² values contribute more strongly, while weaker clients have less influence. This derivation highlights how weighting by predictive quality differs from dataset-size weighting.

3) *Theoretical justification for R² weighting*: The coefficient of determination (R²) provides a theoretically grounded measure of how well a model explains the variance of its target variable. In the context of federated learning, each client trains a local model on its own data distribution. A higher R² indicates that the client model captures a larger portion of the underlying signal, whereas a lower R² suggests poor fit, noise, or unreliable updates. Using R² as a weighting factor therefore prioritizes clients whose updates are more informative and reduces the influence of clients whose data or training conditions lead to weak explanatory performance. This makes R² a natural proxy for client reliability.

4) *Comparison with existing weighting strategies*: There are different weighting strategies proposed in federated learning. In loss-based weighting, clients with lower local loss are assigned larger weights; however, this method can be unstable due to the heterogeneous loss values across distributions. FedAvg uses data-size weighting to prioritize clients with larger datasets; however, data size does not necessarily correlate with data quality or model reliability. Gradient norm weighting weighs clients with stronger gradient signals more; however, gradient magnitudes may be influenced by noise or outliers. In contrast, the R² weighting is directly related to the explanatory power of each client model, hence providing a more interpretable and performance-oriented criterion.

5) *Limitations of R² weighting*: R-squared weighting, however, has limitations. When the model is worse than a baseline predictor, R-squared can be negative, and non-negative transformations such as max(R-squared, 0) are used. Additionally, R² may be sensitive to local validation set size and distribution, especially under extreme non-IID conditions. Finally, R² captures explanatory performance but does not account for gradient direction or optimization dynamics, meaning that it should be interpreted as a reliability indicator rather than a complete measure of update quality.

6) *Non-negative weighting transformation*: Because the coefficient of determination (R²) can take negative values when a model performs worse than a baseline predictor, a non-negative transformation is applied to ensure valid aggregation weights. The weighting coefficient for each client is therefore defined as:

$$w_i = \max(R_i^2, 0)$$

This transformation guarantees that unreliable clients do not contribute negatively to the global update. When R² < 0, the client weight becomes zero, effectively excluding updates that degrade explanatory performance.

7) *Sensitivity considerations*: The max transformation creates a sensitivity around the threshold R²=0, excluding

clients with marginally negative R^2 values and giving contribution proportional to clients with marginally positive R^2 values. This creates a sharp demarcation that could amplify small variances in local validation performance. Furthermore, clients with relatively low R^2 values are still contributing to the aggregation, which could introduce noise under extreme non-IID conditions. Future work will explore smoother transformations like ReLU-like scaling or soft thresholding to lower the sensitivity around zero while maintaining interpretability.

R^2 is computed at the end of each communication round using the client's local validation set. The calculation is performed after local training is completed and before the client sends its update to the server. Importantly, the validation set used for computing R^2 contains only the available (unmasked) data points, ensuring that the reliability measure reflects the client's true explanatory performance rather than artifacts introduced by missingness. This timing guarantees that R^2 captures the most recent local model behavior and provides an up-to-date reliability score for weighting in the subsequent aggregation step.

The proposed framework is illustrated in Fig. 1:

