Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Federated mutual learning enables consistent hierarchical sorting of retired batteries across heterogeneous testing nodes

**Xiankui Wen**

13985410224@139.com

Guizhou Power Grid Co., Ltd

**Qiang Fan**

Guizhou Power Grid Co., Ltd

**Lingrong Pang**

Guizhou Power Grid Co., Ltd

**Huayang Ye**

Guizhou Power Grid Co., Ltd

**Qiangqiang Liao**

Shanghai University of Electric Power

**Article**

**Keywords:** retired batteries, federated mutual learning, hierarchical sorting, non-IID, data silos, consistency

**Posted Date:** November 16th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11100592/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Abstract

Accurate sorting is a prerequisite for the echelon utilization of retired power batteries, yet testing data are scattered across geographically distributed nodes that cannot share raw data, and the incoming battery populations differ substantially among nodes, i.e. the data are non-independent and identically distributed (non-IID). Here we propose a hierarchical sorting method based on federated mutual learning. Each node builds a local hierarchical model that combines prior-feature clustering with cluster-wise decision trees; model-output knowledge is exchanged through a shared unlabeled reference set, and distribution-discrepancy weighting together with confidence filtering enables bidirectional knowledge transfer without raw data leaving any node. Validation on 1,080 simulated retired batteries across four non-IID nodes shows that the global average sorting accuracy reaches 91.10%, 12.83 percentage points above local training alone, with a cross-node consistency of 97.13%; across 11 random-seed repeats the average accuracy is 90.17% ± 2.53%. The proposed method improves sorting accuracy and cross-node consistency under heterogeneous multi-node conditions.


# Introduction

With the rapid development of the electric vehicle industry, the growing stock of new energy vehicles and the gradual retirement of early-generation power batteries are creating an increasing demand for the recycling and echelon utilization of retired power batteries. Improper handling of retired batteries causes serious resource waste and environmental pollution, whereas echelon utilization is an important route to recovering their residual value and supporting the “dual-carbon” goal. The prerequisite of echelon utilization is accurate sorting: retired batteries exhibit significant inconsistency in capacity, internal resistance and self-discharge rate, and must be graded according to their performance before being regrouped for further use.[1,2]

Extensive research has been conducted on the sorting and screening of retired batteries. A quick screening approach based on the fuzzy C-means algorithm has been proposed for the second usage of retired lithium-ion batteries[3]; incremental-capacity-curve feature extraction has been used to regroup retired lithium iron phosphate batteries for echelon utilization[4]; support-vector clustering[5], an improved bisecting K-means algorithm combined with fast pulse tests[6], and Gaussian-mixture-model soft clustering based on electrochemical impedance spectroscopy[7] have been applied to the sorting and regrouping of retired cells; and the performance assessment and classification of retired electric-vehicle batteries for energy-storage applications have been demonstrated[8]. Because accurate sorting relies on the characterization of battery performance, rapid sorting and regrouping based on partial charging curves[9], two-stage sorting combining static and dynamic characteristics[10], feature selection combined with multiple clustering[11], and multi-stage deep sorting based on static and dynamic features[12] have been investigated. At the cell level, data-driven methods can predict battery cycle life before capacity degradation becomes pronounced[13], the state of the art in SOH estimation has been systematically reviewed[14], and public battery-ageing datasets[15] provide a reproducible basis for developing and

evaluating sorting algorithms. All of the above methods assume that the sorting model can centrally access all training samples, which is often unrealistic in actual recycling systems.

In practice, retired-battery recycling channels are highly dispersed: testing data are deposited in recycling outlets and third-party testing institutions in different regions, forming “data silos”. Battery testing data also contain commercially sensitive information about recycling channels, so institutions are reluctant to share raw data. Moreover, because the incoming batteries at each node differ in chemistry composition and historical usage, the data are strongly non-IID. Simply pooling the samples for centralized training would let the model be dominated by the distribution of large-sample nodes, whereas training each node locally leaves small-sample nodes under-accurate and makes cross-node sorting criteria inconsistent. Federated learning provides a technical framework for breaking data silos,[16,17] but the classic FedAvg-style uniform one-way aggregation ignores inter-node distribution discrepancies and is unfavourable to small-sample nodes with large distribution bias.

To address these issues, this paper proposes a hierarchical sorting method based on federated mutual learning for non-IID data across multiple testing nodes. The main contributions are as follows. (1) A local hierarchical sorting model that integrates prior grouping with supervised refinement is constructed to accommodate sample heterogeneity across chemistries and historical usage profiles. (2) A model-output knowledge exchange mechanism based on a unified unlabeled reference set is designed, enabling bidirectional cross-node knowledge transfer while keeping raw testing data within each node. (3) A dynamic knowledge weighting and absorption strategy combining inter-node distribution discrepancy with prediction confidence is proposed to suppress the interference of mismatched knowledge between heterogeneous nodes. (4) The effectiveness and stability of the proposed mechanisms are verified through multi-node non-IID simulation, repeated random-seed experiments, hyperparameter analysis, ablation studies, small-sample experiments and comparison against the centralized performance upper bound.

## Results

# Data distribution and convergence behavior

Figure 1a,b shows a t-SNE visualization of the standardized five-dimensional performance features of the 864 training samples. When colored by testing node (Fig. 1a), the samples exhibit clear group structures that nevertheless overlap each other; when colored by true grade (Fig. 1b), grades T1–T4 are arranged in progressive bands, with natural confusion zones at the boundaries between adjacent grades. This indicates pronounced distribution shifts among nodes in the feature space, while the grades themselves display an overall progressive structure, providing intuitive support for the subsequent hierarchical modeling and cross-node knowledge collaboration.

Figure 2 shows the round-by-round convergence of the two federated schemes. Without federation, the global average accuracy is 78.27% and the cross-node consistency is 73.47%; the proposed method

improves steadily over five rounds of mutual learning to 91.10% and 97.13%, respectively, without oscillation. The FedAvg-style one-way aggregation reaches 90.85% in the first round but falls back to 88.07% in the second round, oscillating between 88.07% and 90.85%. This indicates that, under markedly heterogeneous node distributions, uniform one-way aggregation may expose some nodes to global knowledge that mismatches their local distributions, causing performance fluctuation. In round 5 of the proposed method, the four node accuracies range from 88.89% to 93.18% with no weak node, and the small-sample node D improves most notably (66.67% → 88.89%).

## Sorting performance comparison

To examine the necessity of hierarchical supervised modeling, direct clustering on the five-dimensional performance features was first evaluated as a classical baseline. K-means and fuzzy C-means (FCM, c = 4, m = 2) were trained locally at each node and evaluated after optimal label matching with the Hungarian algorithm; the global average accuracies are only 64.26% and 57.08%, respectively (Table 1). Because the sorting grades are characterized by nested rules jointly determined by capacity retention, internal resistance and self-discharge rate, purely geometric clustering cannot fit such decision boundaries, and unsupervised clusters carry no cross-node semantic consistency; the consistency is therefore not reported (marked “—” in Table 1). This confirms the necessity of the “prior grouping + supervised refinement” modeling route.

Table 1 further compares four supervised training schemes: centralized training (B3), local training only (B1), one-way aggregation (B2) and the proposed method. Centralized training merges the samples of all nodes to train a single model and reaches a global average accuracy of 91.54%, which can be regarded as the performance ceiling when testing data can be freely centralized. The proposed method achieves 91.10%, only 0.44 percentage points below this ceiling, although raw testing data never leave any node, whereas B1 reaches only 78.27%. The cross-node consistency of the proposed method is 97.13%, close to the theoretical ceiling of B3 (all nodes share one model, consistency 100%) and comparable to that of B2 (97.73%); however, the proposed method achieves more balanced node-level accuracies. Under B2, node A reaches only 86.25% (5.00 percentage points below the proposed method) and node D 83.33% (5.56 percentage points below), showing that one-way aggregation sacrifices the distributional “minority” nodes in exchange for superficial consistency.

