SJER Scientific Journal of Engineering Research

futuristech PT. Teknologi Futuristik Indonesia

Date of publication February 24, 2026, date of current version February 24, 2026.
e-ISSN: 3109-1725
Digital Object Identifier 10.64539/sjer.v2i2.2026.413

Article

Federated Temporal Graph Learning for Weakly Supervised Bearing Anomaly Detection

Khagendra Darlami[1,*], Lalit Awasthi[2]

[1] School of Computer Science and Technology, Nanjing University of Information Science and Technology, Nanjing, 210044, China; khagendrad.345@gmail.com
[2] School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, 210044, China
* Correspondence

The authors received no financial support for the research, authorship, and/or publication of this article.

Abstract: Industrial bearing health monitoring is hindered by four interrelated challenges: high class imbalance, the absence of fault-type annotations, stringent data privacy constraints prohibiting centralized aggregation, and non-independent and identically distributed (non-IID) degradation dynamics across geographically dispersed assets. To address these, we propose Fed-TGCN, a novel weakly supervised federated learning framework grounded in temporal graph neural networks. Each client represents a leave-one-bearing-out fold, comprising two training bearings, one validation bearing, and one held-out test bearing, constructs a hybrid spatio-temporal graph from six physics-informed statistical features derived from raw vibration signals; edges encode both sequential dependencies and feature-space similarity via k-nearest neighbors. Pseudo-anomaly labels are generated locally through adaptive thresholding of a degradation score using exponentially weighted moving average, eliminating reliance on expert annotations. Under a strict leave-one-bearing-out protocol on the NASA IMS dataset (12 bearings), local Temporal Graph Convolutional Networks are trained in isolation and aggregated globally via FedAvg. Our method achieves an Average Precision of 0.675 ± 0.276 and Matthews Correlation Coefficient of 0.636 ± 0.285, maintains stronger performance consistency across heterogeneous bearing conditions than isolated and non-graph baselines (ΔMCC = +0.130, $p < 0.01$). Ablation studies confirm the necessity of temporal modeling (MCC drops by 0.069 without GRU). To the best of our knowledge, this is the first work integrating weakly supervised, graph-based federated learning for bearing prognostics under, demonstrating that parameter coordination but not the data sharing which enables degradation-invariant representation learning across heterogeneous assets.

Keywords: Anomaly Learning; Federated learning; Temporal graph neural networks; Weakly supervised anomaly detection; Bearing prognostics; Leave-one-bearing-out.

1. Introduction

Rolling bearings are critical components in rotating machinery, and their unexpected failure can lead to costly downtime and safety hazards [1]. In real-world settings, fault-type labels are rarely available; instead, predictive monitoring must detect early deviations using only operational data. So, we adopt a weakly supervised anomaly learning approach: a pseudo-anomaly, these labels are generated automatically via an adaptive thresholding of a multi-band health indicator derived from physical-statistical features from raw signals [2], [3], by adopting self-supervised training without human annotation.

Modern industrial systems increasingly involve geographically distributed assets whose raw data cannot be centrally shared due to privacy and regulatory constraints [4]. So, federated learning offers a solution, but bearing degradation is highly non-IID across sites and can vary in load, speed, and failure evolution [5]. This requires privacy-preserving models that generalize across unseen assets under weak supervision.

To overcome these limitations, recent efforts explore the construction of weakly-supervised health indicators that adapt to site-specific normal behavior without requiring fault labels [2]. However, integrating such anomaly-

aware models into a privacy-preserving federated framework, while preserving temporal-spectral degradation dynamics and ensuring cross-client generalization, remains largely unexplored.

Although federated learning (FL) has been proposed for industrial, IoT predictive health monitoring and maintenance [4], most existing methods face three key limitations. First, existing federated learning approaches for health monitoring often rely on convolutional or fully connected architectures [6], which may not explicitly model the long-term temporal evolution of degradation dynamics. Second, while some methods reduce reliance on labels through self-supervision, many still assume access to fault-type annotations or task-specific supervision signals that are rarely available in real-world prognostics. In contrast, practical systems require fully unsupervised or weakly supervised anomaly learning using only operational vibration data [2], [3].

In particular, existing federated learning approaches for monitoring bearing health have not yet leveraged graph-based relational modeling to capture dependencies among vibration segments. Although graph neural networks have demonstrated effectiveness in multivariate time-series anomaly detection [7], their integration into federated, privacy-preserving prognostics, particularly under non-IID operating conditions with heterogeneous loads, speeds, and degradation patterns across assets [5], has not been systematically explored. This gap limits generalization in realistic evaluation settings such as leave-one-bearing-out (LOBO), where models must adapt to previously unseen bearing units. To address this, we work on a weakly-supervised, federated graph-based framework that jointly preserves data privacy, models temporal degradation dynamics, and improves cross-asset generalization under operational heterogeneity.

This work makes four key contributions to privacy-preserving bearing health monitoring:

1) We implement a weakly supervised framework that generates pseudo-anomaly labels locally via adaptive thresholding of a multi-band health indicator derived from physics-informed features, calibrated using only normal operational data eliminating reliance on expert annotations.
2) We evaluate under a strict leave-one-bearing-out (LOBO) protocol that simulates real-world asset heterogeneity and data silos, comparing federated and isolated variants of recent baselines to reveal how parameter communication, not data sharing supports generalization under non-IID conditions.
3) We show that federated coordination improves early anomaly detection over isolated models under identical label scarcity and privacy constraints, suggesting federated learning acts as a functional enabler of robustness, not merely a privacy-preserving mechanism.

This paper is structured as follows: In Section 2, we review related work on weakly supervised anomaly learning, graph neural networks for industrial time-series, and federated learning for industrial monitoring. In Section 3, we present our federated temporal graph learning framework, detailing physics-informed feature engineering, hybrid graph construction, adaptive pseudo-label generation via EWMA thresholding, and federated optimization under a strict leave-one-bearing-out (LOBO) protocol. Section 4 describes the experimental setup. Sections 5 and 6 report comprehensive results covering performance analysis, statistical significance testing, ablation studies, cross-condition generalization, and limitations and conclude with implications for trustworthy, privacy-preserving prognostics. Finally, Section 7 outlines future research directions to enhance robustness, scalability, and real-world applicability.

2. Related Work

2.1. Paradigms and Weakly Supervised Anomaly Learning.

Bearing health monitoring has long relied on prognostic-driven paradigms that track degradation without fault-type labels. Early approaches used unsupervised health indicators such as RMS, spectral kurtosis, and envelope energy to model performance decay over time [1], [8].

Although supervised fault classification dominates curated benchmarks such as CWRU [9], it is not suitable for real-world industrial settings where fault annotations are absent, especially during early-stage degradation [10], [11].

Now, three methodological streams coexist:

1) Supervised/few-shot diagnosis, which assumes known fault classes [6], [12];
2) Self-supervised or domain-adaptive learning, which reduces but does not eliminate label dependency [6]; and
3) Weakly supervised anomaly learning, where pseudo-labels are generated automatically from signal-driven health indicators that require only normal operational data for calibration [5].

Among these, weakly supervised anomaly learning, where pseudo-labels are derived from signal-based health indicators using primarily normal operational data, offers greater practical relevance for privacy-constraint and heterogeneous industrial environments [3].

2.2. Deep Learning for Prognostics and Health Management

Deep learning has enabled direct health indicator learning from raw vibration signals, reducing reliance on

hand-crafted features. However, many deep PHM methods both early and current are still depend on fault-type labels or known degradation stages for training or validation, which are rarely available in real industrial settings where failures evolve silently and labeling is costly or infeasible.

To address this, unsupervised and weakly supervised approaches have emerged. Methods like critic-based anomaly detection [5] or contrastive learning with normal-only data [13] avoid explicit labels, yet often struggle under asset heterogeneity (e.g., varying speed, load, or failure modes). In contrast, purely statistical strategies such as EWMA-based thresholding [14] offer greater robustness when fault semantics are unknown. This underscores a key gap: effective PHM models must be both label-free and capable of generalizing across non-IID assets under strict leave-one-asset-out evaluation, a need that motivates privacy-aware, weakly supervised frame-works over supervised alternatives that assume stationary conditions and labeled faults [15], [16].