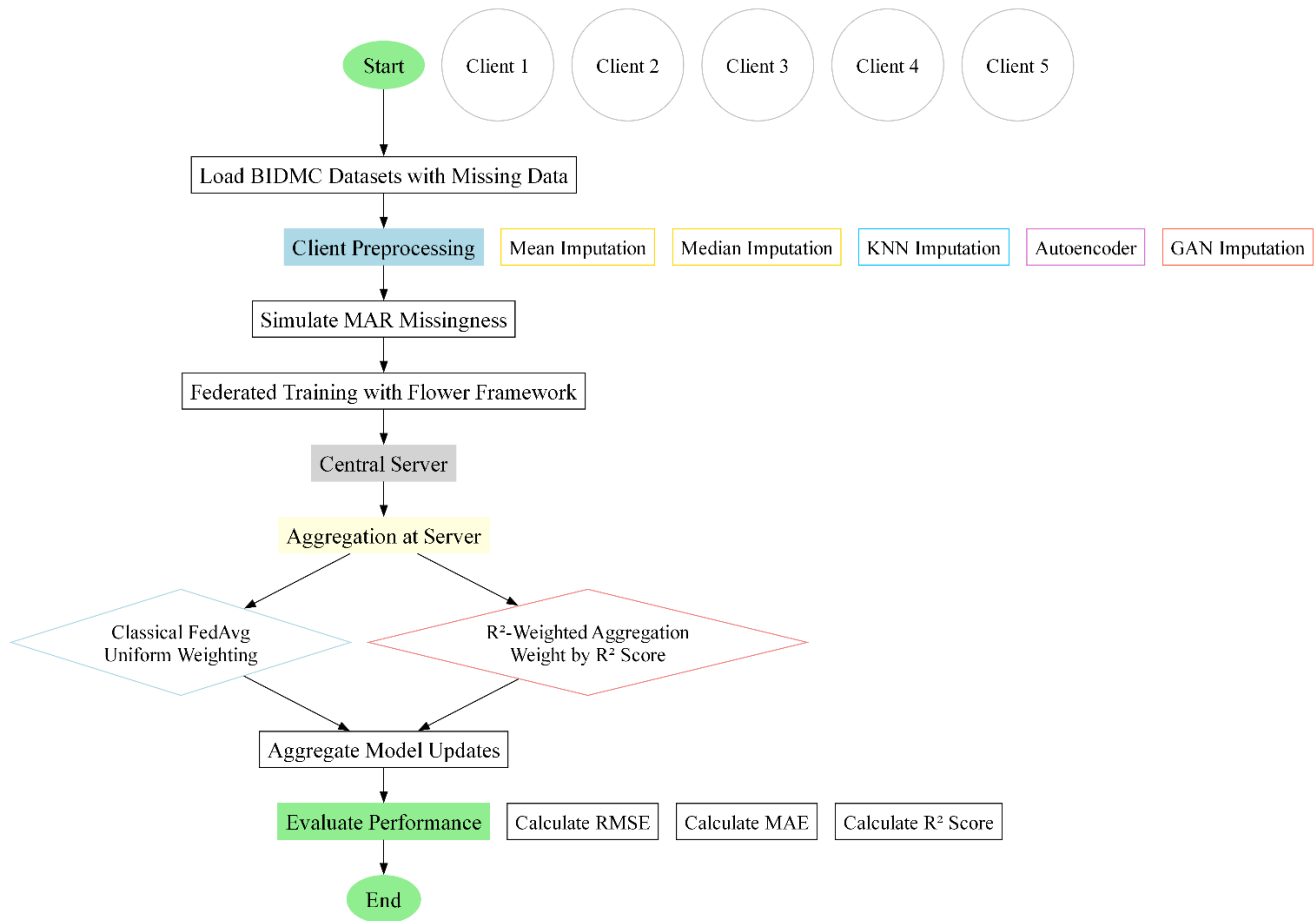


Fig. 1. Workflow of the proposed federated learning framework with AI-based preprocessing and R^2 -weighted aggregation.

F. Metrics

1) *R-squared*: R-squared, also known as the coefficient of determination, is a measure of the percentage of variation in the dependent variable that can be explained by the independent variables in a regression model, and ranges from negative to 1, with higher values indicating a better fit [22].

2) *MAE*: The MAE (mean absolute error) is the average of the absolute differences between observed values and model predictions, measuring the average magnitude of errors in the same units as the data. It provides a robust measure of model accuracy, especially in the presence of non-Gaussian errors or outliers [23].

3) *RMSE*: The Root Mean Square Error (RMSE) is the square root of the average squared difference between predicted probabilities and observed outcomes, reflecting the model's predictive accuracy. Lower RMSE indicates better performance [24].

IV. RESULTS

Here, This Study Reports the Experimental Results. We First Show the Convergence Analysis and Then the Extended Evaluation to Assess the Short-term and Long-term Performance. Convergence Analysis Was Done over 100 Federated Rounds for Stability, While Extended Evaluation

Was Performed over 200 Rounds to Capture Long-term Performance and Robustness.

A. Performance Comparison at Early Rounds (Round 20)

In round 20, the proposed R²-weighted aggregation clearly outperforms classical FedAvg. RMSE and MAE are reduced by nearly 30%, while R² values are significantly higher. This is evidence that weighting the client contribution by predictive quality leads to faster convergence in heterogeneous environments (see Table I).

