Supplementary of "Strategic Integration of Adaptive Sampling and Ensemble Techniques in Federated Learning for Aircraft Engine Remaining Useful Life Prediction"

The supplementary document is organized as follows: Section S1 explores the impact of sliding window sizes on the model's performance; Section S2 examines how sampling proportions affect the model; and Section S3 highlights the long-term prediction capabilities of the proposed model.

S1 Effect of sliding window sizes

To select the optimal sliding window size, we conducted additional computations and analysis under the FedAvg algorithm with 5 clients. Table S1 presents the results of the proposed model's predictive performance under different window sizes. From the results, it is evident that the model achieves the best performance with a window size of 30, where the lowest RMSE (13.25) and the best score (265.81) are observed. As the window size increases beyond 30, the RMSE and score degrade, indicating that a larger window size may introduce unnecessary complexity without improving predictive accuracy.

Table S1: Performance of the Proposed Model with Different Window Sizes in the FedAvg Algorithm when $N = 5$.

Window size	20	25	30	35	40
RMSE	17.16	14.87	**13.25**	14.03	15.82
Score	529.89	338.27	**265.81**	326.55	417.23

S2 Effect of sampling proportions

Table S2 shows how different sampling ratios (q_t) impact the performance of the proposed model under the FedProx algorithm. As the sampling ratio decreases, model performance significantly declines. Specifically, at $q_t = 20\%$, RMSE rises to 17.44, and the Score increases to 502.92. This suggests that selecting only one client per round for training and aggregation results in a lack of data diversity, negatively affecting the model's generalization.

Thus, balancing prediction performance with computational efficiency is crucial when setting the sampling ratio. While lower ratios reduce overhead, they can lead to poor model performance due to insufficient data. An appropriate sampling ratio is key to ensuring both model stability and efficiency.

Table S2: Impact of q_t values on RMSE and Score in the FedProx algorithm when $N = 5$.

q_t	100%(EFL)	80%	60%	40%	20%
RMSE	14.89	14.10	13.62	15.17	17.44
Score	331.27	314.25	279.48	372.18	502.92

S3 Long-term prediction performance

In this section, we evaluate the model's long-term prediction capability using a direct multi-step forecasting strategy. For single-step prediction, the model estimates the RUL at time Y_{n+30} by inputting data from X_n to X_{n+30}. In contrast, the multi-step forecasting approach extends the target label to a series of continuous RUL values. Specifically, when the step size is set to 1, the model predicts Y_{n+30}, followed by Y_{n+31}, and so on, until the forecast reaches the desired number of steps K.

We conduct experiments under the FedAvg framework with 5 clients to assess the model's performance at different forecasting steps (see Table S3). As the prediction steps increases, the RMSE values gradually rise, indicating that the prediction error grows with the length of the forecast. This is due to the increasing uncertainty and error accumulation as the model is required to predict further into the future. Despite this, the multi-step forecasting strategy remains effective in capturing the long-term degradation trends of the equipment.

Table S3: Average long-term predicted performance.

Matrix	Prediction step				
	0	1	2	3	5
RMSE	13.25	13.72	13.86	14.08	14.19