
# Second-level heterogeneous retired battery type identification using pulse-test-enabled federated learning with output-level privacy preservation

eTransportation 29 (2026) 100607

Contents lists available at ScienceDirect

eTransportation

journal homepage: www.journals.elsevier.com/etransportation

# Second-level heterogeneous retired battery type identification using pulse-test-enabled federated learning with output-level privacy preservation

Hang Hu [a], Xinghao Huang [b], Chen Liang [b,c], Ziyang Lyu [c], Lin Su [b], Kaisan Li [b], Zhaoye Qin [a,*], Huadong Mo [d,**], Xuan Zhang [b,***], Changfu Zou [e,****], Shengyu Tao [b,e,*****]

[a] Department of Mechanical Engineering, Tsinghua University, Beijing, 100084, China
[b] Institute of Data and Information, Tsinghua University, Shenzhen, 518055, Guangdong, China
[c] School of Mathematics and Statistics, University of New South Wales, Sydney, 2052, Australia
[d] School of Systems and Computing, University of New South Wales, Canberra, 2610, Australia
[e] Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, 41296, Sweden

ARTICLE INFO

*Keywords:*
Retired batteries
Battery type identification
Federated learning
Pulse test

ABSTRACT

The imminent retirement of large numbers of lithium-ion batteries creates an urgent need for efficient reuse and recycling to realize lifecycle economic and environmental benefits. In this context, reliable battery type information is essential for downstream sorting, reuse evaluation, and recycling process selection, because different cathode chemistries are associated with different material values, processing requirements, and safety considerations. However, such information is often unavailable in practice due to long-term label degradation and restricted access to sensitive battery data, especially under non-independent and identically distributed (non-IID) conditions across stakeholders. To address this challenge, we propose a privacy-preserving federated learning framework for retired battery type identification using only second-level pulse-test data collected locally by clients. The framework incorporates an expert-weighted aggregation mechanism, in which client-specific autoencoders quantify local expertise through reconstruction error, while only classification probabilities and expert indices are transmitted for collaboration. This output-level design reduces privacy exposure by avoiding the exchange of raw data, and model parameters. Experiments on a heterogeneous retired-battery dataset containing 8 battery types, over 600 retired batteries, and 10,184 pulse-test records across LFP, LMO, NMC811, and NMC622 cells with nominal capacities ranging from 10 Ah to 68 Ah show that the proposed framework achieves an average classification accuracy of 96.3% across 100 stochastic non-IID partitioning scenarios. It consistently outperforms distributed aggregation baselines, including Average Aggregation, Class-Count Weighting, and Mahalanobis Distance Weighting, while remaining close to the centralized reference performance. By addressing label-distribution and data-quantity skew under a controlled non-IID setting, the proposed framework provides a methodological proof-of-concept for recovering chemistry-relevant battery type information from historically untraceable retired batteries, while external validation under independently collected datasets remains necessary for deployment-level assessment.

* Corresponding authors.
** Corresponding authors.
*** Corresponding authors.
**** Corresponding author.
***** Corresponding author. Institute of Data and Information, Tsinghua University, Shenzhen, 518055, Guangdong, China.
*E-mail addresses:* h-hu24@mails.tsinghua.edu.cn (H. Hu), huang-xh23@mails.tsinghua.edu.cn (X. Huang), liangc22@tsinghua.org.cn (C. Liang), ziyang.lyu@unsw.edu.au (Z. Lyu), sul24@mails.tsinghua.edu.cn (L. Su), lijiesan24@mails.tsinghua.edu.cn (K. Li), qinzy@mail.tsinghua.edu.cn (Z. Qin), huadong.mo@unsw.edu.au (H. Mo), xuanzhang@sz.tsinghua.edu.cn (X. Zhang), changfu.zou@chalmers.se (C. Zou), shengyu.tao@chalmers.se (S. Tao).

https://doi.org/10.1016/j.etran.2026.100607
Received 21 December 2025; Received in revised form 30 April 2026; Accepted 18 May 2026
Available online 20 May 2026

## 1. Introduction

The rapid global proliferation of electric vehicles has led to an exponential surge in the number of lithium-ion battery (LIB) packs in circulation [1,2]. Given the finite service lifespan of these batteries, the industry is now facing an imminent "retirement wave", precipitating significant reuse and recycling challenges [3–5]. As we confront the impending tide of LIB retirements, the development of reliable sorting technologies is increasingly important, not only for reducing environmental risks but also for supporting a sustainable supply of critical battery materials [6,7]. Unfortunately, many retired batteries enter the downstream recycling stream as "black boxes," with limited traceable historical records and incomplete or degraded labeling information [7]. A major reason is that a large proportion of currently circulating electric vehicles and their batteries were manufactured and deployed before mandatory digital labeling regulations took effect [8], creating a substantial stock of legacy batteries with incomplete or non-standardized labeling that is expected to constitute a large share of the retirement stream in the coming decade [9]. In addition, physical QR codes and labels of batteries may deteriorate over long-term service due to mechanical vibrations, thermal stress, and chemical exposure, further reducing the reliability of externally available battery information [10, 11]. Therefore, accurate identification of retired battery types is important for downstream recycling and reuse, particularly because each battery type carries chemistry-related information relevant to sorting and process selection. Historically, research efforts have predominantly focused on identifying retired LIBs based on state-related indicators, such as remaining capacity and internal resistance [12–16]. In contrast, battery type identification in retired batteries remains relatively underexplored. This limitation is particularly important because recycling technologies such as hydrometallurgy and direct recycling are inherently cathode-specific and therefore depend strongly on reliable chemistry information [17–19]. When such information is unavailable, the effectiveness of downstream sorting, process selection, and value recovery can be substantially constrained.

A major challenge in retired battery sorting is data isolation across stakeholders. As noted above, retired batteries are often treated as "black boxes" with limited traceable historical information, which means that recyclers must rely on external testing to assess their states and characteristics. However, although such tests can provide informative voltage and current responses, these data may also reveal proprietary characteristics of battery design, manufacturing, and degradation behavior [20]. Therefore, battery test data are often retained locally because of their sensitive and proprietary nature, especially in the field of electric vehicles [20–23]. This makes direct data sharing difficult in practice and contributes to the formation of cross-enterprise data silos [24–26], which in turn hinder the development of robust battery type-identification models [27,28]. Another challenge lies in the substantial heterogeneity of retired batteries, whose SOC and SOH states often vary considerably due to diverse service histories [29–31][32], thereby leading to pronounced non-independent and identically distributed (non-IID) data distributions across stakeholders. Therefore, a privacy-preserving collaborative framework is needed to balance data utility with practical privacy constraints while supporting robust battery sorting across distributed stakeholders. Tao et al. [33] pioneered the application of federated learning in the battery cathode chemistry classification to address data isolation concerns. Their framework employs Wasserstein-distance voting to aggregate random forest models, thereby mitigating the adverse effects of statistical heterogeneity among stakeholders. However, under severe non-IID scenarios characterized by both label-distribution skew and data-quantity skew, the effectiveness of distance-based aggregation can be compromised. In addition, conventional privacy-preserving techniques such as differential privacy often introduce a trade-off between privacy protection and model performance [34]. Meanwhile, the lack of historical information, including manufacturing specifications, Battery Management System (BMS) records, and operational cycling history, forces recyclers to rely on conventional characterization methods that are often time-consuming and labor-intensive [33]. Therefore, developing a privacy-preserving collaborative framework for retired battery type identification using fast, field-accessible data remains an important research gap.

In this paper, we propose a novel federated inference framework designed to bridge these critical research gaps by simultaneously addressing the trilemma of rapid data acquisition, severe non-IID heterogeneity, and output-level privacy preservation. The evaluation is conducted on a heterogeneous retired-battery dataset comprising 8 dataset-defined battery types, 640 retired batteries, and 10,184 pulse-test records. The dataset spans three major cathode-chemistry families, including LFP, LMO, and NMC, with the NMC group further covering NMC811 and NMC622 cells. It also covers a wide nominal-capacity range from 10 Ah to 68 Ah, with each battery type containing 52–98 retired batteries and 909–1731 pulse-test records. In addition to material and capacity diversity, the dataset contains heterogeneous SOH and initial SOC distributions, reflecting retired cells with different state and aging conditions. Based on this dataset, 100 stochastic non-IID client partitions are further constructed to evaluate the proposed framework under both label-distribution skew and data-quantity skew. First, to overcome the prohibitive time costs and historical data dependency of conventional methods, a pulse testing scheme is introduced for acquiring battery features. The pulse test enables the retrieval of field-accessible, high-fidelity voltage characteristics within seconds, thereby addressing the "black box" problem without requiring historical cycling records. Second, to address the severe non-IID data distributions among simulated stakeholders in reuse and recycling, we develop an "expert-weighted" aggregation mechanism inspired by the Mixture-of-Experts paradigm. Unlike conventional averaging methods that fail under skewed distributions, this mechanism intelligently identifies and amplifies high-quality predictions from "expert" clients (i.e., the collaborators of the federated model) while suppressing noise from naive ones, improving robustness under the controlled non-IID settings considered in this study. Third, we adopt an output-only architecture to reduce privacy exposure during collaboration. By relying solely on abstract probability vectors and reconstruction error scores (expert index), without exchanging gradients, parameters, or statistical metadata, the proposed framework is designed to reduce exposure to reconstruction attacks while maintaining high classification accuracy. Empirical results demonstrate that this holistic approach achieves an average accuracy exceeding 0.95 across 100 distinct non-IID scenarios, significantly outperforming existing aggregation benchmarks. By successfully recovering critical type information from historically untraceable "black box" batteries, this work provides a methodological basis for future privacy-preserving, cross-enterprise battery sorting studies. Nevertheless, independent external validation across laboratories, manufacturers, and testing platforms remains necessary before deployment-level generalization can be established.

The remainder of this paper is structured as follows: Section 2 describes the dataset and preprocessing. Section 3 details the proposed methodology, followed by an in-depth analysis of the results in Section 4. Finally, Section 5 provides the conclusion. In this study, "chemistry" refers to the cathode material (e.g., LFP, LMO, and NMC), whereas "battery type" refers to the specific classification label in the dataset, defined by the combination of chemistry and nominal capacity (e.g., 35Ah_LFP, 68Ah_LFP, and 21Ah_NMC). Accordingly, the prediction target in this study is the dataset-defined battery type rather than cathode chemistry as an isolated causal variable. The learned decision rules may therefore reflect chemistry-relevant information together with correlated attributes such as nominal capacity.

## 2. Dataset description

A significant challenge in handling retired batteries is the common

unavailability of their historical operating data. In this work, we address the worst-case scenario, assuming no prior information is available. Therefore, a reliable testing method is necessary to determine the current state of these batteries. Tao et al. proposed a pulse testing method that involves consecutively subjecting retired batteries to a matrix of alternating positive and negative current pulses [35]. Using the 5000 ms (5s) pulse width as an example, the retired battery is first injected with a 0.5C positive current pulse for that duration. This is immediately followed by a 75s rest period. After this rest, a 0.5C negative current pulse is injected (also for 5000 ms), followed by another 75s rest period. This entire positive-negative injection cycle is then repeated for the remaining amplitudes in the series: 1C, 1.5C, 2C, and 2.5C. Supplementary Table S1 details the comprehensive testing schedule. The turning points of the resulting voltage response are regarded as features. The same pulse-test schedule was applied to all samples in the present dataset, including the pulse width, current-rate sequence, rest interval, voltage sampling procedure, and turning-point extraction rule. This controlled protocol helps ensure that differences in extracted features mainly reflect differences in the measured cell responses under a consistent excitation condition. Specifically, a total of 41 turning points are extracted from the voltage curves, constituting the 41-dimensional feature vector for the model input. More detailed experiment information can be found in Refs. [35–37]. A key advantage of the pulse testing method is its capacity to rapidly acquire state information from retired batteries without relying on any historical data, such as Battery Management System records or operational cycling history.

The physical rationale of these turning-point features is that the pulse-voltage response reflects the battery's integrated electrochemical dynamics under current excitation. When a current pulse is applied, the terminal voltage response can be broadly decomposed into several coupled components. First, the instantaneous voltage jump at the beginning or end of a pulse is mainly associated with ohmic resistance, including ionic/electronic conduction resistance, electrolyte resistance, contact resistance, and interfacial contributions [38,39]. Second, the gradual voltage evolution during the pulse reflects polarization development, which is related to charge-transfer kinetics, lithium-ion diffusion, electrolyte concentration gradients, and concentration polarization [40,41]. Third, the voltage relaxation during the subsequent rest period reflects the redistribution of lithium concentration gradients and the recovery of electrochemical polarization [38,42]. Therefore, the extracted turning points summarize the transient response of the cell across multiple positive and negative current pulses.

In the present work, these pulse-response features are used for battery type identification rather than direct electrochemical parameter estimation. Different battery types, defined by their cathode chemistry and nominal capacity, can exhibit different voltage-response patterns under the same pulse protocol because of differences in open-circuit-voltage characteristics, polarization behavior, and relaxation dynamics. Therefore, the 41 extracted turning points can be regarded as compact empirical proxies of chemistry-relevant electrochemical response, rather than chemistry-exclusive descriptors. These voltage responses may also be affected by nominal capacity, aging state and SOC/SOH distribution. This is why the present work interprets the model output as battery-type identification under the given dataset distribution, rather than as isolated cathode-chemistry identification.

Accordingly, this work selected the pulse testing method, specifically using a 5000 ms pulse width, to obtain the state of 8 types of retired batteries. Here, the two NMC battery types differ not only in nominal capacity but also in cathode stoichiometry: the 15Ah cells are based on NMC811, whereas the 21Ah cells are based on NMC622. As a result, 41 features can be obtained for each sample. The sample size for each battery type is detailed in Table 1. Furthermore, the state-of-health (SOH) and initial state-of-charge (SOC) distributions of the dataset are shown in the supplementary material (Fig. S1).

**Table 1**
Detailed statistical description of the dataset, including sample size and battery quantity for each battery type.

| Cathode Chemistry | Nominal Capacity (Ah) | Battery Quantity | Data Sample Size |
|---|---|---|---|
| LFP | 35 | 56 | 909 |
| | 68 | 96 | 1461 |
| LMO | 10 | 95 | 1523 |
| | 24 | 64 | 1018 |
| | 25 | 96 | 1396 |
| | 26 | 98 | 1731 |
| NMC 811 | 15 | 83 | 1223 |
| NMC 622 | 21 | 52 | 923 |
| Total | - | 640 | 10,184 |

## 3. Methodologies

As noted above, the classification of retired batteries poses a critical dual challenge: achieving high accuracy while rigorously safeguarding data privacy and security. A Federated Learning (FL) framework is particularly well-suited to address this problem [40–42].