2.3. Graph Neural Networks for Industrial Time-Series

Graph Neural Networks (GNNs) offer a powerful framework for modeling industrial time-series by capturing relational dependencies beyond fixed sequences or grids [17]. While, surveys of GNN [18] demonstrate their broad applicability and structured data representation. In prognostic health monitoring, GNNs can represent degradation dynamics through graphs built from temporal proximity and feature-space similarity, which gives expressive, structure-aware anomaly learning.

Recent work leverages GNNs for unsupervised anomaly detection in multivariate signals. Deng and Hooi [7] propose a dynamic graph model that jointly learns node embeddings and temporal patterns without labels. Similarly, Wang et al. [19] use graph auto-encoders to detect deviations from normal bearing behavior by reconstruction error without fault-type supervision. Following this paradigm, our approach constructs a graph where each node is a time step containing statistical features, and edge combines sequential links with k-NN based similarity. This helps capture both temporal evolution and recurrent degradation states, supporting weakly super-vised anomaly learning in a federated prognostic setting.

2.4. Federated Learning for Industrial Anomaly Detection

Federated learning (FL) powers privacy-preserving predictive health monitoring by training models across distributed assets without sharing raw vibration data, and is critical under data silos, intellectual property, and regulatory constraints [4]. Industrial deployments further challenge FL with non-IID degradation patterns and heterogeneous operating conditions [20].

Existing FL approaches for machinery health often assume supervised or semi-supervised settings requiring fault-type labels [6], or use few-shot/meta-learning that presumes known fault taxonomies [6], [21]. These can be impractical in real-world prognostics, where labels are not available and failure modes evolve.

3. Proposed Method

3.1. Feature Engineering and Graph Construction

Given raw vibration signals $x^{(b)} \in R^{T_b \times S}$ for bearing b, where T_b is the number of time steps (files) and S = 20,480 samples per file recorded at 20 kHz, we follow standard practices in vibration-based diagnostics [22] by segmenting each 1-second file into non-overlapping 40 ms windows (800 samples). From each window, we compute six discriminative time–frequency statistical features such as envelope RMS, kurtosis, RMS, and log-band energies in the BPFO, BPFI, and high frequency ranges as following.

3.1.1. Envelope RMS

$$f_1 = \sqrt{\left(\frac{1}{N}\sum_{n=1}^{N}\left(\left|\left(H(x[n])\right)\right|\right)^2\right)} \quad (1)$$

Where:

- $x[n] \in R^N$: vibration signal segment (800 samples, 40 ms)
- $H(\cdot)$: discrete Hilbert transform
- $H(x[n])$: analytic envelope

It measures the energy of impulsive transients after demodulation; highly sensitive to early-stage bearing faults.

3.1.2. Kurtosis

$$f_2 = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{x[n]-\mu_x}{\sigma_x}\right)^4 \quad (2)$$

Where:

- $\mu_x = \frac{1}{N}\sum x[n]$: mean
- $\sigma_x = \frac{1}{N}\sum (x[n]-\mu_x)^2$: standard deviation

It quantifies impulsiveness beyond Gaussian noise; high values suggest sporadic fault impacts.

3.1.3. Log Energy in BPFO Band

$$f_3 = ln\left(1 + \sum_{f \in B_{BPFO}} |X(f)|^2\right) \quad (3)$$

Where:

- $X(f) = FFT(x)[f]$: Fourier transform at frequency f
- B_{BPFO}= [119–5, 119+5] Hz
- $BPFO \approx 119$ Hz for 2000 RPM (computed via standard bearing geometry)

It captures resonance energy around the theoretical Ball Pass Frequency Outer race; indicative of outer-race defects.

3.1.4. Log Energy in High-Frequency Band

$$f_4 = ln\left(1 + \sum_{f=2000}^{10000} |X(f)|^2\right) \quad (4)$$

Where:

- Frequency range: 2000–10,000 Hz (resonance band of test rig)
- $X(f) = FFT(x)[f]$ denote the discrete Fourier transform of the vibration segment x at frequency f (in Hz).

It monitors broadband high-frequency energy where fault-induced structural resonances typically occur.

3.1.5. Log Energy in BPFI

$$f_5 = ln\left(1 + \sum_{f \in B_{BPFI}} |X(f)|^2\right) \quad (5)$$

Where:

- B_{BPFI}=[181−5, 181+5] Hz
- $BPFI \approx 181$ Hz for 2000 RPM
- $X(f) = FFT(x)[f]$ denote the discrete Fourier transform of the vibration segment x at frequency f (in Hz).

It measures sensitive to inner-race defects through energy concentration near the Ball Pass Frequency Inner race.

3.1.6. Root Mean Square

$$f_6 = \sqrt{\left(\frac{1}{N}\sum_{n=1}^{N} x[n]^2\right)} \quad (6)$$

Where:

- $x[n]$: raw vibration samples

It reflects overall vibration energy but is less specific to incipient faults due to sensitivity to load/speed variations.

This yields a feature matrix $X^{(b)} \in R^{Tb \times 6}$. These window-level features are then aggregated (via mean or max) to produce a single, noise-robust feature vector per file, preserving the original temporal resolution while enhancing sensitivity to early degradation through short-time signal characteristics.

We then construct a dynamic graph $G^{(b)} = (V^{(b)}, E^{(b)})$ for bearing b:

- Each node $v_t \in V^{(b)}$ corresponds to time step t with node feature $x_t = X^{(b)}[t, :]$.
- Edges combine temporal and feature similarity structure:

$$E^{(b)} = T_c \cup F_s(k - NN) \quad (7)$$

where:

$$T_c = (t, t+1) \mid 1 \le t < T_b, \text{ and}$$

$$F_s(k - NN) = (i, j) \mid cos(x_i, x_j) \in top - k, k = 5.$$

This hybrid design follows spatio-temporal graph learning principles [23], preserving both sequential degradation dynamics and recurrent fault states.

Our features such as envelope RMS, spectral kurtosis, and band-limited energies follow the principle of physical-statistical fusion, shown to be effective in unsupervised industrial monitoring [24]. The Figure 1 summarizes the proposed pipeline: six discriminative time–frequency features, the physics-informed health indicator (degradation score), and the generation of pseudo-anomaly labels via adaptive EWMA threshold.

3.2. Temporal Graph Neural Networks

Our model, named Temporal Graph Convolutional Net-work (T-GCN), integrates spatial graph-based message passing with sequential dynamics to capture both inter time-step similarity and bearing degradation evolution. Given node features $X \in R^{T \times 6}$ and edge indices E, T-GCN first applies two Graph Convolutional Network (GCN) layers [17] to encode relational structure:

$$H^{(1)} = \sigma(\hat{A}XW^{(1)}), \quad H^{(2)} = \sigma(\hat{A}H^{(1)}W^{(2)}) \quad (8)$$

where Ã is the normalized adjacency matrix and σ is ReLU. The output embeddings $H^{(2)} \in R^{T \times d}$ are then grouped by bearing and processed by a Gated Recurrent Unit (GRU) to model temporal progression across the asset's lifetime. A final MLP head outputs anomaly logits per time step. This hybrid spatio-temporal design follows principles from traffic forecasting [23] and multivariate anomaly detection [7], which enabled unsupervised, node-level predictions without centralized data access. Finally, unsupervised anomaly labels $y_t \in \{0,1\}$ health indicator is first computed as:

$$h_t^{(b)} = \sum_{m=1}^{6} w_m \cdot max\left(0, \frac{x_{t,m}^{(b)} - \mu_m^{(b)}}{\sigma_m^{(b)}}\right) \quad (9)$$

where $\mu_m^{(b)}$, $\sigma_m^{(b)}$ are the mean and standard deviation over a burn-in window (first 10% of T_b), and weights = [0.30, 0.25, 0.20, 0.15, 0.0, 0.10]. The RMS feature (f_6) is excluded from the degradation score computation because it is highly sensitive to non-degradative operational variations (e.g., transient load changes), which could mislead the EWMA-based pseudo-labeling mechanism in the absence of fault-type supervision.

