FedRepu: A Reputation-Based Aggregation Strategy for Federated Learning in Bearing Fault Diagnosis

Junxian You, *Graduate Student Member, IEEE*, Rui Yang, *Senior Member, IEEE*, Zidong Wang, *Fellow, IEEE*

***Abstract*—Integrating federated learning in bearing fault diagnosis marks a crucial step in industrial operations, showing great potential for enhancing diagnostic accuracy and data privacy. However, the effectiveness of federated learning is often compromised by varying data quality and quantity among clients. FedRepu, a novel reputation-based aggregation strategy for federated learning in bearing fault diagnosis, is introduced in this paper by employing an extended reputation mechanism that scores client trustworthiness based on real-time, historical, and rewarding performance metrics, dynamically adapting model aggregation to quality of contributions. It deals with natural heterogeneity in distributed sources and improves performance uniformly across clients. Comprehensive experiments on the CWRU, PU and HUST datasets, including comparison studies of cross-validation tasks and ablation studies, confirm FedRepu's effectiveness in improving model aggregation efficiency despite reputation discrepancies. Sensitivity analysis further demonstrates its robustness to hyperparameter variations, making FedRepu a promising solution for federated learning in bearing fault diagnosis.**

***Index Terms*—Federated learning, bearing fault diagnosis, reputation-based aggregation strategy, reward-based incentive.**

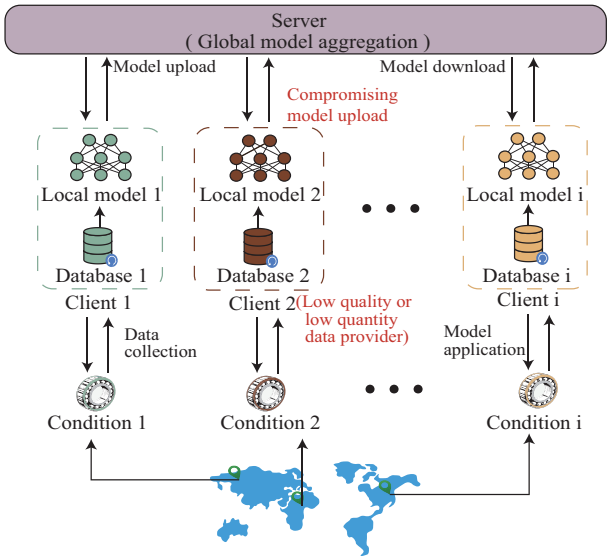


Fig. 1. Basic process of federated learning.

I. Introduction

DEEP learning technologies have significantly advanced bearing fault diagnosis by effectively extracting complex features from vibration signals [1]–[3]. These technologies excel at learning hidden features from raw data and achieve diagnostic accuracies that surpass conventional fault diagnosis methods [4]–[6]. Recent studies have further demonstrated that integrating graph-based representations and domain-constrained learning can effectively enhance the robustness and generalization of diagnostic models across complex industrial scenarios [7], [8]. However, the inherently distributed nature of modern industrial operations inevitably creates isolated "data silos." Models trained exclusively on a single silo are prone to overfitting and often fail to generalize across different working conditions. Therefore, it is imperative to develop a generalized diagnostic model that integrates the collective knowledge from all disparate data silos [9].

Developing a highly generalized diagnostic model requires massive datasets, yet strict data privacy regulations (e.g., GDPR) and the risks of sensitive information exposure prohibit centralized data gathering [10]–[13]. Federated learning offers a promising solution to these challenges by enabling collaborative training through the sharing of model parameters rather than raw data [14]–[16]. In recent years, significant advancements have been made in the field of distributed fault diagnosis [17] and federated learning that emphasizes data privacy [18], [19]. However, as the basic process shown in Fig. 1, training data typically comes from multiple distributed clients, which may belong to different factories, regions, or operational environments with their own data collection and management methods. Such differences can create imbalances in both the quantity and quality of fault data between clients, posing a major challenge for aggregation strategies in federated learning for fault diagnosis [20].

Specifically, some clients may lack advanced data acquisition systems, leading to infrequent or low-quality data collection [21]. Additionally, various operational loads and machine usage patterns can also affect the amount of fault data collected. For instance, clients with frequently used equipment tend to generate larger and more comprehensive datasets compared to intermittently used machinery [22]. Furthermore,

This work is supported by the National Natural Science Foundation of China (62233012) and the Jiangsu Provincial Scientific Research Center of Applied Mathematics (BK20233002). The authors gratefully acknowledge the support from the Jiangsu Provincial Government, the Suzhou Municipal Government, and Xi'an Jiaotong-Liverpool University.

J. You is with the James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, United Kingdom (e-mail: 3163509Y@student.gla.ac.uk);

R. Yang is with School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China (e-mail: R.Yang@xjtlu.edu.cn);

Z. Wang is with the Department of Computer Science, Brunel University London, Uxbridge, Middlesex UB8 3PH, United Kingdom (email: Zidong.Wang@brunel.ac.uk).

Corresponding author: R. Yang.

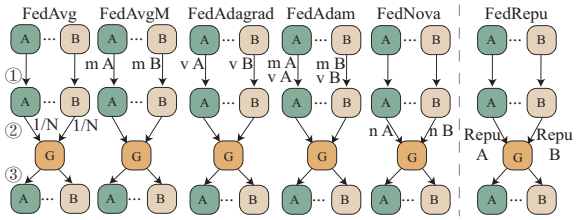


Fig. 2. Comparison of aggregation strategies in federated learning. The figure illustrates five relative strategies handling N local models (such as A & B) of distributed clients, concerning components of the local training phase (①), aggregation mechanisms (②), and global model (G) distribution process (③) across different strategies.

varying maintenance strategies and schedules may also cause inconsistencies in data quality, as some clients may perform proactive maintenance, while others only maintain equipment after a failure occurs. These factors lead to significant differences in both the quantity and quality of bearing fault diagnosis data across clients, affecting the aggregated model performance in federated learning [23].

As demonstrated in Fig. 2, the first five methods represent typical aggregation strategies. These strategies differ in how they handle client updates, global aggregation, and optimization, as detailed below. FedAvg [24] updates parameters in ①, forms a global model (G) by averaging aggregation in ②, and the global model is then distributed back in ③ to start a new iteration. FedAvgM [25] incorporates momentum (m) in ① to improve training stability by considering historical gradients in the current update direction. FedAdagrad [25] includes the sum of squared gradients as the accumulated second-order momentum (v) in ① to adjust learning rates. FedAdam [25] combines momentum and adaptive learning rates, calculating weighted averages of first-order and second-order momentum (m and v) in ①. FedNova [26] focuses on normalizing updates in ② with the number of local training iterations (n). However, these aggregation strategies do not pay ample attention to discrepancies of data quality and quantity between clients.

This limitation arises from a common assumption in these strategies that all clients contribute equally during the aggregation step, resulting in instability in the aggregation process and suboptimal global models [27]. Moreover, many aggregation strategies fail to effectively handle partial client overfitting, where the aggregated model may prioritize clients with large data volumes or with features easy to fit, neglecting the insights from other clients. Therefore, improved aggregation strategies are needed to address the challenges of imbalance in data quality and quantity across clients, enhancing the stability and performance of the aggregated global model [28], [29].

While some recent studies have made significant progress in addressing data imbalance, such as prototype-based contrastive adversarial networks [30] and adversarial transformers [31], these approaches are primarily developed for centralized settings or treat client updates uniformly during aggregation. In federated environments, a solution to avoiding noisy or unreliable updates from dominating the global model should be exploited by considering imbalance not only within each client but also across clients during aggregation.

Fortunately, the reputation-based evaluation mechanisms have attracted increasing attention, primarily using blockchain or smart contracts to transparently record client reputation scores [32]–[34]. The existing reputation-based approaches typically consider simplistic reputation metrics, such as overall accuracy, loss, or model similarity, and are misaligned with safety-critical fault diagnosis, where misclassification costs are asymmetric. Missed alarms are typically far more detrimental than false alarms. Under such conditions, accuracy-based reputation can overestimate clients that achieve high overall accuracy by favoring the majority classes. This gap motivates a cost-aware reputation design that explicitly incorporates imbalance-sensitive measures, enabling aggregation to weight clients by their safety-relevant performance rather than size or accuracy alone.

To better address the imbalance in client data, this paper proposes FedRepu, a novel reputation-based aggregation strategy for federated learning in the field of bearing fault diagnosis, as shown in the final aggregation strategy in Fig. 2. Specifically, FedRepu dynamically assesses the performance of each client and moves beyond the common reputation strategy of simple metric-based weighting by introducing a cost-aware reputation mechanism tailored for safety-critical industrial applications. This helps address the imbalance in data quality and quantity across clients and mitigates the impact of potentially compromising client updates (including faults, outliers, and inconsistent uploads). Additionally, by leveraging cost-aware evaluation, FedRepu addresses the challenge of asymmetric misclassification costs in its aggregation strategies.

The main contributions of this paper are summarized as follows:

1) The concept of reputation is extended within bearing fault diagnosis, incorporating F1-score, false-negative rate, and false-positive rate, addressing the shortcomings of existing reputation approaches under severe imbalance and asymmetric misclassification costs.
2) By incorporating historical reputation factors, FedRepu addresses the issue of forgetfulness during model training, thereby ensuring the stability of the training process.
3) The inclusion of reward factors in the reputation computation enables rapid model convergence to optimal states and enhances the self-balancing capabilities of aggregated models.

