Evaluating the Impact of Data Partitioning and Client Selection on Federated Remaining Useful Life Prediction in Aviation

Faruk Ozdemir, Roy S. Kalawsky, & Mohammed M. Mabkhot

Advanced VR Research Centre, Wolfson School, Loughborough University, Loughborough, LE11 3TU, UK

F.Ozdemir@lboro.ac.uk
R.S.Kalawsky@lboro.ac.uk
M.Mabkhot@lboro.ac.uk

ABSTRACT

Federated Learning (FL) is increasingly explored for Remaining Useful Life (RUL) prediction in aviation, motivated by the distributed nature of operational data across operators and platforms, the need to learn from heterogeneous fleet conditions, and the requirement to preserve data ownership and intellectual property by avoiding raw data sharing. While existing studies report promising results, they rely on subjectively defined benchmarking setups, where non-Independent and Identically Distributed (non-IID) data partitioning, client selection, and comparison criteria are selected without systematic examination of the bias they may introduce. Consequently, it remains unclear whether reported performance differences arise from the learning method itself or from unexamined configuration choices. This paper investigates the bias induced by data partitioning and client selection configurations in FL for aviation RUL prediction. Representative heterogeneity and client selection scenarios, including operating-condition shift, are evaluated under systematic learning settings to isolate their effect on model outcomes. The results show that both partitioning and selection choices can materially influence reported performance independent of the underlying model, demonstrating that selection bias alone can alter fleet-level RUL estimates. These findings highlight the need for harmonised and well-grounded benchmarking practices to support objective comparison and credible evaluation of FL approaches in aviation applications.

1. INTRODUCTION

Prognostics and Health Management (PHM) has become a crucial part of modern aviation maintenance strategies transforming the reactive maintenance into proactive paradigm to ensure operational safety & reliability and cost effectiveness (Fu & Avdelidis, 2023; Kim et al., 2017; Raouf et al., 2025). A key capability within PHM is RUL estimation, which provides maintenance planners with an estimate of the time available before component failure. This information supports timely maintenance scheduling and reduces the risk of catastrophic failures (Boujamza & Elhaq, 2022; Sharma, 2024).

Traditionally, RUL estimation has relied on physics-informed models that provide high interpretability and strong analytical foundations. However, the increasing availability of aircraft sensor data and the need to learn from diverse operational conditions have encouraged the use of data-driven approaches. In particular, Machine Learning (ML) and deep learning methods have demonstrated strong capability in modelling complex temporal degradation patterns in aircraft systems (Fu & Avdelidis, 2023; Lee & Mitici, 2023; Li et al., 2024; Stanton et al., 2023)

Most existing prognostics studies adopt centralised learning architectures in which operational data from multiple sources are aggregated into a single repository to train predictive models (Jeong et al., 2026; Kabashkin, 2024; Kamei & Taghipour, 2023). By pooling data from diverse operational contexts such as flight routes, engine age groups and environments conditions, these approaches allow high-capacity ML architectures, including Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), to learn complex temporal dependencies from large-scale datasets (Du et al., 2023; Vermelin et al., 2024). Such centralised models often achieve strong predictive accuracy and provide a holistic fleet-level learning perspective (Kabashkin, 2024; Vermelin et al., 2024).

Despite these advantages, centralised approaches face practical limitations in real aviation environments. Operational data are typically distributed across multiple organisations, including airlines, original equipment

manufacturers, and maintenance, repair and overhaul providers (Dalgkitsis et al., 2024; Jeong et al., 2026; Kabashkin, 2024). In many cases, regulatory constraints, privacy concerns, intellectual property protection prevent direct sharing of raw operational data between stakeholders (EASA, 2023; Landau et al., 2026; Mabkhot et al., 2025). These constraints have motivated the exploration of decentralised approaches that allow collaborative model training without transferring sensitive data.

FL has emerged as a promising decentralised training paradigm for such environments. In FL, multiple distributed clients collaboratively train a shared global model while keeping their local datasets private. Only model updates are exchanged between participants, enabling collaborative learning without exposing raw data (Dalgkitsis et al., 2024; Kabashkin, 2024; Landau et al., 2026).

Recent studies have investigated the use of FL for aviation prognostics. For example, Rosero et al. (2025) and Landau et al. (2026) demonstrated the potential of FL approaches for training RUL prediction models for aircraft components, highlighting the opportunity for cross-organisational collaboration. Similarly, Kamei & Taghipour (2023) and Dhada et al. (2020) proposed communication-efficient framework showing that federated models can achieve performance comparable to centralised learning while maintaining data security. More recent work has explored the integration of FL with Digital Twin technologies to address sensor drift and data corruption in aviation health monitoring systems (Kabashkin, 2025).

While these studies demonstrate the potential of FL for aviation, several methodological challenges remain insufficiently examined. Existing work typically evaluates proposed FL architectures under fixed experimental configurations, including predefined non-independent and identically distributed (non-IID) data partitions and fixed client selection policies. However, real aviation fleets naturally exhibit heterogeneous data distributions and varying client availability across operators and platforms. When experimental setups do not systematically vary these factors, it becomes difficult to determine whether reported performance improvements arise from the proposed learning method itself or from favourable configuration choices.

In addition, current studies rarely report client selection-related metrics that describe how frequently different clients contribute to training. Measures such as fleet coverage, the Herfindahl-Hirschman Index (HHI), and effective client counts can reveal whether federated training is dominated by a small subset of clients. Without such metrics, it is difficult to assess which fleet assets effectively influence the learning process.

Consequently, existing literature relies on predefined benchmarking setups that overlook the impact of data partitioning and client selection, potentially introducing bias (Li et al., 2024). As noted in recent surveys (Fu et al., 2023; Li et al., 2024; Soltani et al., 2022), real-world FL environments exhibit both statistical and system heterogeneity, which can lead biased models where minority data patterns are underrepresented (Smestad & Li, 2023; Soltani et al., 2022). Although adaptive client selection methods such as resource-aware scheduling (Nishio & Yonetani, 2019) and importance-based selection (Cho et al., 2020), have been proposed, they may introduce additional bias by prioritising high-loss or resource-rich clients and limiting the selection of underrepresented clients (Fu et al., 2023). Therefore, without systematic analysis of how data partitioning and client selection interact and impact on the model performance, it remains unclear whether reported performance improvements reflect realistic fleet conditions or favourable experimental setups.

The current gap lies in the lack of standardised protocols for designing federated environments that reflect aviation realities such as operating condition shifts and varying client availability. As a result, it remains unclear whether reported performance gains in FL studies arises from the proposed algorithms or from favourable data partitioning choices. This paper addresses this gap by investigating the impact of data partitioning and client selection on FL-based RUL prediction in aviation. It examines three questions: i) to what extent do selection policies and data heterogeneity affect global accuracy, ii) whether these factors interact, particularly under high non-IID conditions, iii) whether they lead to a measurable concentration and associated performance degradation.

To answer these questions, a controlled experimental framework is developed using standard FL baselines including FedAvg, FedProx, and SCAFFOLD with an LSTM backbone. The LSTM is selected because RUL prediction involves temporal degradation patterns in multivariate sensor data, and prior studies have shown LSTM models to provide effective baseline performance for this type of task (Ellefsen et al., 2019; Zhang et al., 2018). Experiments are conducted on the NASA C-MAPSS datasets under four partitioning scenarios representing different forms of heterogeneity. Three client selection policies are evaluated: uniform, diversity-oriented, and loss-based. In addition to prediction metrics such as RMSE and Mean Absolute Error (MAE), the study analyses selection bias using coverage rate, participation concentration (HHI), effective client count, and top-20% client selection share.