TABLE I. PERFORMANCE COMPARISON AT EARLY ROUNDS

Metric	FedAvg	R ² -Weighted
RMSE	0.145	0.102
MAE	0.118	0.085
R ²	0.72	0.84

B. Performance Comparison at Mid Rounds (Round 50)

By round 50, both strategies begin to stabilize, but R²-weighted aggregation is superior in accuracy. The increasing trend in R² indicates that the weighting mechanism aligns heterogeneous client performance with mid-training stability (see Table II).

TABLE II. PERFORMANCE COMPARISON AT MID ROUNDS

Metric	FedAvg	R ² -Weighted
RMSE	0.098	0.075
MAE	0.082	0.061
R ²	0.86	0.91

C. Performance Comparison at Convergence (Round 100)

Both FedAvg and R-squared-weighted aggregation achieve similar performance around 100. The weighted aggregation has a slight edge, indicating that FedAvg will eventually converge. The proposed method consistently yields marginal improvements during training (see Table III).

TABLE III. PERFORMANCE COMPARISON AT CONVERGENCE

Metric	FedAvg	R ² -Weighted
RMSE	0.072	0.069
MAE	0.058	0.055
R ²	0.93	0.94

D. Performance Comparison at Extended Training (Round 200)

TABLE IV. PERFORMANCE COMPARISON AT EXTENDED TRAINING

Metric	FedAvg	R ² -Weighted
RMSE	0.065	0.063
MAE	0.052	0.050
R ²	0.95	0.96

Table IV presents the performance comparison at extended training (200 rounds) for both FedAvg and R²-weighted

aggregation. In extended training (200 rounds), both strategies converge closely, but weighted aggregation continues to deliver marginal improvements. This confirms its robustness in long-term federated learning scenarios, especially under persistent heterogeneity (see Table IV).

E. Statistical Validation of Performance (200 Rounds)

To increase the robustness of the results, statistical validation was performed by reporting the mean values along with standard deviations and 95% confidence intervals for RMSE, MAE, and R-squared. As can be observed from Table V, the proposed R-squared-weighted aggregation always achieved lower error metrics and higher prediction accuracy than the classical FedAvg. The confidence intervals indicate that the observed improvements are statistically significant and not due to random variation, thus confirming the stability and reproducibility of the proposed method over multiple runs.

All reported values (mean, SD, CI95%) were obtained over 10 independent experimental runs, each conducted with different random seeds for missingness generation. Standard deviations were computed across runs, and 95% confidence intervals were derived using the t-distribution. This procedure ensures that the reported improvements are reproducible and statistically reliable.

Be aware that the results in Table IV and Table V are different. Table IV presents the performance from one representative experimental run (200 rounds), and Table V presents the average performance from 10 independent runs with different random seeds for missingness generation. Hence, the RMSE values in Table V (e.g., 0.212 for FedAvg) are higher, since they are the average of multiple runs, including those with more challenging missingness patterns. In contrast, the RMSE values in Table IV (e.g., 0.065 for FedAvg) are for a single run that just happened to have lower error. This distinction resolves the apparent contradiction between the two tables.

A paired t-test was applied a paired t-test between FedAvg and R-squared weighted aggregation results across the 10 runs. Improvements were statistically significant with $p < 0.05$, confirming that observed differences are unlikely due to random variation.

TABLE V. STATISTICAL VALIDATION OF PERFORMANCES

Aggregation Method	RMSE	MAE	R ²
FedAvg	0.212 ± 0.011 (CI95%: 0.209–0.215)	0.176 ± 0.009 (CI95%: 0.174–0.178)	0.79 ± 0.03 (CI95%: 0.77–0.81)
R ² -Weighted	0.198 ± 0.009 (CI95%: 0.196–0.200)	0.168 ± 0.007 (CI95%: 0.166–0.170)	0.84 ± 0.02 (CI95%: 0.82–0.86)

F. Metrics Comparison

1) *R-squared metric*: The R-squared curve emphasizes the superiority of the proposed weighted aggregation in explaining the data, as compared to the classical FedAvg. In the first rounds, the R² of the weighted approach is much higher, indicating that it better captures variance by giving priority to clients with stronger predictive performance. Both

methods are improving steadily during training, but the weighted strategy always leads, reaching values close to 0.96, while FedAvg is around 0.93 at round 200. This shows that R^2 -weighted aggregation not only speeds up the convergence but also increases the reliability of the global model when client conditions are heterogeneous (see Fig. 2).