The paper is structured as follows: Section II provides a detailed explanation of the proposed FedRepu strategy; Section III presents experimental results of the cross-validation comparison, ablation studies, and sensitivity analysis that underscore the effectiveness of FedRepu across heterogeneous datasets; Section IV concludes the findings and outlines potential directions for future research.

II. Proposed Reputation-based Federated Learning Aggregation Strategy

This section proposes a novel reputation-based federated learning aggregation strategy for fault diagnosis to address the aforementioned challenges. The detailed process of the proposed FedRepu with related symbols is presented in Fig. 3 and Table I. FedRepu operates in two main stages: computation

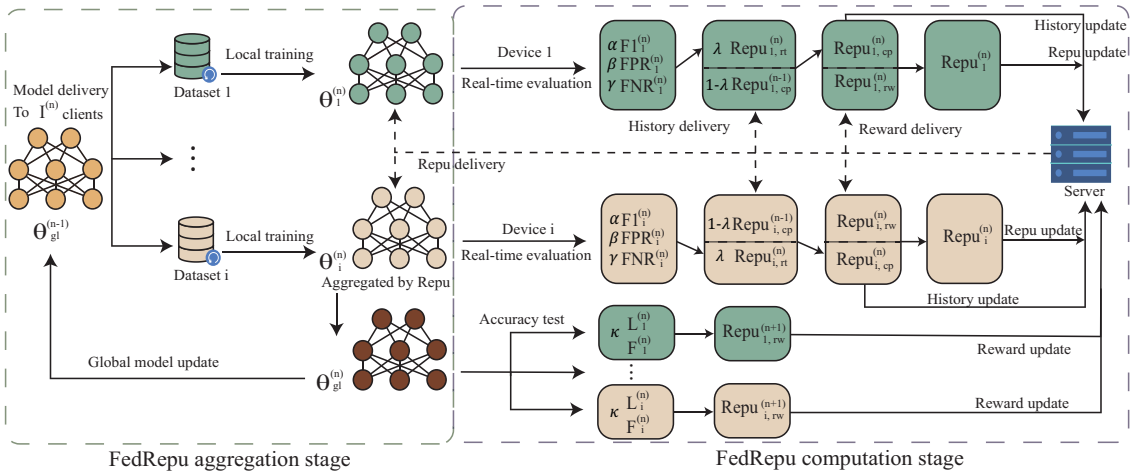


Fig. 3. Detailed process of proposed FedRepu aggregation strategy.

stage and aggregation stage. For each training iteration n from 1 to N, the server sends the current global model parameters to all clients. Each client i trains its local model on local data to obtain updated parameters. The server undergoes a complete FedRepu computation stage, including real-time reputation evaluation (subsection A), historical reputation combination (subsection B), and the reward-based incentive (subsection C). In the FedRepu aggregation stage (subsection D), the global model is updated as a weighted sum of all local models.

A. Reconstruction of Real-Time Reputation Evaluation

Model performance fluctuations in federated learning environments can manifest as varying diagnostic performance across different clients, mainly caused by discrepancies in client data quality and quantity. This reconstructed real-time reputation evaluation aims to grant a high reputation to local models with fewer model performance fluctuations. Generally, FedRepu focuses on reducing model performance fluctuations through continuous iterations, addressing imbalance and misclassification issues. To address the imbalance issue, the $F1$ measure is utilized due to its inherent sensitivity to imbalanced datasets [35]:

$$F1 = \frac{2 \cdot Prec \cdot Rec}{Prec + Rec} = \frac{2P(D|F)}{2P(D|F) + P(D|\neg F) + P(\neg D|F)}, \tag{1}$$

where $P(\cdot)$ denotes the occurrence probability of an event, D and $\neg D$ denote fault diagnosis and non-diagnosis, while F and $\neg F$ denote fault existence and non-existence, respectively.

In addition to evaluating imbalance, real-time evaluation considers model performance fluctuations caused by misclassification. False alarm (corresponding to false positive condition) can be costly in many industrial applications, leading to unnecessary checks or downtime. Conversely, missed alarm (corresponding to false negative condition) can be even more harmful, as faults are left unaddressed and eventually lead to catastrophic equipment failures and severe safety risks. To reduce the cost of misclassification, the concepts of false positive rate (FPR) and false negative rate (FNR) are considered in real-time reputation. FPR and FNR are defined as follows in fault diagnosis:

TABLE I
DEFINITIONS OF SYMBOLS

Symbol	Description
$\theta_i^{(n)}$	Model parameters of client i at iteration n
$\theta_{gl}^{(n)}$	Model parameters of global model at iteration n
$Repu_i^{(n)}$	Full reputation of client i at iteration n
$Repu_{i,rt}^{(n)}$	Real-time reputation of client i at iteration n
$Repu_{i,cp}^{(n)}$	Composite reputation of client i at iteration n
$Repu_{i,rw}^{(n)}$	Reputation reward of client i at iteration n
$F1_i^{(n)}$	F1 score of client i at iteration n
$FPR_i^{(n)}$	False positive rate of client i at iteration n
$FNR_i^{(n)}$	False negative rate of client i at iteration n
$\alpha, \beta, \gamma, \lambda, \kappa$	Weighting coefficients
$L_i^{(n)}$	Local model accuracy tested in client i at iteration n
$F_i^{(n)}$	Fused model accuracy tested in client i at iteration n
n_i	Order of first global iteration where client i takes part
$\overline{Repu_{cp}}^{(n)}$	Average of composite reputations at iteration n
$I^{(n)}$	Number of clients that participate at iteration n
$\phi(\cdot)$	Positive mapping function

$$FPR = P(D|\neg F), \quad FNR = P(\neg D|F). \tag{2}$$

By measuring model performance from imbalance and misclassification using (1) and (2), local models can optimize these metrics to become valuable contributors to the aggregated global model. To ensure the reputation score encapsulates the considerations of model performance fluctuations from imbalance and misclassification, the real-time reputation is reconstructed as follows:

$$Repu_{i,rt}^{(n)} = \alpha F1_i^{(n)} + \beta(1 - FPR_i^{(n)}) + \gamma(1 - FNR_i^{(n)}), \tag{3}$$

where real-time reputation score of i_{th} client during the n_{th} iteration is associated with $F1$, FPR, and FNR. The parameters α, β, $\gamma \in [0, 1]$ allow for fine-tuned importance weighting to emphasize different aspects of current performance.

From the implementation perspective, the three metrics $F1_i^{(n)}$, $FPR_i^{(n)}$, and $FNR_i^{(n)}$ are all computed from the same batch of the local validation that each client has to perform. Therefore, no additional forward/backward passes are introduced on the client side, and the impact on the overall training efficiency is negligible. A detailed explanation of the

misclassification part of real-time reputation score is shown in Appendix A. Besides, Appendix B provides parameter factorisation of the real–time reputation function and the reason for three independent weights.

B. Combination of Historical Reputation Evaluation

Addressing the issue of forgetfulness in federated learning is crucial for maintaining a robust model. To tackle the forgetfulness issue, historical reputation is introduced in this study. Particularly, the historical reputation of a client is combined with the metrics acquired through real-time reputation evaluation to form a dual evaluation of client contributions.

For each client i, let n_i denote the first global iteration where client i takes part. When iteration $n < n_i$, the client i has no track record, and the composite reputation combining the historical reputation of client i is given as $Repu_{i,cp}^{(n)} = 0$. For $n \geq n_i$, consider the average of composite reputations during the previous iteration as:

$$\overline{Repu}_{cp}^{(n-1)} = \frac{1}{I^{(n-1)}} \sum_{j=1}^{I^{(n-1)}} Repu_{j,cp}^{(n-1)}, \tag{4}$$

where $\overline{Repu}_{cp}^{(0)} = 0$ and $I^{(n-1)}$ is the number of clients that participate in iteration $n-1$. $\overline{Repu}_{cp}^{(n-1)}$ will serve as the warm-start baseline for newcomers.

Accordingly, $Repu_{i,cp}^{(n)}$ is updated by:

$$\begin{aligned} Repu_{i,cp}^{(n)} &= \lambda\, Repu_{i,rt}^{(n)} \\ &\quad + (1-\lambda)\big(1-\delta_{n,n_i}\big)\, Repu_{i,cp}^{(n-1)} \\ &\quad + (1-\lambda)\delta_{n,n_i}\, \overline{Repu}_{cp}^{(n_i-1)} \\ &= \lambda \sum_{t=n_i}^{n} (1-\lambda)^{n-t}\, Repu_{i,rt}^{(t)} \\ &\quad + (1-\lambda)^{n-n_i+1}\, \overline{Repu}_{cp}^{(n_i-1)}, \end{aligned} \tag{5}$$

where λ $(0 \leq \lambda \leq 1)$ is a hyperparameter controlling the weight between current and historical reputations and $\delta_{a,b} = 1$ if $a = b$ and 0 otherwise. The first term represents the real-time reputation obtained in the current iteration; the second, active only when $n > n_i$, carries over the previous composite reputation with an exponential decay factor $1-\lambda$; the third, triggered exactly once when $n = n_i$, provides a warm-start by assigning the newcomer the preceding global average instead of an artificially low initial value. In subsequent iterations, this warm-start contribution is weakened by $(1-\lambda)^{n-n_i+1}$ and quickly becomes negligible. By expanding the recursive relationship, the composite reputation is expressed as a weighted sum of all past real-time reputations up to the current iteration, plus a transient warm-start offset. Consequently, a new client delivering consistently good results converges to a high reputation without being unfairly penalised in its initial iteration, fully addressing the potential cold-start concern.