An exponentially weighted moving average (EWMA) then defines a time-varying threshold, yielding:

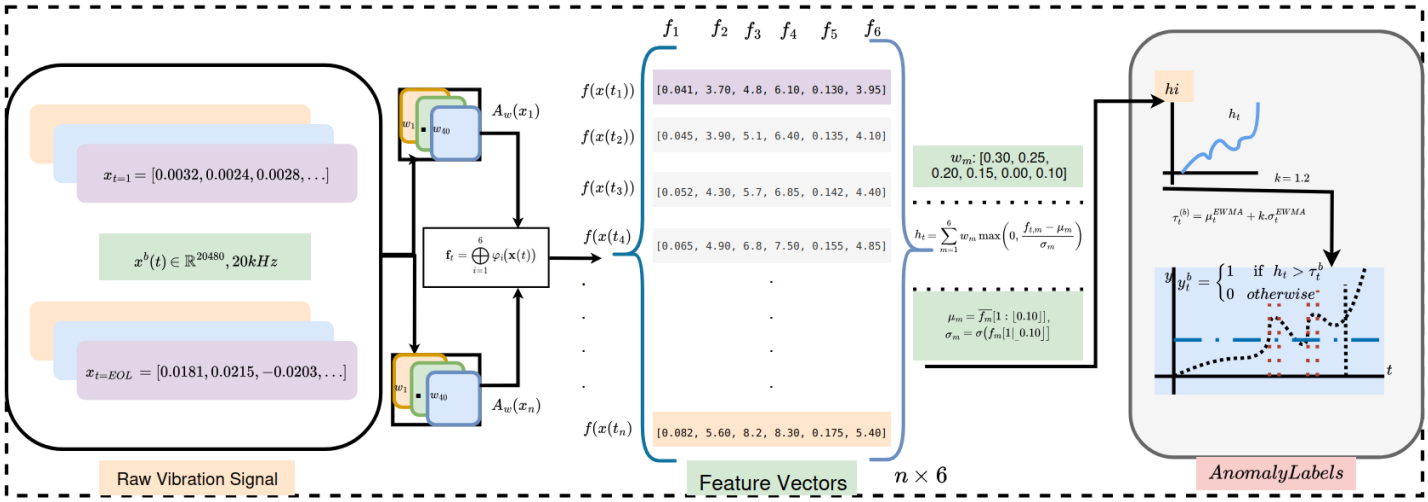

Figure 1. This diagram explain how raw vibration segments are used to extract 6-D physical-statistical features.

$$y_t^{(b)} = I[h_t^{(b)} > \tau_t^{(b)}] \tag{10}$$

which serves as the node-level anomaly label for training.

3.3. Federated Training Strategy

Our federated framework operates under a strict leave-one-bearing-out (LOBO) protocol. The NASA IMS dataset comprises 12 bearings partitioned into three test conditions (1st_test, 2nd_test, 3rd_test), each containing four bearings. For each test, we construct four LOBO folds: in fold k, one bearing serves as test, one as validation, and the remaining two as training ensuring zero data leakage. Evaluation aggregates results over all 12 folds. Each fold k_i acts as an independent FL client, holding local time-series data.

$$D_i = \{(x_t^{(i)}, y_t^{(i)})\}_{t=1}^{T_i} \tag{11}$$

where $x_t \in R^6$ are handcrafted physical-statistical features and $y_t \in \{0,1\}$ are pseudo-anomaly labels generated via adaptive EWMA thresholding of a degradation score. To address severe class imbalance, each client minimizes a weighted binary cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[w_p y_i ln\hat{y}_i + (1-y_i)ln(1-\hat{y}_i)] \tag{12}$$

where the positive weight is set to $w_p = \frac{1-p}{p}$ with $p = thepositiveanomalyratio$, following the standard practices for imbalanced binary classification. Model weight updates are aggregated at the server via Federated Averaging (FedAvg) [25]:

$$\theta^{(r+1)} = \sum_{i=1}^{12}\frac{|D_i|}{\sum_j |D_j|}\theta_i^{(r)} \tag{13}$$

where θ_i is the locally updated TemporalGCN from client i at round r. This preserves raw data privacy while enabling collaborative learning across heterogeneous assets [6].

Figure 2 summarizes the complete federated pipeline: from raw vibration signals to feature extraction, graph construction (temporal + kNN edges), local T-GCN training, and global weight aggregation under LOBO.

4. Experimental Setup

4.1. Dataset Description

We employ the NASA IMS bearing dataset [26], which comprises raw vibration signals from 12 Rexnord ZA-2115 double-row tapered roller bearings distributed across three independent test conditions (1st_test, 2nd_test, and 3rd_test). Data were acquired under constant operating conditions: a radial load of 6,000 lbs applied via a hydraulic actuator and a fixed rotational speed of 2,000 RPM. Vibration signals were sampled at 20 kHz using PCB 353B33 accelerometers mounted on the bearing housing. The 1st_test includes four bearings with dual-channel measurements (Channels 1 and 2), whereas the 2nd_test and 3rd_test each contain four bearings with only Channel 1 available. For consistency and to ensure a fair cross-condition evaluation, we use exclusively Channel 1 across all test conditions. Each bearing's full operational lifecycle from healthy initial state to eventual failure (where observed) is recorded as a sequence of files, with each file containing 20,480 time-domain samples (equivalent to one second of data).

Although our study involves 12 bearings, the effective sample size is substantial; the NASA IMS dataset comprises 37,856 sequential files equivalent to over 15 million 40-ms analysis windows collected under three distinct test conditions with markedly different degradation behaviors (abrupt vs. gradual failure, varying lifespans, and signal characteristics).

