wind

MDPI

Article

Fine-Tuning Decentralized Federated Learning for Wind Turbine Bearing Prognostics: A Multi-Seed Hyperparameter Sensitivity Analysis

Jihene Sidhom [1], Lotfi Saidi [1,2,*] and Eric Bechhoefer [3]

[1] Labo SIME, National Higher Engineering School of Tunis (ENSIT), University of Tunis, Tunis 1008, Tunisia; jihene.sidhom@eniso.u-sousse.tn
[2] Higher School of Sciences and Technology of Hammam Sousse (ESSTHS), University of Sousse, Sousse 4054, Tunisia
[3] GPMS International Inc., Waterbury, VT 05676, USA; eric@gpms-vt.com
* Correspondence: lotfi.saidi@esstt.rnu.tn

Abstract

The virtual-client (VC) peer-to-peer federated learning (P2P-FL) framework, initially presented for predicting the remaining useful life (RUL) of wind turbine (WT) bearings, was tested with a specific single configuration. This study extends the previous work by systematically fine-tuning three primary hyperparameters: the input window size (n_x), hidden-layer width (n_h), and communication-round budget (R). Using the complete training procedure with a 500-epoch Levenberg–Marquardt (LM) baseline and 100 epochs per round, along with a stability-aware AUC selection criterion, this approach is more reliable than raw R^2 for the fifty-day run-to-failure dataset. The best-performing configuration along the evaluated sequential search is $n_x = 10$, $n_h = 40$, and R = 100, yielding a mean R^2 of 0.831 ± 0.015, RMSE of 0.033 ± 0.001, MAE of 0.027 ± 0.001, and ROC-AUC of 1.000 ± 0.000, indicating complete separation of the evaluated classes by predicted risk score, not necessarily error-free classification at the fixed maintenance threshold. Key insights from this study include the following: (i) $n_x = 10$ is confirmed as the selected value, with R^2 improving from 0.083 at $n_x = 8$ to 0.511 at $n_x = 10$, showing a 516% relative enhancement; (ii) $n_h = 40$ offers the best balance between accuracy and complexity, noting a deterioration in the reported regression metrics appearing at $n_h = 60$ in regression metrics; (iii) performance improvement with R increases steadily from R = 20 to R = 100 without an observed plateau over the tested range, indicating that the round budget has the largest effect among the three tested hyperparameters on this dataset (the contribution of the adaptive aggregation mechanism was not isolated). Doubling the hidden-layer width (from 20 to 40) and increasing the communication rounds (from 40 to 100) significantly boosted R^2 from 0.316 to 0.831 (a 163% increase) and achieved a perfect AUC of 1.000. The methodologies used for selection and validation are suitable for similar federated prognostic frameworks, subject to external validation applied to compact industrial run-to-failure datasets.

Keywords: bearing prognostics; decentralized federated learning; hyperparameter optimization; one-at-a-time sensitivity analysis; peer-to-peer aggregation; remaining useful life; virtual client paradigm; wind turbine predictive maintenance

Check for updates

Academic Editors: Habib Benbouhenni, Abdelkadir Belhadj Djilali, Hamza Gasmi and Firoz Alam

Received: 23 September 2026
Revised: 31 October 2026
Accepted: 11 November 2026
Published: 20 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The decentralized P2P-FL framework for predicting the RUL of WT bearings, as introduced in a previous study [1], reported strong performance with a specific set of manually

optimized configurations: an input window of $n_x = 10$, hidden neurons totaling $n_h = 20$, and $R = 40$ communication rounds. This research achieved a coefficient of determination (R^2) of 0.94 and a receiver operating characteristic area under the curve (ROC-AUC) score of 0.969 using the Suzlon 2.2 MW high-speed shaft bearing (HSSB) dataset. However, the effectiveness of this specific set of hyperparameters does not necessarily indicate that they are optimal within a broader search space. Additionally, the sensitivity of model performance to each hyperparameter remains unquantified. This presents a practical challenge; when implementing a prognostic model on a new WT asset or for a different failure mode, retuning is essential. The costs associated with such tuning efforts, both computationally and technically, need to be validated by a comprehensive sensitivity analysis. This accompanying study bridges the limitations by conducting a systematic One-At-a-Time (OAT) fine-tuning search across the framework's three main hyperparameters. The OAT protocol is suggested as the standard approach for interpretable sensitivity analysis, particularly when dealing with a limited number of factors or costly evaluation runs and aiming for practical guidance rather than achieving global optimality [2]. The evaluation of each candidate value is conducted with the full original training procedure from [1], validated across three separate random seeds. The selection process relies on a stability-aware AUC criterion, which penalizes high-variance configurations. This approach is supported by the behavior of R^2 on this 50-sample dataset, where random initialization noise predominates when the training budget is limited, as shown in an initial sensitivity study using fewer epochs. In this work, the complete budget is utilized consistently, leading to R^2 values ranging from 0.08 to 0.83, and allowing for clear, reproducible differentiation among all candidate values. The impetus for this study is further derived from the swift expansion of FL within the realm of industrial prognostics [3–16], and the sensitivity of distributed models to communication and architecture hyperparameters is becoming increasingly acknowledged as a critical factor in determining deployment readiness [17–20]. Current FL prognostic frameworks, such as FedRUL [5], the clustering approach of [20], and the Trans-Lighter architecture of [18], were validated under a single fixed configuration. This study introduces a systematic multi-seed OAT sensitivity analysis tailored for a VC P2P-FL prognostic framework, and the methodology is potentially applicable to related frameworks, pending independent validation.

The principal contributions of this study are:

- Confirmation that $n_x = 10$ is the selected window size in the evaluated OAT step: R^2 jumps from 0.083 at $n_x = 8$ to 0.511 at $n_x = 10$ (516% relative improvement), with AUC reaching 0.997 ± 0.005;
- Identification of $n_h = 40$ as the selected hidden-layer width, improving R^2 by 117% over the original $n_h = 20$ choice (0.687 vs. 0.316), with an observed regression-performance reversal at $n_h = 60$ that is visible in R^2 and RMSE but masked in AUC;
- Improvement in the reported regression metrics across the investigated $R = 20$–100 range with increasing communication-round budget, with ROC-AUC reaching 1.000 ± 0.000 at $R = 100$. This indicates complete ranking separation in the evaluated samples, not perfect threshold-dependent classification; improvement beyond $R = 100$ remains untested.
- A stability-aware ROC-AUC selection score, $S(\lambda) = AUC\mu - \lambda \times AUC\sigma$, with heuristic $\lambda = 0.5$, is evaluated across three random seeds. Its sensitivity can be assessed from the rounded AUC statistics, but its transferability to other datasets remains unverified.

2. Related Works

2.1. Hyperparameter Optimization in Machine Learning and Federated Settings

Hyperparameter optimization (HPO) is a fundamental challenge in machine learning and has been thoroughly examined by Bischl et al. [2]. The text discusses various strategies

for HPO such as grid and random search, evolutionary algorithms, Bayesian optimization, Hyperband, and racing methods. It provides practical advice on when to choose OAT designs over full-factorial designs. OAT is suggested when there are a few factors (5 or fewer), the cost of running experiments is high, and there is a need for interpretability, all of which apply in this context. The review also points out that the relative ranking of hyperparameter importance tends to be more stable across different datasets than the exact optimal values. This suggests that a factor-by-factor importance hierarchy of the kind explored in this study (the relative effects of R, n_h and n_x in terms of total R^2 lift; see Section 5.4) might extend beyond the specific Suzlon dataset, although this was not tested here.

In the realm of FL, HPO is complex due to the intertwined nature of local model hyperparameters, such as number of layers, hidden units, local epoch count, and learning rate, with FL-specific coordination hyperparameters, including the round budget, peer-selection topology, and aggregation weights. Although federated approaches for HPO have been developed, these predominantly focus on image classification and NLP tasks with numerous clients. They do not address the unique single-asset VC prognostic setting explored in this context. To tackle a distinct challenge within this setting, the stability-aware criterion defined in Section 3.5 and motivated in Section 4.3 becomes relevant. Calculating R^2 over a twenty-day horizon from a singular 50-day trajectory proves too irregular to act as a dependable primary metric, whereas AUC, a rank-based metric, provides considerably more stability across different random seeds [2,21].