Remark 1. *The mechanism enables the incorporation of past client contributions into the final decision, avoiding the burden of loading the global model with irrelevant memory. This approach can help obtain a consistent reputation, since rare data presented in a few iterations will not be considered as defining factors. The consistent reputation provides an effective solution to forgetfulness in fault diagnosis. Additionally, the server can make specific decisions by identifying underperforming local models, allowing the continuous update of global model to handle new fault types.*

C. Integration of Reward-Based Incentive

The purpose of integration of reward-based incentives is to improve the global model's ability to self-balance when handling a variety of client datasets. Within the framework of FedRepu, the term $Repu_{i,rw}^{(n)}$ is defined as:

$$Repu_{i,rw}^{(n)} = \kappa(L_i^{(n-1)} - F_i^{(n-1)}), \tag{6}$$

where the computational process is outlined as follows:

1) Local Model Accuracy Computation: At the end of the $(n-1)_{th}$ iteration, the accuracy of the locally trained model for each client i, $L_i^{(n-1)}$, is evaluated using the corresponding validation set, reflecting the local model's ability to predict new data points based on its knowledge of a specific dataset.
2) Global Model Accuracy Computation: When evaluating the aggregated global model on each client's local validation set, the accuracy $F_i^{(n-1)}$ provides insight into the performance of global model across diverse data sources when applied to local individual datasets.
3) Reputation Adjustment Computation: The reward-based reputation, $Repu_{i,rw}^{(n)}$, is computed by applying the scaling factor κ to the difference between the local model accuracy and the global model accuracy, indicating the difference between the client's model and the global standard.

FedRepu introduces this reward term into the global score to comprehensively consider incentives, with the strategy expressed as:

$$Repu_i^{(n)} = Repu_{i,cp}^{(n)} + Repu_{i,rw}^{(n)}. \tag{7}$$

This incentive structure encourages participants to contribute high-quality updates. Additionally, rewarding clients for exceeding the accuracy of the global model prevents the model from relying solely on the most common patterns, thus addressing the challenges of skewed or non-representative local datasets.

Remark 2. *By incentivizing localized improvements beyond the global model, the reward system encourages individual excellence and enhances the overall system's robustness. The integrated model contributes to a robust fault diagnosis network, driven by the clients' motivation to expand their contributions.*

D. Weighted Aggregation of Local Models

Consider a network of $I^{(n)}$ clients, each with a local diagnostic dataset D_i, where i ranges from 1 to $I^{(n)}$. The objective is to collaboratively train a global diagnostic model with parameters $\theta_{gl}^{(N)}$. During each iteration, each client updates its local model based on its dataset using the following rule:

$$\theta_i^{(n)} = \theta_i^{(n-1)} - \eta \nabla L_i(\theta_i^{(n-1)}), \tag{8}$$

where η is the learning rate, and $\nabla L_i(\theta_i^{(n-1)})$ denotes the gradient of the loss function for the model parameters $\theta_i^{(n-1)}$ at client i and iteration $n-1$. The formulation of the aggregated model parameters with reputation weighting can be expressed as follows:

$$\theta_{gl}^{(n)} = \sum_{i=1}^{I^{(n)}} \left(\frac{\phi\big(Repu_i^{(n)}\big)}{\sum_{j=1}^{I^{(n)}} \phi\big(Repu_j^{(n)}\big)} \right) \cdot \theta_i^{(n)}. \tag{9}$$

The core of the aggregation strategy is the computation of a weighted average of client models, where each model's mapped weight is determined by the updated reputation score, $Repu_i^{(n)}$. $\phi(\cdot)$ denotes a positive mapping that converts reputation scores into non-negative aggregation scores before normalization, thereby ensuring valid aggregation weights while preserving the relative ordering of client reputations. Specifically, the global model parameters $\theta_{gl}^{(n)}$ are updated as a weighted sum of all local models $\theta_i^{(n)}$, with weights derived from their respective mapped reputations. Following the local model aggregation, the updated global model $\theta_{gl}^{(n)}$ undergoes evaluation to ascertain its performance across the distributed network. Algorithm 1 provides the pseudocode of the proposed FedRepu aggregation strategy.

From a computational viewpoint, the reputation-based aggregation in (9) has the same asymptotic cost as the standard aggregation in FedAvg. Let $I^{(n)}$ be the number of participating clients and d the number of model parameters; the server needs $\mathcal{O}(I^{(n)} d)$ operations to scale and sum the local models. The extra reputation-related operations only involve combining a constant number of scalars for each client, which is $\mathcal{O}(I^{(n)})$ and therefore dominated by the aggregation of model parameters. As a result, the overall aggregation efficiency is preserved.

Algorithm 1 FedRepu Aggregation Strategy

Input: Number of iterations N, number of clients $I^{(n)}$, initial global model parameters $\theta_{gl}^{(0)}$
Output: Updated final model parameters $\theta_{gl}^{(N)}$

for $n = 1$ **to** N **do**
  Server sends $\theta_{gl}^{(n-1)}$ to all clients
  for $i = 1$ **to** $I^{(n)}$ **do**
    Client i initializes local model $\theta_i^{(n-1)}$ with $\theta_{gl}^{(n-1)}$
    Client i trains $\theta_i^{(n-1)}$ on local data to obtain $\theta_i^{(n)}$
    Client i sends updated model $\theta_i^{(n)}$ back to server
    Client i computes real-time metrics $F1_i^{(n)}$, $FPR_i^{(n)}$, and $FNR_i^{(n)}$ in (1) and (2)
    Client i sends real-time metrics back to server
    Server computes $Repu_{i,rt}^{(n)}$ in (3) with real-time metrics
    Server computes $\overline{Repu}_{cp}^{(n-1)}$ in (4) with previous average composite reputations
    Server computes $Repu_{i,cp}^{(n)}$ in (5) with $Repu_{i,rt}^{(n)}$ and $Repu_{i,cp}^{(n-1)}$
    Server computes $Repu_{i,rw}^{(n)}$ in (6) with $L_i^{(n-1)}$ and $F_i^{(n-1)}$
    Server computes $Repu_i^{(n)}$ in (7) with $Repu_{i,cp}^{(n)}$ and $Repu_{i,rw}^{(n)}$
  end for
  Server aggregates all client local models to update global model $\theta_{gl}^{(n)}$ in (9) and sends to clients
  Client tests client local model and global model to compute $L_i^{(n)}$ and $F_i^{(n)}$ and send to server
end for
return $\theta_{gl}^{(N)}$

III. Experimental Results and Analysis

A. Datasets

In this paper, three open-sourced bearing fault diagnosis datasets from Western Reserve University (CWRU) [36], Paderborn University (PU) [37], and Huazhong University of Science & Technology (HUST) [38] are chosen to conduct a comprehensive evaluation of the proposed FedRepu.

1) For the CWRU bearing dataset, faults were induced into the motor bearings employing the electro-discharge machining (EDM) technique. Data acquisition was performed for bearings in normal condition as well as for bearings with single-point defects on the drive end at the rate of 48,000 samples per second. Vibration data was recorded under four loads and rotational speeds, including A=1797 rpm and 0 HP, B=1772 rpm and 1 HP, C=1750 rpm and 2 HP, and D=1730 rpm and 3 HP. In our federated learning setting, four clients respectively align with these four operational conditions. The classification task, detailed in Table II, assigns 10 categories of data to each client. Thus, the federated learning strategy aims to develop a universal model capable of handling fault diagnosis data irrespective of the operational motor loads or speeds.
2) For the PU bearing dataset, data was collected from a modular test rig by installing ball bearings with different types of damage. The dataset contained vibration signals from seven bearing fault modes sampled at a frequency of 64 kHz. Faults include real damages caused by fatigue pitting and artificial faults induced by EDM, drilling, and manual electric engraving. The fault data was recorded under the following four working conditions of rotational speeds, load torques and radial forces: E=1500 rpm, 0.7 Nm and 1000 N, F=900 rpm, 0.7 Nm and 1000 N, G=1500 rpm, 0.1 Nm and 1000 N, and H=1500 rpm, 0.7 Nm and 400 N. The practical application of this dataset is accomplished by setting four clients corresponding to each condition and seven classification labels, detailed in Table III, corresponding to each bearing fault mode.
3) For the HUST bearing dataset, vibration data was collected from rolling bearings on a Spectra-Quest Mechanical Fault Simulator. The dataset contains vibration signals from nine health states, sampled at a frequency of 25.6 kHz. All fault states in the HUST dataset were artificially introduced, including inner-race, outer-race, ball, and combination faults, where the combination fault denotes the simultaneous presence of inner-race and outer-race defects. In our federated learning setting, four operating conditions are selected to define the federated clients, namely I=20 Hz, J=25 Hz, K=30 Hz, and L=35 Hz. The classification task, detailed in Table IV, assigns nine categories of data to each client.

B. Basic Experiments and Experimental Settings

To maintain consistency across experiments with various aggregation algorithms, a homogeneous computing environment is established. The training, validation, and test datasets

TABLE II
CLASSIFICATION LABELS OF CWRU

Conditional label	Data source	Diameter	Fault type
N	Drive end	Normal	Normal
07IR	Drive end	0.007"	Inner race
07B	Drive end	0.007"	Ball
07OR	Drive end	0.007"	Outer race
14IR	Drive end	0.014"	Inner race
14B	Drive end	0.014"	Ball
14OR	Drive end	0.014"	Outer race
21IR	Drive end	0.021"	Inner race
21B	Drive end	0.021"	Ball
21OR	Drive end	0.021"	Outer race

TABLE III
CLASSIFICATION LABELS OF PU

Conditional label	Fault description
KA01	Outer ring damage of EDM
KA03	Outer ring damage of electric engraver
KA04	Outer ring damage of pitting
KA07	Outer ring damage of drilling
KI01	Inner ring damage of EDM
KI03	Inner ring damage of electric engraver
KI04	Inner ring damage of pitting

contain 70%, 15%, and 15% of the data, respectively. The ResNet1D architecture is selected as the backbone of the learning model due to its proficiency in handling one-dimensional (1D) data. ResNet1D is a direct 1D adaptation of the canonical ResNet-18, where every 2D convolution is replaced by a 1×3 1D kernel. The detailed introduction and usage of ResNet1D can be found in [4].