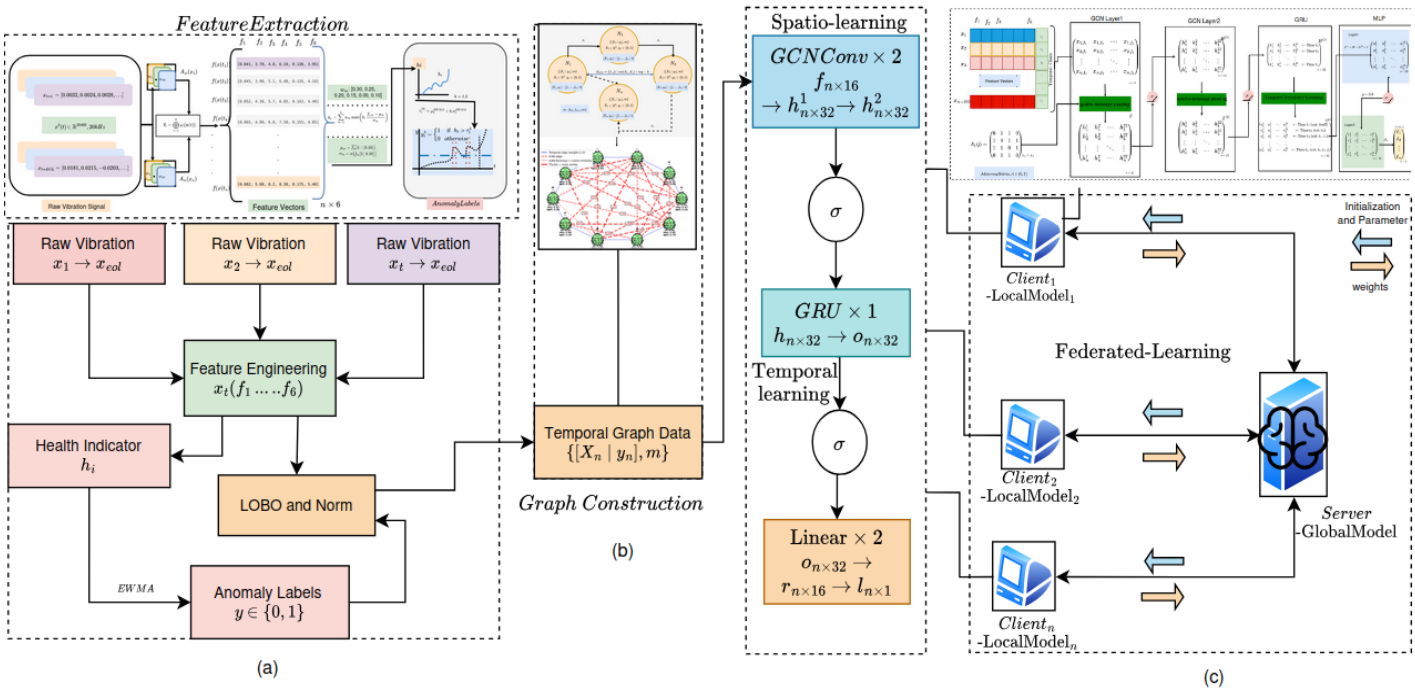

Figure 2. Federated learning pipeline: (a) Raw vibration segments are converted to 6D physical-statistical feature vectors and node-level pseudo-anomaly labels are generated via EWMA-based adaptive thresholding; (b) a hybrid graph (temporal + kNN) is built per bearing; (c) each bearing trains a local T-GCN and a central server aggregates model weights via FedAvg across 12 LOBO folds which never accesses raw data.

Table 1. NASA IMS dataset acquisition parameters and LOBO partitioning. Signal length denotes total sequential files per bearing life-cycle.

Test Condition	Total Bearings	Signal Length	After LOBO (Train/Test Set)
1st_test	4	2156	4 set/Test condition
2nd_test	4	984	4 set/Test condition
3rd_test	4	6324	4 set/Test condition

Table 2. Key hyper-parameters (fixed across all folds).

Parameter	Value
GCN hidden dimension	32
GRU hidden dimension	16
Learning rate	0.001
Optimizer	Adam
Weight decay	1e-5
Local epochs per round	5
EWMA smoothing factor (k)	1.2
Pos Weight	13.3

As summarized in Table 1, this heterogeneity enhances model robustness under the leave-one-bearing-out protocol, ensuring that performance reflects genuine cross-condition generalization rather than over-fitting to a single failure mode.

Also, this setup enables strict leave-one-bearing-out (LOBO) cross-validation, where models are trained on data from three bearings and evaluated on a completely unseen fourth bearing within the same test condition. To mitigate high-frequency sensor noise while preserving fault-sensitive transients, we extract robust statistical features (e.g., envelope RMS, log-band energies, and kurtosis).

Furthermore, Figure 3 shows raw vibration signals (Channel 1) from the first, middle, and last recording files of each bearing across the three NASA IMS test conditions, illustrating the evolution from health

4.2. Experiment Protocols

We adopt a strict leave-one-bearing-out (LOBO) cross-validation protocol across all 12 bearings in the NASA IMS dataset [26]. The bearings are grouped into three test conditions (1st–3rd test), each containing 4 bearings.

This yields 12 total folds, and ensures every bearing is used exactly once as a test set.

In the federated setting, each fold functions as an independent client. During training, clients compute local updates using only their own data and communicate model parameters but not raw signals to a central server, which aggregates them via FedAvg [25]. Training uses early stopping based on global validation performance (aggregated across all validation bearings) and weighted binary cross-entropy to address high label imbalance.

In the Isolated setting, models are trained per fold on the two training bearings and evaluated on the test bearing, which matches the federated protocol in split structure, unlike federated setting, all training occurs single model per fold without exchanging model parameters with other folds, meaning each fold is independent. This

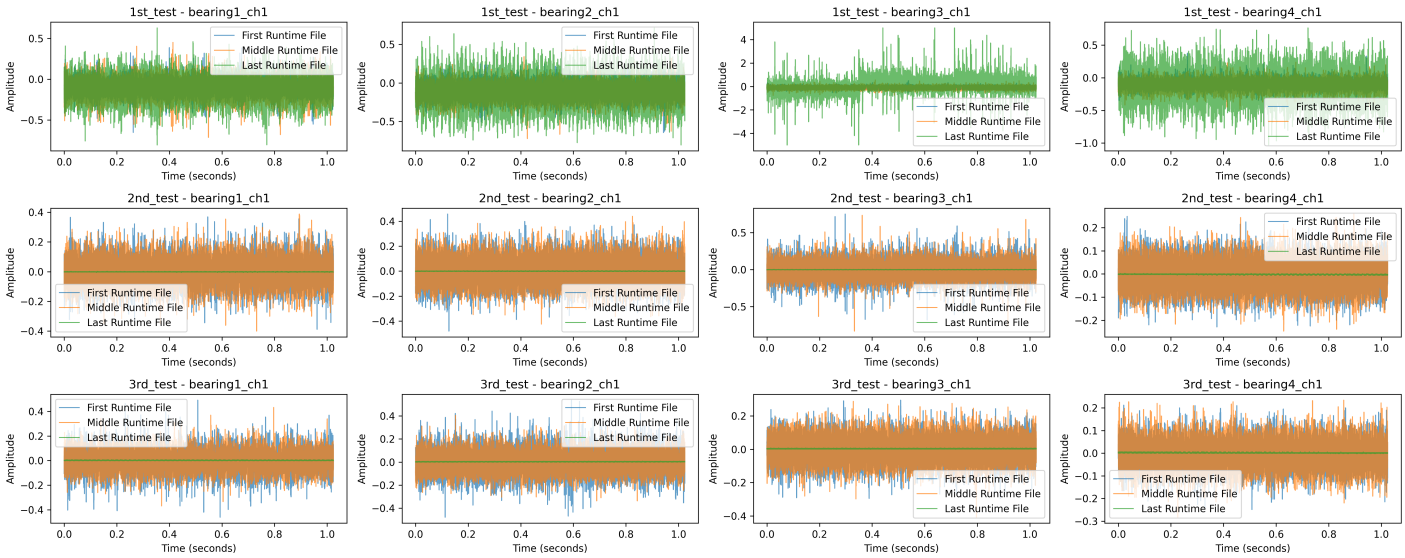

Figure 3. Raw vibration signals (Channel 1) are shown from the first, middle, and last data file of each.

design ensures a fair comparison while respecting the real-world constraints of data silos and privacy [27].

We used a fixed set of hyper-parameters in all experiments: GCN hidden dim = 32, GRU hidden dim = 16, learning rate = 0.001 (Adam), weight decay = 1e-5, local epochs = 5, and EWMA smoothing factor k=1.2, pos weight 13.3. These were selected via grid search on a validation fold and kept constant for all LOBO runs. Table 2 shows all hyper-parameters used throughout the feature engineering and model training.

4.3. Baseline Models

To evaluate the efficacy of privacy-preserving federated learning for bearing prognostics, we compare against five baselines under the same strict leave-one-bearing-out (LOBO) protocol:

1) GNN: a GCN-only variant without temporal modeling [17], [18];
2) LSTM: a recurrent network widely used for sequential degradation tracking in PHM [28];
3) 1D-CNN: a temporal convolutional model commonly adopted in bearing fault diagnosis [29]; and
4) Isolation Forest: a non-deep unsupervised baseline using the same 6 engineered features [30].

In each of the 12 LOBO folds, all models use identical data splits: two training bearings, one validation bearing, and one held-out test bearing, ensuring strict no-leakage evaluation.

Isolated local baselines are trained on the two training bearings of the fold, while our federated approach trains each fold in isolation and communicates only model parameters via FedAvg, preserving raw vibration data privacy. This design reflects real-world deployment constraints: labels are not manually annotated but are instead derived from degradation dynamics via adaptive EWMA thresholding. As a result, very low percentage of the time steps are labeled anomalous, reflecting the sparsity of early degradation events in real-world prognostics, where the goal is to detect deviation before catastrophic failure, not to classify known fault types.