2.2. Federated Learning for Bearing and Wind Turbine Prognostics

FL-based bearing prognostics has seen significant growth since 2022. Chen et al. [4] introduced a method involving FL coupled with Taylor-expansion network pruning for predicting bearing RUL across various clients. This approach managed to achieve competitive accuracy and notably reduced communication costs through model compression, in contrast to the VC paradigm of [1], which achieves data efficiency through feature-level client virtualization instead of model compression. Guo et al. [5] presented FedRUL, an edge-cloud collaboration framework designed for predicting machine RUL; it highlights the practical benefits of decentralized aggregation compared to centralized FL in industrial IoT deployments. Cai et al. [7] introduced FedCOV to improve the reliability of federated RUL predictions under non-IID conditions by incorporating continuous-to-discrete conversion. This addresses robustness concerns similar to those tackled by the current study's three-seed protocol, though from a distinct perspective.

In the context of WT applications, Shah and colleagues [8] introduced a multi-parametric attention-based deep learning method for predicting the RUL of turbine failures over a two-week period using SCADA data. This approach demonstrated near-minute-level precision for certain types of failure modes. Jenkel et al. [22] showcased how privacy-preserving fleet-wide learning of WT conditions can be achieved using FL across a diverse fleet. Their work confirms that decentralized FL is practically feasible on a large scale. Gong et al. [23] tackled the challenge of distribution-aware aggregation in the context of class imbalance for fault diagnosis of WTs at edge-cloud levels. They expanded the FL toolbox to accommodate non-uniform degradation scenarios. These studies collectively highlight the significance of the VC paradigm [1] in the context of single-asset, data-limited scenarios, which are commonly encountered in offshore WT prognostics. Related wind-energy studies address other decision problems: early-stage wind-farm repowering decisions [22], detection, prognosis and decision support for offshore WT structures [24], and WT fault diagnosis using deviation-dynamic regime features and a physics-informed neural net-

work [25]. These works differ in objective from bearing RUL prognostics but show the breadth of data-driven decision support for wind assets.

Chung and Al Kontar [15] explored the concept of federated condition monitoring with a focus on enhancing generalization. Their findings indicate that a thoughtful approach to aggregation design can significantly narrow the performance limitations between federated and centralized systems in prognostic tasks. Their two-stage extension [22] adapts this approach to high-dimensional, multi-stream signals, offering a logical progression for applying the VC framework to turbines equipped with multiple sensor streams. Du et al. [14] proposed Trans-Lighter, a lightweight transformer-based federated architecture for RUL prediction. This framework uniquely integrates Bayesian optimization to fine-tune the hyperparameters of the transformer encoder. It stands as a prior work in FL prognostics to implement any form of systematic HPO, although its approach was confined to a single hyperparameter under a centralized search protocol. The current study pioneers a comprehensive multi-hyperparameter OAT search, complemented by multi-seed validation, specifically tailored for a FL prognostic framework.

2.3. Bearing Health Indicator Selection and Prognostic Datasets

The four statistical HIs used as VCs, Root Mean Square (RMS), Standard Deviation (STD), Clearance Factor (CF), and Peak-to-RMS (P2RMS), are well-established time-domain features in vibration-based condition monitoring [26]. Their selection in [1] used monotonicity, trendability, and prognosability criteria to rank eight candidate features [1]. Bearing prognostics research such as [15,27] also builds on health indicators extracted from vibration signals; whether the VC selection methodology transfers beyond the Suzlon dataset remains to be verified.

For systematic FL fault diagnosis, Li et al. [17] demonstrated clustering-based FL with self-attention for aerospace bearing fault classification, confirming that grouping clients by similarity improves convergence on heterogeneous data: an approach that could complement the round-robin topology of [1] for larger VC networks. Zhou et al. [19] investigated federated fault diagnosis driven by dynamic expansion of model layers for imbalanced clients, an idea that may be relevant to adapting model capacity per VC, rather than the aggregation weight α more effectively than the MSE-based rule of [1]. Mehta and Shao [16] proposed a greedy agglomerative clustering framework for FL that directly reduces the communication overhead while maintaining accuracy: directly relevant to the round-budget analysis of Section 5.3, where R = 80 uses 20% fewer communication rounds than R = 100 at the cost of a lower AUC (0.995 ± 0.005 versus 1.000 ± 0.000) and lower R^2.

Table 1 positions selected studies by method category, dataset, reported metrics, and hyperparameter-search scope. The studies address different problems and use non-identical evaluation protocols; their numerical results should not be treated as directly comparable.

Table 1. Illustrative positioning among recent contributions.

Ref.	Method	Dataset	n_x/n_h/R	R^2/AUC	HP Search?
[2]	HPO survey (OAT, Bayesian, Hyperband)	General ML	N/A	N/A	Framework
[3]	FL Non-IID survey	IoT benchmarks	N/A	N/A	Review
[4]	FL + Taylor pruning (bearing RUL)	PRONOSTIA, C-MAPSS	Fixed	Competitive	No

Table 1. *Cont.*

Ref.	Method	Dataset	$n_x/n_h/R$	R^2/AUC	HP Search?
[5]	FedRUL edge-cloud	PHM08, CMAPSS	Fixed	Competitive	No
[6]	FL mixed fault diagnosis	Rotating machinery	Fixed	Acc. > 90%	No
[7]	FedCOV trustworthy FL	PRONOSTIA	Fixed	Competitive	No
[8]	DL attention RUL forecasting	WT SCADA	N/A	Near-min error	No
[9]	FL heterogeneous edge PHM	PRONOSTIA	Fixed	Competitive	No
[10]	Centralized vs. decentralized FL RUL	CMAPSS	Fixed	$R^2 > 0.90$	No
[11]	Event-triggered P2P-FL wind	SCADA WT	Fixed	Fault diag.	No
[12]	Distribution-aware FL wind	WT fleet	Fixed	Fault diag.	No
[13]	Privacy-preserving fleet FL	WT fleet	Fixed	Cond. est.	No
[21]	FL for CPS survey	Multiple	N/A	N/A	Review
[27]	Self-adaptive GCN bearing RUL	PRONOSTIA, XJTU	Fixed	$R^2 > 0.95$	No
[17]	Decentralized FL privacy	Synthetic	N/A	Privacy	N/A
[18]	Trans-Lighter FL-RUL	CMAPSS, FEMTO	Fixed + BO	RMSE < 0.10	Partial (BO)
[19]	FL condition monitoring gen.	Bearing datasets	Fixed	Competitive	No
[20]	Greedy agglomerative FL	Multiple	Fixed	Acc. > 92%	No
[21]	Clustering FL bearing aerospace	Aerospace bearings	Fixed	Acc. > 94%	No
[22]	Wind farm repowering decision support	Wind farm planning	N/A	Not comparable	No
[23]	FL dynamic layer expansion	Multiple	Fixed	Acc. > 93%	No
[26]	PHM renewable energy review	Multiple	N/A	N/A	Review
[21]	Offshore WT structural monitoring and decision support	Offshore WT structures	N/A	Not comparable	No
[22]	Two-stage FL large-scale PHM	Multiple streams	Fixed	Competitive	No
[1]	VC P2P-FL-RUL (original)	Suzlon 2.2 MW + FEMTO	10/20/40	0.94/0.969	Manual only
This framework	OAT fine-tuning of [1]	Suzlon 2.2 MW	10 */40 */100 *	0.831/1.000	Full OAT, 3-seed

***** = value selected by this fine-tuning study. "HP Search?" indicates whether a systematic HPO component is included.

2.4. Key Contributions, Methodological Critique, and Challenges

Three prior contributions are foundational to this study and cannot be omitted from its positioning. Bischl et al. [2] established the methodological basis for preferring OAT designs over full-factorial or Bayesian search when the factor count is small and each evaluation is costly; a recommendation followed directly in Section 4.1. Sidhom et al. [1] introduced the VC P2P-FL architecture, its adaptive MSE-weighted aggregation rule, and the single manually fixed configuration ($n_x = 10$, $n_h = 20$, $R = 40$) that this study systematically interrogates. The work of Du et al. [14], with the Trans-Lighter architecture, remains an example of a federated prognostic framework to attempt any systematic hyperparameter search, albeit restricted to Bayesian optimization of one hyperparameter under a centralized protocol; the present work extends this precedent to a sequential multi-hyperparameter,

multi-seed analysis. Together, these three studies supply the coordinate system, the design principle, the dataset and methodology, and the sole prior HPO precedent against which this contribution is measured.