The training is conducted on a single GPU of GeForce RTX 3090 to ensure a fair comparison between clients. Synchronization points are established after a fixed number of local training epochs to align the aggregation cycles across all comparative strategies. For the CWRU dataset, the local training stage during each iteration is set to 10 local epochs with 5 aggregation iterations. For PU and HUST datasets, 3 local epochs and 30 aggregation iterations are adopted. Considering the practical performance and repeated hyperparameter tests of FedRepu, the experiments on CWRU and PU are carried out with α=0.4, β=0.3, γ=0.5, λ=0.65, and κ=0.11, while the HUST experiments follow α=0.4, β=0.3, γ=0.5, λ=0.65, and κ=0.5.

The evaluation on the CWRU dataset is conducted with the dataset size varying across clients (ranging from 3,437 to 26,474 samples), with each sample being a one-dimensional tensor of size 1024. The data volume differences inherent in the CWRU working conditions provide a perfect chance to evaluate the efficiency of aggregation strategies when facing serious data imbalance and distribution differences. To visually demonstrate FedRepu's effectiveness in overcoming client data heterogeneity, t-SNE is employed to visualize the feature space learned by the global model across all four CWRU clients (Fig. 4). The top row reveals the inherent challenge of unstructured raw data from each client's different operating conditions. The bottom row demonstrates that the feature representations form well-separated clusters corresponding to the 10 distinct fault classes for every client. This visual evidence directly confirms that FedRepu successfully aggregates diverse client knowledge into an effectively generalized model without being biased towards any specific client data distribution.

TABLE IV
CLASSIFICATION LABELS OF HUST

Conditional label	Fault description
H	Healthy state
0.5X_I	Medium inner race fault
I	Severe inner race fault
0.5X_O	Medium outer race fault
O	Severe outer race fault
0.5X_B	Medium ball fault
B	Severe ball fault
0.5X_C	Medium combination fault
C	Severe combination fault

TABLE V
ACCURACIES OF 5 REPEATED EXPERIMENTS OF FEDREPU ON CWRU

Run	Client 1	Client 2	Client 3	Client 4
1	96.07	98.63	97.92	97.90
2	97.69	98.28	98.31	98.03
3	97.42	98.86	98.68	98.04
4	97.15	97.04	98.15	96.10
5	98.37	98.16	98.03	98.27
Average	97.34	98.19	98.22	97.67

In our basic 5 repeated experiments on CWRU dataset, the FedRepu federated learning aggregation strategy has demonstrated consistently high and stable accuracy across all clients, as shown in Table V. Furthermore, the Welch's t-test results, demonstrated in Table VI, confirm that there is no statistically significant difference ($p > 0.05$) between any pair of clients. The statistical consistency verifies that FedRepu successfully tackles data heterogeneity and imbalance, ensuring balanced and optimal performance across all clients.

C. Comparison Study Results & Analysis with Cross-validation Tasks

To further evaluate the performance of the FedRepu strategy, the comparison study is designed against eight classical or novel federated learning strategies, namely FedAvg [24], FedAvgM [25], FedNova [26], FedRadar [39], MCA [40], Medium-Krum [41], FedLaDO [42] and FedRad [43]. Apart from the methods introduced in Section I, FedRadar introduces an adaptive aggregation mechanism based on gradient radar analysis to capture cross-client relevance in multi-task federated learning, thereby improving client-level personalization; MCA (Multi-dimensional Credibility Aggregation) employs credibility estimation across multiple feature dimensions to resist malicious or low-quality updates under non-IID or adversarial conditions; Medium-Krum is a Byzantine-robust aggregation method that selects client updates with minimal pairwise distance to the majority, effectively filtering out extreme or outlier gradients; FedLaDO emphasizes adaptive global aggregation under heterogeneous local objectives; and FedRad introduces a dynamic aggregation mechanism to rebalance client contributions. Since FedRadar originates from multi-task federated learning, the classifiers of local models are aggregated to satisfy the goal of this experiment.

The cross-validation study tests the effectiveness of FedRepu aggregation strategy under challenging conditions, where the local models are trained using data from multiple source clients, and the global model is then evaluated on the unknown test client (representing a completely different working condition). This setup highlights the aggregation

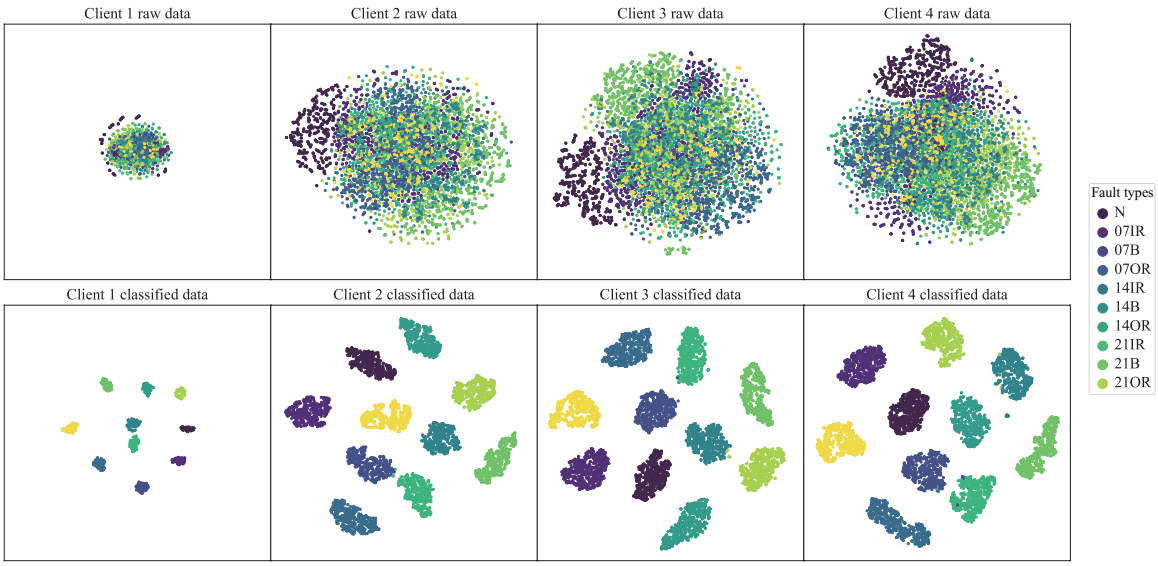


Fig. 4. Visualization of raw data (top row) and classified data (bottom row) in different clients on CWRU dataset using the t-SNE algorithm.

TABLE VI
WELCH'S T-TEST RESULTS BETWEEN DIFFERENT CLIENTS

Comparison	t-statistic	p-value	Consistency? (p > 0.05)
Client 1 vs 2	-1.740915	0.121093	✔
Client 1 vs 3	-2.198617	0.079521	✔
Client 1 vs 4	-0.599689	0.565349	✔
Client 2 vs 3	-0.070397	0.946422	✔
Client 2 vs 4	1.039782	0.330376	✔
Client 3 vs 4	1.315699	0.246677	✔

TABLE VII
TASK ASSIGNMENT OF CROSS-VALIDATION STUDY

CWRU task name	Multiple source clients			Unknown test client
	#1	#2	#3	
T1	B	C	D	A
T2	A	C	D	B
T3	A	B	D	C
T4	A	B	C	D
PU task name	**Multiple source clients**			**Unknown test client**
	#1	#2	#3	
T5	F	G	H	E
T6	E	G	H	F
T7	E	F	H	G
T8	E	F	G	H
HUST task name	**Multiple source clients**			**Unknown test client**
	#1	#2	#3	
T9	J	K	L	I
T10	I	K	L	J
T11	I	J	L	K
T12	I	J	K	L

strategy's ability to generalize to unknown scenarios, offering deeper insights into the robustness of the global model. The specific tasks are listed in Table VII. With the CWRU, PU, and HUST datasets all included, the comparison study contains 12 cross-domain classification tasks in total.

The cross-validation results on the CWRU dataset in Table VIII show a clear performance advantage of FedRepu over the compared baselines. FedRepu achieves the highest average accuracy of 92.26%, compared with 90.68% for the strongest baseline, FedRadar. More importantly, FedRepu is also the most balanced method across the four unseen-condition tasks. Most strategies experience an obvious accuracy drop in T1, while FedRepu still reaches 87.52%, improving on the second-best result of 82.49% by 5.03 percentage points. In T2-T4, some baselines obtain competitive scores on individual tasks, but they also suffer from larger fluctuations. For example, FedAvg falls to 86.59% in T4, and FedLaDO drops to 86.86% in T2. These results indicate that conventional averaging, momentum, or distance-based aggregation is still limited in handling substantial differences in client data quantity and quality. In contrast, FedRepu remains consistently competitive across all tasks, showing strong robustness under cross-condition generalization.