4.4. Evaluation Metrics

Given the high-class imbalance inherent in real-world bearing degradation data where anomalous (degraded) states constitute only approximately 7–10% of all time steps, we adopt four complementary, threshold-robust evaluation metrics that remain informative under severe label skew:

Average Precision (AP): It summarizes the area under the precision–recall curve across all possible classification thresholds, providing a single scalar that reflects model performance without dependence on an arbitrary operating point:

$$AP = \sum_{k=1}^{n} (R_k - R_{k-1}) P_k \tag{14}$$

where P_k and R_k denote precision and recall at the k-th threshold, with $R_0 = 0$. This metric is particularly suitable for early anomaly detection, where recall of rare events is prioritized over overall accuracy.

Matthews Correlation Coefficient (MCC): It offers a balanced measure of classification quality by incorporating all entries of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) and it is widely recommended for imbalanced binary tasks:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

MCC ranges from –1 (total disagreement) to +1 (perfect prediction), with 0 indicating performance equivalent to random guessing. Unlike accuracy or F1, MCC remains reliable even when one class dominates.

Table 3. Performance comparison under strict LOBO evaluation (mean ± std across 12 folds). Higher values are better.

Models	Average Precision (AP)	Matthew Correlation (MCC)	Best F1
Federated T-GCN (Ours)	**0.675 ± 0.276**	**0.636 ± 0.285**	**0.687 ± 0.232**
Cen-T-GCN	0.610 ± 0.255	0.506 ± 0.320	0.573 ± 0.259
GNN	0.582 ± 0.241	0.518 ± 0.262	0.582 ± 0.203
LSTM	0.576 ± 0.294	0.455 ± 0.317	0.521 ± 0.259
CNN	0.578 ± 0.306	0.454 ± 0.347	0.512 ± 0.301
Isolation Forest	0.424 ± 0.293	0.347 ± 0.346	0.469 ± 0.228

Best F1-score: It reports the maximum harmonic mean of precision and recall achievable across all thresholds by $BestF1 = \tau maxF1(\tau)$ where F1:

$$F1(\tau) = 2\frac{Precision(\tau) \times Recall(\tau)}{Precision(\tau) + Recall(\tau)} \quad (16)$$

This captures the optimal trade-off between minimizing false alarms (high precision) and capturing true anomalies (high recall) which is a critical balance in safety-critical prognostics.

All metrics are computed per test bearing and averaged across the 12 LOBO folds. This ensures a realistic evaluation under strict cross-asset generalization without data leakage.

5. Result and Analysis

5.1. Performance Analysis

This evaluation directly addresses two core research questions: Can weakly supervised federated learning improve cross-asset anomaly detection without fault labels or data sharing? and does modeling spatio-temporal degradation via hybrid graphs enhance generalization over non-graph architectures? Our method Fed-TGCN, which combines adaptive EWMA pseudo-labeling, hybrid temporal-kNN graph construction, and FedAvg under strict LOBO that provides affirmative answers.

Table 3 presents a comprehensive comparison under strict LOBO evaluation. Our federated T-GCN achieves an Average Precision (AP) of 0.675 ± 0.276 and Matthews Correlation Coefficient (MCC) of 0.636 ± 0.285, demonstrating strong and balanced anomaly detection performance across diverse bearing conditions. In particular, federated T-GCN exceeds isolated local baselines, despite never exchanging raw vibration data or even normalized features through its federated gradient aggregation and parameter broadcasting.

To the best of our knowledge, no existing study has addressed bearing anomaly detection under the joint constraints of (i) strict data privacy (no raw/feature sharing), (ii) absence of fault-type labels, and (iii) cross-asset generalization via leave-one-bearing-out evaluation using graph-based temporal modeling with graph neural networks. This confirms the novelty of our approach: the first integration of weakly supervised, graph-based federated learning under strict LOBO for bearing prognostics, establishing federated learning not just as a privacy tool but as a functional enabler of robustness under distribution shift and label scarcity.

This advantage also stems from a key distinction in how the two paradigms handle cross-asset heterogeneity under strict LOBO. Isolated-LOBO trains only the two bearings of each fold, whereas Federated-LOBO also trains locally on the same pair yet aggregates their weights from all 12 bearings via FedAvg, yielding a universal degradation representation that generalizes to unseen assets.

Thus, federated LOBO acts as an implicit regularizer: by distilling insights from heterogeneous operating conditions (e.g., different failure modes, speeds, loads), it avoids over-fitting to the idiosyncrasies of any single fold's training pair. This explains its superior MCC and consistent performance across challenging folds (e.g., non-degrading bearings).

The proposed framework also consistently exceeds classical architectures (LSTM, CNN, GNN) and the unsupervised Isolation Forest, demonstrating that our weakly supervised, graph-based federated approach learns more robust degradation representations without centralized data access, making it well-suited for real-world deployment under distribution shift and privacy constraints. Figure 4 shows the validation performance (mean ± standard deviation across the 12 LOBO folds).

Federated T-GCN exhibits relatively higher learning convergence and lower variance, particularly in MCC, compared to Isolated and non-graph alternatives.

Furthermore, Figure 5, the boxplots of the final AP and MCC test on the 12 LOBO folds show that the federated T-GCN not only achieves higher medians but also maintains a tighter inter-quartile range compared to most baselines. It demonstrates robustness to data imbalance and distribution shift, taking advantage of the federated setting. These results confirm that our weakly supervised graph-based federated framework improves performance reliably and consistently across heterogeneous, imbalanced bearing populations.

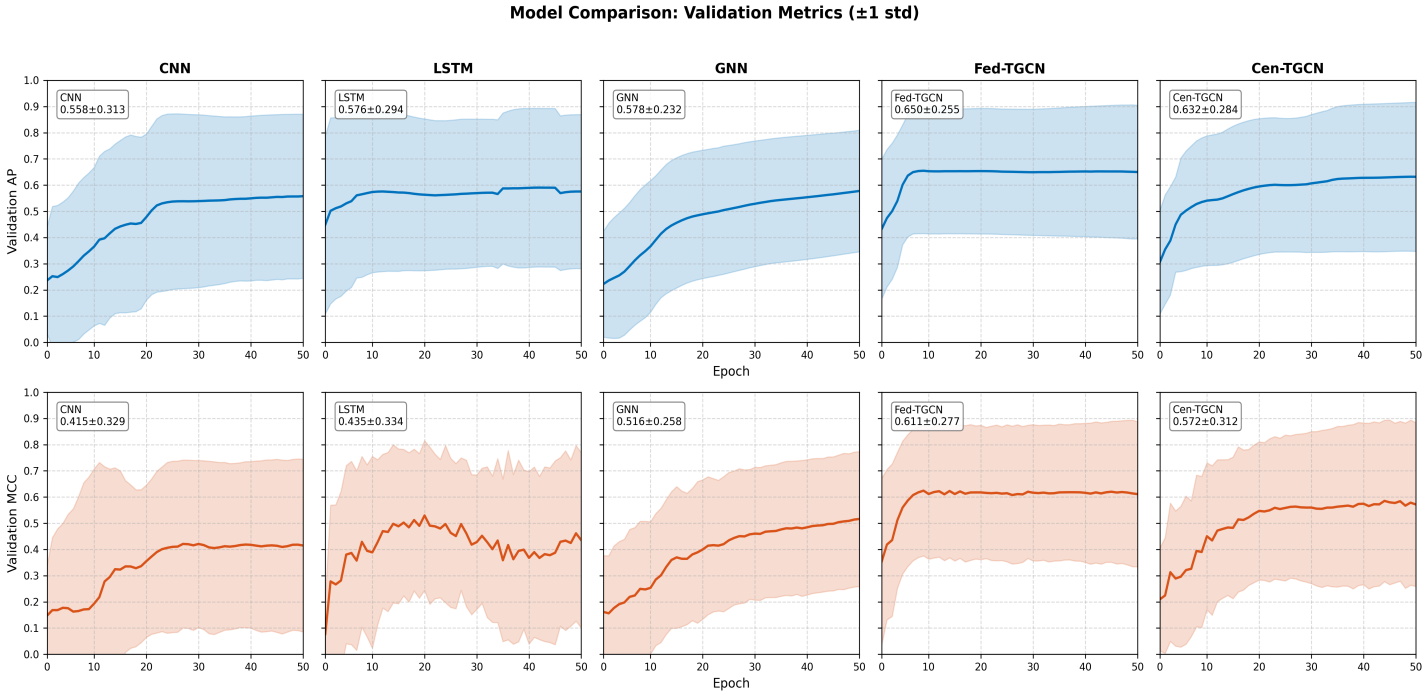

Figure 4. Validation performance across 50 epochs, aggregated (Fed-TGCN) and averaged over 12 LOBO folds (mean ± 1 std). Top: Average Precision (AP); Bottom: Matthews Correlation Coefficient (MCC). Federated T-GCN shows faster convergence and lower variance than baselines.