Table 1
Sorting performance of the six schemes.

| Scheme | Node A (%) | Node B (%) | Node C (%) | Node D (%) | Global average (%) | Cross-node consistency (%) |
|---|---|---|---|---|---|---|
| K-means | 63.75 | 60.71 | 65.91 | 66.67 | 64.26 | — |
| FCM | 53.75 | 57.14 | 59.09 | 58.33 | 57.08 | — |
| B1 Local training only | 87.50 | 83.93 | 75.00 | 66.67 | 78.27 | 73.47 |
| B3 Centralized training | 91.25 | 92.86 | 93.18 | 88.89 | 91.54 | 100.00 |
| B2 One-way aggregation | 86.25 | 89.29 | 97.73 | 83.33 | 89.15 | 97.73 |
| Proposed method | 91.25 | 91.07 | 93.18 | 88.89 | 91.10 | 97.13 |

To examine the robustness of these conclusions to randomness, the dataset was regenerated and independently retrained with ten additional random seeds on top of the default seed, giving 11 independent runs in total (Table 2). Across all 11 runs, the proposed method achieves a global average accuracy of 90.17% ± 2.53%, which is on average 2.35 percentage points higher than one-way aggregation (87.82% ± 1.86%) and only 1.12 percentage points below the centralized ceiling (91.29% ± 1.14%). A paired two-sided Wilcoxon signed-rank test (n = 11) shows that the proposed method significantly outperforms one-way aggregation in global accuracy ($W = 2.0$, $p = 0.0029$, effect size $r = 0.90$) and in weakest-node accuracy ($W = 0.0$, $p = 0.0020$, $r = 0.93$); it wins in 10 of the 11 runs and in 9 of the 10 newly added runs. The weakest-node accuracy of the proposed method is 86.69% ± 2.78%, 4.55 percentage points higher than one-way aggregation (82.14% ± 2.91%), indicating that the improvement is most stable for the small-sample node with the largest distribution bias. It should be noted that the cross-node consistency of the proposed method (95.93% ± 0.84%) is slightly lower than that of one-way aggregation (97.12% ± 0.57%): one-way aggregation constrains all nodes with the same averaged distribution, giving higher superficial consistency at the cost of impaired accuracy at the “minority” nodes (see Table 1). The proposed method trades a reasonable consistency cost of no more than two percentage points for significant gains in global accuracy and node-level balance.

Table 2
Statistics of the repeated experiments with multiple random seeds (n = 11).

| Scheme | Global average accuracy (%) | Weakest-node accuracy (%) | Cross-node consistency (%) |
|---|---|---|---|
| Local training only | 83.45 ± 2.80 | 73.77 ± 6.57 | 76.16 ± 3.11 |
| Centralized training (B3) | 91.29 ± 1.14 | 86.49 ± 2.03 | 100.00 ± 0.00 |
| One-way aggregation (B2) | 87.82 ± 1.86 | 82.14 ± 2.91 | 97.12 ± 0.57 |
| Proposed method | 90.17 ± 2.53 | 86.69 ± 2.78 | 95.93 ± 0.84 |

Figure 3 shows the confusion matrices of the proposed method at each node. The per-row recognition rates of grades T1, T2 and T3 are all above 80%. The main confusions occur between adjacent grades (e.g. 36% of the T4 samples at node A are misjudged as T3, and 20% of the T2 samples at node D are misjudged as T1), which is consistent with the nested structure of the grading rules in Table 3, indicating that the model still has difficulty discriminating samples near adjacent thresholds. In future work, the distance between misclassified samples and the grade thresholds could be analysed to determine whether the errors mainly originate from boundary regions, and the thresholds could be refined or a continuous health score could be introduced to alleviate this problem.

## Hyperparameter sensitivity

Figure 4 examines the influence of the four hyperparameters $\lambda_0$, α, τ and η. (1) For $\lambda_0 \in [0.4, 1.0]$, the global accuracy is never lower than 89.8%; the accuracy peaks at $\lambda_0 = 0.6$ (92.25%), while the consistency is best at $\lambda_0 = 0.8$ (97.13%); considering both metrics, $\lambda_0 = 0.8$ is adopted. (2) For $\alpha \le 0.5$ the performance is stable; increasing α further decreases the monotonically consistency (92.78% at α = 1.6), indicating that excessive decay weakens inter-node knowledge sharing. (3) $\tau = 0.4$ is optimal for both metrics; when $\tau \ge 0.7$, the absorbable pseudo-samples drop sharply and both metrics decline. (4) η = 2–3 is optimal; at η = 1 the knowledge quantity is insufficient (global accuracy 82.62%), whereas an excessively large η introduces low-confidence noise. Within the effective parameter ranges examined here, the method exhibits a broad plateau of stable performance, but degrades noticeably when the knowledge absorption strength is too low (η = 1) or the confidence threshold is too high ($\tau \ge 0.7$); the default configuration lies in the stable plateau.

## Ablation study

To verify the contribution of each mechanism, four ablation variants were designed; the results are shown in Fig. 5 and Table 4. Removing the dynamic consistency weight (fixing $\lambda_k = \lambda_0$) reduces the global accuracy to 87.69%, and the accuracy of node D drops sharply by 8.33 percentage points to 80.56%: the

node with the largest distribution bias is assimilated by the global distribution under a “one-size-fits-all” constraint, which directly demonstrates the protective effect of the dynamic weighting. Removing the confidence filter reduces the accuracy to 87.88%, showing that low-confidence pseudo-labels introduce noise. Keeping only one-way knowledge broadcast gives 89.15%, consistent with the analysis in “Sorting performance comparison”. The full method achieves the most balanced node-level performance among all variants.

Table 4
Results of the ablation study.

| Variant | Global average accuracy (%) | Cross-node consistency (%) | Weakest node (D) accuracy change (p.p.) |
|---|---|---|---|
| Full method | 91.10 | 97.13 | — |
| Fixed consistency weight | 87.69 | 96.81 | −8.33 |
| No confidence filtering | 87.88 | 95.56 | −5.56 |
| One-way knowledge broadcast only | 89.15 | 97.73 | −5.56 |
| No federation (local training) | 78.27 | 73.47 | −22.22 |

## Adaptability to small local samples

The local training-set ratio of each node was scanned from 10% to 70% (with the test set adjusted accordingly); the results are shown in Fig. 6. The accuracy of all three schemes increases with the sample size, but the proposed method has the smallest inter-node variation at every ratio: at a training ratio of 20%, the standard deviation is only 0.68 percentage points (4.66 percentage points for local training only); at a training ratio of 30%, the global accuracy of the proposed method is 77.05%, which is 1.66 and 6.89 percentage points higher than one-way aggregation and local training only, respectively. This shows that federated mutual learning not only improves accuracy but also markedly improves inter-node balance in sample-scarce scenarios, which is of direct engineering significance for retired-battery recycling systems with dispersed outlets and limited samples at each site.

## Discussion

This study targets the data silos of retired-battery testing data and the non-IID nature of inter-node distributions, and proposes a hierarchical sorting method based on federated mutual learning. The results support three main findings.