. Its decentralized architecture enables collaborative training on a robust global model without requiring stakeholders to exchange or centralize their sensitive, raw battery data. In our work, the FL framework is employed to facilitate collaborative classification of retired batteries among multiple recycling stakeholders. The overall flowchart of the classification process is illustrated in Fig. 1. First, the dataset comprising eight battery types was partitioned among the clients to construct a controlled statistical non-IID setting. This setting was designed to evaluate the proposed aggregation mechanism under label-distribution skew and data-quantity skew, rather than to reproduce all sources of real cross-enterprise heterogeneity. This distributed data was treated as proprietary to each recycling stakeholder (i.e., client) and used to train their respective client model. Each client model consists of two sub-models: (1) a classification sub-model that classifies retired batteries using pulse voltage features and outputs probability distributions over the types, and (2) a judgment sub-model that evaluates whether the client is an “Expert Client” and assigns “Expert index” weights. Finally, the classification outputs from all client models are aggregated using the “Expert index” weights. The final decision is determined by selecting the type with the maximum probability from the weighted result. Throughout this process, client models operate entirely within their local data silos. Neither the corresponding statistical properties (e.g., mean, variance) nor the model information (e.g., gradients and weights) is ever exchanged, helping preserve each client's data privacy.

### 3.1. Data distribution and client construction

As detailed in Section 2, each sample in the dataset $D$ is represented by a feature-label pair $(x, y)$. The feature vector $x \in R^{41}$ is a 41-dimensional representation of voltage features obtained via the pulse testing method, and $y \in \mathcal{Y}$ is the corresponding battery type label, where $\mathcal{Y}$ is the set of all eight unique battery types. A key characteristic of this dataset is the absence of any battery historical information. For model training and evaluation, $D$ is partitioned into a training set $D_{train}$ and a testing set $D_{test}$ following an 8:2 ratio for each battery type. To prevent data leakage, this partition is performed at the battery ID level, ensuring that all samples from a single battery are confined exclusively to either $D_{train}$ or $D_{test}$.

To evaluate the proposed framework under controlled client-level statistical heterogeneity, we designed a structured data partitioning methodology to distribute the battery training set $D_{train}$ among $N$ clients, intentionally introducing two primary forms of non-IID data distribution (see Algorithm 1). In practical battery recycling and reuse ecosystems, a stakeholder rarely receives retired batteries from a perfectly homoge-

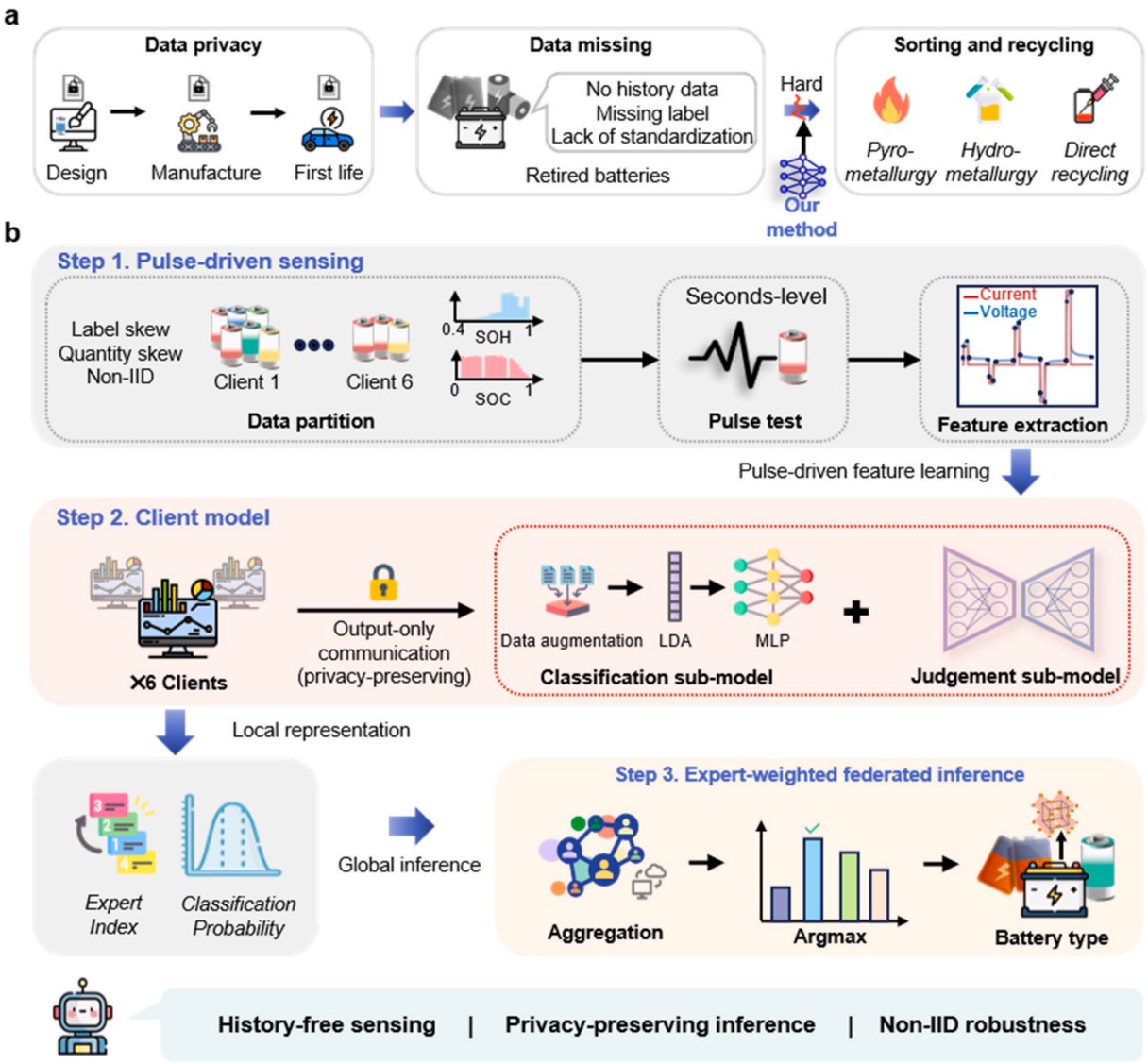


**Fig. 1.** Schematic overview of the proposed privacy-preserving federated framework for retired battery classification. (a) Application context and motivation for privacy-preserving battery type identification from historically untraceable retired batteries. (b) The framework integrates rapid pulse testing for historical-data-free feature acquisition (turning points) with a novel expert-weighted aggregation mechanism. Local clients utilize a dual-stream architecture (comprising classification and judgment models) to generate classification probability and "Expert Index", enabling the central server to achieve robust global decision-making under non-IID conditions with output-level privacy preservation.

neous source. Instead, batteries typically originate from multiple upstream suppliers, vehicle platforms, and collection channels, leading to naturally heterogeneous local inventories. Therefore, the client construction in this work assumes that each client possesses at least three battery types. First, label distribution skew is established. Each client $C_i$ is assigned a variable number of battery types $k_i$, which is drawn from a uniform distribution $U(T_{\min}, T_{\max})$, where $T_{\min}$ and $T_{\max}$ represent the minimum and maximum number of types per client, respectively. Subsequently, $k_i$ unique types are sampled without replacement from the global set of all available battery types and assigned to the client $C_i$. Second, data quantity skew is induced during the sample allocation phase. For each unique battery type $t$, the set of all corresponding data samples $D_t$ is partitioned only among the subset of clients designated to receive that type. The total number of samples $|D_t|$ is divided among the $M_t$ eligible clients using a randomized partitioning algorithm. This algorithm ensures that each eligible client receives a random quantity of samples while enforcing a minimum allocation threshold $size_{min}$ to guarantee minimum participation. Algorithms 1 and 2 outline the pseudo-code for the above data partitioning process. This dual-skew data partitioning methodology ensures that each client's local dataset $D_i$ is heterogeneous, differing from other clients in both its battery type composition and its data volume (i.e., the number of samples per type).

To quantitatively characterize the non-IID data heterogeneity, we first represent each client $C_i$'s local data distribution as a vector $v_i \in R^8$, where each element $v_{i,c}$ corresponds to the number of battery clients $C_i$ that possesses for type $c$. The pairwise similarity $S_{i,j}$ between any two clients $i$ and $j$ is then computed using Cosine Similarity. The formal definition of Cosine Similarity is the cosine of the angle between the two original vectors as presented in Equations (3)–(1):

$$S_{i,j} = \frac{v_i \cdot v_j}{\|v_i\|_2 \|v_j\|_2}, \tag{3.1}$$

This process yields a symmetric $N \times N$ similarity matrix $S$, where $S_{i,i} = 1$ and $S_{i,j} \in [0, 1]$. This matrix provides a comprehensive, quantitative map of inter-client distributional similarity. The results of the partitions are presented in the results section. In practical cross-enterprise deployment, additional heterogeneity may arise from differences in testing equipment, preprocessing protocols, temperature

conditions, initial SOC distributions, contact fixtures, sampling resolution, and battery sources. These factors are beyond the scope of the present partitioning strategy and are further discussed as limitations in Section 4.7.

| Line | Algorithm 1 Dual-Skew Non-IID Data Partitioning |
|---|---|
| | **Input:** $D_{train}$ (Training set), $N$ (Number of clients), $T_{\min}$, $T_{\max}$ (Min/max types per client), $size_{min}$ (Minimum partitioning threshold) |
| | **Output:** $\{D_1, D_2, \ldots, D_N\}$ (A set of $N$ client datasets) |
| 1: | **Procedure:** |
| 2: | Initialize client datasets $D_i \leftarrow \varnothing$ **for** $i = 1$ **to** $N$ |
| 3: | Initialize client type map TypeMap$[i] \leftarrow \varnothing$ **for** $i = 1$ **to** $N$ |
| 4: | $\mathscr{Y} \leftarrow$ GetUniqueTypes$(D_{train})$ *//Get all available battery types* |
| 5: | |
| 6: | *//Step 1: Assign Types (Label Distribution Skew)* |
| 7: | **for** $i = 1$ **to** $N$ **do** |
| 8: | $k_i \leftarrow$ RandomInt$(T_{min}, T_{max})$ |
| 9: | TypeMap$[i] \leftarrow$ SampleWithoutReplacement$(\mathscr{Y}, k_i)$ |
| 10: | **end for** |
| 11: | |
| 12: | *//Step 2: Allocate Samples (Data Quantity Skew)* |
| 13: | **for** each type $t \in \mathscr{Y}$ **do** |
| 14: | $N_t \leftarrow \{i \mid t \in$ TypeMap$[i]\}$ *//Find eligible clients for type t* |
| 15: | $D_t \leftarrow \{(x, y) \in D \mid y = t\}$ *//Get all data for type t* |
| 16: | |
| 17: | *//Handle edge case: type t was not assigned to any client* |
| 18: | **if** $N_t$ is empty **then** |
| 19: | $i_{rand} \leftarrow$ RandomInt$(1, N)$ |
| 20: | $N_t \leftarrow \{i_{rand}\}$*//Randomly assign to one client* |
| 21: | **end if** |
| 22: | |
| 23: | $n_t \leftarrow |N_t|$ |
| 24: | $d_t \leftarrow |D_t|$ |
| 25: | |
| 26: | *//Handle edge case: not enough data for all eligible clients* |
| 27: | **if** $d_t < n_t \times size_{min}$ **then** |
| 28: | $n_t' \leftarrow \max(1, \lfloor d_t / size_{min} \rfloor)$ *//Reduce client count* |
| 29: | $N_t \leftarrow$ SampleWithoutReplacement$(N_t, n_t')$ |
| 30: | $n_t \leftarrow |N_t|$ |
| 31: | **end if** |
| 32: | |
| 33: | *//Get random quantities for the (now finalized) eligible clients* |
| 34: | $Q \leftarrow$ GetRandomPartitions$(d_t, n_t, size_{min})$ **//**See Algorithm 2** ** |
| 35: | Shuffle$(D_t)$ |
| 36: | DataPartitions$\leftarrow$Split$(D_t, Q)$ *//Split $D_t$ into chunks based on Q* |
| 37: | |
| 38: | *//Distribute data to clients* |
| 39: | **for** $j = 1$ **to** $n_t$ **do** |
| 40: | $i \leftarrow N_t[j]$ |
| 41: | $D_i \leftarrow D_i \cup$ DataPartitions$[j]$ |
| 42: | **end for** |
| 43: | **end for** |
| 44: | **return** $\{D_1, D_2, \ldots, D_N\}$ |

## 3.2. Client model

Each client model consists of two components: a classification sub-model and a judgment sub-model. Each client model is trained solely on its respective local dataset. During the classification process, the input features of a sample are fed into every client model. Each client then generates two outputs: a classification probability from its classification sub-model, an "Expert Index" weight from its judgment sub-model.

| Line | Algorithm 2 GetRandomPartitions |
|---|---|
| | **Input:** $S$ (Total sum, $d_t$), $k$ (Number of bins, $n_t$), $m$ (Minimum size per bin, $size_{min}$) |
| | **Output:** $Q$ (A list of $k$ integers summing to $S$, each $\geq m$) |
| 1: | **Procedure:** |
| 2: | |
| 3: | $S_{base} \leftarrow k \times m$ *//The sum allocated for minimums* |
| 4: | $S_{remain} \leftarrow S - S_{base}$ *//The remaining sum to be distributed randomly* |
| 5: | |
| 6: | *//Use "stars and bars" variant to distribute $S_{remain}$ into k bins* |
| 7: | $S_{dist} \leftarrow S_{remain} + k$ |

(*continued on next column*)

(*continued*)

| Line | Algorithm 2 GetRandomPartitions |
|---|---|
| 8: | $P \leftarrow$ SampleWithoutReplacement$(\{1, \ldots, S_{dist} - 1\}, k - 1)$ *//Choose $k-1$ cut points* |
| 9: | $C \leftarrow$ Sort$(\{0\} \cup P \cup \{S_{dist}\})$ |
| 10: | |
| 11: | $A \leftarrow$ Diff$(C) - 1$ *//Calculate sizes of k partitions (summing to $S_{remain}$)* |
| 12: | $Q \leftarrow m + A$ *//Add the minimum size m back to each partition* |
| 13: | |
| 14: | Shuffle$(Q)$ *//Ensure random assignment order* |
| 15: | **return** $Q$ |

### 3.2.1. Classification sub-model

The classification sub-model operates on the 41-dimensional pulse voltage features ($x \in R^{41}$) to generate the corresponding classification probabilities.