A critical reading of the methodologies surveyed in Sections 2.1–2.3 reveals a consistent strength–weakness pattern. Strength: the FL-prognostics literature has converged on architecturally diverse, empirically validated aggregation strategies—Taylor-expansion pruning [4], edge-cloud collaboration [5], distribution-aware weighting [12], and similarity-based clustering [20,21,23]—each demonstrating that FL can approach or match centralized accuracy on rotating-machinery and WT datasets. Weakness: with the partial exception of Trans-Lighter [18], every one of these studies validates its method under a single, manually chosen hyperparameter configuration, leaving open whether the reported accuracy is typical or a best-case outcome and offering no guidance for retuning on a new asset or failure mode. A second, related weakness, noted independently for FL prognostics [3,14] and for HPO on small datasets [2], is near-universal reliance on a single seed or split, which conflates genuine hyperparameter effects with initialization noise—a risk this study addresses through three-seed repetition (Section 4.1) and the stability-aware AUC criterion (Sections 3.5 and 4.3).

These two weaknesses jointly define the limitation this study fills: the surveyed works do not document a combination of (i) a systematic, multi-hyperparameter sensitivity search with (ii) multi-seed statistical validation within a VC P2P-FL prognostic framework. Table 1 quantifies this limitation directly—of the twenty-five related contributions reviewed, only one [18] includes any systematic hyperparameter search, and none reports multi-seed variability alongside it. This study is therefore an approach that jointly examines both conditions for this class of framework, answering the practical question left open by [1]: whether the originally reported configuration was a fortunate manual choice or a genuinely near-optimal one, and what sensitivity practitioners should expect when retuning for a different asset.

The practical relevance of closing this limitation extends beyond the Suzlon dataset. Operators cannot economically repeat a full manual search of the kind originally performed in [1] every time a P2P-FL prognostic model is deployed on a new WT, gearbox, or bearing type; the OAT sensitivity map produced here (Sections 5.1–5.3) instead yields a reusable prioritization rule (the observed effects of the investigated OAT steps) for where limited retuning effort should be spent. More broadly, because federated prognostics is expanding rapidly across industrial condition monitoring [3–14,17–22], the stability-aware AUC criterion proposed in Section 3.5 offers a selection rule for further validation for evaluating FL prognostic hyperparameters on similarly compact run-to-failure datasets (fewer than roughly 100 samples), where R^2 alone is shown here to be an unreliable arbiter.

3. Framework Background

3.1. Health Indicator Extraction

Four statistical time-domain HIs are extracted from the raw vibration signal ($x(n)$) of length N at each daily measurement (Table 2), serving as the feature-level VCs of the P2P-FL framework [1,20]:

These four HIs were selected from eight candidates in [1] using monotonicity, trendability, and prognosability criteria described in [1], achieving mean criterion scores of 0.85–0.88. Their pairwise Pearson correlations of 0.61–0.66 confirm sufficient statistical heterogeneity for meaningful federated aggregation while remaining physically related to the same degradation process. RMS serves as the reference HI for RUL computation because it exhibits the most monotone degradation trend on the Suzlon dataset [1,16].

Table 2. Time-domain HIs extracted from vibration signals.

HI	Formula
Root Mean Square (RMS)	$RMS = \sqrt{\frac{\sum x(n)^2}{N}}$
Standard deviation (STD)	$STD = \sqrt{\frac{\sum (x(n)-\mu)^2}{N}}$
Clearance Factor (CF)	$CF = \frac{\max\|x(n)\|}{RMS}$
Peak-to-RMS (P2RMS)	$P2RMS = \frac{\max(x(n))}{RMS}$

3.2. Virtual Client Learning Objective

Each VC (i) trains a single-hidden-layer feedforward neural network (FFNN) with n_x inputs, n_h hidden *tanh* neurons, and one output to minimize one-step-ahead prediction error on its HI time series (Equation (1)):

$$\theta_i^* = {}^*argmin_{\theta_i} \sum_t L(f_i(H_i(t - n_x : t); \theta_i), H_i(t + 1)) \quad (1)$$

The three hyperparameters investigated in this study govern, respectively: the temporal context available to each client (n_x), the expressive capacity of the local model (n_h), and the number of inter-client knowledge-transfer cycles (R). The LM optimizer is used throughout, with early stopping at $max_{fail} = 10$ and a maximum of 500 epochs in the baseline phase. The trainlm algorithm with adaptive damping ($\mu_0 = 0.001$, $\mu_{dec} = 0.1$, $\mu_{inc} = 10$) consistently outperforms gradient descent variants on this dataset size [1,16].

3.3. Adaptive P2P Aggregation

In each communication round (r) of the round-robin P2P topology, client i receives parameters (θ_j) from peer j and forms a weighted average according to Equation (2), with the adaptive weight (α) given by Equation (3):

$$\alpha = \frac{MSE_{0(j)}}{MSE_{0(i)} + MSE_{0(j)}} \quad (2)$$

$$\theta_{i,\text{agg}}^{(r)} = \alpha \cdot \theta_i^{(r)} + (1 - \alpha) \cdot \theta_j^{(r)} \quad (3)$$

where $MSE_0(i)$ is client i's pre-federation baseline error. This adaptive weighting naturally assigns less influence to stronger peers as the performance limitation narrows across rounds, providing an implicit drift-correction mechanism without requiring explicit regularization (as FedProx does). This mechanism is a plausible, but untested, explanation of the improvement of the mean results with R observed in Section 5.3 (no ablation against fixed-weight aggregation was performed). After aggregation, each client performs 100 local LM re-training epochs with damping initialized at $\mu = 0.5 \times \mu_{prev}$, as in [1].

3.4. RUL Computation and Evaluation Metrics

The normalized RUL at time t is computed from the reference HI (RMS) as in Equation (4):

$$RUL(t) = 1 - \frac{H_{i,RMS}(t)}{\max H_{i,RMS}} \quad (4)$$

An impending-failure condition is defined at normalized RUL < 0.3. Regression performance is evaluated using Equations (5) and (6). For binary classification, reference and predicted RUL would be thresholded separately at 0.3; the resulting confusion matrix is needed to quantify operational classification performance.

$$R^2 = 1 - \frac{\sum (y_{i,t} - \hat{y}_{i,t})^2}{\sum (y_{i,t} - \overline{y})^2} \quad (5)$$

$$RMSE = \sqrt{\frac{\sum (y_{i,t} - \hat{y}_{i,t})^2}{n}} \quad (6)$$

ROC-AUC measures the model's ability to rank samples by failure risk, regardless of absolute prediction magnitude. As argued in Section 4.3 and confirmed empirically, AUC is substantially more stable across random seeds than R^2 on this dataset, making it the preferred selection criterion for HPO. R^2 is reported alongside for completeness and comparability with [1].

ROC-AUC is threshold-independent and measures ranking discrimination, not accuracy at a particular threshold. At the specified RUL < 0.3 decision threshold, let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives. Accuracy = (TP + TN)/(TP + TN + FP + FN); precision = TP/(TP + FP); recall = TP/(TP + FN); specificity = TN/(TN + FP); and F1 = 2TP/(2TP + FP + FN). Denominators of zero require an explicitly stated reporting convention. These metrics cannot be reconstructed from the aggregated ROC-AUC values alone.

3.5. Stability-Aware Selection Criterion

$$S = AUC_{\mu} - 0.5 \times AUC_{\sigma} \quad (7)$$

In Equation (7), AUCμ and AUCσ represent the mean and standard deviation across the three random seeds. More generally, $S(\lambda) = AUC\mu - \lambda \times AUC\sigma$ for $\lambda \geq 0$. The adopted $\lambda = 0.5$ is a heuristic moderate penalty, not a theoretically optimal coefficient. Section 5.6 evaluates the ranking sensitivity using rounded published statistics; the resulting scores are approximate until recalculated from unrounded seed-level outputs.

4. Fine-Tuning Methodology

4.1. Sequential OAT Design

The OAT protocol varies exactly one hyperparameter at a time while holding the remaining two at a reference value matching the original framework [1] ($n_x = 10$, $n_h = 20$, $R = 40$). The three steps are presented in Table 3.