First, bidirectional model-knowledge exchange based on an unlabeled reference set can improve collaborative cross-node sorting performance without directly sharing raw testing samples. In the

simulated scenario of four nodes and 1,080 retired batteries, the global average accuracy rises from 78.27% with local training only to 91.10%, which is 26.84 and 34.02 percentage points higher than the classical clustering baselines (K-means and FCM, respectively), with a cross-node consistency of 97.13%; the node accuracies are balanced (88.89%–93.18%) and the gap to the centralized performance ceiling (91.54%) is only 0.44 percentage points.

Second, distribution-discrepancy weighting and confidence-based knowledge filtering are key to robustness under non-IID conditions. The ablation study shows that removing either mechanism reduces the global accuracy by 2.0–3.4 percentage points, and the small-sample nodes with larger distribution bias are affected more strongly (the largest drop at node D is 8.33 percentage points). This is consistent with the motivation that uniform aggregation schemes such as FedAvg treat statistically dissimilar nodes equally and therefore damage the “minority” nodes, whereas the proposed dynamic weighting and confidence screening suppress mismatched knowledge at both the source and the receiver.

Third, the multi-seed and variable-sample-size experiments show that the proposed method has good stability and small-sample adaptability: with a local training ratio of 20%, it still maintains an average accuracy of 73.74% with an inter-node variation of only 0.68 percentage points. Nevertheless, the performance remains sensitive to the knowledge absorption strength and the confidence threshold, and the choice of the reference set may also influence the quality of the exchanged knowledge; these aspects deserve further investigation.

This study has several limitations. Node dynamics (nodes joining or leaving during operation) are not considered, and although raw data never leave the nodes, model outputs may still leak privacy information, since no formal defence mechanisms such as differential privacy or secure aggregation have been introduced. In addition, the validation is based on simulated data. Future work will validate the method on public battery-ageing datasets and field-measured retired-battery data, and will further investigate secure aggregation, dynamic node management and continuous health scoring mechanisms.

## Methods

# Multi-node non-IID dataset construction

A retired-battery recycling system with four regional testing nodes A–D was simulated. The nodes differ in the chemistry composition of their incoming batteries (lithium iron phosphate LFP / nickel-cobalt-manganese NCM / lithium titanate LTO) and in the composition of historical usage profiles (electric-vehicle power / energy-storage station / low-speed vehicle), with sample sizes of 400, 280, 220 and 180 cells, totalling 1,080 cells and constituting a typical non-IID federated scenario (Table 5). The features of a single cell are organized in two layers: the first-layer prior features consist of nominal voltage, historical depth of discharge and historical average temperature; the second-layer performance features consist of capacity retention rate, internal resistance, monthly self-discharge rate, consistency voltage difference

and predicted cycle life. In the data-generation model, the nominal voltage, cycle-life baseline and resistance baseline of each chemistry follow normal distributions; the capacity retention rate follows a usage-dependent fading distribution, in which severely degraded individuals (capacity retention 12%–34%) account for 10%; in addition, measurement noise is superimposed, including a capacity-repeatability error ($\sigma = 0.02$) and a resistance-instrument indication error ($\sigma = 0.5$ mΩ), to emulate realistic testing noise.

Table 5
Sample composition of the testing nodes.

| Node | Chemistry composition (LFP/NCM/LTO) | Use-profile composition (EV/ESS/LSV) | Sample size (cells) |
|---|---|---|---|
| A | 0.70 / 0.25 / 0.05 | 0.60 / 0.25 / 0.15 | 400 |
| B | 0.15 / 0.75 / 0.10 | 0.75 / 0.10 / 0.15 | 280 |
| C | 0.45 / 0.30 / 0.25 | 0.25 / 0.55 / 0.20 | 220 |
| D | 0.30 / 0.20 / 0.50 | 0.20 / 0.30 / 0.50 | 180 |

At each node, the local data are stratified into a training set and a test set at a ratio of 8:2, giving 864 training cells and 216 test cells in total. The coordination layer additionally constructs an unlabeled reference set R covering the full combination of three chemistries × three usage profiles, with 40 cells per group (360 cells in total), which serves as the vehicle for cross-node knowledge exchange. Figure 7 shows the kernel density estimates of the key performance features of the training samples at each node. The inter-node distribution differences are significant: the capacity retention of node A is concentrated in the range of 0.6–0.85, whereas node D, whose low-speed-vehicle profile accounts for 50%, has a markedly lower distribution centre with a long tail of severely degraded cells. Such pronounced inter-node differences make independent local models prone to performance disparities and also complicate the design of a unified federated model that accommodates all node distributions, which motivates a cross-node collaboration mechanism tailored to non-IID scenarios.

# Sorting grades and evaluation metrics

Referring to industrial grading practice for retired batteries, the retired cells are divided into four sorting grades according to capacity retention rate, internal resistance and monthly self-discharge rate (Table 3). The grading rules are nested: cells failing the additional conditions of T1 or T2 are degraded successively, and cells ultimately failing T3 enter the disassembly and recycling channel.

Table 3
Sorting grade rules for retired batteries.

| Grade | Echelon utilization destination | Capacity retention | Internal resistance | Monthly self-discharge rate |
|---|---|---|---|---|
| T1 | Grid-side energy storage | ≥ 0.60 | ≤ 15 mΩ | ≤ 5% |
| T2 | Frequency-regulation power source | ≥ 0.45 | ≤ 22 mΩ | — |
| T3 | Backup power source | ≥ 0.30 | — | — |
| T4 | Below the above thresholds | — | — | — |

Three evaluation metrics are adopted: (1) node sorting accuracy, the proportion of test samples at a single node whose predicted grade matches the ground-truth grade; (2) global average accuracy, the arithmetic mean of the accuracies of the four nodes; and (3) cross-node consistency, the proportion of reference-set samples for which any two nodes produce identical predictions, which measures the uniformity of the sorting criteria. It should be emphasized that the rules in Table 3 are used to construct unified grade supervision labels; they do not imply that sorting in engineering practice is always performed directly by the complete threshold rules. In actual recycling and testing, each node is affected by differences in testing equipment, testing cost, measurement noise and sample sources, so its characterization of capacity, internal resistance and self-discharge deviates, and the local discriminative models established at different nodes are prone to inconsistency. The focus of this study is therefore not to redefine the grade boundaries, but to investigate collaborative model learning and the harmonization of sorting criteria across non-IID nodes under a unified grading rule.

# Federated mutual learning framework

The architecture of the proposed method is shown in Fig. 8. It consists of K testing nodes and one coordination layer. Node k holds a local dataset $D_k$ and a hierarchical sorting model $M_k$; the coordination layer maintains the unlabeled reference set R and performs knowledge aggregation and weighted distribution. Each round of federated mutual learning comprises three steps: (1) knowledge upload — each node computes the class output probabilities of its local model on R and uploads them; (2) aggregation and weighting — the coordination layer aggregates the knowledge of all nodes weighted by inter-node distribution similarity and forms a weighted knowledge distribution incorporating the dynamic consistency weight of the receiving node; (3) absorption and update — each node filters high-confidence pseudo-label knowledge using a confidence threshold and participates in local retraining with a dynamically weighted replication factor. Throughout the process, only model output probabilities and low-dimensional statistical vectors are exchanged; raw samples never leave their nodes, which protects the data rights of all parties by design.