The contributions of this study are threefold. First, it introduces a reproducible benchmark framework that systematically varies data partitioning and client selection while keeping the model architecture and training budget fixed. Second, it provides statistical evidence showing that data heterogeneity and client selection policies significantly influence predictive performance. Third, it proposes a reporting framework encouraging future FL studies to

explicitly consider partitioning and client selection as key empirical factors.

The remainder of the paper is organised as follows. Section 2 presents the methodology and experimental setup. Section 3 reports the experimental design. Section 4 and 5 present the results and discussion. Section 6 concludes the paper and outlines future work.

2. METHODOLOGY

This study investigates federated RUL prediction under non-IID data partitioning and client selection conditions. The objective is to isolate the effects of (a) data heterogeneity and (b) client selection policy on predictive performance and participation concentration while keeping the model architecture, number of clients, and participation rate fixed across all experiments. A factorial experimental design is adopted using four partitioning scenarios, three heterogeneity levels, four client selection variants, and four C-MAPSS subsets (FD001-FD004), with three random seeds per configuration, as illustrated in Figure 1

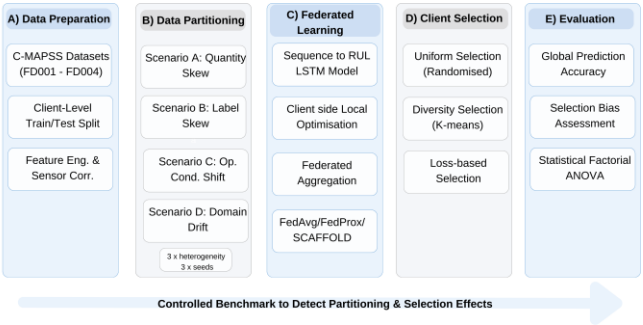

Figure 1 Federated RUL benchmarking methodology

2.1. Data Preparation

NASA C-MAPSS turbofan engine dataset covering all four subsets is used. RUL targets are clipped at 125 cycles following common practice. The official engine-level split between training and test sets is preserved to ensure that train and test engines remain disjoint and to prevent engine-level leakage. All features are standardised locally within each client using training statistics computed only from that client, and the same transformation is applied to the relating test data. Fixed-length sequences are constructed with window length L=30 and stride s=10. Each window is labelled using the RUL value at its final time step.

2.2. Non-IID Partitioning Scenarios

Four partitioning scenarios are defined to represent different non-IID structures, each with three heterogeneity levels (low, mid, high), as depicted in Figure 2 and Figure 3. In all scenarios, each engine is assigned to exactly one client, and assignment occurs before windowing. Scenario A (Quantity skew) introduces client quantity imbalance while keeping label and domain distributions broadly similar across clients. Under low heterogeneity, engines are split evenly across clients. Under mid and high heterogeneity, client proportions are sampled from a symmetric Dirichlet distribution (Dir(α)) over K clients, and engines are allocated without replacement to match these proportions. Values of α=1.0 and α=0.3 are used for mid and high heterogeneity, respectively, while ensuring at least one engine per client.

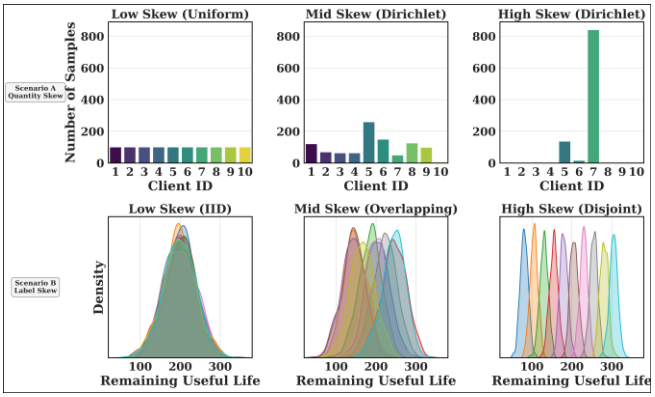

Figure 2 partitioning A & B scenarios

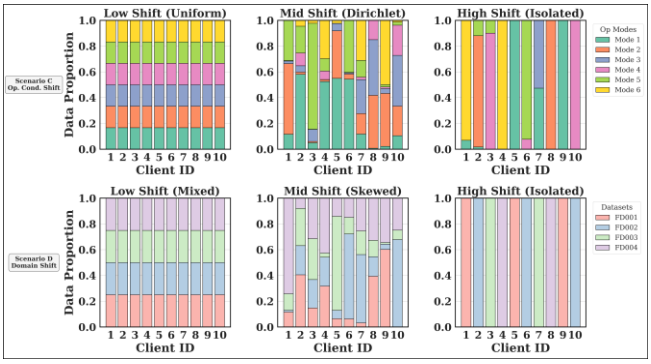

Figure 3 Partitioning C & D scenarios

Scenario B (RUL skew) introduces label skew. RUL values are grouped into B = 5 quantile-based bins, and each engine is assigned a dominant bin based on the mode across its cycles. Engines from each bin are distributed across clients using a Dirichlet allocation over clients. Heterogeneity is controlled through α, with values of 3.0, 1.2, and 0.2 for low, mid, and high heterogeneity, respectively, while ensuring at least one engine per client.

Scenario C (Operating condition shift) introduces domain shift based on operating conditions. Per-engine operating-condition cluster labels are obtained using K-means with k=3 on per-engine means of settings 1, 2, and 3 (random state= 0, $n_{init} = auto$). Engines are allocated using a dominant-mix rule in which each client is associated with a dominant cluster and receives approximately a fraction p_{dom} of its engines from that cluster, with the remainder

drawn from the other clusters. Values of p_{dom}=0.4, 0.7, and 0.95 are used for low, mid, and high heterogeneity, respectively.

Scenario D (General domain shift) represents a broader domain shift using both settings and sensor statistics. Per-engine domain labels are derived using K-means with k=3 (random state= 0, n_{init}=$auto$) on feature vectors containing mean settings 1, 2, and 3 together with per-engine mean and standard deviation of sensor channels. Client assignment follows the same dominant-mix scheme as Scenario C using the same p_{dom} values. Table 1 summarises all scenario configurations.

Table 1 Heterogeneity parameters by scenario.

Scenario	Low	Mid	High
A (Quantity skew)	Even	Dirα = 1.0	Dir α =0.3
B (Rul Skew)	α=3.0	α=1.2	α=0.2
C (OC shift)	p_{dom} =0.4	p_{dom}=0.7	p_{dom}=0.95
D (Domain Shift)	p_{dom}=0.4	p_{dom}=0.7	p_{dom}=0.95

2.3. Federated Learning Setup

FedAvg, FedProx, and SCAFFOLD are adopted using an LSTM-based sequence regressor. The federated learning process is illustrated in Figure 4, where the server broadcasts the global model to participating clients, local training is performed on private engine data, and updated client models are aggregated to produce the next global model. The number of clients is fixed at K=20 for controlled benchmarking. Client participation rate (CPR) is varied as 0.2, 0.4, and 0.6, corresponding to 4, 8, and 12 clients per round, respectively, to evaluate the effect of partial participation on performance and participation concentration. CPR=0.4 is used for the main experiments as a balanced reference setting. A centralised baseline trained on pooled data is included for comparison under the same model and preprocessing setup. Training is performed for 300 communication rounds, and client updates are aggregated using sample-count weighting.