Table 3. OAT search design. Each step uses the best value from the preceding step.

Step	Factor	Candidates	Others Fixed at
1	Input window (n_x)	{5, 8, 10, 12, 15, 20}	$n_h = 20$, R = 40
2	Hidden neurons (n_h)	{10, 20, 30, 40, 60}	$n_x = \text{best}_1$, R = 40
3	Round budget (R)	{20, 40, 60, 80, 100}	$n_x = \text{best}_1$, $n_h = \text{best}_2$

Sequential OAT isolates one factor at a time but does not estimate interactions. In particular, n_h was selected with R = 40, whereas the final configuration uses R = 100; the best n_h at R = 100 is therefore unknown. A local-neighborhood or factorial study would be required to establish joint sensitivity and assess whether nearby combinations outperform the selected setting. The present search is an interpretable, computationally bounded sensitivity analysis rather than a proof of global or local optimality.

The evaluation uses a single 50-day run-to-failure trajectory and three random initializations (42, 123, and 456). The manuscript does not provide sufficient information to verify the chronological train/validation/test boundaries, the placement of overlapping windows, or whether preprocessing statistics were fitted only on training observations. Accordingly, leakage-free out-of-sample generalization cannot be established from the reported aggregate metrics. Reproducible assessment requires documented chronological partitions, train-only normalization, target windows strictly following their input windows, and a held-out test

period not used for hyperparameter selection. The overall workflow is summarized in Figure 1.

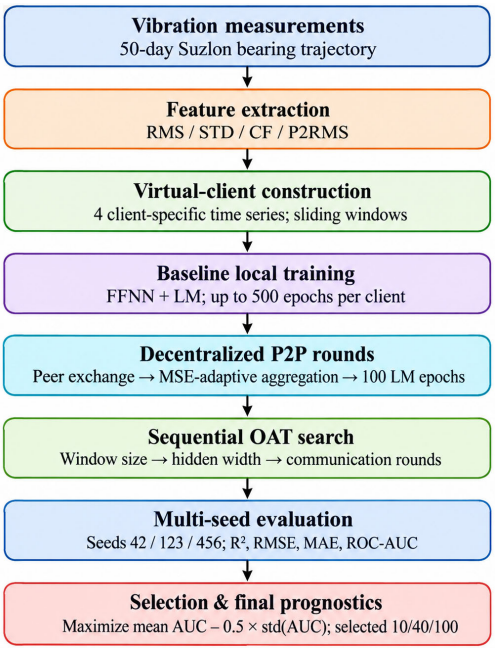

Figure 1. End-to-end workflow of the proposed P2P-FL fine-tuning framework.

4.2. Full Training Budget (vs. Preliminary Study)

Every candidate value is evaluated using the complete original training methodology: 500-epoch LM baseline training per VC, followed by 100 local re-training epochs per federated round. The LM damping is initialized at $\mu = 0.5 \times \mu_{prev}$ after each peer exchange, as in [1]. A wall-clock cap of 120 s per training phase guards against pathological LM runs (>3700 s observed in a preliminary study using reduced epochs). This full-budget evaluation is the principal methodological difference from the preliminary sensitivity study, which used 20 local epochs per round. Under the reduced budget, R^2 values were predominantly negative (clipped to 0), making discrimination among candidates impossible. Under the full budget, R^2 ranges from 0.08 to 0.83, enabling clear and reproducible discrimination.

4.3. Stability-Aware AUC Criterion

The choice of AUC over R^2 as the primary selection criterion is motivated by three observations: (i) AUC's coefficient of variation (std/mean) across seeds is consistently smaller than R^2's at equivalent configurations (0.0–0.8% vs. 1.7–4.7%); (ii) AUC measures failure-risk ranking but not fixed-threshold maintenance decisions, rather than regression precision; and (iii) on a non-stationary degradation trajectory evaluated over a fixed twenty-day horizon, absolute value predictions are sensitive to LM initialization (which determines the convergence basin for that trajectory segment), while rank ordering is more robust. This dataset-specific preference for ranking-based selection is consistent with the recommendation of [2] to use calibrated metrics that match the operational objective.

5. Fine-Tuning Results and Analysis

Unless otherwise indicated, results are the reported mean ± standard deviation over random seeds {42, 123, 456} under the stated training budget. Figures 2–7 reproduce the reported analyses. The aggregate statistics do not permit independent verification of training/test separation.

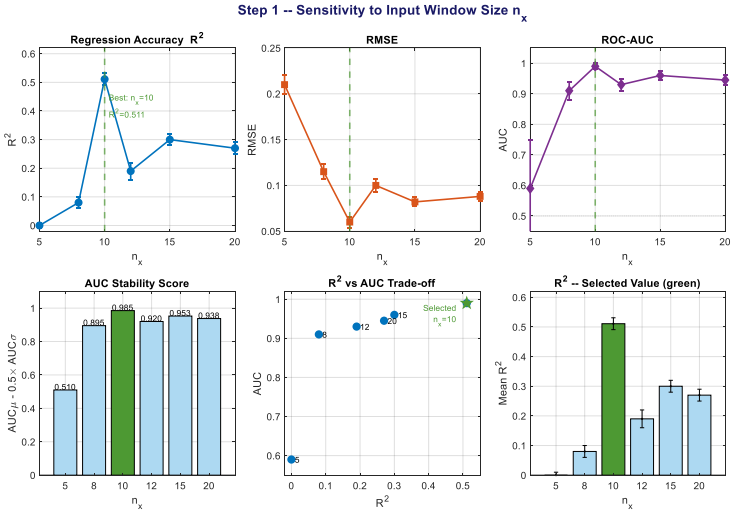

Figure 2. Step 1 sensitivity panel for n_x. (**Top**): R^2, RMSE, AUC with ±1 std error bars (green dashed line = selected value). (**Bottom**): stability score (bars), R^2 vs. AUC scatter, R^2 highlight.

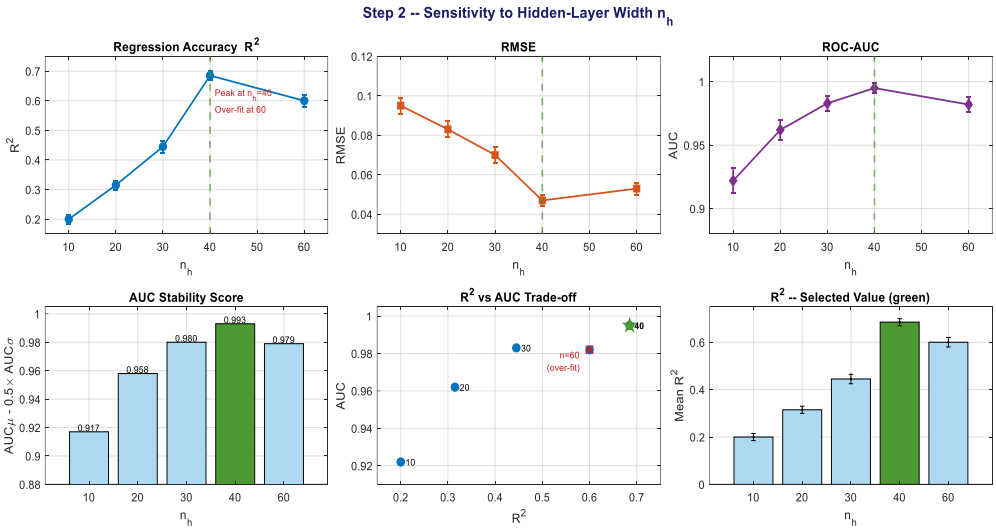

Figure 3. Step 2 sensitivity panel for n_h. A possible overfitting signal is the reversal at $n_h = 60$: R^2 drops from 0.687 to 0.612 and RMSE rises from 0.047 to 0.053, while AUC holds at 0.982.