# Local hierarchical sorting model

The local model adopts a two-layer structure. The first layer standardizes the three-dimensional prior features (nominal voltage, historical depth of discharge and historical average temperature) and applies K-means for coarse-grained grouping. The clusters are not mapped one-to-one onto the nine “chemistry × usage-profile” combinations; instead, the prior features are used to form several sample subdomains with similar chemistry and historical usage characteristics. To determine the number of clusters, the silhouette coefficient, the Calinski–Harabasz (CH) index and the global sorting accuracy after federated training were compared within the range k = 2–6. The sorting performance of k = 2 and k = 3 is close (93.04% versus 91.10%); k = 3 gives a better silhouette coefficient (0.636) and CH index (1358.5) than k = 2 (0.624 and 1260.1), corresponds to the prior structure of the three chemistries and is more interpretable; for k ≥ 4 the clustering becomes fragmented and the global accuracy drops to 87.7%–83.8%. Considering clustering validity, sorting performance and prior interpretability, k = 3 is adopted. The second layer trains a binary decision tree independently within each cluster, using the five-dimensional performance features (capacity retention rate, internal resistance, monthly self-discharge rate, consistency voltage difference and predicted cycle life) as inputs. The split feature of the decision tree is selected with a composite splitting criterion:

$$S(f) = w_1 \cdot G(f) + w_2 \cdot [1 - \mathrm{Gini}(f)] \quad (1)$$

where G(f) is the information gain of splitting on feature f; Gini(f) is the sample-size-weighted Gini index (Gini impurity) of the two child nodes after the split; and $w_1$ and $w_2$ are weighting coefficients, both set to 0.5 in this work. This criterion balances information gain and class purity and is more robust than either single criterion within small-sample clusters. The hierarchical structure brings an additional benefit: the first-layer cluster centres can serve as prior templates, so new samples can be inferred online without retraining.

# Unlabeled reference set and knowledge representation

The coordination layer constructs an unlabeled reference set R covering all chemistries and usage profiles, making the knowledge of all nodes comparable on a unified input. The knowledge representation of node k is the matrix of class output probabilities of its model on R. Compared with directly transmitting raw testing samples, output-probability matrices and low-dimensional statistics substantially reduce the exposure of raw testing data and thus lower the risk of privacy and commercial-information leakage in cross-institution data sharing, while retaining the discriminative-boundary information of the node models, which is sufficient to support cross-node knowledge transfer. It should be noted that this study focuses on the knowledge collaboration mechanism under the constraint that data do not leave their domains; strict privacy-protection techniques such as differential privacy and secure aggregation have not yet been introduced, and the ability of the exchanged outputs to resist attacks such as model inversion remains to be studied.

# Distribution discrepancy and dynamic consistency weight

To characterize inter-node distribution differences, the statistical distribution vector of node k is defined as the concatenation of the mean vector and the standard-deviation vector of its locally normalized second-layer features; the global statistical vector is the sample-size-weighted average over all nodes. The distribution discrepancy of node k is defined as:

$d_k = \|\mu_k - \mu_g\|_2$ (2)

where $\mu_k$ is the statistical distribution vector of node k and $\mu_g$ is the global statistical vector. The dynamic consistency weight is then designed as:

$\lambda_k = \lambda_0 \cdot \exp(-\alpha \cdot d_k)$ (3)

where $\lambda_0$ is the initial weight of the consistency constraint and α is the decay coefficient. The larger $d_k$ is (i.e. the more the incoming distribution deviates from the global distribution), the lower the weight of the knowledge uploaded by that node in the global consensus, which suppresses the interference of nodes with large distribution bias; when receiving knowledge, the node still obtains the weighted fusion of the other nodes, so small-sample nodes are not overwhelmed by large-sample nodes.

# Knowledge absorption and weighted fusion training

For node k, the coordination layer fuses the knowledge of the other nodes weighted by the similarity weight $w_{kj} = \exp(-\alpha \cdot \|\mu_k - \mu_j\|_2)$ to obtain the weighted knowledge distribution targeted at node k. On the node side, a confidence threshold τ is set for knowledge absorption, and only pseudo-labels of reference samples with confidence above τ are retained; the absorbed pseudo-samples participate in local retraining with a replication factor:

$m_k = \text{round}(\lambda_k \cdot c_k \cdot \eta)$ (4)

where $c_k$ is the confidence of the pseudo-sample and η is the knowledge absorption conversion coefficient. The above mechanism is equivalent to a discrete implementation of the overall objective $L = L_{local} + \lambda \cdot L_{KL}$: local samples contribute in full to the local empirical loss, while high-confidence pseudo-samples form the consistency-constraint term with dynamic weights. It should be noted that $\lambda_k$ does not determine whether a node can receive knowledge from other nodes — the received knowledge is formed by the weighted fusion of similarity weights and is available to every node; $\lambda_k$ only controls the relative contribution of high-confidence pseudo-samples in local retraining, preventing nodes with large distribution deviation from being assimilated by the global distribution under an excessively strong global consistency constraint. In this way, federated mutual learning replaces the one-way averaging of FedAvg with a bidirectional “upload–weight–absorb” closed loop, balancing global consistency with node individuality. The default hyperparameters are $\lambda_0 = 0.8$, $\alpha = 0.5$, $\tau = 0.4$ and $\eta = 3$, and the number of federated rounds is $T = 5$.

# Communication and computation overhead

In each round of federated mutual learning, every node uploads the class-probability matrix of its model on the reference set R (360 cells; a 360 × 4 matrix of 32-bit floats, about 5.76 kB) and a 10-dimensional distribution-statistics vector; the weighted knowledge distribution distributed by the coordination layer is likewise a 360 × 4 probability matrix, so the bidirectional communication volume of one node in one round is about 11.5 kB, less than 60 kB cumulatively over five rounds, and the exchange occurs only between the coordination layer and each node without any pairwise node connections. The computation overhead on the node side is dominated by a few decision-tree inferences and local retraining, so the overall computational complexity is low, indicating the potential for deployment and validation on ordinary computing platforms.

# Declarations

### Author contributions

X.W. conceived and supervised the study. Q.F. and L.P. designed the methodology and performed the simulations. H.Y. analysed the results and prepared the figures. Q.L. provided theoretical guidance and revised the manuscript. All authors reviewed and approved the final manuscript.

### Data availability

The simulated datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

### Code availability

The computer code that supports the findings of this study is available from the corresponding author on reasonable request.

### Competing interests

The authors declare no competing interests.

### Funding

This work was supported by the National Key Research and Development Program of China (2024YFE0208100) and the Guizhou Provincial Science and Technology Support Program (Qian Ke He Zhi Cheng [2026] No. 348).

# References

1. Martinez-Laserna, E. et al. Battery second life: Hype, hope or reality? A critical review of the state of the art. *Renew. Sustain. Energy Rev.* **93**, 701–718 (2018).