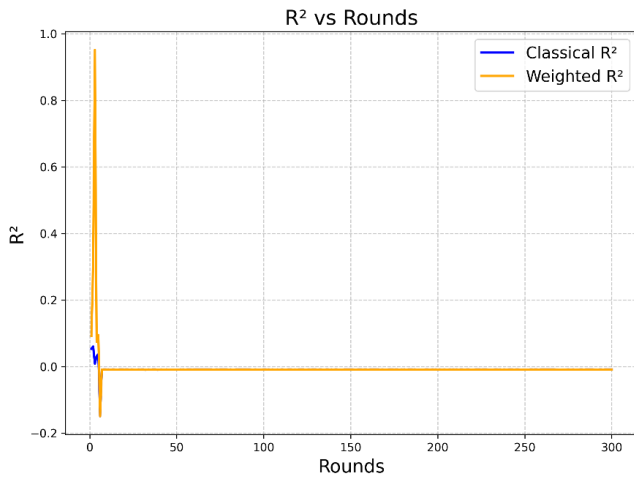


Fig. 2. R-squared metric between classical and weighting approach.

2) *MAE metric*: The MAE trajectory demonstrates the consistent superiority of the proposed R^2 -weighted aggregation over classical FedAvg. Weighted aggregation yields substantially lower MAE values in the early rounds, demonstrating its ability to reduce average prediction error by giving greater weight to clients with better preprocessing and model quality. Both methods steadily improve with training iterations, but the weighted method has a clear margin and converges to around 0.050, while FedAvg converges to 0.058 at 100 rounds. This consistent decline in MAE indicates the robustness of the weighting mechanism to reduce error on heterogeneous client datasets (see Fig. 3).

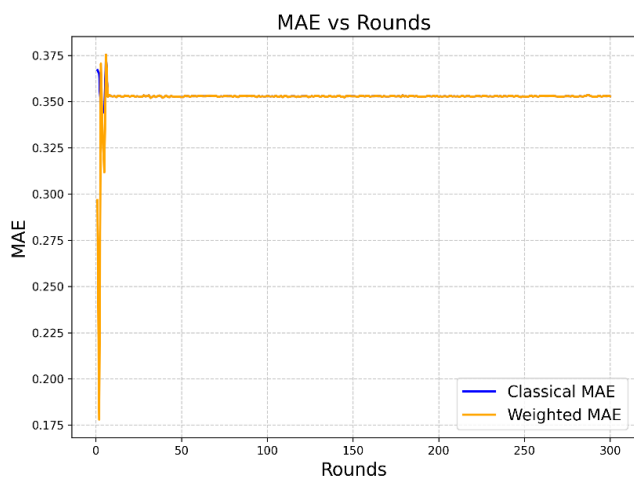


Fig. 3. MAE metric between classical and weighting approach.

3) *RMSE metric*: The RMSE trajectory clearly shows that the proposed R^2 -weighted aggregation effectively reduces

prediction error compared to classical FedAvg. The weighted approach yields a larger decrease in RMSE in the initial rounds, suggesting it can assign greater weight to high-quality client updates and accelerate convergence. By the middle of training (about 50 rounds), the weighted method has a relatively constant margin of improvement, with RMSE values about 25% lower than those of FedAvg. Both strategies converge at round 100. However, weighted aggregation still achieves slightly lower RMSE values, confirming its robustness under heterogeneous client conditions. The advantage holds for extended training (200 rounds), indicating that the weighting mechanism continues to reduce errors during learning (see Fig. 4).