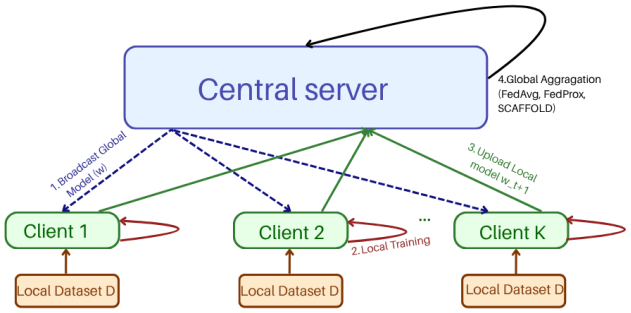

Figure 4 Federated learning process

Local training uses 3 epochs per round with batch size 32 and Adam optimiser (learning rate = 0.001, weight decay = 0.0001). Python, NumPy, and PyTorch random seeds are fixed for reproducibility. Participation concentration metrics are computed from observed client selection frequencies across rounds.

2.4. Client Selection Policies

Three Client Selection (CS) policies are evaluated: uniform, diversity-based, and loss-based selection, as summarised in Figure 5.

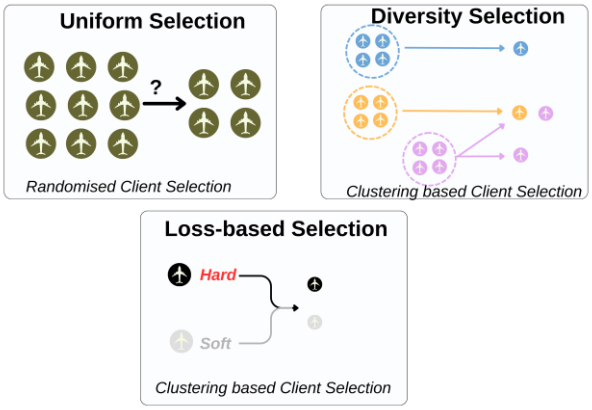

Figure 5 Client selections

In CS-Uniform, each client has equal probability of being selected in a given round, as defined in Eq. (1), where S_t denotes the selected client set at round t, m is the number of participating clients, and K is the total number of clients.

$$\mathbb{P}(i \in \mathcal{S}_t) = \frac{m}{K}, \quad i \in \{1,..,K\} \tag{1}$$

For CS-Diversity-based, client embeddings $\{\mathbf{e}_i\}_{i=1}^{K}$ are computed using the mean feature vector of each client i's dataset and clustered using K-means, as defined in Eq. (2). One client is selected from each non-empty cluster, while remaining slots are filled uniformly from the remaining clients. This policy aims to increase diversity in client set.

$$C_1,..,C_G = KMeans(e_1,..,e_K) \tag{2}$$

In CS-Loss-based, client selection is driven by local training loss from the previous round $l_i^{(t-1)}$. Two variants are considered: soft loss-based selection, where clients are sampled using softmax probabilities defined in Eq. (3), and hard loss-based selection, where the m highest-loss clients are selected according to Eq. (4). For loss-based selection, the first five rounds use CS-Uniform to initialise client loss estimates.

$$p_i^{(t)} = \frac{exp\left(l_i^{(t-1)}/\tau\right)}{\sum_{j=1}^{K} exp\left(l_j^{(t-1)}/\tau\right)} \quad (3)$$

$$\mathcal{S}_t = \arg\max_{|\mathcal{S}|=m} \sum_{i\in\mathcal{S}} l_i^{(t-1)} \quad (4)$$

2.5. Evaluation Metrics

Global RMSE and MAE are reported on the pooled test set, as defined in Eqs. (5) and (6), where N denotes the number of test samples, y_i the true RUL value, and $\hat{y}_i$ the predicted value for sample i. Per-client test RMSE is also computed using the subset of test engines assigned to each client under the corresponding partitioning scheme.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (5)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (6)$$

Client participation concentration is evaluated using four metrics derived from client selection frequencies across rounds. Coverage ratio in Eq. (7) measures the proportion of clients selected at least once, where S denotes the selected client set and K the total number of clients. HHI in Eq. (8) measures selection concentration, where p_i denotes the selection probability of client i. The effective number of clients in Eq. (9) provides an inverse concentration measure, while the top-20% selection share in Eq. (10) quantifies the fraction of selections attributed to the most frequently selected clients, where denotes by $\mathcal{T}_{20}$. Client participation disparity is assessed by comparing the mean per-client RMSE of the top 20% most-selected clients and the bottom 30% least-selected clients. The selected-minus-unselected gap is defined in Eq. (11), where positive values indicate worse performance for frequently selected clients.

$$\text{Coverage} = \frac{|\mathcal{S}|}{K} \quad (7)$$

$$\text{HHI} = \sum_{i=1}^{K} p_i^2 \quad (8)$$

$$N_{\text{eff}} = \frac{1}{\sum_{i=1}^{K} p_i^2} \quad (9)$$

$$\text{Top20Share} = \sum_{i\epsilon\tau_{20}} p_i \quad (10)$$

$$\Delta_{\text{gap}} = RMSE_{sel} - RMSE_{unsel} \quad (11)$$

2.6. Statistical Analysis

Factorial analysis is conducted across scenario (A-D), heterogeneity (low/mid/high), selection policy (CS-Uniform, CS-Diversity, CS-Loss), and dataset (FD001-FD004), with seed treated as a blocking factor. OLS models with HC3 (Heteroskedasticity Consistent) robust standard errors are fitted for RMSE, MAE, and the selected-minus-unselected RMSE gap. Main effects and selection × heterogeneity interactions are analysed using partial η^2 as an effect-size measure, defined in Eq. (12).

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}} \quad (12)$$

Within each (scenario, heterogeneity) combination, paired comparisons are performed between selection policies using matched seeds. Furthermore, to quantify the relationship between client participation bias and global model performance, Pearson and Spearman correlation coefficients are computed between the final RMSE and participation metrics.

3. EXPERIMENTAL DESIGN

Table 2 summarises the factorial experimental design across scenario (A-D), heterogeneity level (low, mid, high), selection policy (CS-Uniform, CS-Diversity, CS-Loss), and dataset (FD001-FD004). Scenarios A-D represent quantity skew, RUL skew, operating-condition shift, and general domain shift, respectively. Model architecture, training hyperparameters, number of clients, and participation rate are fixed across all conditions so that performance differences can be attributed to partitioning and client selection effects rather than variations in training configuration. All experiments use 300 communication rounds and the same LSTM architecture with convergence and baseline comparison analysis. The design yields 576 configurations (4 scenarios × 3 heterogeneity levels × 4 selection policies × 4 datasets × 3 algorithms), each evaluated using three independent random seeds, resulting in 1728 runs. Seeds are treated as a blocking factor in the statistical analysis, and results are aggregated using the median and interquartile range (IQR).