2. Vignesh, S. et al. State of health (SoH) estimation methods for second life lithium-ion battery—Review and challenges. *Appl. Energy*. **369**, 123542 (2024).
3. Zhang, Y. et al. A quick screening approach based on fuzzy C-means algorithm for the second usage of retired lithium-ion batteries. *IEEE Trans. Transp. Electrif*. **7**, 474–484 (2021).
4. Chen, Z. et al. An efficient regrouping method of retired lithium-ion iron phosphate batteries based on incremental capacity curve feature extraction for echelon utilization. *J. Energy Storage*. **56**, 105917 (2022).
5. Li, C. et al. Regrouping and echelon utilization of retired lithium-ion batteries based on a novel support vector clustering approach. *IEEE Trans. Transp. Electrif*. **8**, 3648–3658 (2022).
6. Zhou, Z. et al. A fast screening framework for second-life batteries based on an improved bisecting K-means algorithm combined with fast pulse test. *J. Energy Storage*. **31**, 101739 (2020).
7. Lai, X. et al. Soft clustering of retired lithium-ion batteries for the secondary utilization using Gaussian mixture model based on electrochemical impedance spectroscopy. *J. Clean. Prod.* **339**, 130786 (2022).
8. Liao, Q. et al. Performance assessment and classification of retired lithium ion battery from electric vehicles for energy storage. *Int. J. Hydrog Energy*. **42**, 18817–18823 (2017).
9. Lai, X. et al. Rapid sorting and regrouping of retired lithium-ion battery modules for echelon utilization based on partial charging curves. *IEEE Trans. Veh. Technol.* **70**, 1246–1254 (2021).
10. Yin, H. et al. A two-stage sorting method combining static and dynamic characteristics for retired lithium-ion battery echelon utilization. *J. Energy Storage*. **64**, 107178 (2023).
11. Liu, T. et al. An enhanced sorting method for retired battery with feature selection and multiple clustering. *J. Energy Storage*. **87**, 111422 (2024).
12. Pan, R. et al. Multi-stage deep sorting strategy for retired batteries based on the clustering of static and dynamic features. *J. Energy Storage*. **99**, 113387 (2024).
13. Severson, K. A. et al. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy*. **4**, 383–391 (2019).
14. Berecibar, M. et al. Critical review of state of health estimation methods of Li-ion batteries for real applications. *Renew. Sustain. Energy Rev.* **56**, 572–587 (2016).
15. Saha, B. & Goebel, K. Battery data set. NASA Prognostics Data Repository (2007). Available at: https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/ (Accessed: 20 September 2026).
16. McMahan, B. et al. Communication-efficient learning of deep networks from decentralized data. In Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 54, 1273–1282 (2017).
17. Li, T. et al. Federated optimization in heterogeneous networks. Proc. Mach. Learn. Syst. 2, 429–450 (2020).

# Figures

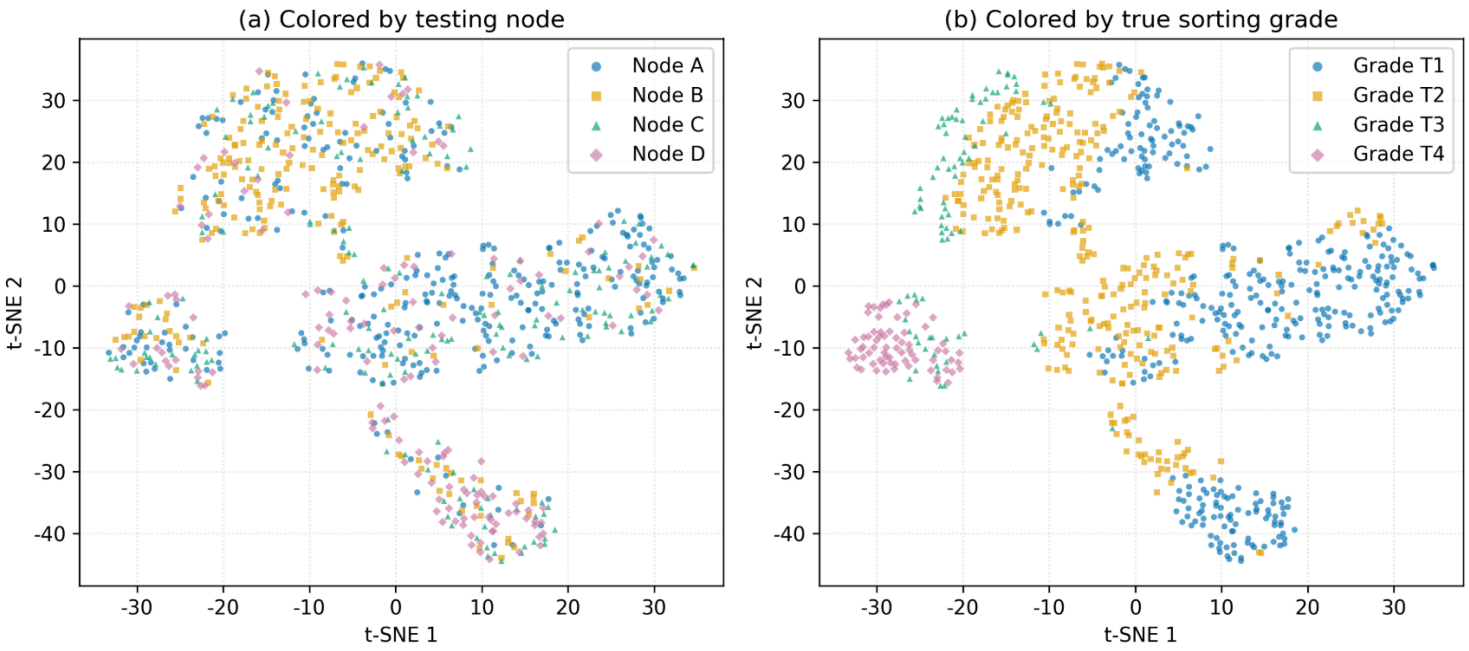



**Figure 1**

**t-SNE visualization of the normalized performance-feature space. (a) Colored by testing node; (b) colored by true sorting grade.**