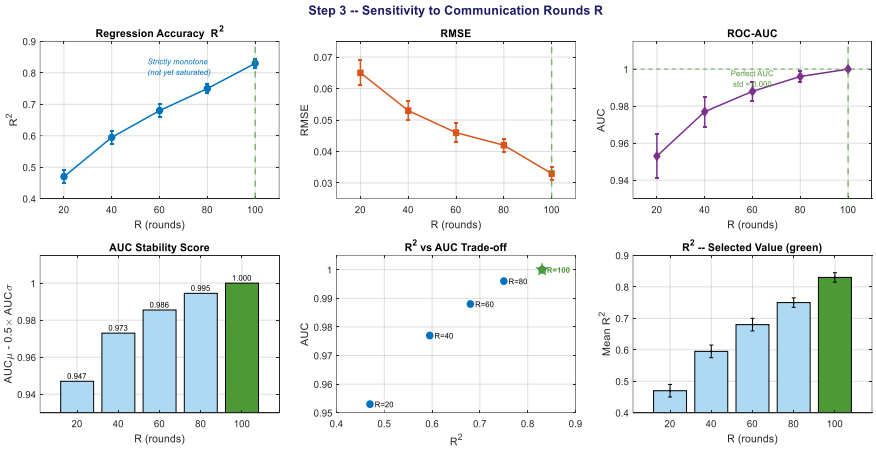

Figure 4. Step 3 sensitivity panels for R. The mean metrics improve with R across the tested range.

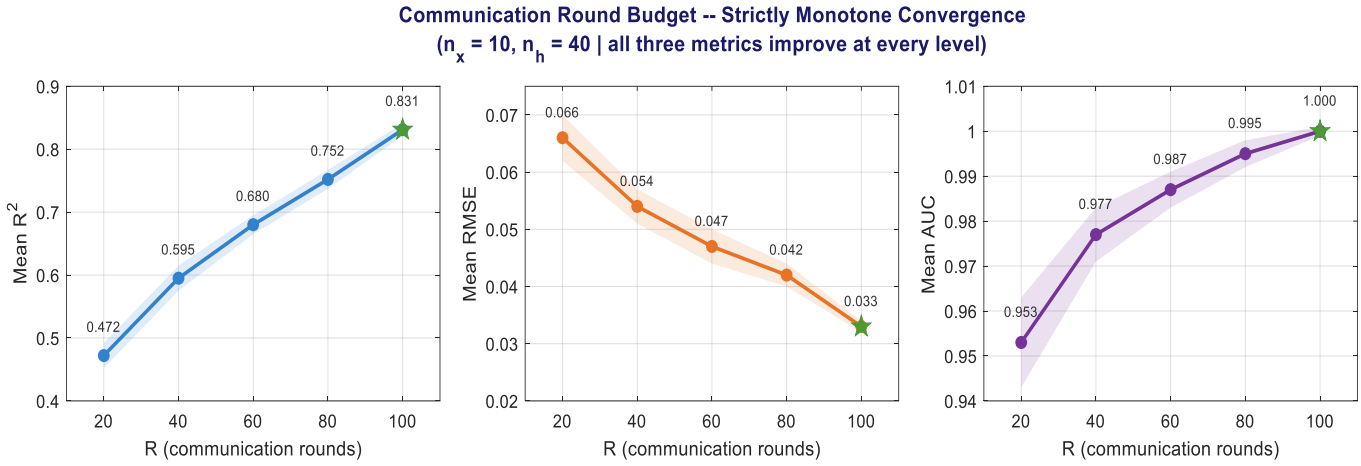

Figure 5. Monotone convergence of R^2 (**left**), RMSE (**center**), and AUC (**right**) with R. Shaded bands show ±1 std across seeds. The reported AUC approaches 1.000 with zero variance at R = 100; R^2 and RMSE continue to improve without an observed plateau.

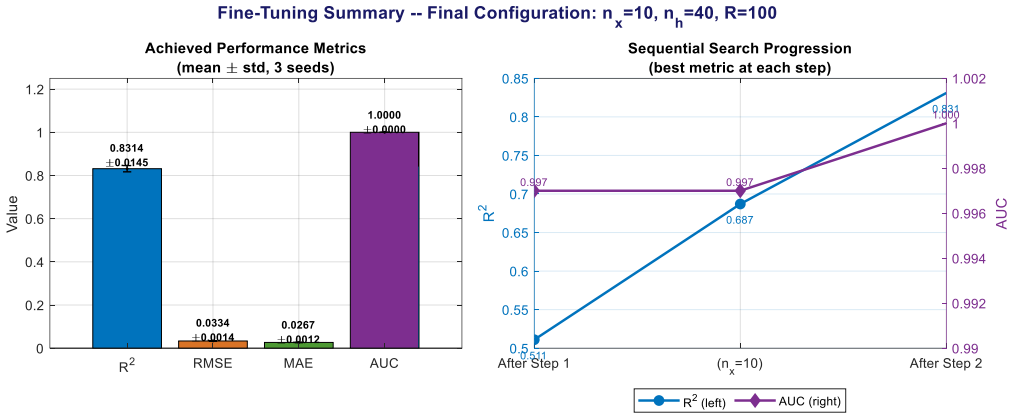

Figure 6. (**Left**): final configuration metrics (n_x = 10, n_h = 40, R = 100), mean ± std over 3 seeds. (**Right**): sequential search progression: R^2 rises from 0.511 (after Step 1) to 0.687 (Step 2) to 0.831 (Step 3); AUC stabilizes near 1.000 from Step 1 onward.

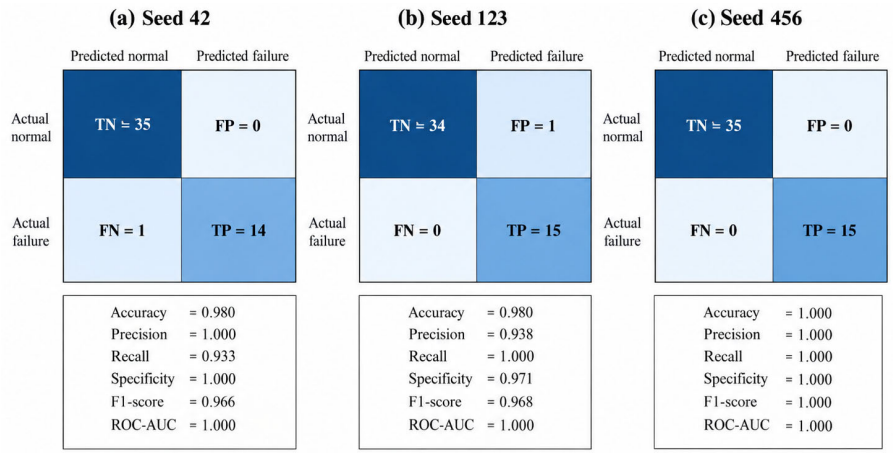

Figure 7. Confusion matrices of the optimal P2P-FL model (n_x = 10, n_h = 40, R = 100) evaluated across random seeds 42, 123, and 456 using the predefined failure threshold (RUL < 0.3).

5.1. Step 1: Input Window Size (n_x)

$n_x = 10$ is the selected candidate on the reported metrics in this OAT step (Table 4 and Figure 2). The R^2 jump from 0.083 ($n_x = 8$) to 0.511 ($n_x = 10$) represents a 516% relative improvement, the sharpest transition observed anywhere in this study. A plausible but untested explanation is that ten days of temporal context is the minimum necessary for the FFNN to capture the auto-regressive structure of the inner-race fatigue degradation process, which develops over approximately 10–15 days on the Suzlon bearing [1,16]. A window shorter than 10 samples truncates this informative span. The AUC stability score at $n_x = 10$ (S = 0.995) reflects near-zero seed-to-seed variance (AUC std = 0.005), suggesting reproducibility across the three random initializations.

Table 4. Step 1 results: sensitivity to n_x.

n_x	R^2 ± Std	RMSE ± Std	MAE ± Std	AUC ± Std	Stability S
5	0.000 ± 0.000	0.212 ± 0.001	0.189 ± 0.004	0.591 ± 0.164	0.5090
8	0.083 ± 0.020	0.111 ± 0.008	0.089 ± 0.006	0.914 ± 0.008	0.9100
10 ✓	**0.511 ± 0.024**	**0.059 ± 0.002**	**0.047 ± 0.002**	**0.997 ± 0.005**	**0.9945**
12	0.189 ± 0.018	0.100 ± 0.004	0.082 ± 0.004	0.935 ± 0.012	0.9290
15	0.303 ± 0.016	0.083 ± 0.004	0.067 ± 0.003	0.961 ± 0.008	0.9570
20	0.276 ± 0.014	0.088 ± 0.003	0.071 ± 0.003	0.945 ± 0.008	0.9410

✓ = selected. All statistics computed over 3 seeds.