Table 2 Summary of the experimental setup

Factor	Levels
Scenario	A: quantity skew; B: RUL; C: operating-condition shift; D: domain shift
Heterogeneity	low, mid, high (Dirichlet α or p_{dom})
Selection policy	CS-Uniform, CS-Diversity, CS-Loss (CS SOFT and CS HARD)
Dataset	FD001, FD002, FD003, FD004
Algorithms	FedAvg, FedProx, SCAFFOLD
Clients / CPR	K = 20; CPR = 0.2, 0.4, 0.6; 300 rounds
Total design	1728
*L, S and RUL clip	30, 10, and125 cycle

*L :Window length; S :stride

4. RESULTS

4.1. Convergence Analysis, Centralised Baseline and FL Comparison

Figure 6 compares convergence behaviour of the federated models and the centralised baseline across the four C-MAPSS datasets. All FL algorithms show stable convergence, with global test RMSE decreasing from initial values above 90-105 to stable plateaus within approximately 60-195 rounds depending on the dataset. FedAvg and FedProx follow closely matched trajectories, while SCAFFOLD exhibits faster initial convergence in some cases (FD003: round 60, FD004: round 160) before reaching comparable final performance.

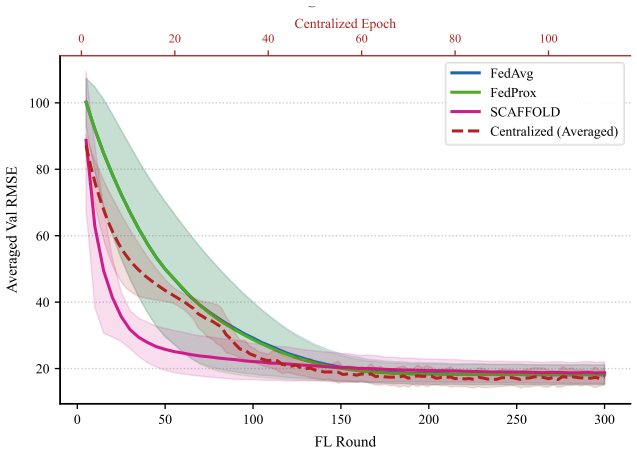

Figure 6 Convergence comparison between Federated and Centralised baseline learning across four datasets

The centralised baseline trained on pooled data converges within approximately 80-120 epochs and consistently achieves lower RMSE than the federated approaches, as summarised in Table 3. For example, the centralised model achieves RMSE values of 14.03 ±0.29 on FD001 and 11.88 ±0.16 on FD003, with the largest performance gap observed for FD003. Overall, the convergence behaviour indicates that 300 communication rounds are sufficient to achieve stable performance across all algorithms.

Table 3 Centralized vs. FL, RMSE and MAE per Dataset (mean ± std)

Method	FD001 RMSE/ MAE	FD002 RMSE/ MAE	FD003 RMSE/ MAE	FD004 RMSE/ MAE
Centralized	**14.03 ± 0.29 /10.48 ± 0.30**	**18.42 ± 0.53 / 13.34 ± 0.18**	**11.88 ± 0.16 / 6.98 ± 0.34**	**19.74 ± 0.25 / 14.70 ± 0.51**
FedAvg	16.47 ± 1.61 / 13.75 ± 2.28	20.71 ± 1.27 / 16.28 ± 1.35	15.02 ± 2.22 / 13.45 ± 3.22	20.21 ± 1.24 / 14.34 ± 1.29
FedProx	16.55 ± 1.62 / 13.80 ± 2.18	20.92 ± 1.81 / 15.82 ± 1.13	15.03 ± 2.16 / 13.56 ± 3.44	20.43 ± 1.50 / 13.53 ± 1.01
SCAFFOLD	17.97 ± 3.75 / 15.15 ± 3.88	21.16 ± 1.46 / 17.49 ± 1.64	16.03 ± 3.21 / 13.84 ± 4.53	19.78 ± 1.21 / 14.95 ± 1.48

4.2. Effect of CPR

To assess sensitivity to communication budget, three client-per-round (CPR) settings (0.2, 0.4, and 0.6) are compared under the same partitioning and client-selection conditions. Figure 7 shows that increasing CPR from 0.2 to 0.6 reduces median global test RMSE from 17.4 to 15.8 while also reducing variability across experimental conditions. The effect differs across selection policies as demonstrated in Figure 8. Uniform and diversity-based selection remain relatively stable across CPR values, whereas hard loss-based selection shows the greatest sensitivity, with mean RMSE decreasing from approximately 22 at CPR = 0.2 to approximately 16 at CPR = 0.6.

Higher participation rates therefore partially reduce the effect of concentrated client selection. Based on this behaviour, CPR = 0.4 is selected for the remaining experiments as a balanced participation setting that preserves clear performance differences between selection policies.

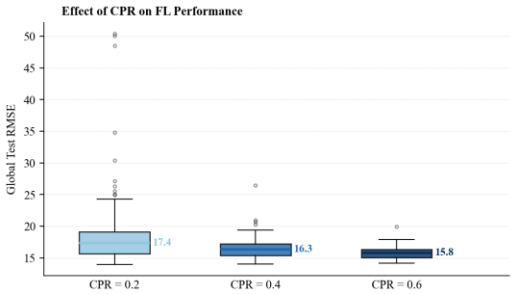

Figure 7 Effect of CPR on FL Performance

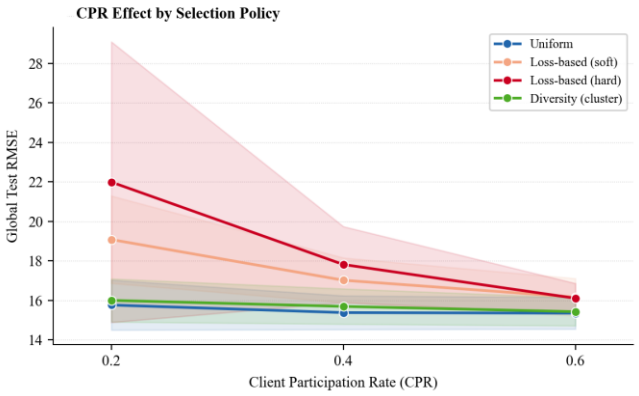

Figure 8 Effect of client participation rate on FL performance

4.3. Main Effects and Interaction

Figure 9 summarises RMSE distributions pooled across FD001-FD004 for FedAvg, FedProx, and SCAFFOLD under the four partitioning regimes and client selection policies. Partitioning shows a clear effect on both error magnitude and variability. Domain shift (D) is generally the most challenging regime, whereas operating-condition shift (C) produces the lowest dispersion and the smallest differences across selection policies. RUL skew (B) occupies an intermediate position but shows greater sensitivity to unstable participation patterns.

Client selection also influences RMSE. Uniform and diversity-based selection generally produce the lowest median RMSE, followed by soft loss-based selection, while hard loss-based selection produces the highest RMSE and widest spread. FedAvg and FedProx show broadly similar behaviour, whereas SCAFFOLD often achieves competitive median performance but can produce high-RMSE outliers under RUL skew and domain shift, particularly when high-loss clients are repeatedly selected. These results indicate an interaction between partitioning and client selection. The effect of loss-based participation becomes more pronounced under RUL skew and domain shift, but is weaker under operating-condition shift, where RMSE is less sensitive to the selection policy.