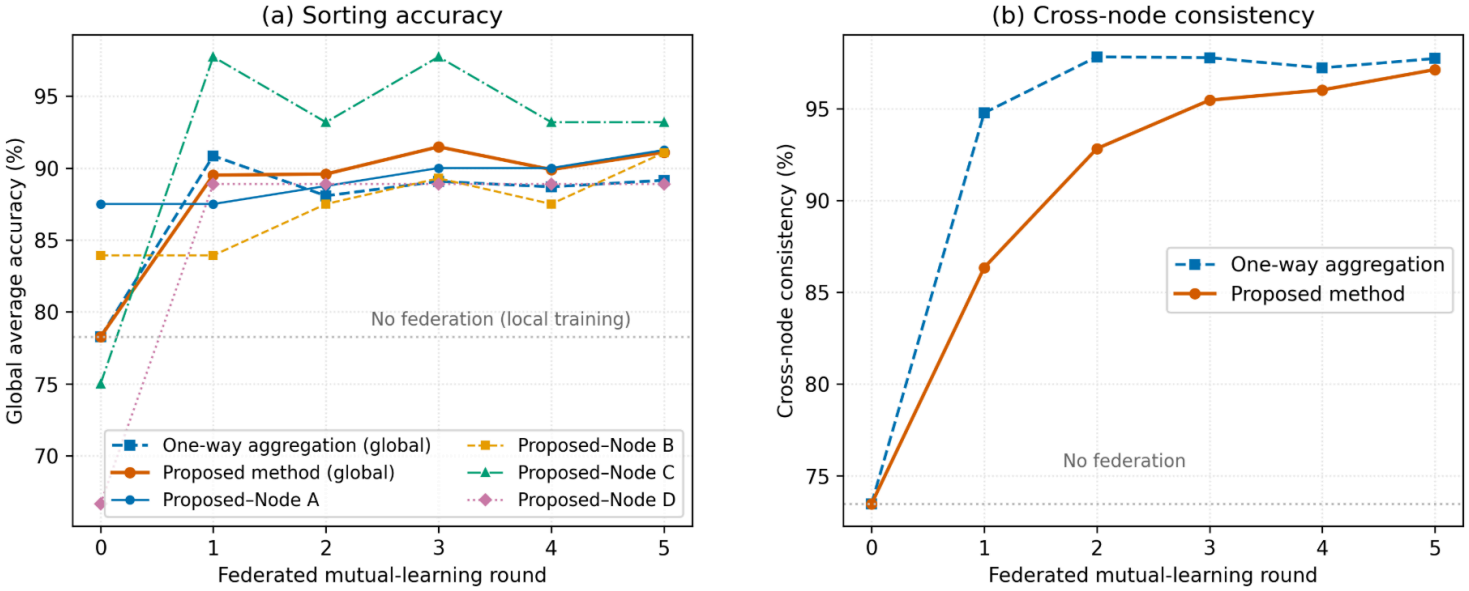
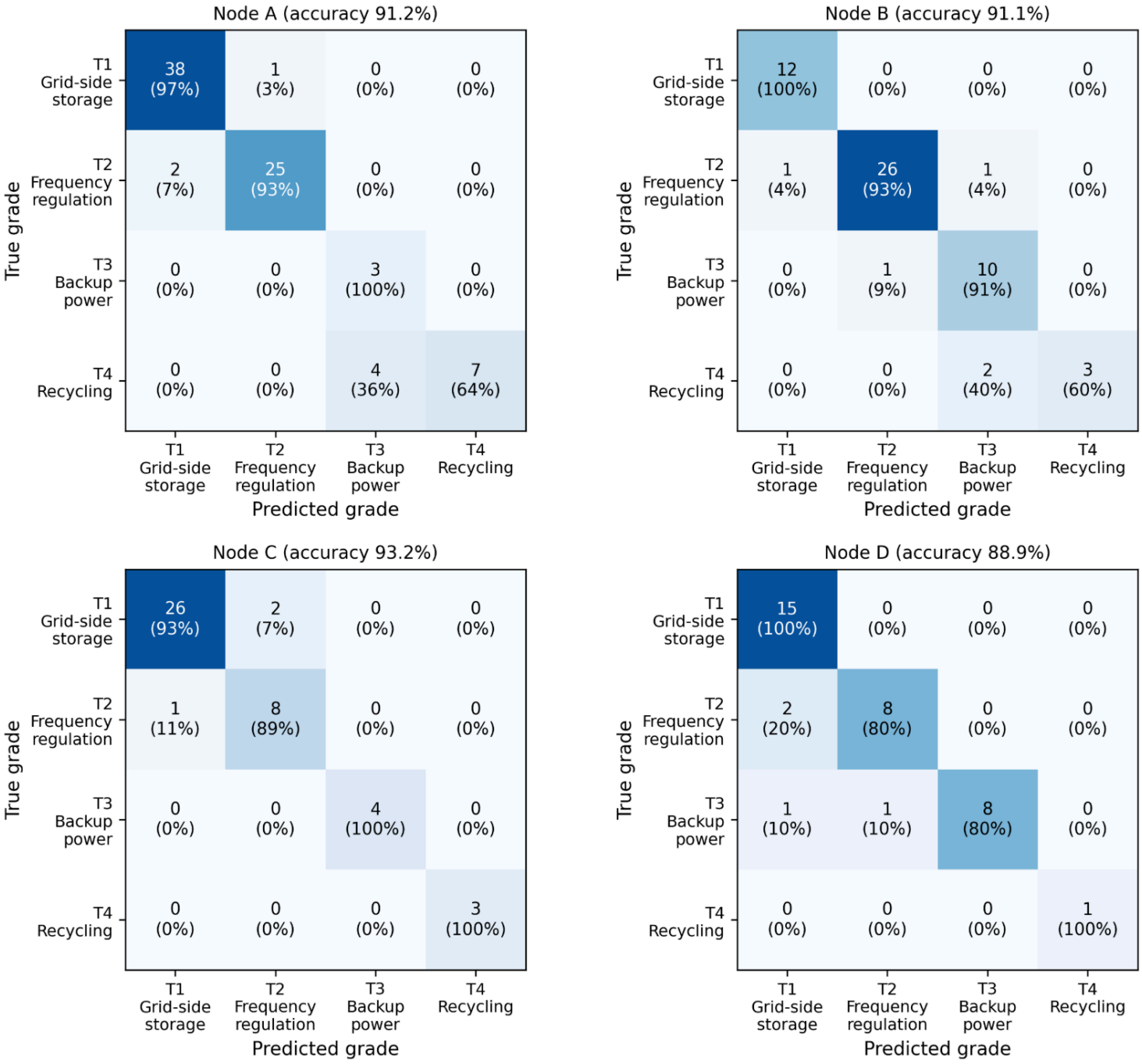


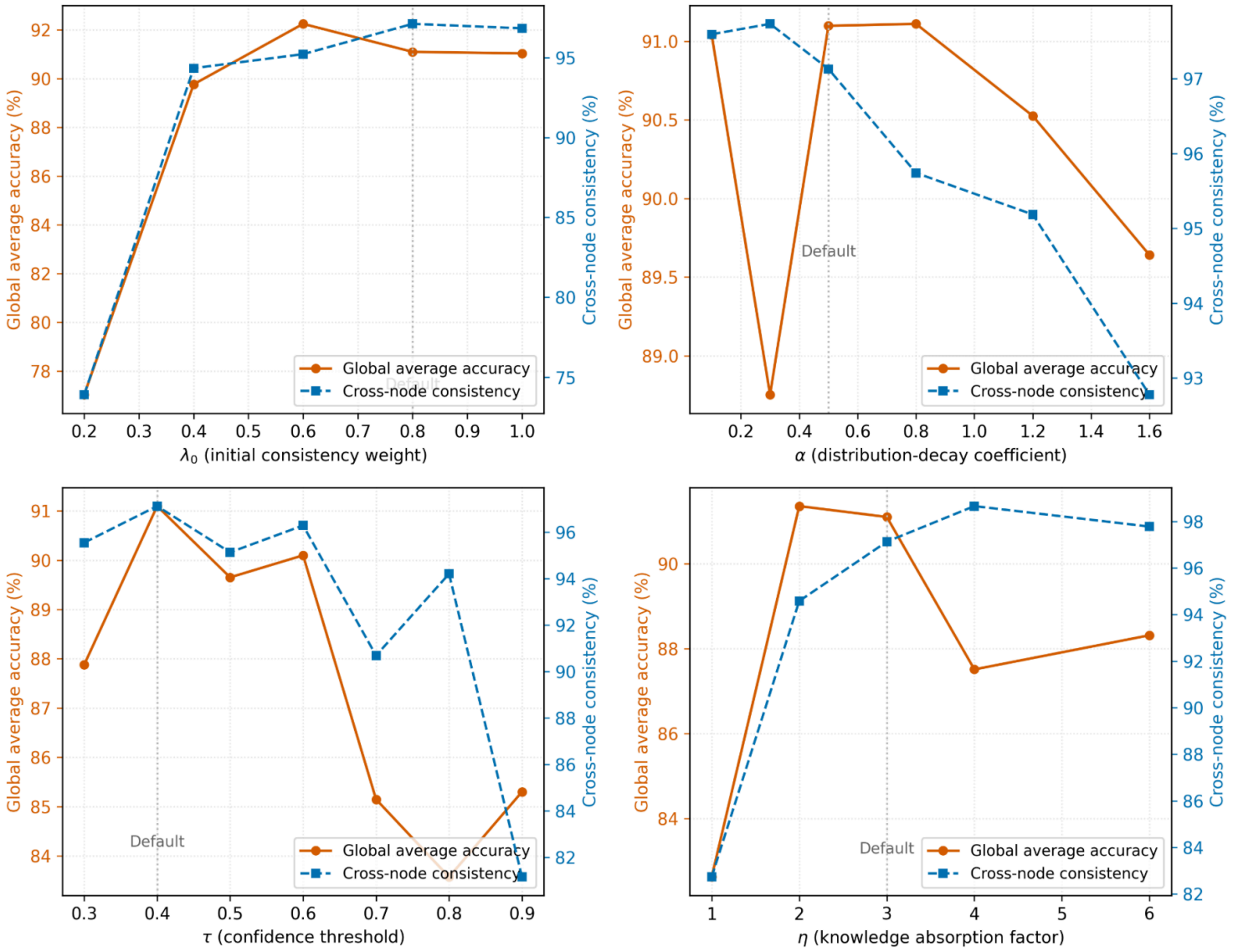

**Figure 2**

**Convergence of the two federated schemes over five mutual-learning rounds. (a) Global average sorting accuracy; (b) cross-node consistency.**