The PU dataset is more challenging because the differences in rotational speed, load torque, and radial force create much stronger distribution shifts between clients. On this dataset, FedRepu attains an average accuracy of 80.12%, clearly surpassing the strongest baseline FedLaDO at 70.07%. The gap becomes particularly striking in T6, where the unseen condition F is difficult for all baseline methods, and no competitor exceeds 54.60%. FedRepu reaches 91.93%, improving upon the best baseline by 37.33 percentage points. This result suggests that the proposed reputation mechanism is particularly effective when only a subset of client updates remains informative for the held-out domain. Although FedRepu is not the best method in T7, where FedLaDO reaches 79.95%, it still dominates T5, T6, and T8 and therefore achieves the highest overall average with the strongest cross-task stability.

The HUST results further confirm the effectiveness of FedRepu in a relatively stable yet still heterogeneous cross-domain setting. FedRepu achieves the highest average accuracy of 96.05%, exceeding the strongest baseline FedLaDO at 94.60%, and records the best results in Tasks T9, T10, and T12. Compared with PU, the performance gaps on HUST are more moderate, suggesting that inter-client discrepancy is less disruptive to cross-domain transfer in this dataset. Even under such conditions, FedRepu still delivers the best overall result, showing that the proposed aggregation mechanism remains

effective not only in strongly shifted environments but also in comparatively stable task settings.

Taken together, the cross-validation studies on CWRU, PU, and HUST consistently show that FedRepu achieves the highest average accuracies, namely 92.26%, 80.12%, and 96.05%, respectively. The advantage becomes more pronounced as the domain discrepancy grows, especially on the PU dataset. On CWRU and HUST datasets, FedRepu still maintains the most stable overall performance among the compared methods. Since FedRepu evaluates client contributions from multiple perspectives, including real-time predictive quality, historical consistency, and reward-based adjustment, it can distinguish reliable updates from less informative ones more effectively than conventional aggregation strategies. Consequently, the influence of data imbalance and distribution shift can be reduced during aggregation, leading to stronger and more stable cross-domain generalization for federated bearing fault diagnosis.

TABLE VIII
COMPARISON RESULTS OF CROSS-VALIDATION ON CWRU, PU, AND HUST DATASETS WITH DIFFERENT STRATEGIES (%)

CWRU

Strategy	T1	T2	T3	T4	Avg
Medium-Krum	78.24±1.47	94.93±2.21	95.12±3.17	89.18±1.04	89.37
FedNova	80.21±2.42	94.73±1.13	96.01±1.02	**90.35±0.42**	90.32
FedAvg	80.60±2.20	96.46±1.04	95.31±0.67	86.59±0.80	89.74
FedAvgM	80.69±2.04	96.15±0.22	95.97±0.61	87.01±1.84	89.95
MCA	80.24±2.24	96.01±0.95	95.63±0.47	88.87±1.13	90.19
FedRadar	82.23±1.47	**96.60±1.47**	95.62±1.47	88.28±1.47	90.68
FedRad	82.49±2.80	90.27±1.13	94.37±1.04	79.92±1.14	86.76
FedLaDO	80.51±1.50	86.86±3.19	**96.77±0.66**	81.57±2.48	86.43
FedRepu	**87.52±3.78**	95.92±0.49	96.46±0.73	89.12±1.94	**92.26**

PU

Strategy	T5	T6	T7	T8	Avg
Medium-Krum	83.10±2.77	54.60±0.91	56.27±2.60	76.33±5.49	67.57
FedNova	77.14±2.22	48.47±4.93	49.82±8.95	78.29±5.73	63.43
FedAvg	80.40±2.50	51.79±1.30	54.98±1.67	81.48±1.49	67.16
FedAvgM	76.04±5.28	45.85±3.57	53.09±6.73	76.12±5.36	62.78
MCA	66.95±5.93	39.27±3.81	31.97±9.38	59.36±6.11	49.39
FedRadar	76.98±4.17	47.15±4.60	54.32±8.15	81.48±6.57	64.98
FedRad	72.61±5.08	38.96±3.78	74.39±1.49	81.87±4.92	66.96
FedLaDO	80.40±2.97	39.33±2.30	**79.95±2.20**	80.59±7.69	70.07
FedRepu	**86.54±2.64**	**91.93±4.78**	53.96±2.09	**88.04±2.93**	**80.12**

HUST

Strategy	T9	T10	T11	T12	Avg
Medium-Krum	87.78±2.07	93.72±1.91	95.22±2.26	96.52±0.74	93.31
FedNova	89.76±3.46	95.12±0.25	96.33±1.32	96.86±1.47	94.52
FedAvg	87.97±1.98	94.49±0.93	96.38±1.42	97.20±1.20	94.01
FedAvgM	84.98±3.68	94.40±0.71	94.93±2.40	96.09±0.52	92.60
MCA	88.55±2.09	94.25±1.70	95.99±1.66	96.76±1.41	93.89
FedRadar	86.18±4.68	93.04±2.46	94.40±2.29	97.83±0.31	92.86
FedRad	86.23±0.94	96.43±0.45	**97.44±0.86**	96.77±0.72	94.22
FedLaDO	89.61±2.55	94.68±2.38	96.86±0.75	97.25±0.94	94.60
FedRepu	**91.40±3.51**	**97.15±2.19**	97.10±0.94	**98.55±0.66**	**96.05**

D. Ablation Study Results & Analysis

As described in Section II, the FedRepu computation stage consists of three key components: real-time reputation evaluation, historical reputation fusion, and a reward-based incentive mechanism. The ablation results in Fig. 5, Fig. 6, and Fig. 7 provide direct evidence of how these components contribute to the final performance. The real-time reputation module establishes a strong baseline on all three datasets. On CWRU, this variant maintains accuracies above 91% for all four clients. On PU, despite the much stronger heterogeneity across operating conditions, it still achieves competitive results for several clients, with the best accuracy reaching 93.10%. On HUST, the real-time-only variant attains an average accuracy of 95.47%, and all four clients remain above 90%. These results indicate that the real-time reputation mechanism can identify the immediate quality of local updates and suppress the influence of less reliable client contributions during aggregation.

When historical reputation is incorporated, the aggregation process becomes more robust to short-term fluctuations in local model behaviour. On CWRU, historical fusion improves the weakest client from 91.04% to 97.96%, pushing all clients beyond the 97% level and clearly enhancing inter-client consistency. On PU, the effect is more differentiated under the stronger distribution shift, where Client 1 improves markedly from 71.84% to 86.21%. On HUST, the average accuracy rises from 95.47% to 96.88%, while all four clients remain within a narrow high-accuracy range between 93.48% and 98.26%. Taken together, these results show that historical reputation effectively stabilizes aggregation by incorporating temporal consistency into the evaluation of client contributions.

The reward mechanism provides a further complementary contribution by explicitly strengthening local updates that offer additional value relative to the current global model. On CWRU, the reward-enhanced variant raises Client 1 from 91.04% to 96.61%, while all client accuracies remain above 96%, confirming that the reward term can improve the quality of aggregation without sacrificing overall balance. On PU, the reward term substantially reshapes the client-wise performance pattern, where Client 1 and Client 2 improve to 86.02% and 89.85%. On HUST, the reward-enhanced variant reaches an average accuracy of 96.96%, higher than the real-time-only variant and very close to the history-enhanced result. These results verify that the reward mechanism is not redundant. Instead, it further improves the discrimination of informative client updates and enhances the adaptive aggregation ability of FedRepu.

The best results are achieved when all three components are integrated into the complete FedRepu framework. Across all three datasets, the full model delivers the strongest overall performance and the clearest evidence of component complementarity. On CWRU, the complete FedRepu pushes all clients above 98%. On PU, although the dataset remains the most challenging due to severe heterogeneity and cross-condition discrepancy, the full model still raises all clients above 75%, with the best result reaching 89.66%. On HUST, the complete FedRepu attains the highest average accuracy of 97.97%, with all clients above 97%. These results strongly validate the design of FedRepu: real-time reputation captures the current reliability of local models, historical reputation improves temporal stability, and reward-based incentives further emphasize updates that contribute more effectively to the global objective. Their joint effect enables FedRepu to better address client heterogeneity and to produce a more robust, accurate, and well-balanced global model for federated bearing fault diagnosis.

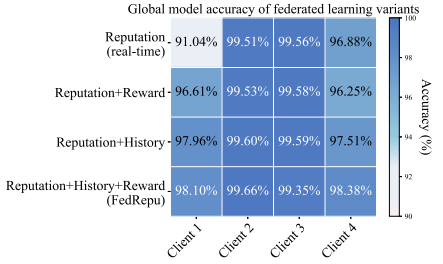


Fig. 5. Ablation study of FedRepu aggregation strategy on CWRU dataset.

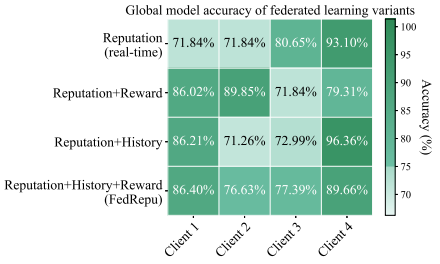


Fig. 6. Ablation study of FedRepu aggregation strategy on PU dataset.

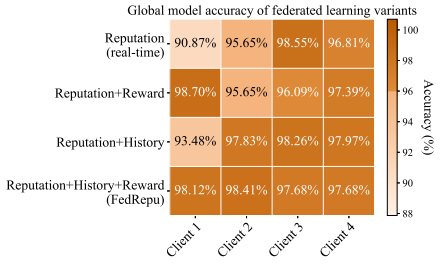


Fig. 7. Ablation study of FedRepu aggregation strategy on HUST dataset.