The training of this sub-model comprises a two-stage pipeline designed to handle the inherent data imbalance and high-dimensional feature space of each client's local data. First, the non-IID nature of the client data introduces a significant skew in the quantity distribution, resulting in a severe class imbalance. The number of samples per battery type may vary significantly, resulting in local datasets that are heavily biased toward the majority classes. This imbalance may cause the model to overlook the minority types, resulting in poor performance and generalization for those classes. Thus, a data augmentation strategy is introduced. This method oversamples all minority battery types to match the sample count of the majority type. New samples are synthesized by adding correlated noise to randomly selected instances. This noise is drawn from a multivariate normal distribution ($\mathcal{N}(0, \Sigma)$), with a covariance matrix $\Sigma$ empirically calculated from the features of that specific battery type. This approach is intended to preserve the inherent correlations among the 41 pulse test features, thereby helping ensure that the synthetic samples retain the physical characteristics of the original voltage pulses. Second, the raw 41-dimensional features $x \in R^{41}$ may be subject to the curse of dimensionality. This high dimensionality often introduces significant redundancy and uninformative noise, which can impede model performance by increasing the risk of overfitting and obscuring the underlying discriminative patterns. Therefore, a Linear Discriminant Analysis (LDA) model is trained on the augmented dataset to project $x$ into a lower-dimensional, class-separative feature space. The LDA model is a supervised dimensionality reduction technique. Its core objective is to find a lower-dimensional feature subspace that maximizes separability between classes. It achieves this by finding a linear projection that simultaneously maximizes the variance between different classes while minimizing the variance within each class. In this work, the dimensionality of this latent space is set to $k_i - 1$, where $k_i$ is the number of unique battery types present in the client $i$'s local dataset. This design is consistent with the practical scenario considered in this work, where each client is assumed to hold a heterogeneous battery inventory collected from multiple sources. In the rare extreme case of a single-type client, where LDA is not applicable, the dimensionality-reduction step can be omitted and the original feature representation can be used directly. Finally, a Multi-Layer Perceptron classifier is trained exclusively on these $k_i - 1$ LDA-transformed features. The classification sub-model is optimized by minimizing the Cross-Entropy Loss, which is the standard objective function for multi-class classification tasks. During inference, this trained classifier then processes the transformed features to yield the final classification probabilities $P_i$ for each client. The parameter of the classification sub-model is detailed in Table 2. This two-stage approach allows that the classifier operates on a balanced and highly discriminative feature set, optimized for the specific data distribution of its client.

### 3.2.2. Judgment sub-model

Due to the significant skew in label distribution, a client's model may be entirely naive to certain battery types encountered during inference.

**Table 2**
The parameters of the classification sub-model.

| Component | Parameter | Value |
|---|---|---|
| Dimensionality Reduction (LDA) | n_components | $k_i - 1$ |
| | solver | SVD |
| Classifier (MLP) | hidden_layer_sizes | (100, 100, 100, 100) |
| | activation | RELU |
| | solver | ADAM |
| | learning_rate_init | 0.01 |
| | max_iter | 300 |

**Table 3**
The parameters of the judgement sub-model.

| Component | Parameter | Value |
|---|---|---|
| Kernel PCA (KPCA) | n_components | 20 |
| | kernel | RBF |
| | gamma | None |
| Auto-encoder | Encoder_layer_sizes | (32,16) |
| | Decoder_layer_sizes | (16,32) |
| | activation | RELU |
| | solver | AdamW |
| | learning_rate_init | 0.01 |
| | max_iter | 300 |

In such cases, the client's output for that sample is unreliable. Therefore, conventional aggregation methods are suboptimal because they fail to account for heterogeneous client-level expertise, resulting in degraded classification performance. To overcome this limitation, we propose a judgment sub-model that operates in parallel with the classification sub-model. Its function is to quantify the reliability of the client's output for a given input by generating a weight, the 'Expert Index'. The 'Expert Index' is derived from the reconstruction error of a client-specific autoencoder. The underlying principle is that an autoencoder, trained exclusively on a client's local data distribution, will learn to reconstruct familiar, in-distribution samples (i.e., battery types present in the client's dataset) efficiently, resulting in low reconstruction error. Conversely, when presented with an out-of-distribution sample (i.e., battery types absent from the client's dataset), the autoencoder will struggle to reconstruct it, yielding a high reconstruction error. Therefore, the 'Expert Index' is formulated as the inverse of this reconstruction error, ensuring that a higher index value directly quantifies greater client 'expertise' on a given sample, making it an appropriate reliability weight for the final aggregation procedure.

The judgement sub-model also operates on the 41-dimensional pulse voltage features $x \in R^{41}$. In fact, the judgment sub-model is an ensemble of type-specific autoencoders, with one autoencoder trained for each unique battery type present in the client's local dataset. The training pipeline for this ensemble is a multi-stage process designed to create robust representations. First, the client's dataset is balanced via a simple oversampling strategy (i.e., random duplication), ensuring each minority type has sufficient samples. This augmentation method is intentionally distinct from the Gaussian noise-based approach used for the classification sub-model. The rationale for this choice is that the objective for each autoencoder is not generalization (unlike the classification sub-model), but rather to intentionally overfit to the precise latent structure of the assigned battery type. This induced overfitting is advantageous here, as it enhances the autoencoder's discriminability. A model that is "over-specialized" on its own data will be more "brittle" and will produce a significantly higher reconstruction error when confronted with out-of-distribution samples, thus making it a more effective novelty detector. Second, the 41-D features $x$ are projected into a 20-dimensional non-linear feature space using Kernel Principal Component Analysis (KPCA) with an RBF kernel. Following this projection, the features for each type are standardized using the Z-score method. Finally, a separate autoencoder with a latent bottleneck dimension of 10 is trained on the processed data for each battery type by minimizing the L1 reconstruction loss (Mean Absolute Error). This architecture ensures that each autoencoder becomes a specialist on the latent structure of its assigned type. During inference, a sample's reconstruction error is evaluated against all autoencoders in the ensemble, and the minimum of these errors is used to derive the final 'Expert Index', thus quantifying the sample's familiarity with the client's overall data distribution. The subsequent aggregation procedure is detailed in the sub-section that follows. The parameters of the judgment sub-model are detailed in Table 3.

### 3.2.3. Aggregation method

The fundamental objective of the aggregation procedure is to intelligently fuse the heterogeneous and specialized expertise from each client model, achieving a robust global consensus that transcends individual limitations, all while strictly preserving the data privacy of each client.

The final aggregation procedure is a weighted-averaging scheme, executed on a central server, that leverages the 'Expert Index' to produce a single, robust global prediction. For a given test sample $x_{test}$, each of the $N$ clients generates and transmits two distinct outputs to this central aggregator: (1) a classification probability vector $P_i(x_{test}) \in R^{\mathrm{K}}$ from its classification sub-model, where K is the total number of battery types, and (2) a scalar reliability score $e_i(x_{test})$ from its judgment sub-model. This score $e_i(x_{test})$ is formally defined as the minimum reconstruction error generated by the client's ensemble of $k_i$ type-specific autoencoders, where $k_i$ is the number of unique battery types present in that client's local dataset.

Upon receiving these data at the central server, the set of client-level reliability scores $\{e_1(x_{test}), \cdots, e_N(x_{test})\}$ is transformed into a normalized vector of 'Expert Index' weights $\{w_1, \cdots, w_N\}$. This transformation is achieved using the temperature-controlled softmax function detailed in Equations (3)–(2). The use of an extremely low temperature ($T = 0.05$) is a critical design choice. It "sharpens" the softmax output, causing the function to assign sufficiently large weights to those "expert" clients with small reconstruction errors, while effectively silencing the unreliable, non-expert clients.

$$w_i = \frac{\exp\left(-\frac{e_i(x_{test})}{T}\right)}{\sum_{j=1}^{k_i} -\frac{e_j(x_{test})}{T}}, \tag{3.2}$$

Finally, the global probability vector $P_{global}(x_{test})$ is computed as the weighted sum of all individual client probability vectors, detailed in Equations (3)–(3). The final predicted class is the one corresponding to the argmax of this aggregated global probability vector $P_{global}(x_{test})$.

$$P_{global}(x_{test}) = \sum_{i=1}^{N} w_i P_i(x_{test}). \tag{3.3}$$

The key objective of the aggregation procedure is to reduce privacy exposure by limiting exchanged information to output-level signals. This is achieved because the entire communication relies solely on the transmission of abstract, high-level model outputs, i. e, the probability vector $P_i(x_{test})$ and the scalar error score $e_i(x_{test})$. Crucially, this stands in stark contrast to other common paradigms, both of which introduce significant privacy risks. Standard federated training (e.g., FedAvg) mandates the sharing of model parameters or gradients, which are known to be vulnerable to data reconstruction attacks. Similarly, conventional aggregation methods (e.g., distance-based approaches) often require clients to transmit intermediate feature representations or distance metrics, which can also be exploited to infer properties of the local data. In our approach, at no point are the local raw data, their underlying statistical properties, or any model internals (parameters or gradients) disclosed to the central server. This "output-only" architecture

reduces the server's reliance on local data by restricting communication to abstract model outputs, thereby helping preserve client privacy and reducing exposure to privacy attacks such as reconstruction and inversion attacks.

### 3.3. Evaluation metrics

To evaluate the performance of the proposed framework, a comprehensive set of metrics is employed, including accuracy, precision, recall, and the F1-score. The metrics are detailed below:

$$\text{Accuracy} = \frac{\sum_{i=1}^{m} TP_i}{Nm}, \tag{3.4}$$

$$\text{Precision}_{\text{macro}} = \frac{1}{m}\sum_{i=1}^{m} \text{Precision}_i, \tag{3.5}$$

$$\text{Precision}_{\text{weighted}} = \sum_{i=1}^{m} (w_i \times \text{Precision}_i), \tag{3.6}$$

$$\text{Recall}_{\text{macro}} = \frac{1}{m}\sum_{i=1}^{m} \text{Recall}_i, \tag{3.7}$$

$$\text{Recall}_{\text{weighted}} = \sum_{i=1}^{m} (w_i \times \text{Recall}_i), \tag{3.8}$$

$$F1_{\text{macro}} = \frac{1}{m}\sum_{i=1}^{m} F1_i, \tag{3.9}$$

$$F1_{\text{weighted}} = \sum_{i=1}^{m} (w_i \times F1_i). \tag{3.10}$$

where, $\text{Precision}_i = \frac{TP_i}{TP_i+FP_i}$, $\text{Recall}_i = \frac{TP_i}{TP_i+FN_i}$, $F1_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$, and $w_i = \frac{n_i}{N_{test}}$.

Specifically, *TP* denotes the number of positive samples correctly classified, while *FP* represents the number of negative samples mistakenly identified as positive. Conversely, *TN* are the negatives correctly rejected, and *FN* are the positives incorrectly missed. Here, $m$ represents the total number of battery types, which is equal to 8 in this work, $i$ corresponds to a battery type, $n_i$ is the number of samples for that type, and $N_{test}$ is the size of the test set.

### 3.4. Implementation details

The proposed federated learning framework and all associated algorithms were implemented in Python 3.9 and PyTorch 2.9.1. The hardware configuration consisted of an Intel Core i7-11800H CPU, 32 GB of RAM, and an NVIDIA GeForce RTX 3050 Ti Laptop GPU. As noted, to guarantee the reproducibility of the experimental results, all stochastic processes were controlled using fixed random seeds.

## 4. Results and discussion

This section presents the experimental results and discussion of the proposed framework. We first characterize the simulated non-IID client distributions, then evaluate the client-level and aggregated classification performance, compare the proposed method with multiple baselines, and finally discuss its privacy-preserving properties and broader implications.

### 4.1. Client data distribution result

To assess the robustness of the proposed framework, we simulated 100 distinct data-partitioning scenarios across 100 random seeds. In our primary experimental setup, the total number of clients was set to $N = 6$. Each client corresponds to a distinct recycling stakeholder. Following the methodology described in Section 3, each client was assigned a minimum of 3 ($T_{min} = 3$) and a maximum of 5 ($T_{max} = 5$) unique battery types, with a guaranteed minimum of 5 batteries for each assigned type ($size_{min} = 5$). As a representative example, Fig. 2(a) illustrates the quantitative distribution (i.e., total number of batteries) for each client, which is used to analyze the network's quantity skew. Concurrently, Fig. 2(b) visualizes the pairwise label distribution skew via a Cosine Similarity heatmap.

Fig. 2(a) provides a clear visualization of the severe quantity skew simulated in the federated environment (random seed 42). The local dataset size for each client is highly imbalanced. We can observe a stark division between “data-rich” and “data-poor” clients. Clients 2, 4, and 5 possess the vast majority of the data, with local dataset sizes of 131, 142, and 130 batteries, respectively. In sharp contrast, Clients 1 and 0 are “data-poor,” holding only 27 and 32 batteries. This disparity is significant: the largest client (Client 4) is more than 5.2 times the size of the smallest client (Client 1).

To quantitatively visualize the label distribution skew, the pairwise Cosine Similarity matrix $S$ (as defined in Section 3.1) is presented as an annotated heatmap in Fig. 2(b). This matrix confirms a highly heterogeneous federated environment, as the similarity scores span the entire possible range from 0.00 (perfectly dissimilar) to 0.85 (highly similar). The heatmap immediately reveals distinct client “clusters” that share similar data structures: Clients 0 and 2 show an exceptionally high similarity ($S_{0,2} = 0.85$), indicating their local datasets are composed of nearly the same proportion of battery types, while Clients 3 and 5 form a second, strong cluster ($S_{3,5} = 0.71$). Conversely, the most extreme examples of label skew are evidenced by pairs with zero similarity, such as $S_{0,4} = 0.00$ and $S_{1,3} = 0.00$. This indicates that their distribution vectors are orthogonal, implying that their local datasets are completely disjoint with respect to the battery types they contain. The above analysis provides a clear quantitative map of the complex skew in the quantitative and label distributions, validating the challenging, non-IID environment used to test our proposed framework.

### 4.2. Client model performance

Fig. 3 depicts the client's classification performance for each type of battery under the random seed 42. This figure illustrates the local validation accuracy of each client's classification sub-model, tested exclusively on the battery types present in its own local dataset.

The results demonstrate that each client model has become a high performing “specialist” on its own data distribution. Across all eight subplots, clients that were trained on a specific type (indicated by a colored bar) achieve exceptionally high classification accuracy, with the minimum accuracy exceeding 0.9 and many reaching 1.

Conversely, the absence of a bar (e.g., for Clients 0, 1, 2, 3, and 5 in the ‘21Ah_NMC’ subplot) visually confirms that these clients are ‘naive’ to this battery type, as it was absent from their local dataset. This figure establishes a critical premise for our work: the challenge lies not in the quality of individual client models but in the extreme heterogeneity of their expertise. This is precisely why conventional aggregation methods (discussed in subsection 4.4) fail, as they are incapable of distinguishing between an expert's high-performing prediction and a naive client's unreliable guess.

To directly validate the core assumption underlying the Expert Index, the relationship between reconstruction error and classification reliability was analyzed at the client level. Specifically, we evaluated all six client models under the representative Seed 42 partition on the full test set. Across these client models, correctly classified samples consistently exhibited lower reconstruction errors than incorrectly classified samples. The corresponding violin plots and Mann–Whitney $U$ test results are provided in the Supplementary Material (Fig. S2 and Table S2). These results provide direct support for the underlying assumption of the Expert Index, namely that lower reconstruction error is associated with higher classification reliability, thereby supporting the effectiveness of using reconstruction-error-based expert weighting in the proposed framework.