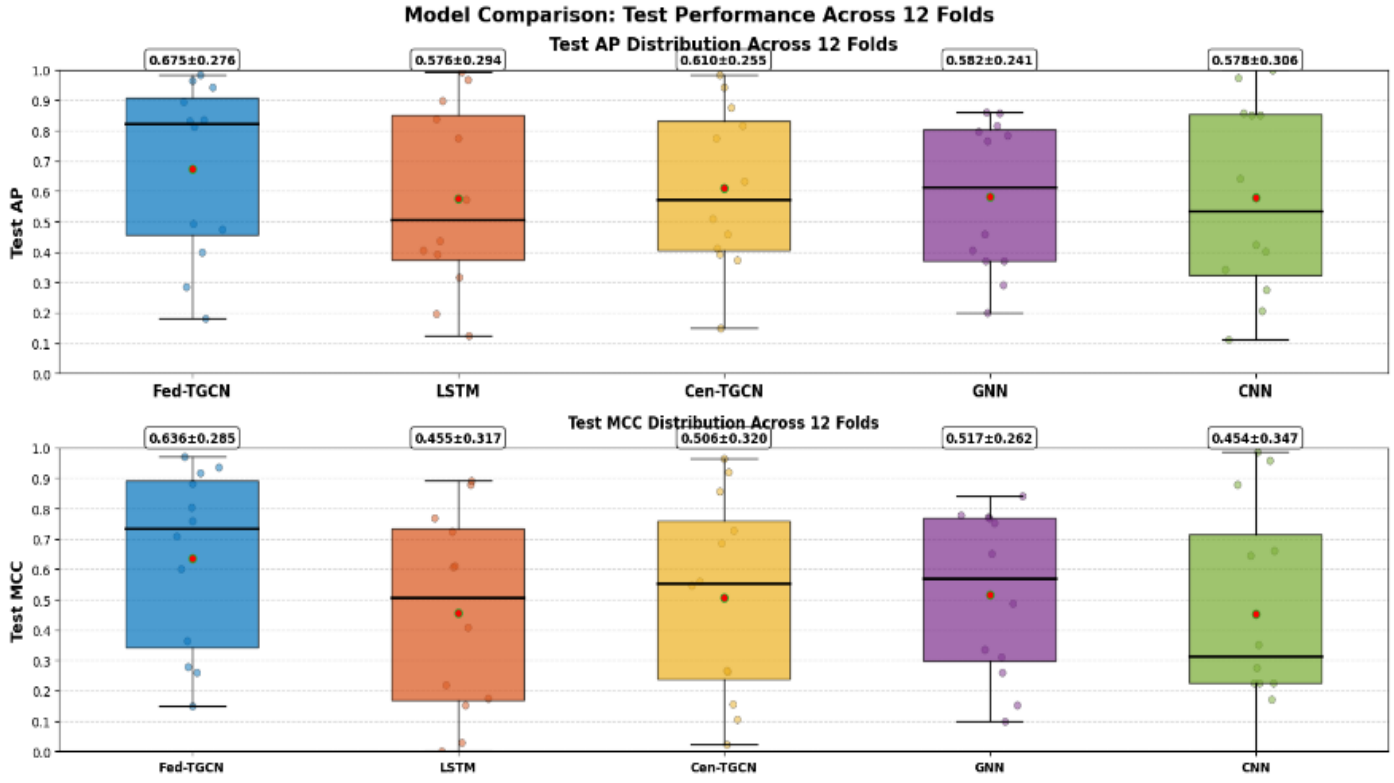

Figure 5. Final test performance across the same 12 LOBO folds for all models. Top: Average Precision (AP); Bottom: Matthews Correlation Coefficient (MCC). Box-plots show median (line), inter-quartile range (IQR, box), whiskers (1.5×IQR), and outliers.

And, Figure 6 presents the aggregated confusion matrix across all 12 leave-one-bearing-out (LOBO) test folds, summarizing predictions over 33,646 time-step windows. The model correctly identifies 2,265 true anomalies (TP) and 30,858 normal states (TN), while yielding 4,211 false positives (FP), primarily due to transient operational variations misclassified as early faults and only 522 false negatives (FN), reflecting strong recall in detecting genuine degradation events.

This low FN rate is critical in safety-sensitive prognostics, where missed anomalies pose greater risk than false alarms. The imbalance-aware design of our weakly

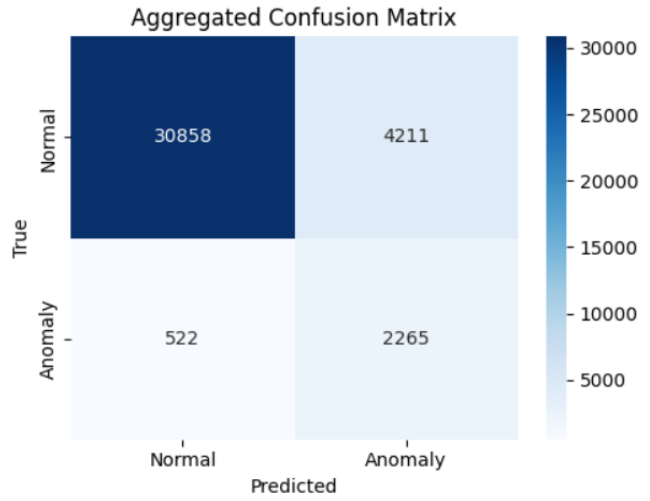

Figure 6. Confusion matrix aggregated across all 12 LOBO test folds (33,646 time-step windows). The model achieves low false negatives (522) and high true positives (2,265).

Table 4. Wilcoxon signed-rank test results (12 LOBO folds). $p < 0.05$ indicates statistically significant.

Comparison	AP	MCC
Fed-TGCN vs. LSTM	0.092	0.021*
Fed-TGCN vs. Cen-TGCN	0.034*	0.003**
Fed-TGCN vs. GNN	0.021*	0.001***
Fed-TGCN vs. CNN	0.092	0.012

*Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.*

Table 5. Fed-TGCN AP by test condition (mean ± std over 4 folds).

Test Condition	AP
1st_test	0.208 ± 0.068
2nd_test	0.456 ± 0.237
3rd_test	0.863 ± 0.064

supervised framework combining adaptive EWMA pseudo-labeling with federated temporal graph learning enables this favorable trade-off: high sensitivity to incipient faults without excessive false alarms, even under very high-class skew. Notably, many FPs may correspond to unannotated but real degradation precursors, a known limitation of pseudo-labeling in label-scarce settings. Overall, the confusion matrix validates that Fed-TGCN achieves robust, operationally meaningful anomaly detection across heterogeneous bearing assets under strict privacy constraints.