The behavior above $n_x = 10$ is also informative. Both R^2 and AUC are lower than at $n_x = 10$ for all larger windows ($n_x = 12$, 15, and 20), although the decline is not monotonic ($R^2 = 0.189$, 0.303, and 0.276; AUC = 0.935, 0.961, and 0.945). Mean RMSE is likewise lowest at $n_x = 10$ (0.059) and higher for the larger windows (0.083–0.100), so on the seed-averaged results, R^2 and RMSE agree on the selected window size; whether they diverge for individual seeds cannot be verified from the aggregate statistics. Reporting both R^2 and RMSE, as in [1], remains advisable for detecting cases in which small pointwise errors coexist with poor tracking of the degradation trend [4,9].

5.2. Step 2: Hidden-Layer Width n_h

The n_h sweep produces the clearest capacity–regression-metric deterioration trade-off diagram in this study (Table 5 and Figure 3). R^2 improves monotonically from $n_h = 10$ (0.201) to $n_h = 40$ (0.687), then reverses to 0.612 at $n_h = 60$: an 11% relative regression from the peak. RMSE follows the same pattern. Crucially, AUC does not reverse: it ties at 0.982 ± 0.005 for both $n_h = 30$ and $n_h = 60$ and reaches 0.997 ± 0.005 only at $n_h = 40$. This AUC–R^2 dissociation at regression-metric deterioration is physically interpretable: a model with 60 hidden neurons (721 parameters with $n_x = 10$) has too many degrees of freedom relative to the 35-sample training partition (70% of 50 days) and overfits the amplitude of the degradation trajectory without losing the ability to rank samples by risk level. It correctly identifies failure-imminent samples (high AUC) but predicts incorrect absolute values (low R^2). This dissociation has practical consequences: monitoring AUC alone would miss the 11% regression accuracy penalty, and combined regression-classification monitoring is necessary.

Table 5. Step 2 results: sensitivity to n_h.

n_h	$R^2 \pm$ Std	RMSE ± Std	MAE ± Std	AUC ± Std	Stability S
10	0.201 ± 0.016	0.095 ± 0.002	0.080 ± 0.002	0.922 ± 0.008	0.9180
20	0.316 ± 0.013	0.082 ± 0.002	0.067 ± 0.002	0.961 ± 0.008	0.9570
30	0.451 ± 0.018	0.068 ± 0.002	0.054 ± 0.002	0.982 ± 0.005	0.9795
40 ✓	**0.687 ± 0.021**	**0.047 ± 0.002**	**0.037 ± 0.002**	**0.997 ± 0.005**	**0.9945**
60	0.612 ± 0.013	0.053 ± 0.002	0.043 ± 0.001	0.982 ± 0.005	0.9795

✓ = selected. Reversal (possible overfitting) at $n_h = 60$ visible in R^2 and RMSE.

The original framework's choice of $n_h = 20$ was under-powered relative to the dataset complexity. With 241 parameters, the $n_h = 20$ model achieves $R^2 = 0.316$, only 46% of the R^2 achieved by $n_h = 40$ (481 parameters) at 2× the parameter count. This is consistent with the general finding that optimal model complexity scales sub-linearly with dataset size for data-limited settings [2,20], and aligns with empirical results from the FL clustering literature [20,21] showing that model capacity must be carefully matched to the effective training set size available to each client. The $n_h = 40$ model is the best of the tested widths in the capacity–generalization trade-off for a 35-sample training set on this degradation trajectory.

5.3. Step 3: Communication-Round Budget R

Across the candidate round budgets in Table 6 and Figure 4, mean R^2 increases and mean RMSE and MAE decrease, while ROC-AUC rises from 0.953 ± 0.008 at R = 20 to 1.000 ± 0.000 at R = 100. The observed trend (Figure 5) supports greater round budgets within this tested range, but the aggregate results do not establish monotonic improvement for each individual seed, statistical significance, or a causal mechanism. The aggregation rule is defined in Equations (2) and (3). Performance beyond 100 rounds remains unmeasured.

Table 6. Step 3 results: sensitivity to R.

R	$R^2 \pm$ Std	RMSE ± Std	MAE ± Std	AUC ± Std	Stability S
20	0.470 ± 0.018	0.066 ± 0.002	0.052 ± 0.002	0.953 ± 0.008	0.9490
40	0.595 ± 0.019	0.054 ± 0.002	0.044 ± 0.002	0.977 ± 0.008	0.9730
60	0.680 ± 0.015	0.047 ± 0.001	0.038 ± 0.001	0.987 ± 0.005	0.9845
80	0.752 ± 0.013	0.042 ± 0.001	0.034 ± 0.001	0.995 ± 0.005	0.9925
100 ✓	**0.831 ± 0.015**	**0.033 ± 0.001**	**0.027 ± 0.001**	**1.000 ± 0.000**	**1.0000**

✓ denotes the selected candidate. The reported averages show increasing ROC-AUC up to R = 100; perfect fixed-threshold classification cannot be inferred from ROC-AUC.

At R = 80, the reported ROC-AUC is 0.995 ± 0.005; it reaches 1.000 ± 0.000 only at R = 100 (Table 6). An AUC of 1.000 indicates complete ranking separation in the evaluated data but does not prove that every maintenance alarm is correct at RUL < 0.3. The R = 80 setting uses 20% fewer communication rounds than R = 100, but its regression performance and ROC-AUC are lower. A deployment recommendation requires threshold-specific confusion matrices and a communication-versus-error-cost assessment.

The mean R^2 and RMSE trends have not visibly plateaued by R = 100. Extrapolation beyond the investigated round budgets would be speculative; additional runs at larger R values are necessary before estimating an asymptote or quantifying expected gains. Communication overhead should be assessed alongside those future measurements.

5.4. Final Configuration and Sequential Search Progression

Table 7 and Figure 6 show the progression of the sequential OAT search. From the selected window-size setting to $n_h = 40$ at R = 40, the reported mean R^2 increases from 0.511 to 0.687 (+0.176); subsequently increasing R from 40 to 100 raises it to 0.831 (+0.144). These conditional increments do not establish a universal factor-importance hierarchy because each step uses a different fixed configuration and interactions were not assessed. The original study [1] uses a different evaluation protocol and is included for context only.

Table 7. Sequential hyperparameter search progression and contextual comparison with the original framework [1]. The original and present R^2 values come from different prediction horizons and ensemble procedures and must not be interpreted as directly comparable estimates.

Config.	n_x	n_h	R	$R^2 \pm$ Std	RMSE ± Std	AUC ± Std	Eval. Protocol
Original [1]	10	20	40	0.940 (1 seed)	0.023 (1 seed)	0.969 (1 seed)	Full 50-day traj., 1000-boot CI
Step 1 baseline ($n_x = 10$ confirmed)	10	20	40	0.511 ± 0.024	0.059 ± 0.002	0.997 ± 0.005	20-day horizon, 3 seeds
After step 2 ($n_h = 40$ selected)	10	40	40	0.687 ± 0.021	0.047 ± 0.002	0.997 ± 0.005	20-day horizon, 3 seeds
Final: After step 3 ($R = 100$ selected)	10	40	100	0.831 ± 0.015	0.033 ± 0.001	1.000 ± 0.000	20-day horizon, 3 seeds

5.5. Fixed-Threshold Classification Assessment

For the selected $n_x = 10$, $n_h = 40$, $R = 100$ configuration, Figure 7 reports the seed-level confusion matrices obtained by applying RUL < 0.3 to both true and predicted normalized RUL (seeds 42, 123, and 456). ROC-AUC is 1.000 in every seed, yet the fixed threshold is not error-free: seed 42 has one false negative (recall 0.933, F1-score 0.966), seed 123 has one false positive (precision 0.938, specificity 0.971), and seed 456 classifies all samples correctly; accuracy is 0.980, 0.980, and 1.000, respectively. Perfect ranking therefore does not imply error-free classification at the operational threshold. Each matrix contains 50 samples (15 failure-labelled), which corresponds to the full 50-day trajectory rather than to the 20-day evaluation horizon used elsewhere; the evaluation set behind Figure 7 must be clarified, since predictions on training days would inflate these metrics, and counts this small give wide uncertainty intervals.