## Figure 3

**Confusion matrices of the proposed method at the four testing nodes.**




## Figure 4

**Sensitivity analysis of the four hyperparameters.**

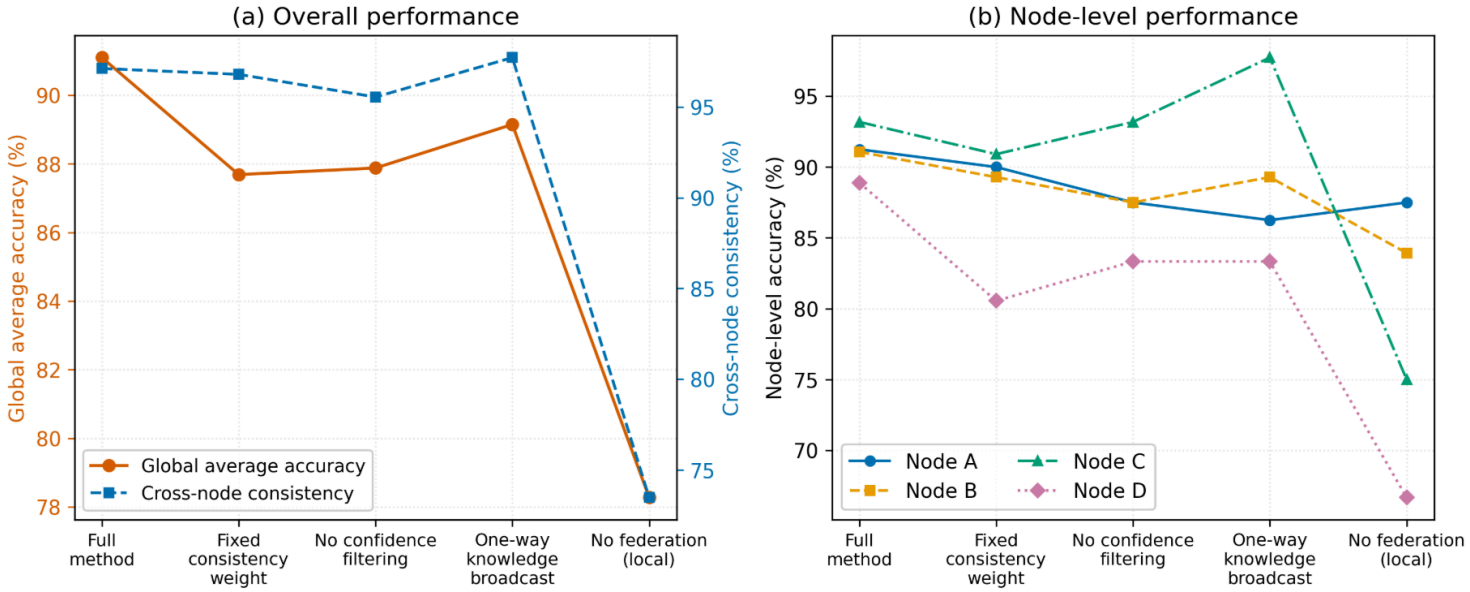



**Figure 5**

**Ablation study results. (a) Overall performance; (b) node-level performance.**

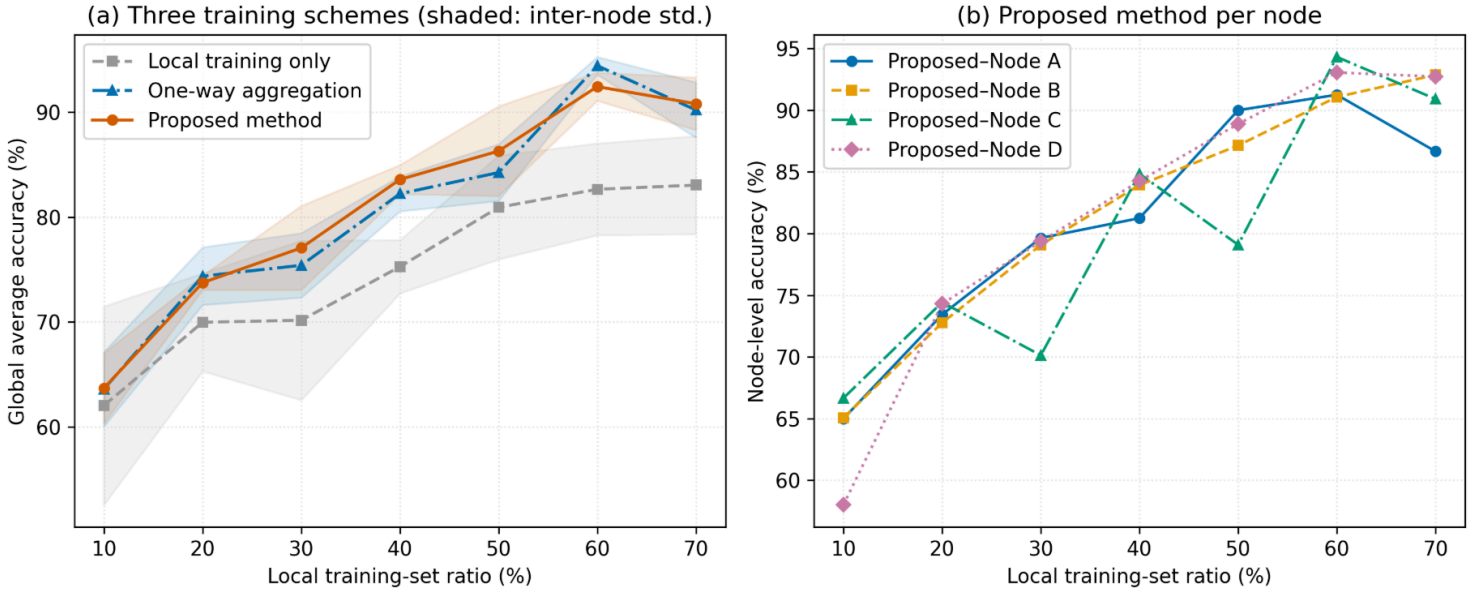



**Figure 6**

**Learning curves under different local training-set ratios. (a) Three training schemes (shaded bands denote the inter-node standard deviation); (b) the proposed method at each node.**