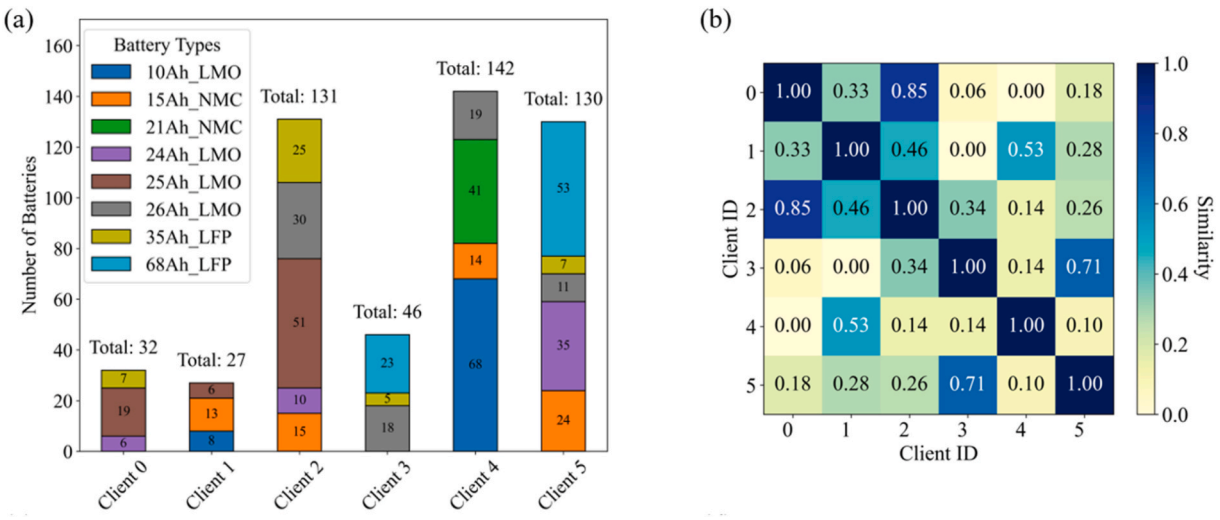


**Fig. 2.** Visualization of the severe non-IID heterogeneity under a representative scenario (Random Seed 42). (a) The quantitative distribution of batteries across clients, illustrating the Data Quantity Skew. (b) The pairwise Cosine similarity heatmap of client label distributions, quantifying the Label Distribution Skew.

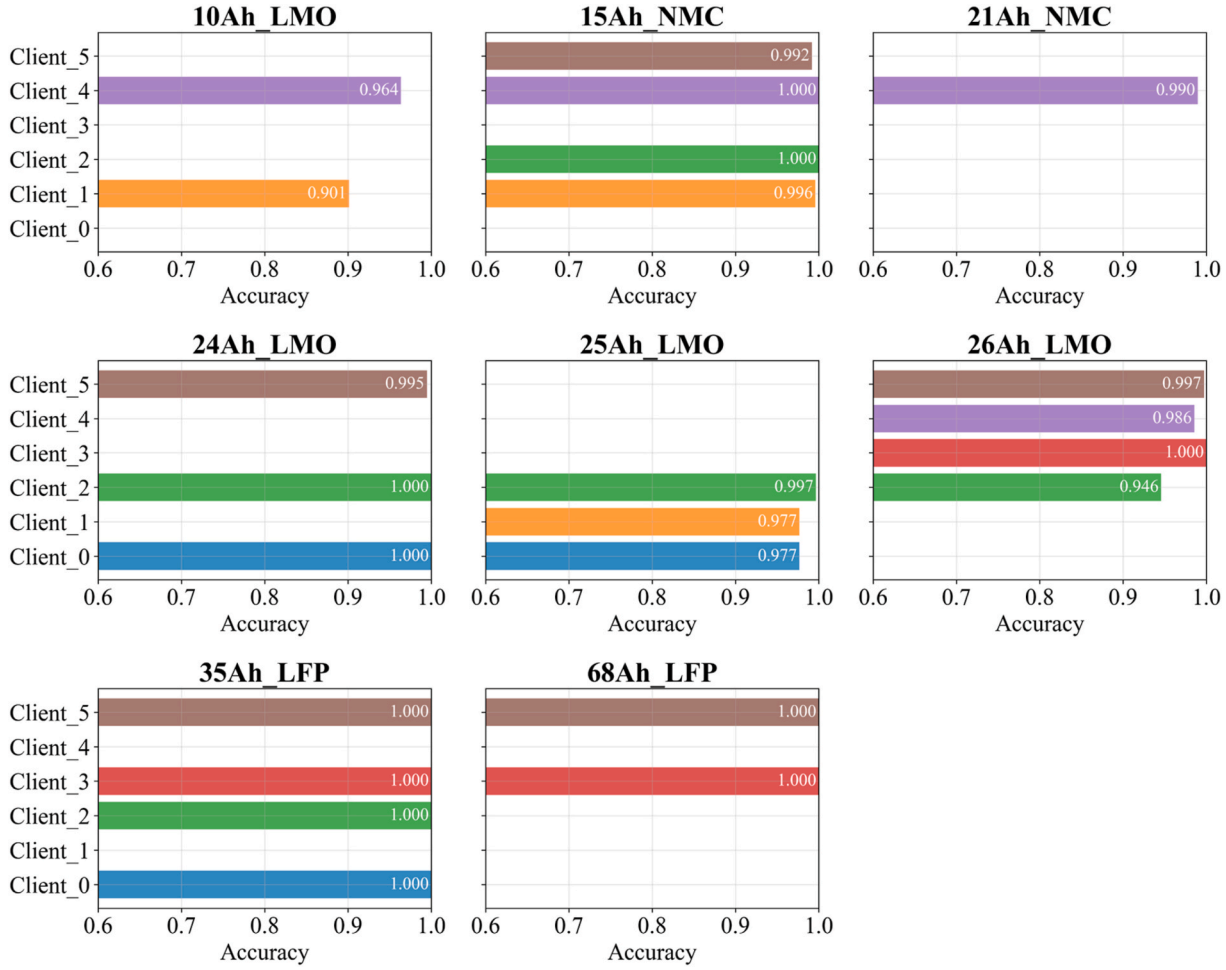


**Fig. 3.** Client-wise classification accuracy for each battery type under random seed 42. The absence of a bar indicates that the corresponding battery type was not present in the client's local dataset.

### *4.3. Aggregation performance under 100 different random seeds*

To comprehensively evaluate the robustness of the proposed framework under varying non-IID conditions, this section analyzes the aggregation performance across all 100 random seeds from two perspectives. First, the overall classification performance is examined using global metrics. Second, the results are further analyzed by battery type to reveal class-specific robustness.

#### 4.3.1. Overall classification performance of the proposed framework

To evaluate robustness of the proposed framework across different non-IID scenarios, we evaluated its performance using all 100 distinct data partitioning seeds, as discussed in Section 4.1. Fig. 4 plots the results for classification metrics.

The results demonstrate that the framework's overall classification performance is consistently strong. Across all 100 distinct non-IID partitioning scenarios, the proposed method consistently achieves high mean scores for all seven key average metrics. All average metrics consistently approach 0.97. These values are high in absolute terms and demonstrate the framework's effectiveness in accurately classifying battery types. This high classification performance directly demonstrates that the expert-weighted aggregation mechanism is effective, as it amplifies the results of "expert" clients while ignoring disturbances from unreliable, "naive" clients.

The framework's high average performance is accompanied by strong robustness and stability. This robustness is visually confirmed by the tight clustering of individual data points (blue dots) around their respective means, indicating that the framework's performance is not sensitive to the specifics of any single random seed. This visual stability is then supported by the low standard deviation (Std.) for all key metrics. Indeed, the Std. for all metrics is below 0.014; specifically, the Std. for Accuracy is 0.0133, and for Weighted Precision it is even lower at 0.0107. Crucially, this robustness extends to worst-case scenarios. The framework proves its resilience by avoiding catastrophic performance failure. Even in the most challenging partitioning scenario (Seed 32), the minimum recorded Accuracy (Min: 0.9073) and Macro F1 (Min: 0.9179) remain exceptionally high, well above the 0.90 threshold. Another critical observation from Fig. 4 is the close convergence of the Macro-Averaged metrics and their Weighted-Averaged counterparts (e.g., Macro-Averaged metrics: Precision = 0.9672, Recall = 0.9648, F1 = 0.9647; Weighted-Averaged metrics: Precision = 0.9661, Recall = 0.9633, F1 = 0.9633). The Macro Average assigns equal weight to all classes, regardless of sample size, making it highly sensitive to poor performance on minority classes. The Weighted Average biases its score towards the majority classes. The fact that our Macro F1 score is as high as our Weighted F1 score provides strong evidence that our proposed "expert-weighted" aggregation method has effectively mitigated the impact of non-IID data heterogeneity. This convergence demonstrates that our framework does not sacrifice minority-class performance; rather, by effectively identifying and amplifying "expert" clients, it achieves remarkably consistent, high performance across all classes, regardless of their prevalence in the test data.

In conclusion, the above analysis confirms the framework's robust classification performance under complex scenarios. The combination of an exceptionally high mean (signifying effectiveness), a resilient worst-case minimum (proving it avoids catastrophic failure), and a near-zero variance (signifying robustness) demonstrates that performance across all 100 challenging non-IID scenarios is virtually identical. This demonstrates that our proposed framework is a statistically significant, effective, and robust solution for retired battery classification.

#### 4.3.2. Aggregation performance under different battery types

Having established the framework's high overall robustness across 100 random seeds, we now disaggregate this performance by battery type. This analysis reveals that the framework's exceptional

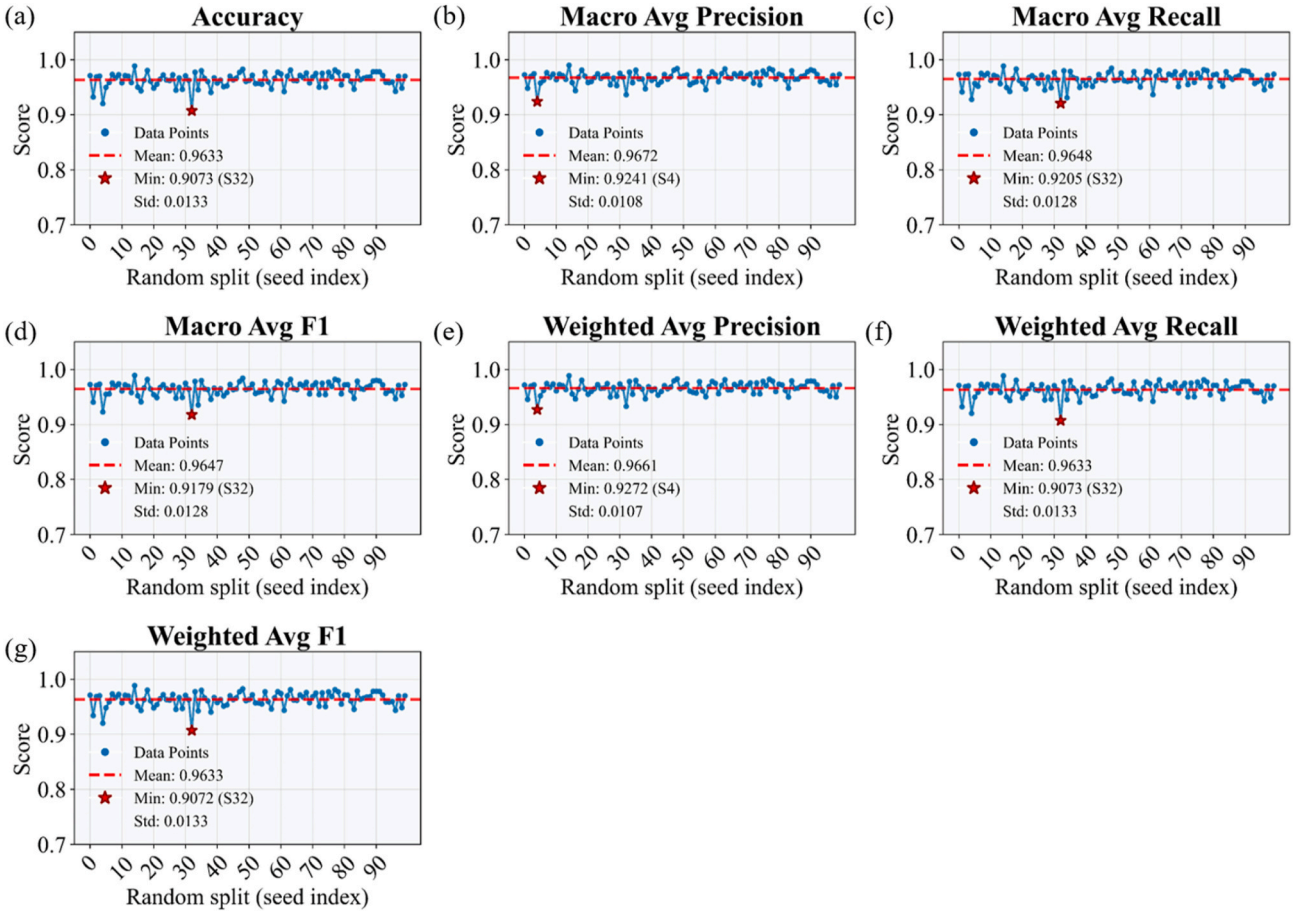


**Fig. 4.** Robustness assessment of the proposed framework across multiple random data splits. The figure illustrates the variations in seven performance metrics over 100 independent trials with different random seeds. The red dashed lines indicate the mean values, while the red stars highlight the worst-case scenarios (minimum scores). The notation (e.g., S32) indicates the specific random seed index (i.e., Random seed 32) corresponding to the worst-case performance.

performance is not uniform, allowing for a stratification into three distinct groups. This grouping is based on a profile of key F1-Score statistics—specifically, the Mean (Mean), the Standard Deviation (Std.), and the Minimum value (Min). By grouping results, a comprehensive landscape of the framework's behavior is provided across different battery types, isolating robustly identified types from those that remain sensitive to data heterogeneity. This stratification is driven by clear quantitative divides in the F1-Score statistics: a "high-variance" group is distinguished by relative instability (e.g., Std. $>$ 0.03 and Min. F1 $<$ 0.80). The remaining stable types are then further separated into a "high-performing" group (defined by Mean F1 $>$ 0.98 and Std. $<$ 0.02) and a "stable workhorse" group (defined by Mean F1 $<$ 0.98 and Std. $<$ 0.02).

Fig. 5 plots the performance metrics for the most stable cohort, defined as the "high-performing, highly robust" group. This group, which includes 35Ah_LFP (a), 68Ah_LFP (b), and 15Ah_NMC (c), quantitatively satisfies the most stringent performance criteria: a mean F1-Score above 0.98 and a standard deviation below 0.02 under 100 different scenarios.