E. Hyperparameter Sensitivity Analysis

To validate the rationality of the selected hyperparameters and further examine the robustness of FedRepu, a comprehensive sensitivity analysis is conducted on five key coefficients. The results on the CWRU, PU, and HUST datasets are summarized in Table IX. In each group of experiments, one hyperparameter is varied while the remaining four are fixed at the proposed settings. For CWRU and PU, the adopted baseline configuration is $(\alpha = 0.4, \beta = 0.3, \gamma = 0.5, \lambda = 0.65, \kappa = 0.11)$. For HUST, the retained baseline configuration is $(\alpha = 0.4, \beta = 0.3, \gamma = 0.5, \lambda = 0.65, \kappa = 0.5)$, which is consistent with the strongest cross-domain setting used in the previous experiments.

The experimental results on the CWRU dataset demonstrate the inherent stability of the proposed method. The proposed configuration achieves the highest average accuracy of 98.87%, serving as a local optimum. Notably, even when the hyperparameters deviate from these values, the model maintains a high performance level, with average accuracies consistently remaining above 96%. This phenomenon suggests that FedRepu is resilient to hyperparameter fluctuations in scenarios with relatively distinct fault features.

On the more challenging PU dataset, the sensitivity analysis reveals the necessity of precise parameter tuning to handle complex data distributions. Unlike the broad robustness observed on CWRU, the performance on PU is more sensitive to parameter changes, yet the proposed settings yield the peak performance for α, β, λ, and κ. For instance, varying the historical decay factor λ from the proposed 0.65 to 0.75 results in a significant accuracy drop from 82.52% to 77.35%, confirming that the chosen value appropriately balances historical memory against current updates. Regarding the weight of the false-negative component γ, although a slightly lower value of 0.4 yields a marginal improvement on PU (83.81%) compared to the chosen 0.5 (82.52%), this setting leads to a performance decline on the CWRU dataset (dropping to 97.74%). Consequently, $\gamma = 0.5$ is selected as the final setting to ensure a robust generalization capability across diverse datasets.

The HUST results show a comparatively stable sensitivity pattern. The retained baseline setting achieves an average accuracy of 97.97%, and most nearby parameter changes remain within a narrow high-performance range. A local optimum appears at $\gamma = 0.4$, where the average accuracy reaches 98.30%, while the retained baseline with $\gamma = 0.5$ remains close to this level. These results suggest that the HUST task provides a relatively smooth response surface around the retained configuration, and that the parameter setting used in this analysis already lies in a stable and effective region rather than depending on a narrowly tuned point.

Overall, the sensitivity analysis provides strong support for the practicality of the selected hyperparameter settings. The proposed configurations are not arbitrary empirical choices, but the result of balancing accuracy, robustness, and cross-dataset consistency under different levels of heterogeneity. These findings further demonstrate that the proposed reputation-based aggregation strategy maintains both effectiveness and stability in federated bearing fault diagnosis.

TABLE IX
SENSITIVITY ANALYSIS OF HYPERPARAMETERS ON THE CWRU, PU, AND HUST DATASETS (THE PROPOSED SETTINGS ARE MARKED AS OURS.)

Dataset	Parameter	Value	Client 1	Client 2	Client 3	Client 4	Avg
CWRU	α	0.3	98.10	97.65	97.46	96.65	97.47
		0.5	99.19	98.21	97.32	96.83	97.89
		0.4 (Ours)	98.10	99.66	99.35	98.38	98.87
	β	0.2	96.70	97.45	97.62	96.95	97.02
		0.4	96.61	98.12	97.69	96.81	97.31
		0.3 (Ours)	98.10	99.66	99.35	98.38	98.87
	γ	0.4	96.88	97.51	98.70	97.88	97.74
		0.6	98.10	98.59	97.85	96.42	97.74
		0.5 (Ours)	98.10	99.66	99.35	98.38	98.87
	λ	0.55	96.47	96.39	97.25	96.62	96.68
		0.75	98.37	98.10	97.96	97.55	97.99
		0.65 (Ours)	98.10	99.66	99.35	98.38	98.87
	κ	0.08	97.01	96.26	97.51	96.93	96.93
		0.14	97.29	97.07	97.11	96.76	97.05
		0.11 (Ours)	98.10	99.66	99.35	98.38	98.87
PU	α	0.3	73.18	71.07	80.08	89.46	78.45
		0.5	86.21	70.88	79.12	89.46	81.42
		0.4 (Ours)	86.40	76.63	77.39	89.66	82.52
	β	0.2	70.50	75.48	81.61	99.04	81.66
		0.4	81.80	81.42	80.65	72.22	79.02
		0.3 (Ours)	86.40	76.63	77.39	89.66	82.52
	γ	0.4	86.02	78.74	79.69	90.80	83.81
		0.6	84.87	70.88	71.84	87.55	78.78
		0.5 (Ours)	86.40	76.63	77.39	89.66	82.52
	λ	0.55	87.74	73.56	74.33	80.65	79.07
		0.75	70.11	80.65	76.05	82.57	77.35
		0.65 (Ours)	86.40	76.63	77.39	89.66	82.52
	κ	0.08	84.67	71.84	70.88	90.80	79.55
		0.14	75.29	73.75	70.31	78.54	74.47
		0.11 (Ours)	86.40	76.63	77.39	89.66	82.52
HUST	α	0.3	85.22	91.01	99.57	98.55	93.59
		0.5	81.59	86.81	95.22	98.26	90.47
		0.4 (Ours)	98.12	98.41	97.68	97.68	97.97
	β	0.2	95.36	97.83	99.28	99.13	97.90
		0.4	97.68	95.94	94.93	95.94	96.12
		0.3 (Ours)	98.12	98.41	97.68	97.68	97.97
	γ	0.4	97.68	98.99	98.12	98.41	98.30
		0.6	84.06	85.80	97.25	98.26	91.34
		0.5 (Ours)	98.12	98.41	97.68	97.68	97.97
	λ	0.55	85.51	91.16	95.80	97.97	92.61
		0.75	93.48	96.96	99.42	98.26	97.03
		0.65 (Ours)	98.12	98.41	97.68	97.68	97.97
	κ	0.3	97.97	94.78	91.45	90.87	93.77
		0.7	96.52	96.23	93.04	95.07	95.22
		0.5 (Ours)	98.12	98.41	97.68	97.68	97.97

IV. Conclusion

In this work, we focus on the area of bearing fault diagnosis and exploit recent advances in federated learning to improve diagnostic accuracy and guarantee data privacy. The integration of FedRepu represents a significant advancement in diagnosis models, demonstrating improved effectiveness and promising performance. The experiments evaluate the detailed implementation of FedRepu and demonstrate its capability in self-balancing across clients. For future work, there is a need to further refine FedRepu in compressed and encrypted federated learning environments to uphold data privacy while maintaining the usage of reputation assessment. In addition, extending FedRepu to multi-sensor fusion will enable the integration of heterogeneous sensing modalities (e.g., vibration, temperature, and acoustic signals) to improve performance under diverse operating conditions. Furthermore, incorporating multi-task diagnosis capabilities will allow the framework to jointly learn related fault patterns or health indicators, thereby improving adaptability in complex industrial systems. These efforts will contribute to the development of more efficient and privacy-conscious fault diagnosis methods in industrial maintenance and operational management.

References

[1] B. R. Nayana, R. Subha, R. Radhakrishnan, and P. Geethanjali, "Scalable Bearing Fault Diagnosis Using Metaheuristic Feature Selection and Machine Learning for Diverse Operating Conditions," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2469606, 2025.

[2] Y. Zhan, R. Yang, J. You, M. Huang, W. Liu, and X. Liu, "A Systematic Literature Review on Incomplete Multimodal Learning: Techniques and Challenges," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2467083, 2025.

[3] L. Wen, X. Li, L. Gao, and Y. Zhang, "A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 7, pp. 5990–5998, 2017.

[4] X. Chen, R. Yang, Y. Xue, M. Huang, R. Ferrero, and Z. Wang, "Deep Transfer Learning for Bearing Fault Diagnosis: A Systematic Review Since 2016," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[5] A. Guo, Z. Zhao, J. Sun, and H. Ge, "Fault Classification of Meta-Action Unit Using CEEMDAN Double-Layer Decomposition and COA-SVM," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2469602, 2025.

[6] V. Kwao and I. Raptis, "Sensor Fault Diagnosis in Nonlinear Stochastic Systems: A Hybrid System Formulation," *International Journal of Systems Science*, vol. 56, no. 8, pp. 1755–1773, 2025.

[7] X. Zhang, J. Liu, X. Zhang, and Y. Lu, "Self-Supervised Graph Feature Enhancement and Scale Attention for Mechanical Signal Node-Level Representation and Diagnosis," *Advanced Engineering Informatics*, vol. 65, p. 103197, 2025.

[8] Y. Xu, S. Li, K. Feng, R. Huang, B. Sun, X. Yang, Z. Zhao, and G. Q. Huang, "Domain Constrained Cascadic Multireceptive Learning Networks for Machine Health Monitoring in Complex Manufacturing Systems," *Journal of Manufacturing Systems*, vol. 80, pp. 563–577, 2025.

[9] R. Huang, J. Li, Y. Liao, J. Chen, Z. Wang, and W. Li, "Deep Adversarial Capsule Network for Compound Fault Diagnosis of Machinery toward Multidomain Generalization Task," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2020.

[10] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly Accurate Machine Fault Diagnosis Using Deep Transfer Learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2446–2455, 2018.

[11] S. Lu, Z. Gao, Q. Xu, C. Jiang, A. Zhang, and X. Wang, "Class-Imbalance Privacy-Preserving Federated Learning for Decentralized Fault Diagnosis With Biometric Authentication," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 9101–9111, 2022.