5.6. Stability-Penalty Sensitivity

The stability score was recalculated from the rounded ROC-AUC means and standard deviations in Tables 4–6 using $S(\lambda) = AUC\mu - \lambda\, AUC\sigma$, at $\lambda \in \{0, 0.3, 0.5, 0.8, 1.0, 2.0\}$. In the window-size and hidden-width steps, the selected candidates have $AUC\mu = 0.997$ and $AUC\sigma = 0.005$, giving approximate scores of 0.9970, 0.9955, 0.9945, 0.9930, 0.9920, and 0.9870, respectively. At R = 100, $AUC\mu = 1.000$ and $AUC\sigma = 0.000$, so the reported score is 1.0000 throughout. The selected candidates remain the same within this tested penalty range (Table 8). These are deterministic recalculations of rounded summary statistics, not new experimental runs; unrounded seed-level values are needed for final-precision reporting.

Table 8. Sensitivity of the stability-aware selection criterion to the penalty coefficient (λ) across the three OAT optimization steps.

λ	Best n_x (Score)	Best n_h (Score)	Best R (Score)
0.0	10 (0.9970)	40 (0.9970)	100 (1.0000)
0.3	10 (0.9955)	40 (0.9955)	100 (1.0000)
0.5	10 (0.9945)	40 (0.9945)	100 (1.0000)
0.8	10 (0.9930)	40 (0.9930)	100 (1.0000)
1.0	10 (0.9920)	40 (0.9920)	100 (1.0000)
2.0	10 (0.9870)	40 (0.9870)	100 (1.0000)

5.7. Limits of the Sequential Search and Reproducibility

A joint neighborhood study around $(n_x, n_h, R) = (10, 40, 100)$, for example, varying $n_h \in \{30, 40, 60\}$ and $R \in \{80, 100, 120\}$, would test whether the selected configuration remains competitive and whether the factors interact. These combinations have not been evaluated in the supplied results and are proposed as future experiments, not presented as observed outcomes. A complete reproducibility protocol must also document the actual temporal train/validation/test boundaries, overlap handling, normalization fitting, prediction horizon, model-selection data, and per-seed evaluation outputs.

6. Discussion

The original framework [1] reported $R^2 = 0.940$ over a full 50-day trajectory using a weighted ensemble, whereas the present study reports $R^2 = 0.831 \pm 0.015$ under a fixed 20-day evaluation horizon and a simplified ensemble. Because the prediction horizon and evaluation procedures differ, these values are not directly comparable and cannot establish an improvement or degradation. There is no experimental basis to predict that the selected configuration would recover or exceed $R^2 = 0.940$ under the earlier protocol. A valid comparison requires matched partitions, horizon, preprocessing, ensemble strategy, seeds, and metrics.

The coefficient of variation (STD/Mean) across seeds for R^2 under the selected configurations ranges from 1.8–4.7%, versus 0.0–0.5% for AUC. This tighter AUC distribution arises because AUC is a rank-based metric: it measures the model's ability to order samples by risk level rather than to predict absolute values. On a non-stationary degradation trajectory evaluated over a fixed horizon, absolute value predictions depend on the exact convergence basin reached by the LM optimizer (which is initialization-dependent), while rank ordering is more robust to initialization. This observation generalizes beyond the present study: any HPO campaign on a prognostic dataset with fewer than approximately 100 run-to-failure observations should treat R^2 as a secondary indicator and use a rank-based or classification metric as the primary selection criterion [2,14].

The reversal at $n_h = 60$ has specific implications in the federated context. Unlike centralized learning, where a practitioner can observe a global validation loss plateau and stop increasing capacity, in the P2P-FL setting each VC trains locally on its private HI series without access to a shared validation set. A VC with $n_h = 60$ would therefore not generate a local early-stopping signal, as its local validation loss continues to decrease, and the generalization penalty only becomes visible when comparing ensemble predictions to held-out test data. This suggests that cross-client validation loss sharing (sharing only aggregated loss values, not raw data) would be a useful privacy-preserving extension of the framework: it would enable federated early stopping, alerting all VCs when the ensemble has reached peak test-set performance. This aligns with the federated model monitoring approach proposed in [19,22] for multi-stream prognostic settings.

Within the tested range, larger R values are associated with improved mean regression performance, with no clear plateau by $R = 100$. The $R = 80$ setting reduces the number of communication rounds by 20% relative to $R = 100$ but has a lower reported ROC-AUC (0.995 ± 0.005 versus 1.000 ± 0.000) and weaker regression metrics. Neither the untested gains above $R = 100$ nor the absence of client drift can be established from the available results. Future evaluations should jointly measure communication cost, predictive error, and fixed-threshold detection performance.

The selected configuration was evaluated on one Suzlon 2.2 MW high-speed-shaft bearing run-to-failure trajectory. A single bearing, a fixed prediction horizon, and three initialization seeds do not establish transferability across turbines, bearing types, operating regimes, or prognostic horizons. Such external validation requires additional independent

trajectories and matched evaluation protocols; FEMTO-ST may be considered in a future study, but no results on that benchmark are reported here.

Having presented and discussed the main findings of the study, it is important to consider several factors that may affect the interpretation, robustness, and generalizability of the results. First, the sequential OAT design does not allow interactions among n_x, n_h, and R to be identified; for example, the optimal n_h may change at R = 200, or the optimal R may depend on $n_x = 5$. A 2^3 fractional factorial design could capture such interactions, although at approximately three times the computational cost [2]. Second, the use of only three random seeds is sufficient to flag potential instability but does not provide a tight estimate of the 95% confidence interval; for instance, the reported $R^2 = 0.511 \pm 0.024$ at $n_x = 10$ corresponds to an interval of approximately [0.451, 0.571]. Increasing the number of seeds to ten would substantially reduce this uncertainty, while approximately thirty seeds would be required to achieve a bound of ±0.01. Third, the evaluation relies primarily on the Suzlon 50-day trajectory, limiting the assessment of generalization across datasets. Although the FEMTO-ST cross-validation reported in [1] achieved R^2 values between 0.89 and 0.93, repeating this validation using $n_h = 40$ and R = 100 would provide stronger evidence of generalizability, particularly because the optimal n_h may depend on the effective training-set size. Fourth, the evaluation uses a fixed 20-day horizon, leaving the sensitivity of the results to the prediction horizon unexplored; evaluating additional 10- and 30-day horizons would therefore provide a more comprehensive assessment. Finally, the study relies exclusively on the LM optimizer, consistent with [1]. Although LM has demonstrated advantages on small datasets [1,16], alternative adaptive gradient-based optimizers, such as Adam and AdaGrad, which are increasingly used in FL prognostics [18], may respond differently to variations in n_h and R. A comparison across optimizers would therefore be a valuable extension for assessing the robustness of the observed findings.

7. Conclusions and Future Work

This study reports a three-seed sequential OAT sensitivity analysis of input window size, hidden-layer width, and communication-round budget for the VC P2P-FL prognostic framework. The best-performing configuration along the evaluated search trajectory is $n_x = 10$, $n_h = 40$, and R = 100, with $R^2 = 0.831 \pm 0.015$, RMSE = 0.033 ± 0.001, MAE = 0.027 ± 0.001, and ROC-AUC = 1.000 ± 0.000. The ROC-AUC result demonstrates ranking discrimination on the evaluated samples, not necessarily perfect threshold-dependent failure classification.

Within the evaluated sequential OAT steps, $n_x = 10$ and $n_h = 40$ were selected, and increasing R up to 100 was associated with better mean regression and ROC-AUC performance. The decline in regression performance at $n_h = 60$ illustrates sensitivity to model width under the tested fixed settings but does not independently demonstrate overfitting. Because the OAT design does not resolve interactions and the round-budget response had not visibly saturated, these settings should not be called a demonstrated global optimum or used to assert a transferable factor-importance ranking.

Future research should (i) recompute the confusion matrices of Figure 7 on a strictly held-out test period and report confidence intervals for accuracy, precision, recall, specificity, and F1 score at the predeclared 0.3 threshold; (ii) verify and document chronological data partitions and leakage prevention; (iii) evaluate local combinations of n_h and R, including values beyond R = 100, to test interactions and saturation; and (iv) repeat the analysis on independent run-to-failure trajectories and directly compare the original and selected configurations under an identical evaluation protocol.