The results for this group are exceptionally high and stable. The mean F1-Scores for these three types are 0.9923, 0.9858, and 0.9848, respectively, indicating a notable average classification effectiveness. This high mean performance is matched by strong stability, as evidenced by negligible standard deviations, all consistently at or below 0.015. Visually, this stability is confirmed by the extremely tight, "flat" clustering of the data points near the 1.0 mark across all 100 seeds. Furthermore, the framework's robustness is underscored by its worst-case performance; even the lowest F1-Score recorded in this group was an exceptionally high 0.9139 (for 35Ah_LFP at Seed 4). As detailed in the confusion matrix for this specific seed (see Supplementary Material Fig. S3), this performance drop is not a systemic failure. Still, it is driven almost entirely by a single, specific confusion pattern: the misclassification of 40 68Ah_LFP samples as 35Ah_LFP. This large False Positive (FP) count is the sole reason for the drop in the 35Ah_LFP class's Precision (to 0.8414), which in turn lowers the F1-Score. This analysis confirms that, even in the worst-case scenario, the framework's only

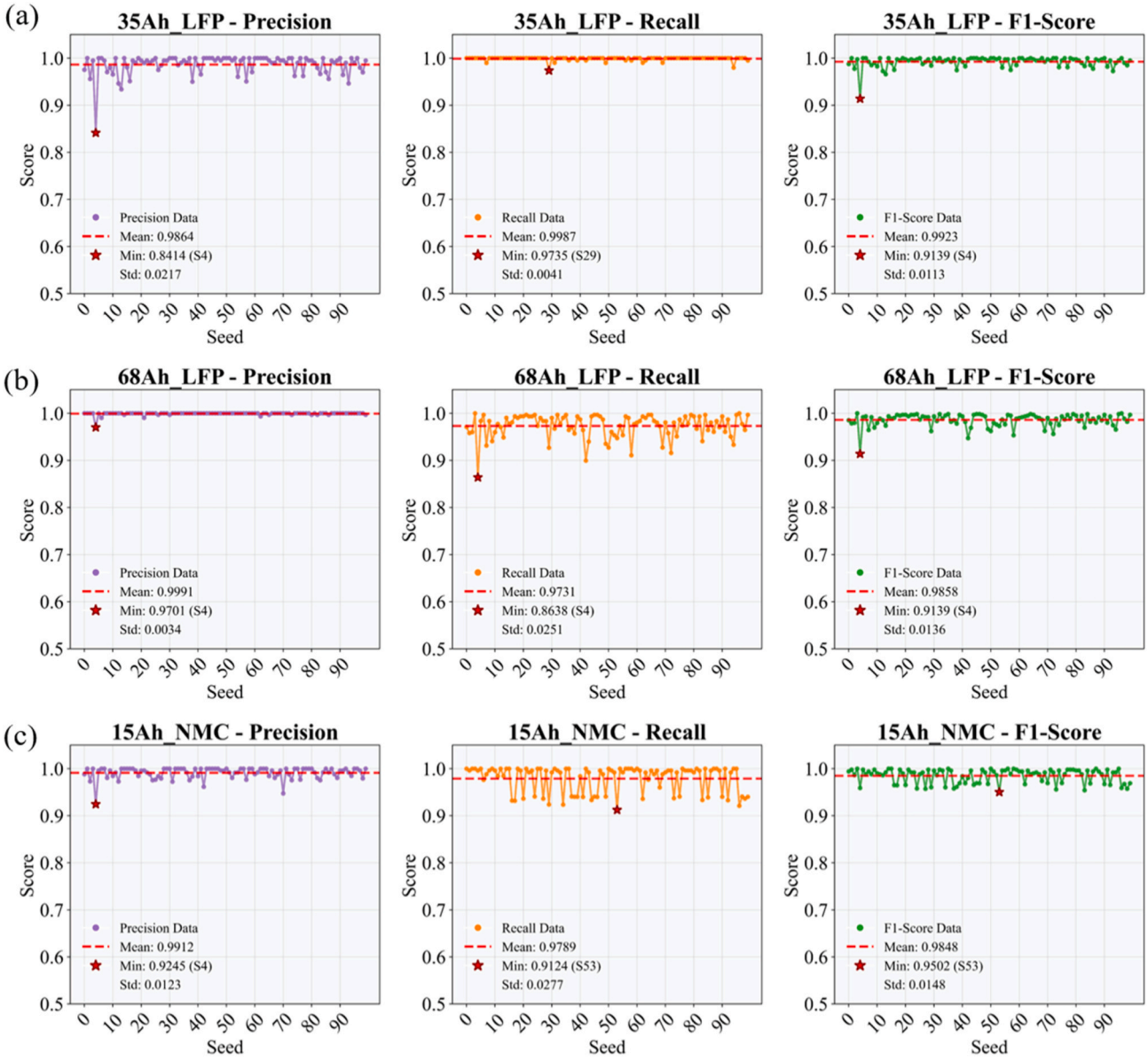


**Fig. 5.** Performance metrics (Precision, Recall, and F1-Score) across 100 random seeds for the 'Group 1' (High-Performing, Highly Robust) battery types: (a) 35Ah_LFP, (b) 68Ah_LFP, and (c) 15Ah_NMC. Low Std indicates high stability. The minimum scores are highlighted with red stars, accompanied by labels such as '(S4)' to denote the specific random seed index (Random seed 4) corresponding to the lowest value.

significant error is in distinguishing between two highly similar battery types sharing the same LFP chemistry. In contrast, its ability to distinguish LFP from all other chemistries remains strong.

The framework's highly stable "workhorse" performance is detailed in Fig. 6, which plots the results for 10Ah_LMO (a) and 24Ah_LMO (b). This group is quantitatively defined by its combination of high mean performance and low variance, satisfying our criteria of Mean F1 $< 0.98$ and Std. $< 0.02$.

Both types demonstrate exceptional reliability. They maintain nearly identical mean F1-Scores—0.9682 for 10Ah_LMO and 0.9690 for 24Ah_LMO. This high effectiveness is complemented by excellent stability, as evidenced by their low Std values (0.0183 and 0.0182, respectively). Visually, the data points in all plots remain tightly clustered around their means, similar to the "high-performing" group.

This group's distinction from the "high-performing" group lies in its response to different scenarios. Although performance varies slightly across challenging non-IID scenarios, the overall degradation remains limited. For instance, the worst-case scenario for 10Ah_LMO (Seed 38) still resulted in a high F1-Score of 0.9132. This demonstrates that the framework exhibits limited performance degradation across these types, unlike the more obvious degradation observed in the final group, reinforcing the framework's overall robustness.

Finally, Fig. 7 illustrates the results for the "high-variance, stress-test" group. This group, comprising 21Ah_NMC (a), 25Ah_LMO (b), and 26Ah_LMO (c), is clearly distinguished by its significant instability, which quantitatively satisfies our criteria for Std. $> 0.03$ and Min. F1 $< 0.80$. This volatility is immediately visible in their plots, which show wider "spreads" and deeper, more frequent performance drops compared to the previous two groups. This is confirmed by their comparatively large standard deviations, such as 0.0433 for 21Ah_NMC's F1-Score and 0.0358 for 25Ah_LMO's. This instability suggests that these specific NMC/LMO types are more ambiguous, likely sharing voltage features that are highly similar to each other. This makes them harder to distinguish when a specific non-IID partitioning (i.e., a random seed) creates a particularly challenging "knowledge gap" for the federated network. However, this high-variance group further highlights the value of the proposed framework. Despite this volatility, the mean F1-Scores remain exceptionally high (e.g., 0.9479 for 21Ah_NMC and 0.9503 for 26Ah_LMO). This high variance demonstrates that some random partitioning seeds can indeed confuse the network. However, it also proves the framework's robustness. Even when confronted with these most challenging scenarios (e.g., Seed 34 for 21Ah_NMC), the F1-Score does not exhibit a complete performance breakdown. Instead, it achieves a robust F1 score of 0.7285, the lowest observed across all groups and scenarios. This proves the framework's fundamental ability to 'find the expert' is always active, successfully mitigating what would otherwise be a total failure.

In summary, this three-tiered analysis of the per-type robustness validates the framework's overall effectiveness. The method achieves high performance on the more easily separable battery types, maintains stable performance on the reliable "workhorse" types, and retains strong average performance even for the more ambiguous, high-variance "stress-test" types. This comprehensive validation proves that our expert-weighted aggregation is not only effective on average but is also a promising and robust solution for managing the complex, heterogeneous, and unpredictable nature of controlled non-IID environments considered in this study.

### 4.4. Comparison with other frameworks

To validate our framework, it was benchmarked against four aggregation strategies, representing a progression from naive to sophisticated mechanisms: the Average Aggregation (AA), the Class Count Weighted (CCW), and two Mahalanobis Distance Weighted (MDW) variants [43]. AA and CCW serve as baselines to quantify the performance deficit when client heterogeneity is ignored or when data volume

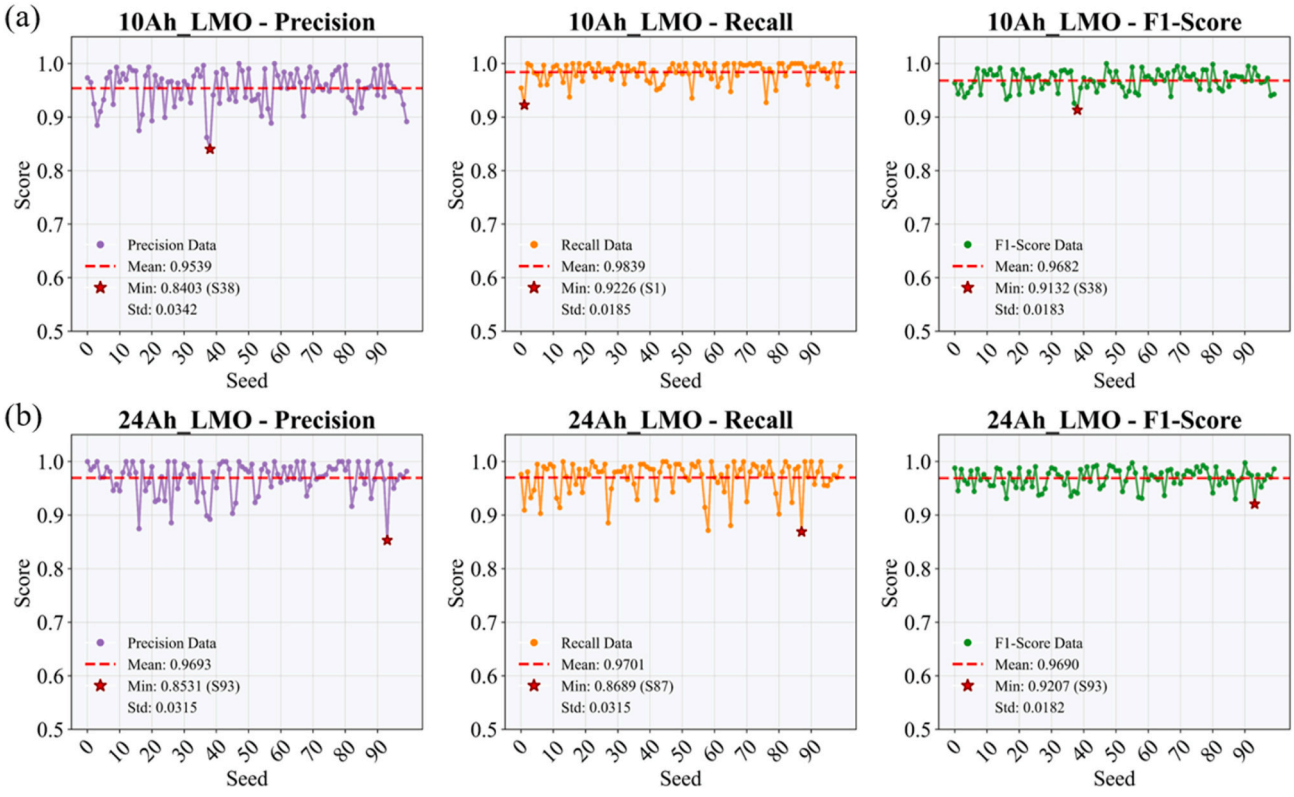


**Fig. 6.** Performance metrics (Precision, Recall, and F1-Score) across 100 random seeds for the 'Group 2' (Stable workhorse) battery types: (a) 10Ah_LMP and (b) 24Ah_LMO. The results demonstrate that the framework avoids catastrophic failure even in worst-case splits. The worst-case is annotated, for instance, where 'S38' denotes seed 38, indicating the scenario that yielded the minimum observed value.

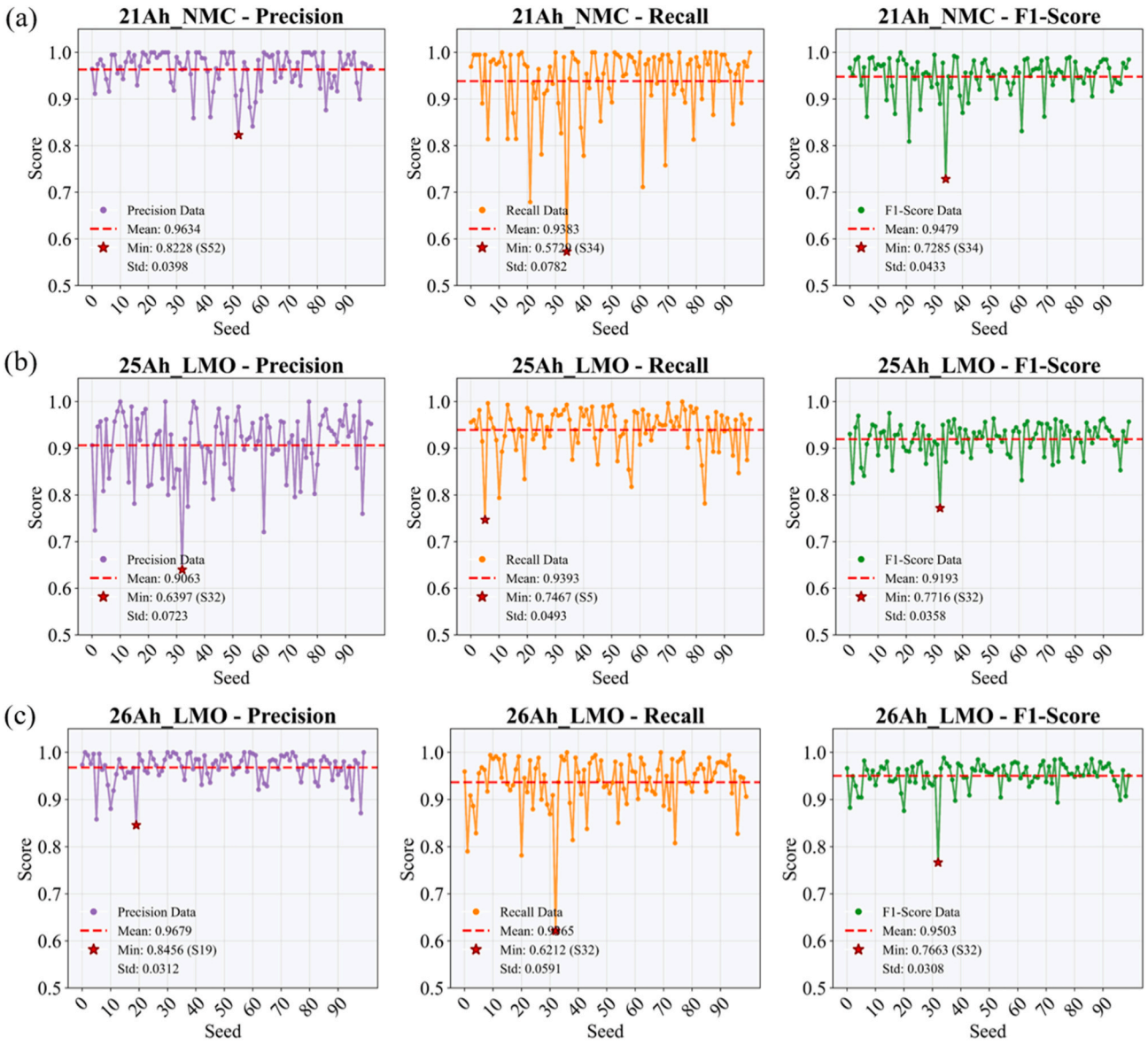


**Fig. 7.** Performance metrics (Precision, Recall, and F1-Score) across 100 random seeds for the 'Group 3' (High-variance) battery types: (a) 21Ah_NMC, (b) 25Ah_LMO, and (c) 26Ah_LMO. In contrast to the highly clustered results observed previously, these datasets exhibit sporadic drops in Recall and Precision. The worst-case instances are annotated, e.g., 'S34′ represents seed 34.