[12] Q. Zhu and C.-X. Shi, "Data-Driven Fault-Tolerant Consensus Control for Nonlinear Multi-Agent Systems under Sensor and Actuator Faults," *International Journal of Systems Science*, vol. 57, no. 7, pp. 2006–2023, 2026.

[13] C. Jia, Z. Yang, F. Jia, and X. He, "An Integrated Framework of Fault-Tolerant Control and Diagnosis for Multi-Agent Systems Based on A Novel Evaluation Function and Topology Switching Mechanism," *International Journal of Systems Science*, pp. 1–16, 2025.

[14] Q. Wu, C. Dong, F. Guo, L. Wang, X. Wu, and C. Wen, "Privacy-Preserving Federated Learning for Power Transformer Fault Diagnosis With Unbalanced Data," *IEEE Transactions on Industrial Informatics*, 2023.

[15] L. Ma, J. He, K. Lu, D. Wang, L. Yin, and Z. Li, "A Contrastive Learning and Knowledge Distillation-Based Framework for Efficient Federated Intrusion Detection in IoT," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2518963, 2025.

[16] X. Yan, C. Wang, and Y. Jin, "Federated Bimodal Graph Neural Networks for Text-Image Retrieval," *International Journal of Network Dynamics and Intelligence*, vol. 4, no. 2, p. 100009, 2025.

[17] S. Lu, Z. Gao, P. Zhang, Q. Xu, T. Xie, and A. Zhang, "Event-Triggered Federated Learning for Fault Diagnosis of Offshore Wind Turbines with Decentralized Data," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 2, pp. 1271–1283, 2023.

[18] G.-Y. Huang and C.-H. Lee, "Federated Learning Architecture for Bearing Fault Diagnosis," in *2021 International Conference on System Science and Engineering (ICSSE)*. IEEE, 2021, pp. 408–411.

[19] Y. Hu, C. Zhang, and S. Liu, "Privacy-Preserving Distributed Recursive Filtering for State-Saturated Systems with Quantization Effects," *International Journal of Network Dynamics and Intelligence*, vol. 4, no. 2, p. 100012, 2025.

[20] G. Rigatos, Z. Gao, K. Busawon, P. Siano, M. Al-Numay, and M. Abbaszadeh, "Fault Diagnosis for VSC-HVDC Transmission Systems Using Kalman Filter-Based Disturbance Observers," *International Journal of Systems Science*, pp. 1–25, 2025.

[21] C. Mao and C. Wen, "Fault Diagnosis Method for Rolling Bearings in High-Noise Environments Based on COA-FMD-ITEO," *International Journal of Network Dynamics and Intelligence*, vol. 4, no. 3, p. 100016, 2025.

[22] R. Saha, S. Misra, A. Chakraborty, C. Chatterjee, and P. K. Deb, "Data-Centric Client Selection for Federated Learning Over Distributed Edge Networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 2, pp. 675–686, 2022.

[23] X. Yin, X. Li, Y. Zhang, W. Zhou, and Z. Liu, "Multi-Task Network Based Health Status Assessment of Cutting Tools," *International Journal of Network Dynamics and Intelligence*, vol. 4, no. 2, p. 100008, 2025.

[24] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

[25] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan, "Adaptive Federated Optimization," *arXiv preprint arXiv:2003.00295*, 2020.

[26] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7611–7623, 2020.

[27] V. Hassija, V. Chawla, V. Chamola, and B. Sikdar, "Incentivization and Aggregation Schemes for Federated Learning Applications," *IEEE Transactions on Machine Learning in Communications and Networking*, 2023.

[28] Y. Xue, R. Yang, X. Chen, Z. Tian, and Z. Wang, "A Novel Local Binary Temporal Convolutional Neural Network for Bearing Fault Diagnosis," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[29] D. Li, J. Chen, R. Huang, Z. Chen, and W. Li, "Sensor-Aware CapsNet: Towards Trustworthy Multisensory Fusion for Remaining Useful Life Prediction," *Journal of Manufacturing Systems*, vol. 72, pp. 26–37, 2024.

[30] C. Wang, Z. Wang, and H. Dong, "A Novel Prototype-Assisted Contrastive Adversarial Network for Weak-Shot Learning with Applications: Handling Weakly Labeled Data," *IEEE/ASME Transactions on Mechatronics*, 2023.

[31] C. Wang, Z. Wang, H. Dong, S. Lauria, W. Liu, Y. Wang, F. Fadzil, and X. Liu, "Fusionformer: A Novel Adversarial Transformer Utilizing Fusion Attention for Multivariate Anomaly Detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[32] D. C. Nguyen, M. Ding, Q.-V. Pham, P. N. Pathirana, L. B. Le, A. Seneviratne, J. Li, D. Niyato, and H. V. Poor, "Federated Learning Meets Blockchain in Edge Computing: Opportunities and Challenges," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12 806–12 825, 2021.

[33] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive Mechanism for Reliable Federated Learning: A Joint Optimization Approach to Combining Reputation and Contract Theory," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 700–10 714, 2019.

[34] J. Qi, F. Lin, Z. Chen, C. Tang, R. Jia, and M. Li, "High-Quality Model Aggregation for Blockchain-Based Federated Learning via Reputation-Motivated Task Participation," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 18 378–18 391, 2022.

[35] J. Miao and W. Zhu, "Precision–Recall Curve (PRC) Classification Trees," *Evolutionary Intelligence*, vol. 15, no. 3, pp. 1545–1569, 2022.

[36] W. A. Smith and R. B. Randall, "Rolling Element Bearing Diagnostics Using The Case Western Reserve University Data: A Benchmark Study," *Mechanical Systems and Signal Processing*, vol. 64, pp. 100–131, 2015.

[37] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "Condition Monitoring of Bearing Damage in Electromechanical Drive Systems by Using Motor Current Signals of Electric Motors: A Benchmark Data Set for Data-Driven Classification," *Proceedings of the 3rd European Conference of the Prognostics and Health Management Society*, vol. 2016, pp. 152–156, 2016.

[38] C. Zhao, E. Zio, and W. Shen, "Domain Generalization for Cross-Domain Fault Diagnosis: An Application-Oriented Perspective and A Benchmark Study," *Reliability Engineering & System Safety*, vol. 245, p. 109964, 2024.

[39] X. Jiang, J. Zhang, and L. Zhang, "FedRadar: Federated Multi-Task Transfer Learning for Radar-Based Internet of Medical Things," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 1459–1469, 2023.

[40] Z. Luan, W. Li, M. Liu, and B. Chen, "Robust Federated Learning: Maximum Correntropy Aggregation against Byzantine Attacks," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[41] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[42] J. Bai, D. Wu, S. Zeng, Y. Zhao, Y. Qu, and S. Yu, "Non-IID Free Federated Learning with Fuzzy Optimization for Consumer Electronics Systems," *IEEE Transactions on Consumer Electronics*, 2025.

[43] D. Xiao, J. Diao, J. Ding, L. Jiang, and C. Zhao, "FedRad: A Dynamic Low-Cost Federated Learning for Fault Diagnosis of Rotating Machinery," *IEEE Transactions on Instrumentation and Measurement*, 2025.

APPENDIX

APPENDIX A
DETAILED EXPLANATION OF MISCLASSIFICATION PART OF REAL-TIME REPUTATION SCORE

1. Assumptions and Objectives

- **Assumption 1**: Each false positive situation incurs a fixed fault diagnosis cost $C_{FP} > 0$, and each false negative situation incurs a fixed fault diagnosis cost $C_{FN} > 0$.
- **Assumption 2**: $FPR, FNR \in [0, 1]$.
- **Assumption 3**: The model's expected misclassification cost $E[R]$ follows a linear form:

$$E[R] = C_{FP} \cdot FPR + C_{FN} \cdot FNR$$

- **Objective**: Construct a scoring function $\Phi(E[R])$ of misclassification part such that:
  - $\Phi(E[R])$ is a linear, monotonically decreasing function of $E[R]$;
  - $\Phi(E[R]) \in [0, 1]$ to maintain consistency with normalized metrics F1.

2. From Fault Diagnosis Cost to Reputation Score

Let:

$$\underline{R} = \min E[R] = 0 \quad (\text{when } FPR = FNR = 0)$$

and

$$\overline{R} = \max E[R] = C_{FP} + C_{FN} \quad (\text{when } FPR = FNR = 1)$$

Thus,

$$E[R] \in [\underline{R}, \overline{R}] = [0, C_{FP} + C_{FN}]$$

To map $[\underline{R}, \overline{R}]$ to $[0, 1]$ in a linear, monotonically decreasing manner, we require:

$$\Phi(\underline{R}) = 1, \quad \Phi(\overline{R}) = 0$$

Suppose:

$$\Phi(E[R]) = aE[R] + b$$

subject to the following constraints:

$$\begin{cases} \Phi(\underline{R}) = a\underline{R} + b = 1 \\ \Phi(\overline{R}) = a\overline{R} + b = 0 \end{cases}$$

Solving yields:

$$a = -\frac{1}{\overline{R}}, \quad b = 1$$

Hence:

$$\begin{aligned} \Phi(E[R]) &= 1 - \frac{E[R]}{\overline{R}} \\ &= 1 - \frac{C_{FP} \cdot FPR + C_{FN} \cdot FNR}{C_{FP} + C_{FN}} \\ &= \frac{C_{FP} + C_{FN} - (C_{FP} \cdot FPR + C_{FN} \cdot FNR)}{C_{FP} + C_{FN}} \\ &= \frac{C_{FP} \cdot (1 - FPR) + C_{FN} \cdot (1 - FNR)}{C_{FP} + C_{FN}} \\ &\in \left[1 - \frac{\overline{R}}{\overline{R}}, 1 - \frac{\underline{R}}{\overline{R}}\right] = [0, 1] \end{aligned}$$

This shows that misclassification part of real-time score can be viewed as the weighted average of $(1 - FPR)$ and $(1 - FNR)$, with weights $\frac{C_{FP}}{C_{FP}+C_{FN}}$ and $\frac{C_{FN}}{C_{FP}+C_{FN}}$, respectively.