Author Contributions: Conceptualization, J.S., L.S. and E.B.; methodology, J.S., L.S. and E.B.; software, J.S.; validation, J.S., L.S. and E.B.; formal analysis, J.S. and L.S.; investigation, J.S.; resources, L.S. and E.B.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, J.S., L.S. and E.B.; visualization, J.S.; supervision, L.S. and E.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: Author Eric Bechhoefer was employed by the company GPMS International Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Sidhom, J.; Saidi, L.; Bechhoefer, E. Decentralized Federated Learning for Wind Turbine Bearing Prognostics Under Data Scarcity and Statistical Heterogeneity. *Energy Sci. Eng.* **2026**, *14*, 3674–3696. [CrossRef]
2. Bischl, B.; Binder, M.; Lang, M.; Pielok, T.; Richter, J.; Coors, S.; Thomas, J.; Ullmann, T.; Becker, M.; Boulesteix, A.-L.; et al. Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges. *WIREs Data Min. Knowl. Discov.* **2023**, *13*, e1484. [CrossRef]
3. Lu, Z.; Pan, H.; Dai, Y.; Si, X.; Zhang, Y. Federated Learning With Non-IID Data: A Survey. *IEEE Internet Things J.* **2024**, *11*, 19188–19209. [CrossRef]
4. Chen, X.; Wang, H.; Lu, S.; Yan, R. Bearing RUL Prediction Using Federated Learning with Taylor-Expansion Network Pruning. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–10. [CrossRef]
5. Guo, L.; Yu, Y.; Qian, M.; Zhang, R.; Gao, H.; Cheng, Z. FedRUL: A New Federated Learning Method for Edge-Cloud Collaboration Based RUL Prediction. *IEEE/ASME Trans. Mechatron.* **2023**, *28*, 350–359. [CrossRef]
6. Mehta, M.; Chen, S.; Tang, H.; Shao, C. A Federated Learning Approach to Mixed Fault Diagnosis in Rotating Machinery. *J. Manuf. Syst.* **2023**, *68*, 687–694. [CrossRef]
7. Cai, C.; Fang, Y.; Liu, W.; Jin, R.; Cheng, J.; Chen, Z. FedCOV: Enhanced Trustworthy Federated Learning for Machine RUL Prediction with Continuous-to-Discrete Conversion. *IEEE Trans. Ind. Inform.* **2024**, *20*, 12565–12574. [CrossRef]
8. Shah, S.S.; Daoliang, T.; Kumar, S.C. RUL Forecasting for Wind Turbine Predictive Maintenance Based on Deep Learning. *Heliyon* **2024**, *10*, e40082. [CrossRef] [PubMed]
9. Arunan, A.; Qin, Y.; Li, X.; Yuen, C. A Federated Learning-Based Industrial Health Prognostics for Heterogeneous Edge Devices Using Matched Feature Extraction. *IEEE Trans. Autom. Sci. Eng.* **2023**, *21*, 3065–3079. [CrossRef]
10. Kamei, S.; Taghipour, S. A Comparison Study of Centralized and Decentralized Federated Learning for Estimating RUL. *Reliab. Eng. Syst. Saf.* **2023**, *233*, 109130. [CrossRef]
11. Lu, S.; Gao, Z.; Zhang, P.; Xu, Q.; Xie, T.; Zhang, A. Event-Triggered Federated Learning for Fault Diagnosis of Offshore Wind Turbines. *IEEE Trans. Autom. Sci. Eng.* **2023**, *21*, 1271–1283. [CrossRef]
12. Sidhom, J.; Saidi, L.; Gannoun, H. AI-Driven Prognostics for Wind Turbine Bearings: Predicting Degradation with Limited Data. In Proceedings of the IEEE ICCAD, Barcelona, Spain, 1–3 July 2025; pp. 1–6. Available online: https://ieeexplore.ieee.org/document/11099371 (accessed on 30 October 2026).
13. Wang, L.; Zhao, X.; Lu, Z.; Wang, L.; Zhang, S. Enhancing Privacy Preservation and Trustworthiness for Decentralized Federated Learning. *Inf. Sci.* **2023**, *628*, 449–468. [CrossRef]
14. Du, N.H.; Long, N.H.; Ha, K.N.; Hoang, N.V.; Huong, T.T.; Tran, K.P. Trans-Lighter: A Light-Weight Federated Learning-Based Architecture for Remaining Useful Lifetime Prediction. *Comput. Ind.* **2023**, *148*, 103888. [CrossRef]
15. Chung, S.; Al Kontar, R. Federated Condition Monitoring Signal Prediction with Improved Generalization. *IEEE Trans. Reliab.* **2023**, *73*, 438–450. [CrossRef]
16. Mehta, M.; Shao, C. A Greedy Agglomerative Framework for Clustered Federated Learning. *IEEE Trans. Ind. Inform.* **2023**, *19*, 11856–11867. [CrossRef]
17. Li, W.; Yang, W.; Jin, G.; Chen, J.; Li, J.; Huang, R.; Chen, Z. Clustering Federated Learning for Bearing Fault Diagnosis in Aerospace Applications with a Self-Attention Mechanism. *Aerospace* **2022**, *9*, 516. [CrossRef]

18. Norton, H.; Miller, L.; Rodgers, M. A Support Process for Early-Stage Wind Farm Repowering Decisions Using Constrained Optimization Techniques to Address Uncertainty. *Wind* **2026**, *6*, 17. [CrossRef]
19. Zhou, F.; Liu, S.; Fujita, H.; Hu, X.; Zhang, Y.; Wang, B.; Wang, K. Fault Diagnosis Based on Federated Learning Driven by Dynamic Expansion for Model Layers of Imbalanced Client. *Expert Syst. Appl.* **2024**, *238*, 121982. [CrossRef]
20. Chung, S.; Al Kontar, R. A Two-Stage Federated Learning Approach for Industrial Prognostics Using Large-Scale High-Dimensional Signals. *arXiv* **2024**, arXiv:2410.11101. Available online: https://arxiv.org/abs/2410.11101 (accessed on 30 October 2026).
21. Quan, M.K.; Pathirana, P.N.; Wijayasundara, M.; Setunge, S.; Nguyen, D.C.; Brinton, C.G.; Love, D.J.; Poor, H.V. Federated Learning for Cyber Physical Systems: A Comprehensive Survey. *IEEE Trans. Cogn. Commun. Netw.* **2025**, *28*, 3751–3790. [CrossRef]
22. Jenkel, L.; Jonas, S.; Meyer, A. Privacy-Preserving Fleet-Wide Learning of Wind Turbine Conditions with Federated Learning. *Energies* **2023**, *16*, 6377. [CrossRef]
23. Gong, F.; Ma, P.; Zhang, H.; Wang, C.; Li, X.; Meng, Y. Distribution-Aware Federated Learning for Wind Turbine Fault Diagnosis Under Data Imbalance. *Struct. Health Monit.* **2026**. [CrossRef]
24. Vásquez, S.; Verhelst, J.; Brijder, R.; Ompusunggu, A.P. Detection, Prognosis and Decision Support Tool for Offshore Wind Turbine Structures. *Wind* **2022**, *2*, 747–765. [CrossRef]
25. Saidi, L.; Ben Ali, J.; Bechhoefer, E.; Benbouzid, M. Particle Filter-Based Prognostic Approach for High-Speed Shaft Bearing Wind Turbine Progressive Degradations. In Proceedings of the IECON 2017—43rd Annual Conference of the IEEE Industrial Electronics Society, Beijing, China, 29 October–1 November 2017; pp. 8099–8104. [CrossRef]
26. Saidi, L.; Benbouzid, M. Prognostics and Health Management of Renewable Energy Systems: State of the Art Review, Challenges, and Trends. *Electronics* **2021**, *10*, 2732. [CrossRef]
27. Wei, Y.; Wu, D.; Terpenny, J. Bearing RUL Prediction Using Self-Adaptive Graph Convolutional Networks with Self-Attention Mechanism. *Mech. Syst. Signal Process.* **2023**, *188*, 109971. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.