is used as a sole indicator for expertise, and the MDW baseline was evaluated in two variants: a version without an integrated Differential Privacy mechanism (MDW_WO) and a version with an integrated Differential Privacy mechanism (MDW_W). This specific comparison is instrumental in demonstrating how conventional distance-based methods suffer from a severe privacy-utility trade-off. Detailed implementation specifications for these aggregation baselines are provided in Supplementary Material Note S1. Furthermore, a fully centralized model (trained on all data) was included to serve as a theoretical performance upper bound. The statistical results for all 100 random seeds are presented in Table 2, with the distributions for Accuracy and Avg F1 visualized in the box plots in Fig. 8. We further

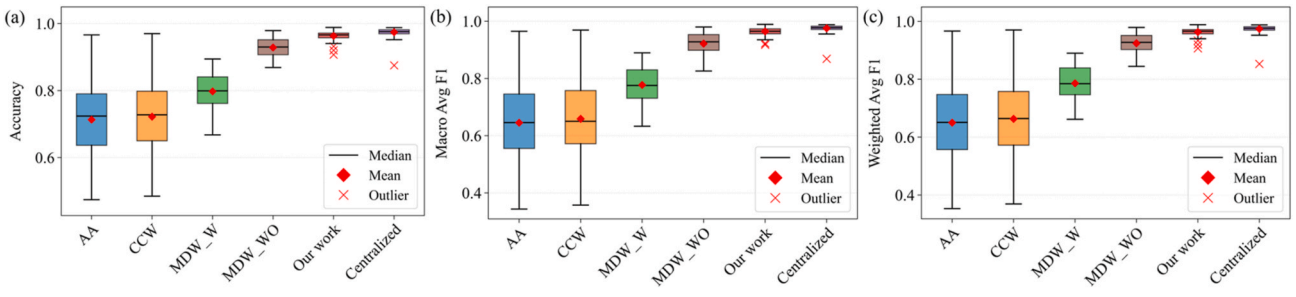


**Fig. 8.** The performance comparison of five methods. (a) Accuracy, (b) Macro average F1, and (c) Weighted average F1.

conducted pairwise statistical significance tests based on the 100 matched random seeds, and the Wilcoxon signed-rank test results with Holm–Bonferroni correction are reported in the Supplementary Material (Table S3–S5). The analysis reveals three main findings.

First, our proposed framework achieves consistently higher performance than all other aggregation baselines. As shown in Table 4, our method's mean Accuracy (0.963) and mean Macro F1 (0.965) outperform the next-best baseline by a clear margin, MDW_WO (0.929 Accuracy, 0.922 Macro F1). This corresponds to absolute improvements of 0.034 in Accuracy and 0.043 in Macro F1. Compared with the privacy-protected baseline MDW_W, the proposed framework further improves Accuracy and Macro F1 by 0.165 and 0.188, respectively. These improvements are further supported by the pairwise statistical significance tests reported in the Supplementary Material (Table S2–S4), which show that our method significantly outperforms all distributed baselines across the main metrics. The most critical insight comes from the Macro Avg F1 metric, which is highly sensitive to minority-class failure. The conventional baseline methods suffer a performance collapse on this metric (e.g., AA: 0.645, CCW: 0.659), proving their inability to handle the severe label skew. In sharp contrast, our method's Macro F1 (0.965) is almost identical to its Accuracy (0.963), providing the strong evidence that our expert-weighted mechanism effectively addresses the non-IID challenge by properly amplifying expert clients and silencing non-experts.

Second, the framework is not only more effective but also substantially more stable and robust. This is immediately evident in the box plots of Fig. 8. The plots for the baseline methods (AA, CCW, MDW_W, MDW_WO) are characterized by large interquartile ranges (IQRs) and wide whiskers across all three-evaluation metrics, indicating severe performance variance. This visual instability is quantitatively confirmed in Table 4 by the high standard deviations (e.g., AA Std.: 0.111, CCW Std.: 0.107) in accuracy. However, our proposed method is highly stable, with a very low Std. around 0.013 for accuracy, macro F1 score, and weighted F1 score. This stability also extends to challenging scenarios: although performance declines can still occur under the most challenging partitions, the proposed framework maintains a relatively high worst-case performance floor, with minimum values of 0.907 in Accuracy and 0.918 in Macro F1.

Finally, the results show that our privacy-preserving framework remains close to the centralized upper bound. As shown in Table 4, the centralized model achieves the highest mean scores across all key metrics, with 0.974 in Accuracy and 0.976 in Macro Avg F1. Our proposed framework tracks this performance closely, achieving a mean Macro Avg F1 of 0.965 and an Accuracy of 0.963, corresponding to absolute gaps of only 0.011 in Macro Avg F1 and 0.011 in Accuracy. This suggests that the proposed framework can retain most of the benefit of centralized learning under privacy-preserving collaboration.

**Table 4**
Statistical evaluation of Average Aggregation, Class Count Weighted, Mahalanobis Distance Weighted, the Proposed Framework, and the Centralized Baseline regarding Accuracy, Macro Average F1-Score, and Weighted Average F1-Score.

| Metric | Model | Mean | Std | Min | Range | CV |
|---|---|---|---|---|---|---|
| Accuracy | AA | 0.713 | 0.111 | 0.474 | 0.492 | 15.631 |
| | CCW | 0.722 | 0.107 | 0.485 | 0.486 | 14.862 |
| | MDW_W | 0.798 | 0.056 | 0.668 | 0.227 | 6.963 |
| | MDW_WO | 0.929 | 0.029 | 0.869 | 0.110 | 3.119 |
| | Our work | 0.963 | 0.013 | **0.907** | **0.081** | 1.383 |
| | Centralized | **0.974** | **0.012** | 0.875 | 0.113 | **1.279** |
| Macro Avg F1 | AA | 0.645 | 0.134 | 0.343 | 0.622 | 20.768 |
| | CCW | 0.659 | 0.129 | 0.357 | 0.612 | 19.586 |
| | MDW_W | 0.777 | 0.064 | 0.633 | 0.257 | 8.277 |
| | MDW_WO | 0.922 | 0.040 | 0.826 | 0.154 | 4.293 |
| | Our work | 0.965 | **0.013** | **0.918** | **0.071** | 1.325 |
| | Centralized | **0.976** | **0.013** | 0.869 | 0.119 | **1.320** |
| Weighted Avg F1 | AA | 0.650 | 0.136 | 0.353 | 0.614 | 20.881 |
| | CCW | 0.663 | 0.130 | 0.369 | 0.601 | 19.656 |
| | MDW_W | 0.786 | 0.061 | 0.662 | 0.229 | 7.704 |
| | MDW_WO | 0.925 | 0.034 | 0.845 | 0.135 | 3.682 |
| | Our work | 0.963 | **0.013** | **0.907** | **0.081** | **1.379** |
| | Centralized | **0.974** | 0.014 | 0.853 | 0.135 | 1.469 |

Moreover, this comparison reveals a noteworthy finding regarding worst-case robustness. While the Centralized model is superior on average, it is more vulnerable in some scenarios to "pathological" data partitioning scenarios. This is evidenced by its Minimum (Min) scores. In its most challenging scenario, the Centralized model's performance dips to a Macro Avg F1 of 0.869 and a Weighted Avg F1 of 0.853. In contrast, our federated 'expert-weighted' mechanism appears more robust in the worst-case scenarios considered here. Our framework establishes a significantly higher performance floor across all key metrics. Its minimum Accuracy (0.907), Macro Avg F1 (0.918), and Weighted Avg F1 (0.907) are all higher than the worst-case minimums achieved by the centralized model. This stronger worst-case performance suggests a potential architectural advantage: a single centralized model acts as a "single point of failure," making it sensitive to the specific composition of its randomly selected 80% training set. Our federated approach, by functioning as a robust ensemble of specialists, leverages the 'expert-weighted' aggregation to provide an additional layer of robustness, effectively mitigating the impact of a "poor" data partition and thus proving more robust in its absolute worst-case scenario.

Subsequently, we present results from four selected random seeds: one that yielded the highest accuracy, representing the best-case scenario (Seed 14); one that resulted in the lowest accuracy, representing the most challenging case (Seed 32); and two others chosen at random (Seeds 42 and 82). The corresponding data partitioning visualizations for these representative scenarios are detailed in Supplementary Fig. S4–S11. The performance comparison across these seeds is shown in a radar chart in Fig. 9, with detailed numerical values in Table 5.

The radar charts further support the overall strong performance of the proposed framework. As observed in Fig. 9, the polygonal area enclosed by our method's performance curve (labeled "This work") consistently envelops those of the baseline aggregation methods (AA, CCW, MDW_W and MDW_WO) across all seven evaluation axes. This geometric dominance indicates that our framework does not trade off precision for recall or sacrifice minority-class performance; rather, it achieves a simultaneous improvement across all metrics.

We first examine the worst-case scenario (Seed 32), which represents the lower bound of our framework's performance. Even in this most challenging environment, where our method's Accuracy drops to 0.9073, the framework still maintains a relatively high-performance floor. In contrast, the naive aggregation baselines perform much worse, with AA and CCW achieving only 0.7344 and 0.7436, respectively. The MDW variants perform better than these naive baselines, reaching 0.7552 for MDW_W and 0.8899 for MDW_WO, but both remain below our method. These results suggest that the proposed framework remains comparatively robustness even under the most difficult partitioning scenario considered here.

The analysis of the randomly selected seeds (e.g., Seed 42) further illustrates the framework's robustness under severe data skew. In this scenario, the naive aggregation baselines suffer substantial performance degradation: as shown in Table 5, the Accuracy of AA and CCW drops to 0.5758 and 0.5772, respectively. By contrast, our proposed framework maintains a high Accuracy of 0.9509 under the same partition. This corresponds to an absolute advantage of about 0.37 in Accuracy over the naive baselines, supporting the view that the proposed Expert Index mechanism can more effectively suppress unreliable client contributions and preserve useful local knowledge under highly non-IID conditions.

Furthermore, the analysis of Seed 14 (the best-case scenario) reveals that in some individual random seeds, our method can achieve performance comparable to or even slightly higher than that of the centralized model. However, the statistical significance analysis across all 100 matched random seeds shows that the centralized model remains significantly better overall. While the Centralized model achieves a stable Accuracy of 0.9745, our method achieves 0.9886, with a

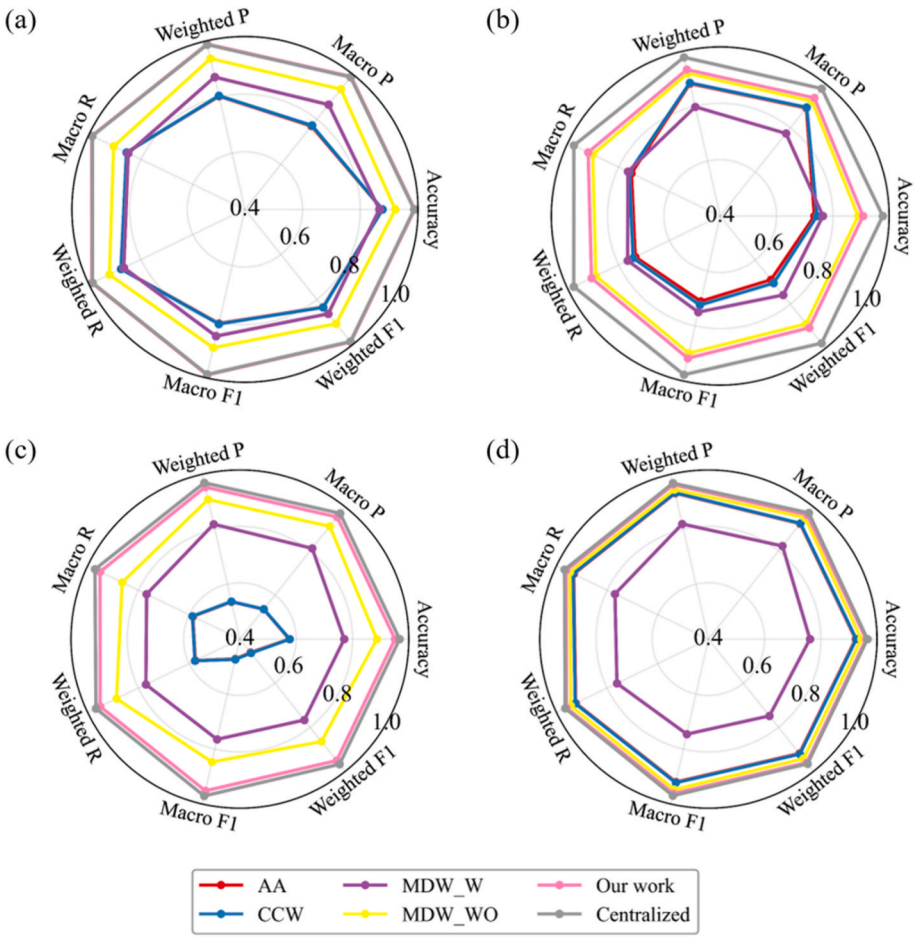


**Fig. 9.** Multi-metric performance comparison of classification models. The radar charts illustrate the comprehensive evaluation of seven metrics (Accuracy, Macro/Weighted Precision, Recall, and F1-Score) utilizing pulse test voltage features. The subplots represent four distinct random partitioning scenarios: (a) Seed 14 (Best-case scenario), (b) Seed 32 (Worst-case scenario), (c) Seed 42 (Random scenario), and (d) Seed 82 (Random scenario).