5.2. Significance Test

Bearing degradation data exhibit significant class imbalance and strong non-IID characteristics across operating conditions. This leads to high performance variance across the 12 LOBO folds, evident in the standard deviations of all models. To assess whether the performance gains of our federated T-GCN are statistically significant, we performed Wilcoxon signed-rank tests (non-parametric, paired, two-sided) comparing its fold-wise AP and MCC scores against all baselines.

As shown in Table 4, our method achieves statistically significant improvements in MCC in all baselines under strict lobo (all $p < 0.05$), with significant gains against GNN ($p = 0.001$) and Cen-TGCN ($p = 0.003$). In AP, the federated T-GCN significantly exceeds the Cen-GCN ($p = 0.034$) and GNN ($p = 0.021$), although the differences with LSTM and CNN are not significant ($p > 0.05$), likely due to the high variance in folds with sparse degradation signals.

5.3. Ablation and Cross-Condition Analysis

We analyze the contribution of key components through ablation. Removing the GRU module (i.e., using a GCN only architecture) reduces global MCC from 0.636 to 0.567 and AP from 0.675 to 0.632, with the largest degradation in low-signal folds (e.g., 1st_test_fold_3: MCC drops to 0.181).

This confirms that modeling temporal evolution is essential for robust anomaly learning, especially when degradation signals are weak or noisy.

Table 5 shows that performance also varies across test conditions reflecting inherent data characteristics:

- 3rd_test achieved high AP (0.863 ± 0.064) due to long, stable degradation trends.
- 1st_test (AP = 0.208 ± 0.068) and 2nd_test (AP = 0.456 ± 0.237) show lower performance due to abrupt failures or highly unstable signals in bearings.

This shows that our framework's output is faithful to the underlying degradation physics, it does not force anomalies where none exist, a crucial property for trustworthy prognostics. Rather than over-fitting to transient noise or operational artifacts, the model's weakly supervised design, grounded in adaptive EWMA thresholding of a multi-band health indicator, ensures that predicted anomalies align with meaningful, sustained deviations from normal behavior.

5.4. Limitation

While our framework demonstrates strong performance in detecting anomalies based on degradation, several limitations arise from both data and protocol constraints. First, bearing performance declines with minimal stability and highly fluctuating degradation signals, such as two units in 1st_test that exhibit unstable drift of the health indicator. In these cases, our EWMA based pseudo-labeling might generated high, but in irregular pattern, the irregular temporal structure of these labels leads to under-constrained local learning. This reflects a fundamental challenge in weakly supervised prognostics: effective anomaly learning requires a measurable, monotonic degradation trend, not merely operational data [5]. As noted

in recent studies, early-stage anomalies often lack consistent statistical signatures, making them indistinguishable from noise without prior knowledge of fault evolution [2]. Second, our method assumes uniformly sampled and temporally aligned vibration segments, a condition not fully satisfied by the NASA IMS dataset and often violated in real-world industrial deployments with missing, asynchronous, or multi-rate sensor streams [1].

Although our windowing and feature aggregation mitigate minor misalignments, significant temporal gaps or sampling heterogeneity would require explicit handling (e.g., via interpolation or event-based modeling), which is beyond the current scope. Third, and notably, the LOBO evaluation protocol interacts asymmetrically with federated and isolated local learning paradigms. While LOBO ensures realistic cross-asset generalization, it restricts isolated local models to training on only two bearings per fold, severely limiting their exposure to degradation diversity. In contrast, federated LOBO aggregates knowledge from all 12 heterogeneous clients, effectively learning a more universal representation of degradation dynamics through weight averaging. This structural advantage is amplified in small-data, high-heterogeneity regimes, precisely the setting where federated learning was originally motivated [4], [20].

Furthermore, due to data heterogeneity, such as varying degradation patterns, signal characteristics, and fault modes, the results are highly variable and inconsistent, often failing to generalize across test conditions. This underscores the limitations of methods that treat time steps independently and ignore temporal structure and asset-specific dynamics in unsupervised prognostics [11], [22].

Thus, LOBO in a federated setting not only preserves data privacy but also inherently favors federated learning by enabling broader experience sharing across clients, whereas isolated local models are constrained by the very data isolation LOBO enforces. This nuance should be acknowledged when interpreting federated vs. isolated local comparisons under strict LOBO: the protocol, while realistic, introduces an architectural bias that benefits collaborative learning paradigms.

6. Conclusion

We have presented a federated, privacy-preserving framework based on graph neural networks for bearing health monitoring, operating under realistic industrial constraints: weak supervision, severe class imbalance (approximately 7% anomalies), strict data silos, and cross-asset heterogeneity. By integrating adaptive EWMA-based pseudo-labeling with a Temporal Graph Convolutional Network trained under a strict leave-one-bearing-out (LOBO) protocol, our method achieves balanced performance across all 12 NASA IMS bearings (AP = 0.675 ± 0.276, MCC = 0.636 ± 0.285), outperforming both isolated local models and non-graph baselines with statistically significant gains in MCC ($p<0.01$).

Crucially, the LOBO protocol by design requires models to generalize to previously unseen assets. In this setting, **federated learning does not gain an advantage from increased data volume**, but rather from **cross-client parameter communication**, which enables the transfer of degradation-invariant patterns across heterogeneous operating conditions. In contrast, isolated models, despite having access to the same amount of local data per fold, remain confined to asset-specific representations and struggle to capture universal degradation dynamics. This suggests that federated coordination functions as a **representation sharing mechanism** that enhances robustness under heterogeneity complementing the LOBO evaluation's emphasis on real-world generalizability. Together, these properties make federated graph learning particularly well-suited for industrial prognostics, where data cannot be shared, yet reliable early-warning capability is essential.

Collectively, our results support a new principle for industrial AI: Federated Representation Harmonization (FRH), the idea that in non-IID, label-scarce prognostics, federated averaging does not merely preserve privacy but actively harmonizes heterogeneous degradation manifolds into a shared latent space where universal fault precursors emerge. This reframes federated learning from a compliance mechanism into a generalization engine, with profound implications: in real-world PHM, collaboration across data silos is not optional but essential for reliability.

7. Future Work

Several promising directions emerge from this work. First, to improve robustness on non-degrading or slowly degrading assets, we plan to integrate self-supervised contrastive learning [27], which can capture subtle deviation patterns without relying solely on threshold-based pseudo-labels. Second, our current graph construction assumes fully sampled, complete time-series segments. Extending the framework to handle missing data, irregular sampling, or asynchronous sensor streams common in real industrial settings would improve practical applicability.

Third, we aim to scale the approach to multi-component systems (e.g., gearboxes, motors) by modeling cross-component dependencies via heterogeneous temporal graphs, enabling system-level health monitoring within a federated paradigm. Fourth, while our evaluation uses the NASA IMS dataset, validation on larger, multi-site industrial datasets with diverse operating conditions, loads, and bearing types is essential to assess real-world generalizability. Finally, to support edge deployment, we will explore communication-efficient federated strategies, such as gradient sparsification or model pruning, to reduce bandwidth overhead while preserving anomaly detection fidelity.

8. Declarations

8.1. Author Contributions

Khagendra Darlami: Conceptualization, Methodology, Software, Data Curation, Formal analysis, Investigation, Resources, Writing - Original Draft, Supervision, Project administration; **Lalit Awasthi:** Writing - Review & Editing, Software, Visualization, Formal analysis, Investigation, Validation, Resources.

8.2. Institutional Review Board Statement

Not applicable.

8.3. Informed Consent Statement

Not applicable.

8.4. Data Availability Statement

The dataset used in this study is the publicly available as NASA IMS Bearing Dataset, which can be accessed from the NASA Prognostics Data Repository at: https://data.nasa.gov/dataset/ims-bearings# or PHM Society, NASA Prognostics Center of Excellence Data Set Repository [Mirror] https://data.phmsociety.org/nasa/; Direct download: https://phm-datasets.s3.amazonaws.com/NASA/4.+Bearings.zip.