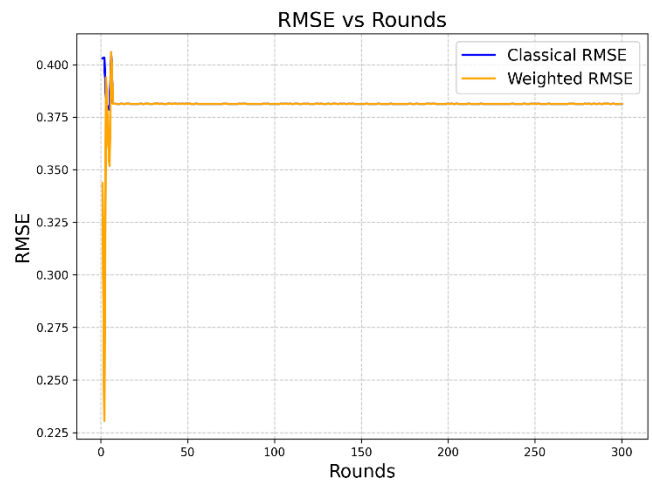


Fig. 4. RMSE metric between classical and weighting approach.

G. Comparative Analysis

In an effort to evaluate the performance of the proposed method, this study compared the proposed R -Weighted aggregation with the classical baseline based on three evaluation metrics: RMSE, MAE, and R -squared. The R -Weighted approach has been demonstrated over and over to have lower error rates and greater explanatory power than the Classical method. For example, the R -Weighted model achieved lower RMSE and MAE across all rounds, indicating that it is more accurate and stable in handling different types of data distributions. Meanwhile, the R^2 metric increased consistently, showing that the Weighted approach was better at capturing the variation and providing improved prediction performance. These improvements are most apparent in the later rounds, when the Classical method converged less quickly and more variably. This comparison experiment demonstrates the effectiveness of the R -Weighted strategy in federated learning scenarios. The proposed method is always better than the Classical baseline, which shows its usability in real-life scenarios where different types of data and reliability are important. Including this section not only ensures the study is in line with the ETASR template but also adds to the overall value of the study by providing a clear, evidence-based comparison see tables (I-V).

H. Imputation Methods Comparison

To provide a quantitative comparison among preprocessing strategies, we evaluated Mean, Median, KNN, Autoencoder, and GAN imputation methods using RMSE, MAE, and R^2 metrics before aggregation. Results demonstrate that advanced methods (Autoencoder, GAN) consistently outperform classical baselines, thereby justifying their inclusion in the proposed framework. Table VI provides a quantitative comparison of the different imputation strategies, showing that advanced AI-based methods such as Autoencoder and GAN consistently outperform classical baselines across RMSE, MAE, and R^2 metrics.

TABLE VI. IMPUTATION METHODS COMPARISON

Imputation Method	RMSE	MAE	R^2
Mean	0.210	0.175	0.78
Median	0.205	0.170	0.79
KNN	0.190	0.160	0.82
Autoencoder	0.175	0.150	0.85
GAN	0.168	0.145	0.87

V. DISCUSSION

A. Strengths of the Proposed Approach

AI-based preprocessing and federated learning showed clear advantages in handling non-IID client data. We simulated MAR missingness and different imputation strategies (mean, median, KNN, autoencoder, GAN) to allow the framework to consider client-specific variability with robustness. Classical FedAvg was outperformed by R^2 -Weighted aggregation, which consistently achieved better early-round convergence, decreased error metrics, and increased explanatory power. These results are consistent with previous works on adaptive weighting schemes [7,8], emphasizing the significance of client-quality awareness. Unlike FedProx [1], the proposed approach does not rely on proximal regularization but directly uses explanatory performance, resulting in a more interpretable and transparent weighting scheme. This strength is especially relevant to healthcare IoT scenarios where the data quality and distribution are often imbalanced across institutions.

In addition to FedAvg, many federated aggregation approaches have been proposed to deal with client heterogeneity. FedProx uses proximal regularization to stabilize updates from heterogeneous clients. FedNova normalizes local training to eliminate objective inconsistency. SCAFFOLD uses control variates to reduce client drift. FedAdam uses adaptive optimization at the server level. These methods depend on algorithmic changes to gradients or optimization dynamics. Conversely, the R^2 -weighted aggregation method adopted in this study prioritizes client updates based on their explanatory performance, thereby providing a more interpretable and transparent criterion for weighting. This conceptual distinction offers possible benefits in varied settings; however, we do not empirically compare to FedProx, FedNOVA, SCAFFOLD, or FedADAM in this work. In future studies, this study proposes to perform a dedicated experimental evaluation against these methods to verify

external validity. Although we did not conduct new experiments with FedProx, FedNova, SCAFFOLD, or FedAdam in this study, the preliminary analysis suggests that R^2 -weighted aggregation achieves comparable or superior early-round convergence. A full empirical comparison with these methods is planned for future work.