**Table 5**
Robustness testing of the proposed framework and baselines under representative non-IID data partitioning settings (including best, worst, and random scenarios).

| Seed | Model | Accuracy | Macro Avg Precision | Weighted Avg Precision | Macro Avg Recall | Weighted Avg Recall | Macro Avg F1 | Weighted Avg F1 |
|---|---|---|---|---|---|---|---|---|
| 14 | AA | 0.8765 | 0.7702 | 0.8034 | 0.8552 | 0.8765 | 0.8063 | 0.8342 |
| | CCW | 0.8770 | 0.7728 | 0.8047 | 0.8555 | 0.8770 | 0.8077 | 0.8349 |
| | MDW_W | 0.7780 | 0.8097 | 0.8133 | 0.7663 | 0.7780 | 0.7497 | 0.7622 |
| | MDW_WO | 0.9214 | 0.9336 | 0.9372 | 0.9050 | 0.9214 | 0.8924 | 0.9073 |
| | This work | **0.9886** | **0.9900** | **0.9888** | **0.9888** | **0.9886** | **0.9893** | **0.9886** |
| | Centralized | 0.9745 | 0.9767 | 0.9754 | 0.9755 | 0.9745 | 0.9755 | 0.9742 |
| 32 | AA | 0.7344 | 0.8905 | 0.8822 | 0.7496 | 0.7344 | 0.7101 | 0.6893 |
| | CCW | 0.7436 | 0.8933 | 0.8849 | 0.7582 | 0.7436 | 0.7243 | 0.7046 |
| | MDW_W | 0.7552 | 0.7533 | 0.7754 | 0.7443 | 0.7552 | 0.7311 | 0.7473 |
| | MDW_WO | 0.8899 | 0.9219 | 0.9193 | 0.9023 | 0.8899 | 0.8992 | 0.8993 |
| | This work | 0.9073 | 0.9363 | 0.9331 | 0.9205 | 0.9073 | 0.9179 | 0.9072 |
| | Centralized | **0.9745** | **0.9767** | **0.9754** | **0.9755** | **0.9745** | **0.9755** | **0.9742** |
| 42 | AA | 0.5758 | 0.5357 | 0.5359 | 0.5857 | 0.5758 | 0.4722 | 0.4610 |
| | CCW | 0.5772 | 0.5361 | 0.5370 | 0.5872 | 0.5772 | 0.4737 | 0.4631 |
| | MDW_W | 0.7667 | 0.7984 | 0.8034 | 0.7554 | 0.7667 | 0.7578 | 0.7622 |
| | MDW_WO | 0.8865 | 0.9120 | 0.9081 | 0.8641 | 0.8865 | 0.8470 | 0.8647 |
| | This work | 0.9509 | 0.9544 | 0.9534 | 0.9512 | 0.9509 | 0.9519 | 0.9513 |
| | Centralized | **0.9745** | **0.9767** | **0.9754** | **0.9755** | **0.9745** | **0.9755** | **0.9742** |
| 82 | AA | 0.9241 | 0.9241 | 0.9311 | 0.9323 | 0.9241 | 0.9201 | 0.9193 |
| | CCW | 0.9261 | 0.9257 | 0.9328 | 0.9347 | 0.9261 | 0.9226 | 0.9216 |
| | MDW_W | 0.7772 | 0.8161 | 0.8087 | 0.7854 | 0.7772 | 0.7775 | 0.7737 |
| | MDW_WO | 0.9439 | 0.9513 | 0.9451 | 0.9476 | 0.9439 | 0.9475 | 0.9424 |
| | This work | 0.9609 | 0.9637 | 0.9620 | 0.9617 | 0.9609 | 0.9617 | 0.9604 |
| | Centralized | **0.9745** | **0.9767** | **0.9754** | **0.9755** | **0.9745** | **0.9755** | **0.9742** |

corresponding Macro F1 of 0.9893. This suggests that selective aggregation of expert predictions may help mitigate the influence of less reliable client outputs under certain non-IID scenarios.

In summary, conventional aggregation methods are highly sensitive

to the randomness of data partitioning, and their performance can degrade substantially under severe data skew, as illustrated by the results for Seed 42. In contrast, our proposed framework maintains consistently strong performance across the representative scenarios considered here. Whether in the best-case scenario (Seed 14) or the worst-case scenario (Seed 32), it demonstrates the level of stability and robustness required for real-world deployment. More broadly, the supplementary statistical significance analysis across the 100 matched random seeds shows that the proposed method significantly outperforms all distributed baselines. Although the centralized model remains significantly better on Accuracy, Macro Avg F1, and Weighted Avg F1, our framework remains close to the centralized upper bound in magnitude and shows a higher observed performance floor under the most challenging partitioning scenarios.

### 4.5. Privacy preservation analysis

While the preceding subsections have established the framework's strong classification performance and robustness, a critical advantage of our proposed architecture lies in its inherent capability to preserve data privacy. Unlike performance metrics, privacy is structural; it is determined by what information is exchanged. In this subsection, we analyze the privacy-preserving properties of our framework from a "Privacy-by-Design" perspective, contrasting it with standard federated training paradigms.

#### 4.5.1. Resistance to privacy attacks

Standard federated learning methods (e.g., FedAvg) rely on the iterative exchange of model updates (e.g., gradients or weight parameters) between clients and the server [44,45]. Recent studies have demonstrated that sharing gradients is not secure; adversaries can exploit these updates to reconstruct the original private training data via Inversion Attacks (e.g., Deep Leakage from Gradients) [46]. In sharp contrast, our framework adopts a "Federated Inference" (or "Output-Only") architecture. The client models are trained locally and frozen. During the aggregation phase, no model internals, including gradients, weights, or Hessians, are ever transmitted. The central server receives only the final inference results. By mathematically blocking access to the gradient information flow, our framework substantially reduces the feasibility of gradient-based reconstruction attacks in the proposed setting.

#### 4.5.2. Obfuscation of feature representations

Another class of aggregation methods, exemplified by the aforementioned MDW approach, relies on computing statistical distances in feature space. To mitigate the privacy risks associated with sharing the necessary statistics (e.g., covariance matrices), these methods typically incorporate Differential Privacy (DP) mechanisms, which inject calibrated noise into the transmission. However, this introduces a clear trade-off: the added noise inevitably distorts the signal, leading to the significant performance degradation observed in our baseline comparisons [43].

In stark contrast, our method alleviates this privacy-utility trade-off. It minimizes privacy risks without the need for performance-degrading noise by transmitting only highly abstract, task-specific outputs: the probability vector $P_i(x)$ and the scalar reliability score $e_i(x)$. The Probability Vector $P_i(x)$ effectively compresses the 41-dimensional raw input into a simple distribution over 8 classes, thereby substantially reducing the information available for reconstructing the specific voltage curve. The score $e_i(x)$ is even more abstract: it is a single number representing "familiarity" based on reconstruction error. It reveals how well a client knows a sample but contains very-limited information about what the sample actually looks like.

#### 4.5.3. Black-box nature and minimal information leakage

Consequently, from the perspective of the central server (or a malicious eavesdropper), each client operates as a black-box local model. The link between the raw data and the transmitted message is substantially weakened by the complex, non-linear transformations of the local Neural Networks. The only theoretical risk remaining is "Membership Inference", i.e., deducing if a specific sample type exists in a client's dataset based on a uniquely low reconstruction error. Crucially, however, this residual leakage is deemed operationally benign within the battery recycling industry. In this industrial context, the categorical identity (i.e., battery type/chemistry) is typically treated as non-sensitive inventory information. The primary privacy concern in this sector is the protection of raw historical data and usage profiles, which constitute proprietary intellectual property. Since our framework exposes only the former (metadata) while without directly exposing the latter (content), it aligns with real-world industrial privacy standards.

In summary, our framework achieves a favorable privacy-performance trade-off. It matches or exceeds the performance of centralized upper bounds (as shown in Section 4.4) while restricting collaboration to output-level information from local data. By limiting communication to abstract model outputs, the framework substantially reduces exposure to gradient-based reconstruction attacks and limits model information leakage relative to parameter-sharing paradigms.

### 4.6. Discussion

The core innovation of this study lies in the formulation of a privacy-native "Expert-Weighted" federated learning paradigm, which addresses the "data silo" dilemma in the retired battery recycling industry. In contrast to traditional paradigms that perceive privacy and utility as conflicting objectives, typically compromising accuracy through differential privacy noise, our framework reveals that these goals can be jointly improved in this setting. By leveraging the reconstruction error of local autoencoders, we decoupled model reliability assessment from access to raw data. This "Output-Only" architecture ensures that the aggregation server can identify and amplify high-quality predictions from expert clients without direct access to the underlying data distribution and model information, thereby reducing exposure to various privacy attacks (e.g., reconstruction and inversion attacks) [47].

An interesting observation is that our framework maintains a higher performance floor than the centralized model in some pathological non-IID scenarios. Although the centralized model still performs better on average, its performance can drop more noticeably under particularly unfavorable partitions, such as Seed 32. By contrast, our federated framework behaves more like an ensemble of specialists: when one client is unreliable in an unfamiliar region, another client with more relevant local knowledge can compensate. This may partly explain why the proposed framework appears more resilient in the most challenging non-IID cases.

Furthermore, this work has important implications for the circular economy of lithium-ion batteries. The prevailing barrier to efficient recycling has not been a lack of identification technology, but rather substantial proprietary barriers that prevent the sharing of historical cycling data. By demonstrating high-accuracy classification (Accuracy $>$96%) using only snapshot pulse features without any historical dependency, we provide a viable technical pathway to unlock the value of "Black Box" batteries. This suggests that second-level pulse testing, combined with expert-weighted aggregation, may provide a promising methodological bias for future privacy-preserving battery type sorting studies, while further validation with independently collected datasets and standardized testing protocols remains necessary before deployment-oriented conclusions can be drawn.

### 4.7. Limitations and deployment considerations

It should first be noted that the present dataset contains multiple levels of internal heterogeneity. At the material and capacity levels, it includes 8 dataset-defined battery types across LFP, LMO, NMC811, and

NMC622 cells, with nominal capacities ranging from 10 Ah to 68 Ah. At the retired-cell state level, the samples exhibit heterogeneous SOH and initial SOC distributions. At the federated-client level, the 100 stochastic non-IID partitions introduce both label-distribution skew and data-quantity skew; for example, in the representative Seed 42 partition, local client sizes range from 27 to 142 batteries and pairwise label-distribution similarity ranges from 0.00 to 0.85. These factors provide a meaningful basis for evaluating internal robustness under heterogeneous retired-battery and simulated non-IID conditions.

Nevertheless, this internal heterogeneity does not fully represent external cross-enterprise domain shifts caused by different laboratories, manufacturers, testing platforms, equipment calibration, temperature-control conditions, preprocessing protocols, contact fixtures, and battery sources. Therefore, the present results should be interpreted as controlled internal validation rather than deployment-level external validation. Future work should collect independent external datasets and investigate adaptive or personalized pulse-test protocols and continual-learning strategies for broader deployment scenarios.

A further limitation is the potential influence of confounding factors on the learned decision rules. The present study focuses on dataset-defined battery type identification rather than the isolation of cathode chemistry as a pure causal factor. In the dataset, each battery type label is defined by both cathode chemistry and nominal capacity. Therefore, the classifier may exploit not only cathode-related electrochemical signatures, but also correlated attributes such as nominal capacity, cell format, manufacturer-specific design, aging state, SOC/SOH distribution, and test-response amplitude. These factors can influence the pulse-voltage turning-point features and may contribute to the classification decision. Accordingly, the current results should be interpreted as battery-type identification under the given dataset distribution, rather than definitive evidence that the model exclusively learns cathode-material information. Future studies using controlled datasets, in which chemistry, capacity, manufacturer, cell format, aging condition, and testing protocol are independently varied, will be needed to disentangle these effects.

As an auxiliary practical-risk analysis, chemistry-level confusion matrices were further provided in the Supplementary Material Fig. S12 by aggregating the eight battery-type labels into LMO, NMC, and LFP groups. This analysis helps distinguish intra-chemistry errors from more consequential inter-chemistry errors. The results show that most classification results remain correct at the chemistry level across representative non-IID partitions. For example, the diagonal values reach 99.4%–99.8% in Seed 14 and remain 92.2%–99.8% even in the most challenging representative partition, Seed 32. These results indicate that many battery-type-level errors do not necessarily lead to high-risk chemistry-level misclassification. Nevertheless, inter-chemistry errors are not fully eliminated, such as the 5.8% LFP-to-NMC confusion in Seed 42 and the 0.7% NMC-to-LFP confusion in Seed 82. Therefore, future deployment should incorporate risk-aware evaluation, confidence calibration, and rejection mechanisms for high-consequence misclassification pairs.

The practical implementation of the pulse-test protocol also requires further standardization. The extracted turning-point features may be affected by initial SOC, temperature, rest time, current magnitude, contact resistance, sampling resolution, and preprocessing procedures. Initial SOC determines the local OCV-SOC region [29–31], temperature affects ohmic resistance and diffusion kinetics [38–41], rest time influences voltage relaxation [42], current magnitude changes polarization amplitude [40,41], contact resistance distorts the instantaneous voltage jump [38], and sampling resolution affects turning-point extraction accuracy [35–37]. These effects can be more pronounced in retired cells because of uncertain useable capacity, elevated internal resistance, and heterogeneous aging histories [39,48]. Therefore, practical deployment requires consistent control or recording of SOC, temperature, rest time, current magnitude, sampling frequency, voltage filtering, and turning-point extraction rules, together with fixture calibration to reduce contact-resistance effects.

Finally, the present study adopts a closed-set classification setting, where each test sample is assumed to belong to one of the predefined eight battery types. This assumption may underestimate the complexity of practical recycling scenarios. In real recycling streams, unknown battery types, damaged cells, and mixed batches may be encountered [49–51]. In such cases, a closed-set classifier may be forced to assign an unknown or abnormal sample to one of the known categories, leading to potentially unreliable sorting decisions. Therefore, future work should extend the proposed framework toward open-set or open-world battery sorting, including unknown-type detection, confidence calibration, and rejection mechanisms. In addition, adaptive learning and continual learning strategies should be explored to enable the model to update its knowledge when new battery types, new manufacturers, or new aging patterns are encountered over time, while avoiding catastrophic forgetting of previously learned types. The reconstruction-error-based Expert Index may provide a potential basis for such extension, because samples unfamiliar to all client models may yield consistently high reconstruction errors. Such signals could be further used to trigger rejection, expert review, or incremental model updating in a future adaptive battery-sorting framework [52].

## 5. Conclusion

This paper proposes a novel, privacy-preserving federated inference framework designed to address the three critical barriers currently impeding the circular economy of retired batteries: the prevalent unavailability of historical cycling data, the statistical heterogeneity inherent in non-IID data silos, and the stringent privacy constraints prohibiting the sharing of raw data or model parameters.

To address the first barrier, our framework operates exclusively on the 41-dimensional voltage features obtained via the pulse testing method, enabling accurate classification without any dependency on historical battery information. To resolve the latter two challenges, we introduce a dual-stream architecture comprising a classification sub-model and a judgment sub-model. This design facilitates the precise quantification of client-level expertise (via the ‘Expert Index’), enabling an ‘expert-weighted’ aggregation mechanism that effectively mitigates the impact of non-IID heterogeneity. More importantly, this entire process operates under an output-only collaboration architecture, which supports data privacy preservation by precluding the transmission of raw data or model parameters.