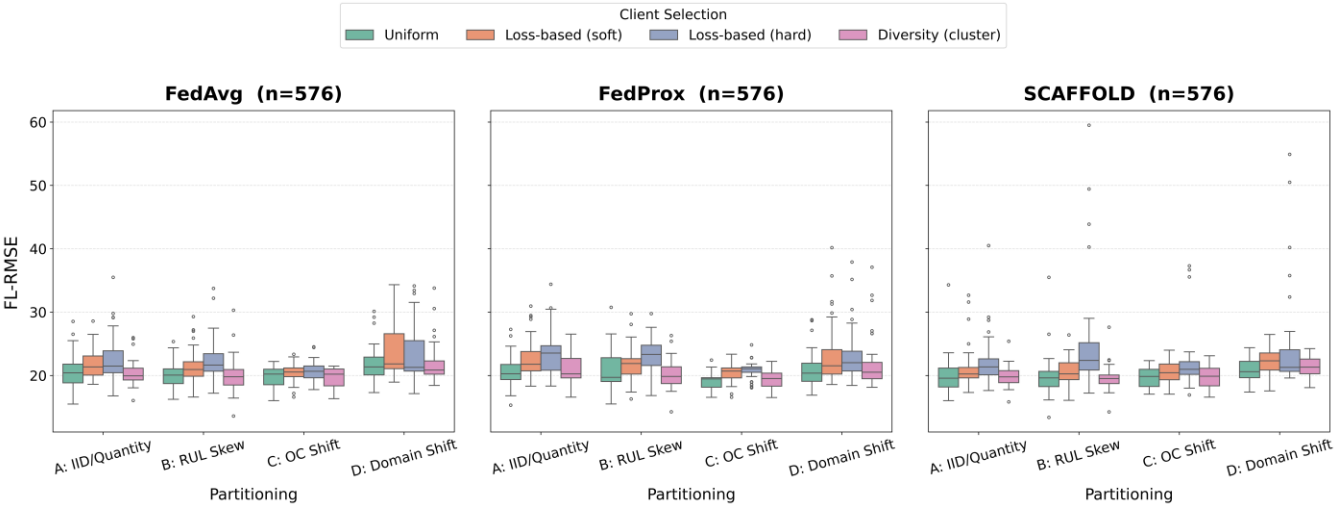

Figure 9 RMSE distribution across partitioning and client selection policies

Figure 10 shows that partitioning strategy is the strongest factor affecting RMSE, accounting for 11.1% of the variance (partial $\eta^2 = 0.111$, $p < 0.001$). Client selection follows closely at 9.9% ($p < 0.001$), indicating that participation policy has an effect comparable to the underlying data distribution. Dataset identity explains a smaller but statistically significant 5.7% ($p < 0.001$), reflecting baseline differences across the four C-MAPSS subsets.

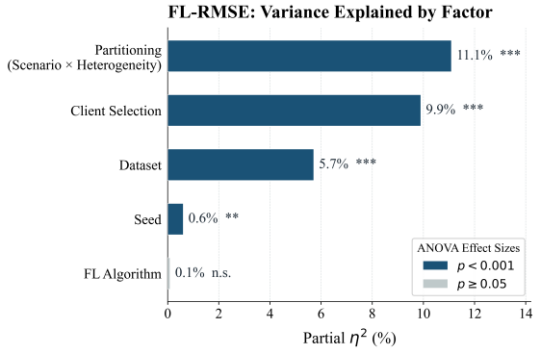

Figure 10 RMSE variance explained by experimental factor.

In contrast, FL algorithm explains only 0.1% of the variance and is not statistically significant ($p = 0.45$), suggesting that the evaluated algorithms reach broadly similar error levels once partitioning and selection conditions are fixed. Seed contributes less than 1% of the variance ($p < 0.01$), indicating stable behaviour across runs. Overall, partitioning and client selection emerge as the dominant factors influencing benchmark performance.

4.4. Participation Concentration

Client selection policy is the dominant driver of participation concentration in Figure 11, whereas partitioning scenario (A-D) has minimal influence across the four metrics. Coverage ratio remains 1.0 across all policies and scenarios, confirming that all clients are selected at least once during training. The clearest separation appears in HHI and effective number of clients. Soft loss-based selection produces the highest concentration (HHI ≈ 0.08-0.09, effective clients ≈12-13), whereas uniform selection yields the lowest concentration (HHI ≈0.05, effective clients ≈ 20) and the lowest top-20% selection share (≈0.23). Diversity-based selection occupies an intermediate position, with HHI≈0.063 and effective clients ≈16 across all scenarios. Hard loss-based selection shows concentration levels comparable to uniform selection in HHI and effective-client metrics. This observation is consistent with the broader RMSE trends, while also suggesting that selection frequency and client composition may influence performance in different ways.