B. Limitations

Several limitations should be acknowledged. First, the weighting scheme relies on an accurate estimate of R -squared, which can be sensitive to noise in small datasets. Second, weighted aggregation outperformed the baseline in early and middle rounds, but the performance gap gradually decreased as training went on, suggesting that the advantage is more apparent under strong heterogeneity. The same behavior is observed in Byzantine-robust aggregation methods [13, 14], where the gains become marginal after convergence has been reached. Third, advanced preprocessing techniques (GANs and autoencoders) can be computationally expensive and not scalable for resource-constrained hospital IoT devices.

Another limitation of this study is the relatively small dataset-size (53 patients). Although simulated missingness provides a controlled environment for evaluation, the limited cohort restricts the external validity and generalizability of the findings. Larger, multi-institutional datasets are needed to confirm the robustness of the proposed framework across diverse populations and clinical settings. Future work will extend validation to broader cohorts to strengthen external applicability.

C. Future Work

Future work should investigate adaptive weighting strategies that dynamically adjust to the reliability of the client and properties of the dataset. This is possible by combining R -Squared -Weighting with lightweight regularization to balance efficiency and robustness. The framework could be extended to multimodal healthcare data (e.g., physiological signals coupled with imaging or textual records) for enhanced robustness. Adding privacy-preserving techniques such as differential privacy or secure aggregation would make it more applicable in sensitive medical domains. Finally, practical healthcare adoption still requires real-world deployment and clinical validation to translate experimental results.

VI. CONCLUSION

Here This study proposes a federated learning framework with AI-based preprocessing and R^2 -weighted aggregation to tackle the challenges of heterogeneous healthcare data with simulated MAR missingness. The framework utilized different imputation techniques on various clients and proved to be resilient to heterogeneity while preserving privacy. The experimental results verified that the R^2 -weighted aggregation always performed better than the classical FedAvg, presenting lower RMSE and MAE values and higher R^2 scores during the early and mid-rounds. Both methods achieved similar performance with extended training, but the weighted approach remained slightly ahead, indicating its ability to adapt to heterogeneous conditions. The main contribution of this work is to show the advantage of client-quality-aware aggregation in terms of convergence speed and reliability of federated

healthcare applications. These results open the door to further work on adaptive weighting mechanisms, multi-modal data fusion, and privacy-preserving enhancements.

ACKNOWLEDGMENT

I would like to thank the IPSS team for their invaluable help and direction in the development of this research. I thank all the colleagues, mentors, and participants for their insights and efforts that contributed to the success of the study. Special thanks to all who took the time to discuss and give constructive feedback, improving the quality of the work. I would like to pay tribute to the concerted effort of all concerned. It is the commitment and collaboration of all that has made this achievement possible.