Extensive experiments across 100 distinct random partitioning scenarios support the effectiveness of the framework. The method achieves a high average accuracy of 0.963 and a Macro F1 of 0.965. The close convergence of these metrics provides evidence that the framework mitigates label-distribution and data-quantity skew without historical data. Beyond average performance, the results reveal a notable finding regarding robustness: our federated framework demonstrates higher resilience than the theoretical centralized upper bound. While the monolithic centralized model suffered performance dips in “pathological” partitions (dropping to an accuracy of 0.875), our framework established a significantly higher performance floor (0.907). These results indicate that the robust ensemble of specialists provides greater robustness against severe performance degradation, effectively mitigating the “single point of failure” risk inherent in centralized training.

Crucially, these performance gains are achieved within a “Privacy-by-Design” architecture. By transmitting only abstract probability vectors and scalar error scores, the framework eliminates the need to share raw data, statistical properties, or model gradients, thereby effectively reducing exposure to reconstruction and inversion attacks, such as gradient inversion. In addition, the current results should be interpreted as dataset-defined battery type identification rather than isolated cathode-chemistry identification, because the learned latent features may also reflect correlated factors such as nominal capacity, aging state, cell design, and test-response amplitude. In summary, this work

provides a scalable solution for the battery recycling industry, enabling stakeholders to overcome data barriers and collaboratively leverage industry-wide expertise, regardless of the batteries' historical traceability or proprietary restrictions. Future work will extend this snapshot-based, expert-weighted paradigm to regression tasks, such as State-of-Health (SOH) estimation, to further support the circular economy of electric vehicle batteries.

## CRediT authorship contribution statement

**Hang Hu:** Formal analysis. **Xinghao Huang:** Formal analysis. **Chen Liang:** Formal analysis. **Ziyang Lyu:** Formal analysis. **Lin Su:** Formal analysis. **Kaisan Li:** Formal analysis. **Zhaoye Qin:** Funding acquisition. **Huadong Mo:** Funding acquisition. **Xuan Zhang:** Funding acquisition. **Changfu Zou:** Funding acquisition. **Shengyu Tao:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

This work was supported by the Swedish Research Council through a Project Grant (Grant No. 2023-04314), the European Union's Horizon Europe programme through the Marie Skłodowska-Curie Actions (Grant No. 101131278 and Grant No. 101283078), Australian Economic Accelerator Ignite Grant 'Health status estimation and resilient closed-loop supply chain for retired electric vehicle batteries' (IG240100338) and the ACT Government under the Energy Innovation Fund 'Online health monitoring and anomaly detection in li-ion batteries via trustworthy AI'.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.etran.2026.100607.

## Data availability

Data will be made available on request. Code can be found in https://github.com/AkrhamHorrorComing/FL-Retired-Battery-Sorting.

## References

[1] Harper G, Sommerville R, Kendrick E, Driscoll L, Slater P, Stolkin R, et al. Recycling lithium-ion batteries from electric vehicles. Nature 2019;575:75–86. https://doi.org/10.1038/s41586-019-1682-5.
[2] International Energy Agency. Global EV outlook 2024. Paris: International Energy Agency; 2024.
[3] Chen Q, Hou Y, Lai X, Shen K, Gu H, Wang Y, et al. Evaluating environmental impacts of different hydrometallurgical recycling technologies of the retired nickel-manganese-cobalt batteries from electric vehicles in China. Separ Purif Technol 2023;311:123277. https://doi.org/10.1016/j.seppur.2023.123277.
[4] Fan E, Li L, Wang Z, Lin J, Huang Y, Yao Y, et al. Sustainable recycling technology for Li-Ion batteries and beyond: challenges and future prospects. Chem Rev 2020;120:7020–63. https://doi.org/10.1021/acs.chemrev.9b00535.
[5] Arshad F, Azam MU, Manurkar N, Zhang F, Idrees BS, Ahmad A, et al. Life cycle assessment of lithium-ion and secondary batteries: a comparative analysis on environmental impacts and graphite recycling. eTransportation 2026;27:100514. https://doi.org/10.1016/j.etran.2025.100514.
[6] Harper G, Sommerville R, Kendrick E, Driscoll L, Slater P, Stolkin R, et al. Recycling lithium-ion batteries from electric vehicles. Nature 2019;575:75–86. https://doi.org/10.1038/s41586-019-1682-5.
[7] Olivetti EA, Ceder G, Gaustad GG, Fu X. Lithium-ion battery supply chain considerations: analysis of potential bottlenecks in critical metals. Joule 2017;1:229–43. https://doi.org/10.1016/j.joule.2017.08.019.
[8] Global EVOutlook 2025. n.d.
[9] Morse I. A dead battery dilemma. Science 2021;372:780–3. https://doi.org/10.1126/science.372.6544.780.
[10] White paper summarizing the findings from the U.S. environmental protection agency's battery labeling outreach and research (2021 to 2024). n.d.
[11] Why battery DPPs need RAIN RFID data carriers: addressing the limitations of QR codes. n.d.
[12] Wu Z, Zhao Z, Fang Z, Bao H, Liu P, Su X, et al. Rapid sorting of retired lithium-ion batteries using novel sorting feature extraction and a two-step classification method. J Power Sources 2025;654:237845. https://doi.org/10.1016/j.jpowsour.2025.237845.
[13] Wang Y-C, Chen K-C. Classification of aged batteries based on capacity and/or resistance through machine learning models with aging features as input: a comparative study. J Clean Prod 2024;471:143431. https://doi.org/10.1016/j.jclepro.2024.143431.
[14] Liu X, Tang Q, Feng Y, Lin M, Meng J, Wu J. Fast sorting method of retired batteries based on multi-feature extraction from partial charging segment. Appl Energy 2023;351:121930. https://doi.org/10.1016/j.apenergy.2023.121930.
[15] Gu X, Li J, Zhu Y, Wang Y, Mao Z, Shang Y. A quick and intelligent screening method for large-scale retired batteries based on cloud-edge collaborative architecture. Energy 2023;285:129342. https://doi.org/10.1016/j.energy.2023.129342.
[16] Liu T, Chen X, Peng Q, Peng J, Meng J. An enhanced sorting method for retired battery with feature selection and multiple clustering. J Energy Storage 2024;87:111422. https://doi.org/10.1016/j.est.2024.111422.
[17] Sulzer V, Mohtat P, Aitio A, Lee S, Yeh YT, Steinbacher F, et al. The challenge and opportunity of battery lifetime prediction from field data. Joule 2021;5:1934–55. https://doi.org/10.1016/j.joule.2021.06.005.
[18] Aitio A, Howey DA. Predicting battery end of life from solar off-grid system field data using machine learning. Joule 2021;5:3204–20. https://doi.org/10.1016/j.joule.2021.11.006.
[19] Fan E, Lin J, Zhang X, Yang J, Chen R, Wu F, et al. Anode materials sustainable recycling from spent lithium-ion batteries: an edge-selectively nitrogen-repaired graphene nanoplatelets. eTransportation 2022;14:100205. https://doi.org/10.1016/j.etran.2022.100205.
[20] Naresh VS, Sriram VS, Krishna VJ, Chandini VD, Sri RN, Durga KJ, et al. Privacy-preserving state of health prediction for electric vehicle batteries: a comprehensive review. Comput Electr Eng 2024;118:109416. https://doi.org/10.1016/j.compeleceng.2024.109416.
[21] Kröger T, Belnarsch A, Bilfinger P, Ratzke W, Lienkamp M. Collaborative training of deep neural networks for the lithium-ion battery aging prediction with federated learning. eTransportation 2023;18:100294. https://doi.org/10.1016/j.etran.2023.100294.
[22] Kim T, Ochoa J, Faika T, Mantooth HA, Di J, Li Q, et al. An overview of cyber-physical security of battery management systems and adoption of blockchain technology. IEEE J Emerg Sel Top Power Electron 2022;10:1270–81. https://doi.org/10.1109/JESTPE.2020.2968490.
[23] Li J, Khalatbarisoltani A, Che Y, Yang Y, Hu X. Data-secure and privacy-protected electric vehicle battery fault detection using decentralized federated learning with differential privacy. Energy Storage Mater 2025;83:104681. https://doi.org/10.1016/j.ensm.2025.104681.
[24] Ma L, Tian J, Zhang T, Guo Q, Chung C. Privacy-preserving federated semi-supervised learning for battery life prediction amid data scarcity. J Energy Storage 2025;128:117152. https://doi.org/10.1016/j.est.2025.117152.
[25] Lai R, Wang J, Tian Y, Tian J. FedCBE: a federated-learning-based collaborative battery estimation system with non-IID data. Appl Energy 2024;368:123534. https://doi.org/10.1016/j.apenergy.2024.123534.
[26] Chen X, Wang X, Deng Y. Federated learning-based prediction of electric vehicle battery pack capacity using time-domain and frequency-domain feature extraction. Energy 2025;319:135002. https://doi.org/10.1016/j.energy.2025.135002.
[27] Hu X, Xu L, Lin X, Pecht M. Battery lifetime prognostics. Joule 2020;4:310–46. https://doi.org/10.1016/j.joule.2019.11.018.
[28] Thelen A, Huan X, Paulson N, Onori S, Hu Z, Hu C. Probabilistic machine learning for battery health diagnostics and prognostics—review and perspectives. Npj Mater Sustain 2024;2:14. https://doi.org/10.1038/s44296-024-00011-1.
[29] Ye S, An D, Wang C, Zhang T, Xi H. Towards fast multi-scale state estimation for retired battery reusing via Pareto-efficient. Energy 2025;319:134848. https://doi.org/10.1016/j.energy.2025.134848.
[30] Liu Z, Meng B, Pan R, Zhou J. An enhanced sorting framework for retired batteries based on multi-dimensional features and an integrated clustering approach. Energy AI 2025;22:100612. https://doi.org/10.1016/j.egyai.2025.100612.
[31] Jiang T, Sun J, Wang T, Tang Y, Chen S, Qiu S, et al. Sorting and grouping optimization method for second-use batteries considering aging mechanism. J Energy Storage 2021;44:103264. https://doi.org/10.1016/j.est.2021.103264.
[32] Jiang S, Tao S, Lee J, Moura S. Defining an accuracy limit in battery state estimation. Joule 2026;10:102342. https://doi.org/10.1016/j.joule.2026.102342.
[33] Tao S, Liu H, Sun C, Ji H, Ji G, Han Z, et al. Collaborative and privacy-preserving retired battery sorting for profitable direct recycling via federated machine learning. Nat Commun 2023;14:8032. https://doi.org/10.1038/s41467-023-43883-y.
[34] Bagdasaryan E, Poursaeed O, Shmatikov V. Differential privacy has disparate impact on model accuracy. Adv Neural Inf Process Syst 2019;32.

[35] Tao S, Guo R, Lee J, Moura S, Casals LC, Jiang S, et al. Immediate remaining capacity estimation of heterogeneous second-life lithium-ion batteries via deep generative transfer learning. Energy Environ Sci 2025;18:7413–26. https://doi.org/10.1039/D5EE02217G.
[36] Tao S, Ma R, Chen Y, Liang Z, Ji H, Han Z, et al. Rapid and sustainable battery health diagnosis for recycling pretreatment using fast pulse test and random forest machine learning. J Power Sources 2024;597:234156. https://doi.org/10.1016/j.jpowsour.2024.234156.
[37] Tao S, Ma R, Zhao Z, Ma G, Su L, Chang H, et al. Generative learning assisted state-of-health estimation for sustainable battery recycling with random retirement conditions. Nat Commun 2024;15:10154. https://doi.org/10.1038/s41467-024-54454-0.
[38] Bernardi DM, Go J-Y. Analysis of pulse and relaxation behavior in lithium-ion batteries. J Power Sources 2011;196:412–27. https://doi.org/10.1016/j.jpowsour.2010.06.107.
[39] Waag W, Käbitz S, Sauer DU. Experimental investigation of the lithium-ion battery impedance characteristic at various conditions and aging states and its influence on the application. Appl Energy 2013;102:885–97. https://doi.org/10.1016/j.apenergy.2012.09.030.
[40] Kim J, Park S, Hwang S, Yoon W-S. Principles and applications of galvanostatic intermittent titration technique for lithium-ion batteries. J Electrochem Sci Technol 2021;13:19–31. https://doi.org/10.33961/jecst.2021.00836.
[41] Chien Y-C, Liu H, Menon AS, Brant WR, Brandell D, Lacey MJ. Rapid determination of solid-state diffusion coefficients in Li-based batteries via intermittent current interruption method. Nat Commun 2023;14:2289. https://doi.org/10.1038/s41467-023-37989-6.
[42] Smith K, Wang C-Y. Power and thermal characterization of a lithium-ion battery pack for hybrid-electric vehicles. J Power Sources 2006;160:662–73. https://doi.org/10.1016/j.jpowsour.2006.01.038.
[43] Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 2020;37:50–60. https://doi.org/10.1109/MSP.2020.2975749.
[44] Fair evaluation of federated learning algorithms for automated breast density classification: the results of the 2022 ACR-NCI-NVIDIA federated learning challenge. Med Image Anal 2024;95:103206. https://doi.org/10.1016/j.media.2024.103206.
[45] Mcmahan HB, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-efficient learning of deep networks from decentralized data. 2016.
[46] Grataloup A, Jonas S, Meyer A. A review of federated learning in renewable energy applications: potential, challenges, and future directions. Energy AI 2024;17:100375. https://doi.org/10.1016/j.egyai.2024.100375.
[47] Yang Q, Liu Y, Cheng Y, Kang Y, Chen T, Yu H. Federated transfer learning. In: Yang Q, Liu Y, Cheng Y, Kang Y, Chen T, Yu H, editors. Federated learning. Cham: Springer International Publishing; 2020. p. 83–93. https://doi.org/10.1007/978-3-031-01585-4_6.
[48] Barai A, Uddin K, Widanage WD, McGordon A, Jennings P. A study of the influence of measurement timescale on internal resistance characterisation methodologies for lithium-ion cells. Sci Rep 2018;8:21. https://doi.org/10.1038/s41598-017-18424-5.
[49] Tao S, Moura S, Brandell D, et al. Artificial intelligence for battery reuse, recycling and remanufacturing. Nat. Rev. Clean Technol. 2026;2:366–81. https://doi.org/10.1038/s44359-026-00167-0.
[50] Assran M, Duval Q, Misra I, Bojanowski P, Vincent P, Rabbat M, et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive. Architecture 2023.
[51] He K, Tao S, Fu S, Fan H, Tao Y, Wang Y, et al. A Novel Quick Screening Method for the Second Usage of Parallel-connected Lithium-ion Cells Based on the Current Distribution. J Electrochem Soc 2023;170:030514. https://doi.org/10.1149/1945-7111/acbf7e.
[52] Huang X, Tao S, Liang C, et al. iMOE: prediction of second-life battery degradation trajectory using interpretable mixture of experts. Nat Commun 2026;17:2549. https://doi.org/10.1038/s41467-026-69369-1.