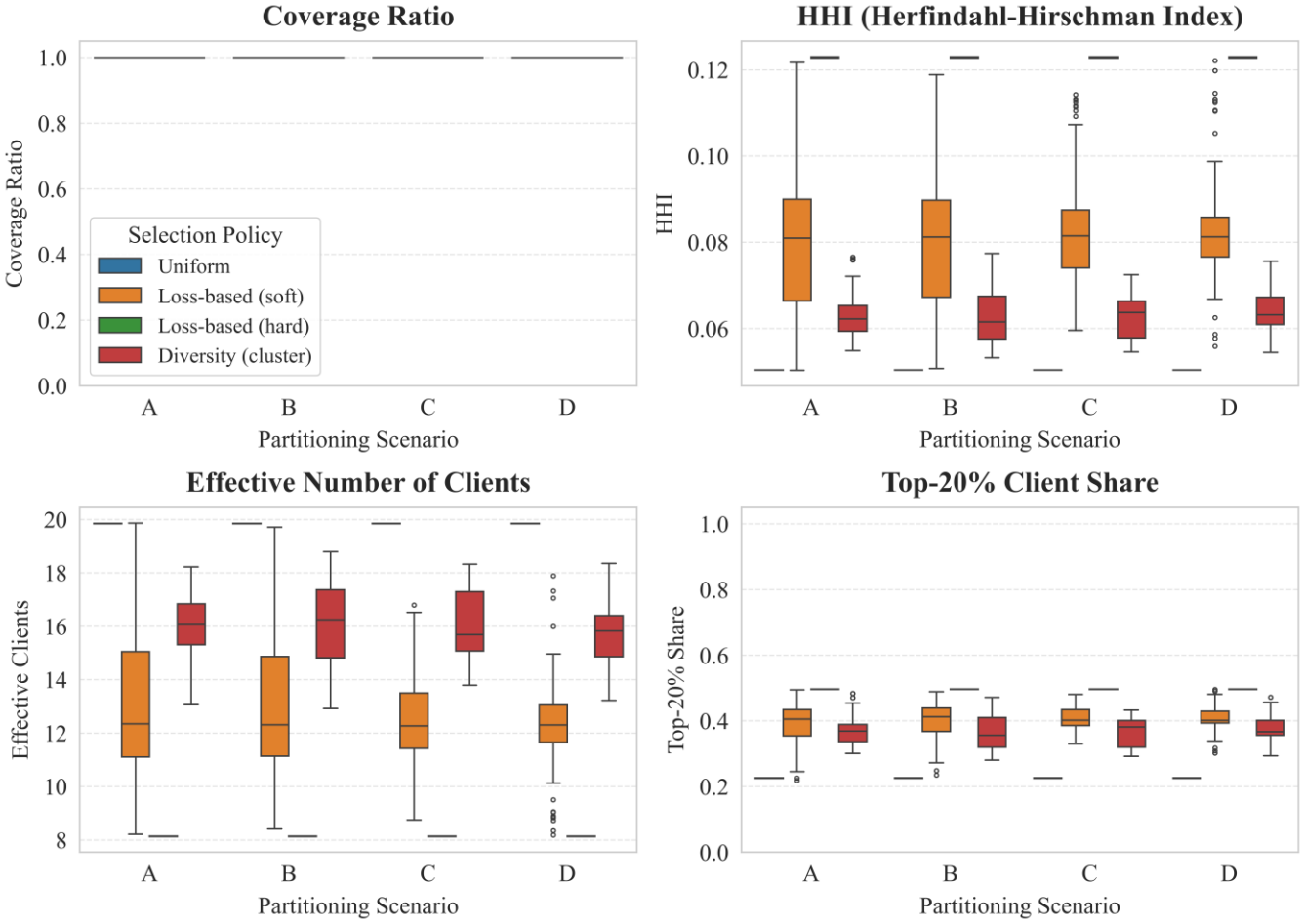

Figure 11 Participation concentration by selection policy.

4.5. Selected vs Unselected Disparity

Figure 12 shows performance disparity between highly selected and rarely selected clients across selection policies, heterogeneity levels, and partitioning scenarios. Absolute RMSE gaps increase with heterogeneity across all policies (Figure 12a &12b), with gaps exceeding 10 RMSE units under high heterogeneity in Scenario A (Quantity skew). In contrast, Scenario C (Operating-condition shift) consistently produces the smallest gaps across all policies (Figure 12c). Diversity-based selection exhibits the highest variability in gap magnitude, whereas uniform selection remains the most stable.

Figure 12d reveals a policy-specific pattern. Under diversity-based and hard loss-based selection, highly selected clients exhibit higher mean RMSE than rarely selected clients, indicating repeated participation of more difficult client distributions. In contrast, uniform and soft loss-based selection show slightly higher RMSE among rarely selected clients. Hard loss-based selection produces the highest mean RMSE among selected clients (≈23.5), suggesting that prioritising high-loss clients can concentrate difficult optimisation cases within the active training set.

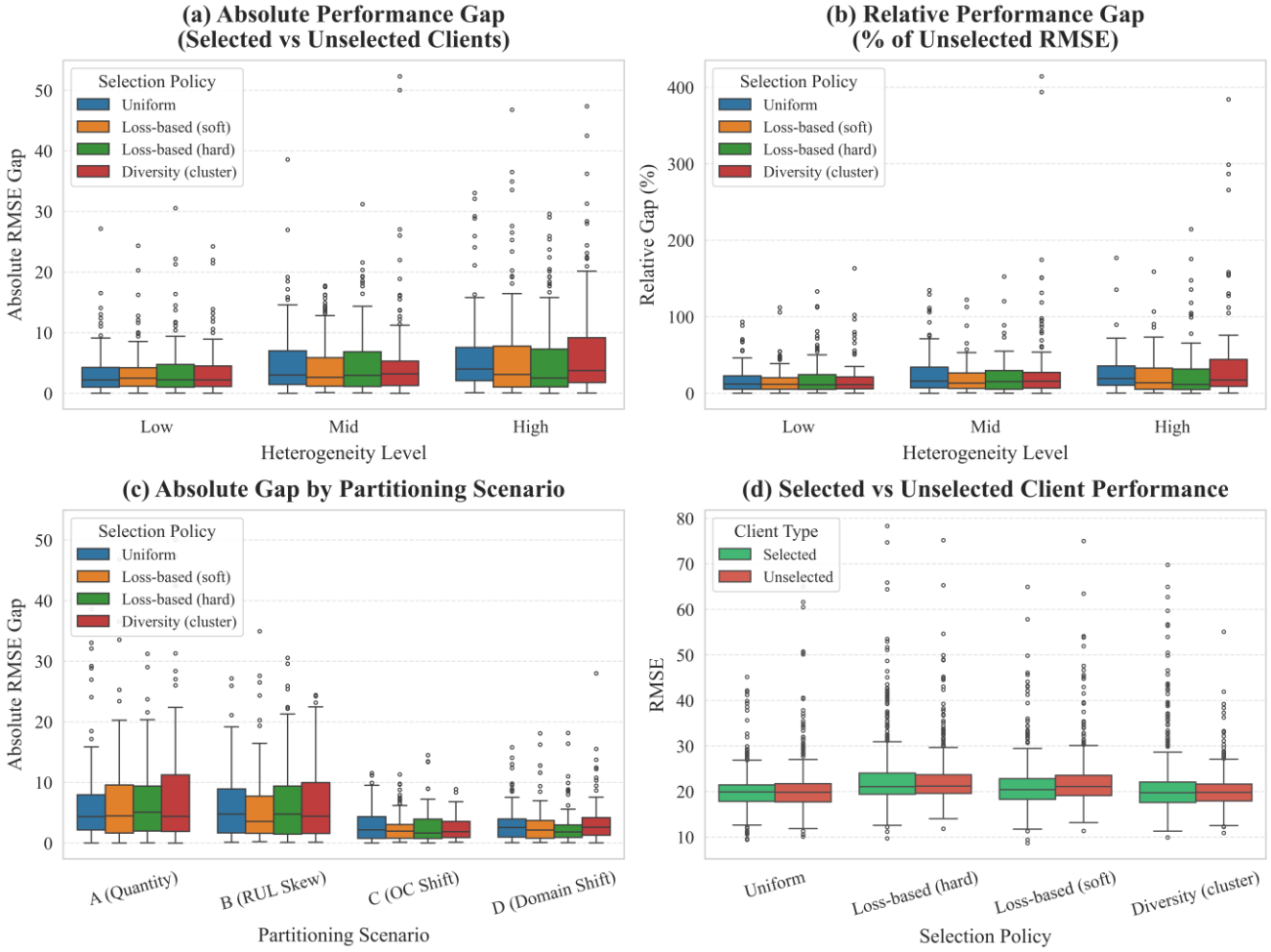

Figure 12 Selected versus unselected performance disparity.

4.6. Mechanism Evidence

Figure 13 shows the relationship between participation concentration and RMSE across all runs. Higher HHI and lower effective client counts generally coincide with higher RMSE, indicating that more concentrated participation is associated with poorer federated performance. However, this relationship is weaker within individual selection policies, suggesting that the pooled association is driven primarily by differences between policies rather than by run-to-run variability within a fixed policy.

Hard loss-based selection exhibits both the highest concentration and the poorest RMSE behaviour, consistent with the ANOVA results showing a strong main effect of client selection on performance. The relationship between concentration and selected-versus-unselected disparity is weaker overall, although diversity-based selection shows a noticeable positive association in which higher concentration coincides with larger RMSE gaps. The results suggest that FL performance is influenced more strongly by the structure of the client selection strategy than by small variations in participation concentration within a given policy.