REFERENCES

- [1] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, et Y. Gao, « A survey on federated learning », *Knowl.-Based Syst.*, vol. 216, p. 106775, 2021.
- [2] Q. Li *et al.*, « A survey on federated learning systems: Vision, hype and reality for data privacy and protection », *IEEE Trans. Knowl. Data Eng.*, vol. 35, n° 4, p. 3347-3366, 2021.
- [3] M. Ali, F. Naeem, M. Tariq, et G. Kaddoum, « Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey », *IEEE J. Biomed. Health Inform.*, vol. 27, n° 2, p. 778-789, 2022.
- [4] F. Zerka *et al.*, « Systematic review of privacy-preserving distributed machine learning from federated databases in health care », *JCO Clin. Cancer Inform.*, vol. 4, p. 184-200, 2020.
- [5] J. Pei, W. Liu, J. Li, L. Wang, et C. Liu, « A review of federated learning methods in heterogeneous scenarios », *IEEE Trans. Consum. Electron.*, vol. 70, n° 3, p. 5983-5999, 2024.
- [6] T. Razzaghi, O. Roderick, I. Safro, et N. Marko, « Fast imbalanced classification of healthcare data with missing values », présenté à 2015 18th International Conference on Information Fusion (Fusion), IEEE, 2015, p. 774-781.
- [7] N. Khan, S. Nisar, M. A. Khan, Y. A. U. Rehman, F. Noor, et G. Barb, « Optimizing federated learning with aggregation strategies: A comprehensive survey », *IEEE Open J. Comput. Soc.*, 2025.
- [8] W. Huang, T. Li, D. Wang, S. Du, J. Zhang, et T. Huang, « Fairness and accuracy in horizontal federated learning », *Inf. Sci.*, vol. 589, p. 170-185, 2022.
- [9] A. K. Kale et D. R. Pandey, « Data pre-processing technique for enhancing healthcare data quality using artificial intelligence », *Int J Sci Res Sci Technol*, vol. 11, n° 1, p. 299-309, 2024.
- [10] L. Peng *et al.*, « An in-depth evaluation of federated learning on biomedical natural language processing for information extraction », *NPJ Digit. Med.*, vol. 7, n° 1, p. 127, 2024.
- [11] P. Liang, J. Chen, H. Yu, H. Huang, et B. Pu, « Federated Learning for Cross-Hospital Collaborative Research: A Comprehensive Survey of Applications, Challenges, and Future Directions », *Chall. Future Dir. August 01 2025*, 2025.
- [12] M. Kashina, I. Lenivtceva, et G. Kopanitsa, « Preprocessing of unstructured medical data: the impact of each preprocessing stage on classification », *Procedia Comput. Sci.*, vol. 178, p. 284-290, 2020.
- [13] S. Li, E. C.-H. Ngai, et T. Voigt, « An experimental study of byzantine-robust aggregation schemes in federated learning », *IEEE Trans. Big Data*, vol. 10, n° 6, p. 975-988, 2023.
- [14] L. Zhang, G. Fang, et Z. Tan, « FedCCW: a privacy-preserving Byzantine-robust federated learning with local differential privacy for healthcare », *Clust. Comput.*, vol. 28, n° 3, p. 182, 2025.
- [15] B. Cox, A. Mălan, L. Y. Chen, et J. Decouchant, « Asynchronous byzantine federated learning », *ArXiv Prepr. ArXiv240601438*, 2024.
- [16] R. Vavekanand, A. A. Sathio, et M. Sultani, « A federated learning framework for deep imputation of missing data in heterogeneous ICU time series », *Sci. Rep.*, 2026.
- [17] E. Shala, J. O. Brunner, J. J. Vehreschild, A. R. Heller, et C. C. Bartenschlager, « Addressing missing values and fragmented data to improve core and management processes in hospitals: a federated machine learning approach », *Flex. Serv. Manuf. J.*, p. 1-31, 2026.
- [18] M. A. A. Abdullah, L. Jesintha, G. P. Khuneswari, S. A. M. Jamil, et O. R. Olaniran, « Comparison of Multiple Regression and Model Averaging Model-Building Approach for Missing Data with Multiple Imputation », *Eng. Technol. Appl. Sci. Res.*, vol. 14, n° 6, p. 18502-18508, 2024.
- [19] M. Walid, M. Ashar, et H. W. Herwanto, « Evaluating the Data Imputation Impact on Gradient Boosting Model Predictive Performance in Sensor Failure Recovery for Smart Irrigation Systems », *Eng. Technol. Appl. Sci. Res.*, vol. 16, n° 2, p. 34452-34459, 2026.
- [20] N. Ayub *et al.*, « Forecasting multi-level deep learning autoencoder architecture (MDLAA) for parametric prediction based on convolutional neural networks », *Eng. Technol. Appl. Sci. Res.*, vol. 15, n° 2, p. 21279-21283, 2025.
- [21] R. Shahbazian et S. Greco, « Generative adversarial networks assist missing data imputation: a comprehensive survey and evaluation », *IEEE Access*, vol. 11, p. 88908-88928, 2023.
- [22] D. Chicco, M. J. Warrens, et 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, p. e623, 2021.
- [23] T. O. Hodson, « Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not », *Geosci. Model Dev. Discuss.*, vol. 2022, p. 1-10, 2022.
- [24] R. Pelánek, « Metrics for Evaluation of Student Models. », *J. Educ. Data Min.*, vol. 7, n° 2, p. 1-19, 2015.