3. Complete Real-Time Reputation Score

To ensure the real-time reputation score fully captures the information of the metrics, $Repu_{i,rt}$ is defined as a linear weighted sum:

$$Repu_{i,rt} = \alpha F_1 + \varepsilon \Phi(E[R])$$

where $\alpha, \varepsilon \in [0, 1]$ allow adjusting the importance of imbalance metric $F1$ and the misclassification cost $E[R]$. Since $\Phi(E[R]) \in [0, 1]$, the range of the score satisfies

$$0 \ \leq \ Repu_{i,rt} \ \leq \ \alpha + \varepsilon \ \leq \ 2$$

The above equation can be expanded as:

$$\begin{aligned} Repu_{i,rt} = \alpha F1_i \\ + \frac{\varepsilon C_{FP_i} \cdot (1 - FPR_i) + \varepsilon C_{FN_i} \cdot (1 - FNR_i)}{C_{FP_i} + C_{FN_i}} \end{aligned}$$

In practice, one may directly replace the fraction with two separate adjustable coefficients β and γ, each controlling the importance of $(1 - FPR_i)$ and $(1 - FNR_i)$:

$$Repu_{i,rt} = \alpha F1_i + \beta(1 - FPR_i) + \gamma(1 - FNR_i)$$

This approach provides additional flexibility in weighting different aspects of model performance, thereby enabling a more tailored real-time reputation score.

APPENDIX B
PARAMETER FACTORISATION OF THE REAL–TIME REPUTATION FUNCTION AND THE REASON FOR THREE INDEPENDENT WEIGHTS

1. Notation and Parameter-Space Factorisation

To facilitate a rigorous mathematical analysis and clear parameter-space interpretation, we first define the metric vector:

$$\xi = (F1, \ 1 - FPR, \ 1 - FNR)^\top \in [0, 1]^3$$

and the corresponding hyperparameter vector:

$$\mathbf{p} = (\alpha, \beta, \gamma)^\top \in [0,1]^3$$

Then the real-time reputation score introduced can be explicitly expressed as a bilinear mapping:

$$\mathcal{T}(\mathbf{p}, \xi) = \alpha\, F1 + \beta\,(1 - FPR) + \gamma\,(1 - FNR)$$

Recall from Appendix A that we define the misclassification unit costs:

$$C_{FP} > 0, \quad C_{FN} > 0$$

We further introduce the cost ratio:

$$w = \frac{C_{FP}}{C_{FP} + C_{FN}} \in (0,1)$$

Given a fixed cost ratio w, the mapping $\mathcal{T}$ admits a simpler factorisation form:

$$\mathcal{T}(\mathbf{p}, \xi) = \alpha \cdot F1 + \varepsilon \cdot \big[w(1 - FPR) + (1 - w)(1 - FNR)\big]$$

where the parameter ε relates linearly to β and γ:

$$\varepsilon = \beta + \gamma, \quad \beta = \varepsilon w, \quad \gamma = \varepsilon(1 - w), \quad \frac{\beta}{\beta + \gamma} = w$$

For each fixed w, the three-dimensional parameter space (α, β, γ) reduces to a two-dimensional manifold, $\mathcal{S}_w$, clearly defined by:

$$\mathcal{S}_w = \left\{ (\alpha, \beta, \gamma) \;\middle|\; \frac{\beta}{\beta + \gamma} = w,\ \alpha > 0,\ \beta + \gamma > 0 \right\}$$

All manifolds share the common α-axis ($\beta = \gamma = 0$) and differ in their slope in the (β, γ)-plane. Therefore, the full parameter space $[0,1]^3$ can be seen as a continuous family of these two-dimensional manifolds indexed by the cost ratio w.

2. *Cost–Ratio Update and Parameter Perturbation*

In industrial scenarios, the cost ratio w often changes periodically according to evolving business requirements. We analyze the complexity of updating parameters under two implementations:

a) Two–Parameter Implementation: In the two-parameter scheme, parameters are stored as (α, ε), with the cost ratio w hard-coded. When w is updated, the scale of the reputation function $\mathcal{T}$ changes accordingly. Differentiating $\mathbf{p} = (\alpha, \beta, \gamma)$ with respect to w yields the sensitivity

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}w} = (0,\ \varepsilon,\ -\varepsilon)^\top$$

and the corresponding directional derivative of the reputation score

$$\frac{\partial \mathcal{T}}{\partial w} = \varepsilon\big[(1 - \mathrm{FPR}) - (1 - \mathrm{FNR})\big]$$

Thus, to maintain a smooth, real-time update of $\mathcal{T}$ without abrupt numerical shifts, one must recalibrate ε. In practice, this recalibration necessitates conducting a hyperparameter search, leading to a complexity of at least $\mathcal{O}(MK)$, where M is the number of candidate values evaluated, and K is the number of validation runs per candidate.

b) Three–Parameter Implementation: In the three-parameter implementation, the cost ratio w is external to the hyperparameters. When updating w from w_0 to w_1, the parameter vector $\mathbf{p}_0 = (\alpha_0, \beta_0, \gamma_0)$ is updated via a deterministic linear re-scaling:

$$\mathbf{p}_1 = \big(\alpha_0,\ k\,w_1,\ k\,(1 - w_1)\big)$$

where

$$k = \beta_0 + \gamma_0 = \beta_1 + \gamma_1$$

is chosen to preserve the total weight assigned to the second and third terms, and the overall scale of $\mathcal{T}$ changes smoothly under the new cost ratio, avoiding abrupt shifts in the reputation scores. The parameter-level sensitivity to w is a constant

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}w} = (0,\ k,\ -k)^\top$$

independent of data and other hyperparameters. The sensitivity of the reputation score is

$$\frac{\partial \mathcal{T}}{\partial w} = k\big[(1 - FPR) - (1 - FNR)\big]$$

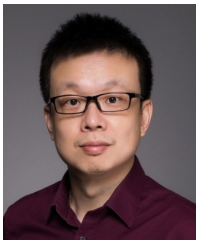
The parameter adjustment is completed in constant time with complexity $\mathcal{O}(1)$, enabling an instantaneous hot-swap when w varies.

3. *Rationale for Exposing β, γ*

Although a two-parameter form is sufficient for a fixed cost ratio, making (β, γ) tunable provides the following advantages:

- **Flexibility:** The cost ratio w varies across sites and phases; keeping β, γ explicit obviates the repeated re-search of ε.
- **Continuity:** Deterministic re-embedding (β, γ) with $(kw, k(1 - w))$ keeps the reputation score $\mathcal{T}$ numerically smooth.
- **Efficiency:** Parameter updates degrade from $\mathcal{O}(MK)$ to a single $\mathcal{O}(1)$ assignment on the server; no client training or communication is affected.

Junxian You (Graduate Student Member, IEEE) received the B.S. degree in automation from Yanshan University, China, in 2023, and M.Res. degree with distinction in computer science from University of Liverpool, U.K., in 2025. He is currently pursuing the Ph.D. degree with University of Glasgow, U.K. He is an active reviewer for international journals and conferences, including IEEE Transactions on Network Science and Engineering, IEEE Internet of Things Journal, and IEEE International Conference on Communications. His research interests include federated learning, reinforcement learning and multi-agent systems.



Rui Yang (Senior Member, IEEE) received the B.Eng. degree in computer engineering and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2008 and 2013, respectively. He is currently a Senior Associate Professor with the School of Advanced Technology, Xi'an Jiaotong–Liverpool University, Suzhou, China, and an Honorary Lecturer with the Department of Computer Science, University of Liverpool, Liverpool, U.K. He is the author or co-author of several technical papers and is also a very active reviewer for many international journals and conferences. His research interests include machine-learning-based data analysis and applications. Dr. Yang serves as an Associate Editor for Neurocomputing, Cognitive Computation, and IEEE Transactions on Instrumentation and Measurement.

Zidong Wang (Fellow, IEEE) received the B.Sc. degree in mathematics from Suzhou University in 1986, and the M.Sc. degree in applied mathematics and the Ph.D. degree in electrical engineering from Nanjing University of Science and Technology, in 1990 and 1994, respectively. He is currently Professor of dynamical systems and computing in the Department of Computer Science, Brunel University London, U.K. From 1990 to 2002, he held teaching and research appointments in universities in China, Germany, and the U.K. His research interests include dynamical systems, signal processing, bioinformatics, control theory and applications. Prof. Wang has published a number of papers in international journals. He is a holder of the Alexander von Humboldt Research Fellowship of Germany, the JSPS Research Fellowship of Japan, William Mong Visiting Research Fellowship of Hong Kong. He serves (or has served) as the Editor-in-Chief for International Journal of Systems Science, Neurocomputing, Systems Science & Control Engineering, and an Associate Editor for 12 international journals including IEEE Transactions on Automatic Control, IEEE Transactions on Control Systems Technology, IEEE Transactions on Neural Networks, IEEE Transactions on Signal Processing, and IEEE Transactions on Systems, Man, and Cybernetics-Part C. He is a Member of the Academia Europaea, a Member of the European Academy of Sciences and Arts, an Academician of the International Academy for Systems and Cybernetic Sciences, a Fellow of the Royal Statistical Society and a member of program committee for many international conferences.