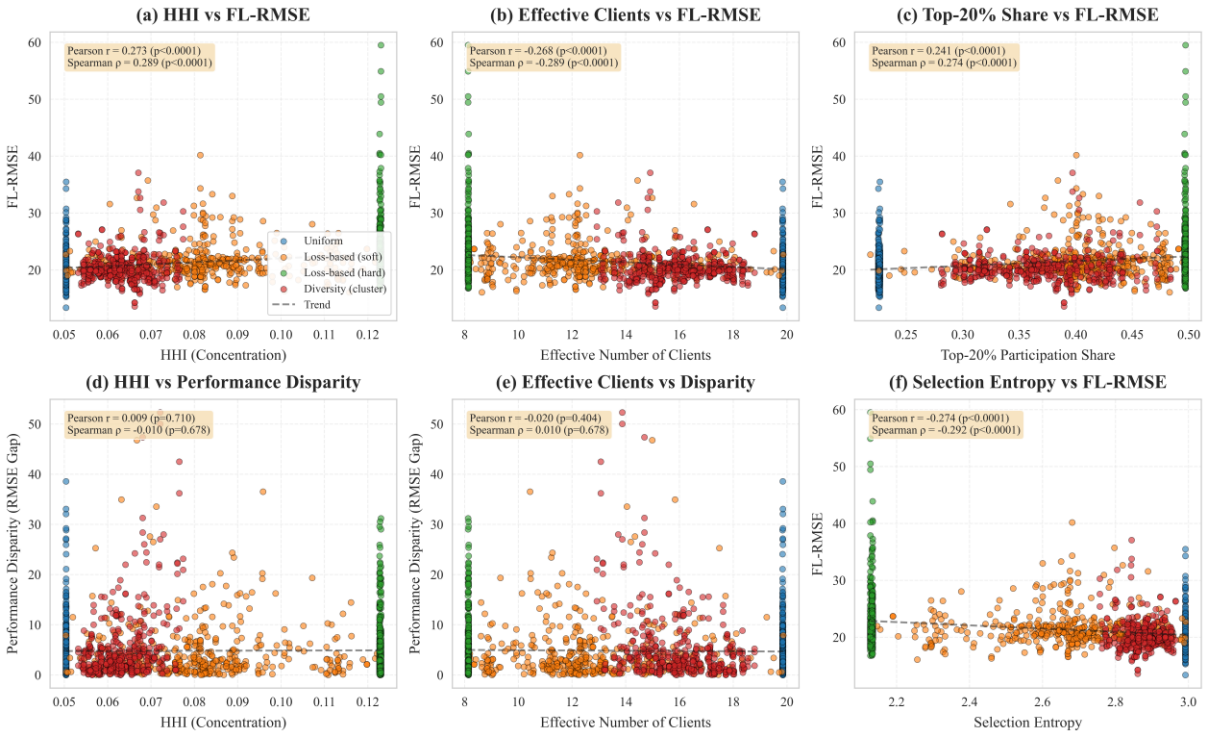

Figure 13 Relationship between participation concentration and RMSE.

5. DISCUSSION

The main finding of this study is that FL performance for aviation RUL prediction is strongly influenced by the interaction between data partitioning and client selection. Across the evaluated settings, the same LSTM architecture produced broadly similar behaviour under FedAvg, FedProx, and SCAFFOLD once partitioning regime and selection policy were fixed, while the largest RMSE differences emerged across heterogeneity structures and participation patterns. These findings are consistent with recent FL studies reporting strong sensitivity to non-IID data and participation imbalance. The results therefore suggest that partitioning and client selection should be treated as important benchmarking factors, as they can substantially influence the reported performance. The limited differences between FedAvg, FedProx, and SCAFFOLD further indicate that, under the evaluated conditions, heterogeneity structure had greater influence on performance than the aggregation correction mechanism itself.

The participation analysis provides additional insight into client selection bias. Loss-based selection, particularly the hard variant, concentrated participation on a narrower subset of clients and was generally associated with higher prediction error. The selected-versus-unselected analysis further supports this interpretation. Under diversity-based and hard loss-based selection, selected clients frequently exhibited higher RMSE than unselected clients, indicating repeated participation of more difficult or error-prone client distributions. In contrast, uniform and soft loss-based selection produced smaller or reversed gaps, suggesting that different participation patterns affect client groups differently. This is relevant in aviation settings, where a policy that improves average performance may still produce uneven utility across operators or fleet segments.

The comparison with the centralised baseline provides an additional reference point for interpreting the federated results. As expected, pooled-data training achieved lower RMSE than the federated models, confirming the performance trade-off associated with decentralised training under the same architecture and preprocessing conditions. At the same time, the federated models remained competitive under milder heterogeneity levels, suggesting that decentralised collaboration remains feasible when client distributions are less extreme. The performance gap increased under more challenging partitioning regimes, particularly domain shift, reinforcing the importance of reporting federated results relative to a pooled upper-bound reference rather than in isolation.

The client-per-round analysis also provides a practical observation. Increasing participation improved stability and reduced error, although this remains constrained by communication and client-availability limitations. Uniform and diversity-based selection were relatively stable across participation levels, whereas hard loss-based selection benefited most from increased participation. This indicates that communication budget and selection strategy should be

considered jointly, since broader participation can partially reduce the negative effects of concentrated client selection.

One limitation of this study is generalisability, as the empirical evaluation is conducted only on C-MAPSS. Nevertheless, the proposed framework is dataset-agnostic and can be extended to other aircraft subsystems or prognostics datasets by adapting the partitioning and client assignment rules. The broader contribution lies in the benchmarking protocol itself, which supports systematic and reproducible analysis of partitioning and client selection effects across different federated prognostics settings.

These results suggest that future aviation FL studies should treat partitioning, selection policy, and communication budget as important experimental factors rather than secondary implementation details. Benchmarks that report only algorithmic differences without controlling for these factors risk overstating the effect of the learning rule itself. More broadly, meaningful FL evaluation in aviation requires benchmark designs that reflect operational heterogeneity, partial participation, and the practical constraints of fleet collaboration.

6. CONCLUSION AND FUTURE WORK

This study investigated the influence of data partitioning and client selection on federated RUL prediction for aviation using a controlled benchmarking framework. Using an LSTM-based architecture with FedAvg, FedProx, and SCAFFOLD, the results showed that performance is influenced more strongly by partitioning regime and participation pattern than by the FL aggregation algorithm itself. Heterogeneity structure and client selection policy affected both global prediction error and participation concentration, demonstrating that these factors should be treated as important experimental variables in federated aviation studies.

The proposed framework provides a reproducible benchmarking protocol that can be extended to other aircraft subsystems or prognostics datasets through modified partitioning and client assignment rules. The results suggest that future aviation FL studies should report partitioning structure, selection policy, and participation settings explicitly rather than treating them as secondary implementation details. Future work will focus on more realistic heterogeneity formulations and client selection mechanisms that better reflect operational constraints and reduce participation bias.

ACKNOWLEDGEMENT

The first author, Faruk Ozdemir, gratefully acknowledges the financial sponsorship and PhD scholarship provided by the Ministry of National Education of the Republic of Turkey.