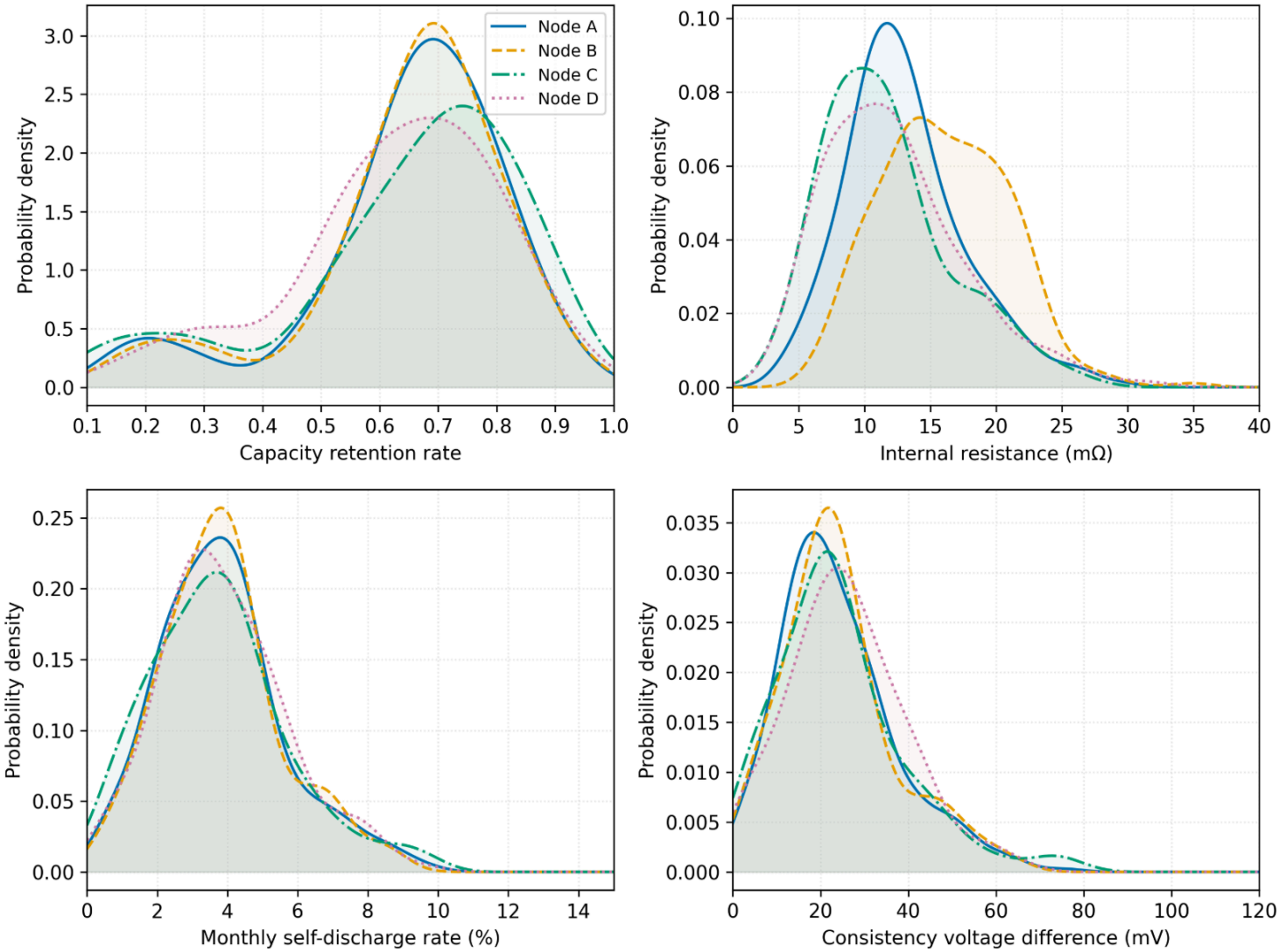



**Figure 7**

**Kernel density estimates of the key performance features of the training samples at each node.**

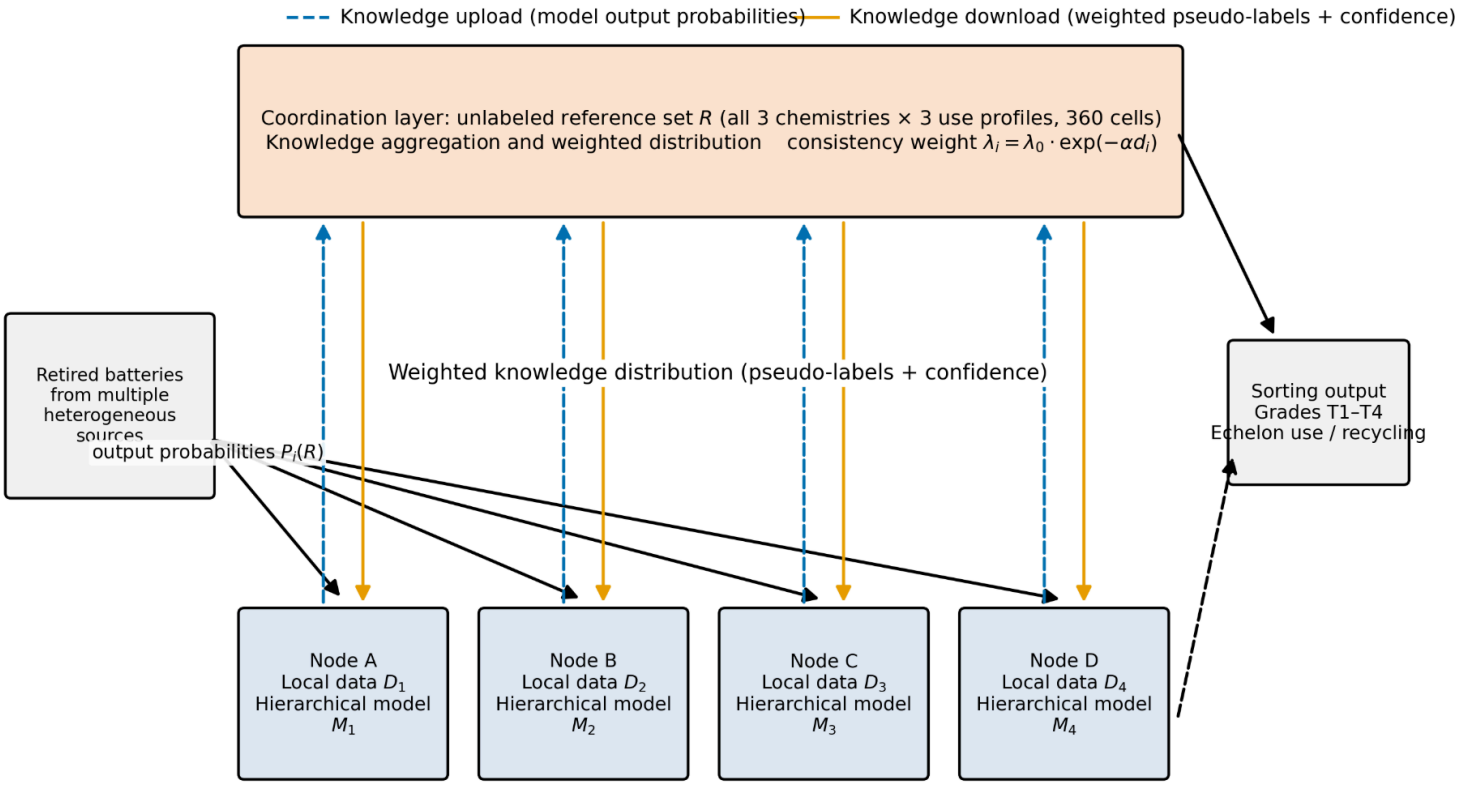



**Figure 8**

**Overall framework of the proposed federated mutual learning hierarchical sorting method.**