NOMENCLATURE

ANOVA	Analysis of Variance
C-MAPSS	Commercial Modular Aero-Propulsion System Simulation
CNN	Convolutional Neural Network
CPR	Clients-Per-Round
CS	Client Selection
FL	Federated Learning
HHI	Herfindahl-Hirschman Index
HC	Heteroskedasticity Consistent
IID	Independent and Identically Distributed
IQR	Interquartile Range
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
OLS	Ordinary Least Squares
PHM	Prognostics and Health Management
RMSE	Root Mean Square Error
RUL	Remaining Useful Life

REFERENCES

Boujamza, A., & Lissane Elhaq, S. (2022). Attention-based LSTM for Remaining Useful Life Estimation of Aircraft Engines. *IFAC-PapersOnLine*, *55*(12), 450–455. https://doi.org/10.1016/j.ifacol.2022.07.353

Cho, Y. J., Wang, J., & Joshi, G. (2020). *Client Selection in Federated Learning: Convergence Analysis and Power-of-Choice Selection Strategies*.

Dalgkitsis, A., Koufakis, A., Stutterheim, J., Mifsud, A., Atwani, P., Gommans, L., De Laat, C., Papagianni, C., & Oprescu, A. (2024). Secure Collaborative Model Training with Dynamic Federated Learning in Multi-Domain Environments. *Proceedings of SC 2024-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 755–759. https://doi.org/10.1109/SCW63240.2024.00107

Dhada, M., Jain, A. K., Herrera, M., Perez Hernandez, M., & Parlikad, A. K. (2020). Secure and communications-efficient collaborative prognosis. *IET Collaborative Intelligent Manufacturing*, *2*(4), 164–173. https://doi.org/10.1049/iet-cim.2020.0035

Du, N. H., Long, N. H., Ha, K. N., Hoang, N. V., Huong, T. T., & Tran, K. P. (2023). Trans-Lighter: A light-weight federated learning-based architecture for Remaining Useful Lifetime prediction. *Computers in Industry*, *148*. https://doi.org/10.1016/j.compind.2023.103888

EASA. (2023). *Artificial intelligence Roadmap 2.0*.

Fu, L., Zhang, H., Gao, G., Zhang, M., & Liu, X. (2023). Client Selection in Federated Learning: Principles, Challenges, and Opportunities. *IEEE Internet of Things Journal*, *10*(24), 21811–21819. https://doi.org/10.1109/JIOT.2023.3299573

Fu, S., & Avdelidis, N. P. (2023). Prognostic and Health Management of Critical Aircraft Systems and Components: An Overview. In *Sensors* (Vol. 23, Number 19). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/s23198124

Jeong, C., Yue, X., & Chung, S. (2026). Fed-Joint: Joint modeling of nonlinear degradation signals and failure events for remaining useful life prediction using federated learning. *Reliability Engineering and System Safety*, *267*. https://doi.org/10.1016/j.ress.2025.111833

Kabashkin, I. (2024). Integration of Foundation Models and Federated Learning in AIoT-Based Aircraft Health Monitoring Systems. *Mathematics*, *12*(21). https://doi.org/10.3390/math12213428

Kabashkin, I. (2025). Framework for Addressing Imbalanced Data in Aviation with Federated Learning. *Information*, *16*(2), 147. https://doi.org/10.3390/info16020147

Kamei, S., & Taghipour, S. (2023). A comparison study of centralized and decentralized federated learning approaches utilizing the transformer architecture for estimating remaining useful life. *Reliability Engineering and System Safety*, *233*. https://doi.org/10.1016/j.ress.2023.109130

Kim, N.-H., Dawn, A., & Joo-Ho, C. (2017). *Prognostics and health management of engineering systems.* Switzerland: Springer International Publishing.

Landau, D., de Pater, I., Mitici, M., & Saurabh, N. (2026). Federated learning framework for collaborative remaining useful life prognostics: An aircraft engine case study. *Future Generation Computer Systems*, *174*. https://doi.org/10.1016/j.future.2025.107945

Lee, J., & Mitici, M. (2023). Deep reinforcement learning for predictive aircraft maintenance using probabilistic Remaining-Useful-Life prognostics. *Reliability Engineering and System Safety*, *230*. https://doi.org/10.1016/j.ress.2022.108908

Li, J., Chen, T., & Teng, S. (2024). A comprehensive survey on client selection strategies in federated learning. In *Computer Networks* (Vol. 251). Elsevier B.V. https://doi.org/10.1016/j.comnet.2024.110663

Li, X., Zhang, W., Li, X., & Hao, H. (2024). Partial Domain Adaptation in Remaining Useful Life Prediction With Incomplete Target Data. *IEEE/ASME Transactions on Mechatronics*, *29*(3), 1903–1913. https://doi.org/10.1109/TMECH.2023.3325538

Listou Ellefsen, A., Bjørlykhaug, E., Æsøy, V., Ushakov, S., & Zhang, H. (2019). Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. *Reliability Engineering & System Safety*, *183*, 240–251. https://doi.org/10.1016/J.RESS.2018.11.027

Llasag Rosero, R., Silva, C., Ribeiro, B., Albisser, M., Brutsche, M., & Arias Chao, M. (2025). Label synchronization strategies for hybrid federated learning. *Reliability Engineering and System Safety*, *256*. https://doi.org/10.1016/j.ress.2024.110751

Mabkhot, M. M., Kalawsky, R. S., & Liaqat, A. (2025). Introducing the Manufacturing Digital Passport (MDP): A New Concept for Realising Digital Thread Data Sharing in Aerospace and Complex Manufacturing. *Systems*, *13*(8), 700. https://doi.org/10.3390/systems13080700

Nishio, T., & Yonetani, R. (2019). Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge. *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 1–7. https://doi.org/10.1109/ICC.2019.8761315

Raouf, I., Kumar, P., Cheon, Y., Tanveer, M., Jo, S.-H., & Kim, H. S. (2025). Advances in Prognostics and Health Management for Aircraft Landing Gear—Progress, Challenges, and Future Possibilities. *International Journal of Precision Engineering and Manufacturing-Green Technology*, *12*(1), 301–320. https://doi.org/10.1007/s40684-024-00646-4

Sharma, R. K. (2024). Framework Based on Machine Learning Approach for Prediction of the Remaining Useful Life: A Case Study of an Aviation Engine. *Journal of Failure Analysis and Prevention*, *24*(3), 1333–1350. https://doi.org/10.1007/s11668-024-01922-w

Smestad, C., & Li, J. (2023). A Systematic Literature Review on Client Selection in Federated Learning. *ACM International Conference Proceeding Series*, 2–11. https://doi.org/10.1145/3593434.3593438

Soltani, B., Haghighi, V., Mahmood, A., Sheng, Q. Z., & Yao, L. (2022). A survey on participant selection for federated learning in mobile networks. *Proceedings of the 17th ACM Workshop on Mobility in the Evolving Internet Architecture, MobiArch 2022*, 19–24. https://doi.org/10.1145/3556548.3559633

Stanton, I., Munir, K., Ikram, A., & El-Bakry, M. (2023). Predictive maintenance analytics and implementation for aircraft: Challenges and opportunities. *Systems Engineering*, *26*(2), 216–237. https://doi.org/10.1002/sys.21651

Vermelin, W. S., Mishra, M., Eng, M. P., Andersson, D., & Kyprianidis, K. (2024). Collaborative Training of Data-Driven Remaining Useful Life Prediction Models Using Federated Learning. *International Journal of Prognostics and Health Management*, *15*(2). https://doi.org/10.36001/ijphm.2024.v15i2.3821

Zhang, J., Wang, P., Yan, R., & Gao, R. X. (2018). Long short-term memory for machine remaining life prediction. *Journal of Manufacturing Systems*, *48*, 78–86. https://doi.org/10.1016/J.JMSY.2018.05.011