8.5. Acknowledgment

The authors would like to express their deepest gratitude to their families and friends for their unwavering support and encouragement throughout this research. We also gratefully acknowledge the NASA Prognostics Data Repository for making the IMS bearing dataset publicly available, which enabled the experimental validation of this work.

8.6. Conflicts of Interest

The authors declare no conflicts of interest.

9. References

[1] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mechanical Systems and Signal Processing*, vol. 104, pp. 799–834, May 2018, https://doi.org/10.1016/j.ymssp.2017.11.016.

[2] T. Sun, C. Yin, H. Zheng, and Y. Dong, "An unsupervised framework for dynamic health indicator construction and its application in rolling bearing prognostics," *Reliability Engineering & System Safety*, vol. 260, pp. 111039–111039, Mar. 2025, https://doi.org/10.1016/j.ress.2025.111039.

[3] X. Li, C. Cheng, and Z. Peng, "Unsupervised construction of health indicator for rotating machinery via multi-criterion feature selection and attentive variational autoencoder," *Science China Technological Sciences*, vol. 67, no. 5, pp. 1524–1537, Apr. 2024, https://doi.org/10.1007/s11431-023-2610-4.

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May 2020, https://doi.org/10.1109/msp.2020.2975749.

[5] S. Wang, Y. Vidal, and F. Pozo, "An unsupervised approach to early fault detection and performance degradation assessment in bearings," *Advanced Engineering Informatics*, vol. 68, pp. 103620–103620, Jul. 2025, https://doi.org/10.1016/j.aei.2025.103620.

[6] S. Zhang, F. Ye, B. Wang, and T. Habetler, "Few-Shot Bearing Fault Diagnosis Based on Model-Agnostic Meta-Learning," *IEEE Transactions on Industry Applications*, pp. 1–1, 2021, https://doi.org/10.1109/tia.2021.3091958.

[7] A. Deng and B. Hooi, "Graph Neural Network-Based Anomaly Detection in Multivariate Time Series," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, pp. 4027–4035, May 2021, https://doi.org/10.1609/aaai.v35i5.16523.

[8] W. Caesarendra, G. Niu, and B.-S. Yang, "Machine condition prognosis based on sequential Monte Carlo method," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2412–2420, Mar. 2010, https://doi.org/10.1016/j.eswa.2009.07.014.

[9] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213–237, Jan. 2019, https://doi.org/10.1016/j.ymssp.2018.05.050.

[10] An, N. Ho Kim, and J.-H. Choi, "Options for Prognostics Methods: A review of data-driven and physics- based prognostics," *Annual Conference of the PHM Society*, vol. 5, no. 1, Oct. 2013, https://doi.org/10.36001/phmconf.2013.v5i1.2184.

[11] P. Jieyang, A. Kimmig, W. Dongkun, Z. Niu, F. Zhi, W. Jiahai, X. Liu, J. Ovtcharova, "A systematic review of data-driven approaches to fault diagnosis and early warning," *Journal of Intelligent Manufacturing*, Sep. 2022, https://doi.org/10.1007/s10845-022-02020-0.

[12] H. Wang, J. Wang, Y. Zhao, Q. Liu, M. Liu, and W. Shen, "Few-Shot Learning for Fault Diagnosis With a Dual Graph Neural Network," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1559–1568, Feb. 2023, https://doi.org/10.1109/tii.2022.3205373.

[13] Q. Zhou, W. Ma, Y. Zhang, and J. Guo, "Bearing fault diagnosis for variable working conditions via lightweight transformer and homogeneous generalized contrastive learning with inter-class repulsive discriminant," *Engineering Applications of Artificial Intelligence*, vol. 139, pp. 109548–109548, Oct. 2024, https://doi.org/10.1016/j.engappai.2024.109548.

[14] V. Jorry, Z.-S. Duma, Tuomas Sihvonen, Satu-Pia Reinikainen, and Lassi Roininen, "Statistical batch-based bearing fault detection," *Journal of Mathematics in Industry*, vol. 15, no. 1, Feb. 2025, https://doi.org/10.1186/s13362-025-00169-w.

[15] Z. Wang, Z. Xu, C. Cai, X. Wang, J. Xu, K. Shi, X. Zhong, Z. Liao, Q. Li, "Rolling bearing fault diagnosis method using time-frequency information integration and multi-scale TransFusion network," *Knowledge-Based Systems*, vol. 284, p. 111344, Jan. 2024, https://doi.org/10.1016/j.knosys.2023.111344.

[16] A. A. Soomro, M. B. Muhammad, A. A. Mokhtar, M. H. M. Saad, N. Lashari, M. Hussain, U. Sarwar, A. S. Palli, "Insights into modern machine learning approaches for bearing fault classification: A systematic literature review," *Results in Engineering*, vol. 23, p. 102700, Sep. 2024, https://doi.org/10.1016/j.rineng.2024.102700.

[17] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *ArXiv (Cornell University)*, Jan. 2016, https://doi.org/10.48550/arxiv.1609.02907.

[18] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 1–21, 2020, https://doi.org/10.1109/TNNLS.2020.2978386.

[19] M. Wang, J. Yu, H. Leng, X. Du, and Y. Liu, "Bearing fault detection by using graph autoencoder and ensemble learning," *Scientific Reports*, vol. 14, no. 1, p. 5206, Mar. 2024, https://doi.org/10.1038/s41598-024-55620-6.

[20] P. Kairouz and H. B. McMahan, "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1, 2021, https://doi.org/10.1561/2200000083.

[21] W. Song, D. Wu, W. Shen, and B. Boulet, "Early fault detection for rolling bearings: A meta-learning approach," *IET Collaborative Intelligent Manufacturing*, vol. 6, no. 2, May 2024, https://doi.org/10.1049/cim2.12103.

[22] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—A tutorial," *Mechanical Systems and Signal Processing*, vol. 25, no. 2, pp. 485–520, Feb. 2011, https://doi.org/10.1016/j.ymssp.2010.07.017.

[23] J. Ji and H. Dong, "Spatio-Temporal Graph Convolutional Networks for Traffic Prediction Considering Multiple Spatio-Temporal Information," *2024 20th International Conference on Mobility, Sensing and Networking (MSN)*, pp. 730–737, Dec. 2024, https://doi.org/10.1109/msn63567.2024.00103.

[24] C. Feng, C. Liu, and D. Jiang, "Unsupervised anomaly detection using graph neural networks integrated with physical-statistical feature fusion and local-global learning," *Renewable Energy*, vol. 206, pp. 309–323, Apr. 2023, https://doi.org/10.1016/j.renene.2023.02.053.

[25] J. Wang, B. Liang, Z. Zhu, E. Thepie Fapi, and H. Dalal, "Communication-Efficient Network Topology in Decentralized Learning: A Joint Design of Consensus Matrix and Resource Allocation," *IEEE Transactions on Networking*, vol. 33, no. 2, pp. 761–776, Apr. 2025, https://doi.org/10.1109/tnet.2024.3511333.

[26] J. Lee, H. Qiu, G. Yu, J. Lin, and Rexnord Technical Services, "Bearing data set," NASA Ames Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA, USA, 2007.

[27] G. Li, M. Wei, D. Wu, Y. Cheng, J. Wu, and J. Yan, "Zero-Sample fault diagnosis of rolling bearings via fault spectrum knowledge and autonomous contrastive learning," *Expert Systems with Applications*, vol. 275, p. 127080, May 2025, https://doi.org/10.1016/j.eswa.2025.127080.

[28] Yasir Saleem Afridi, L. Hasan, R. Ullah, Z. Ahmad, and J.-M. Kim, "LSTM-Based Condition Monitoring and Fault Prognostics of Rolling Element Bearings Using Raw Vibrational Data," *Machines*, vol. 11, no. 5, pp. 531–531, May 2023, https://doi.org/10.3390/machines11050531.

[29] M. Xu, P. Guan, X. Shi, R. Jiang, J. Tian, J. Geng, G. Xiong, "Research on Bearing Fault Diagnosis Methods Based on Various Convolutional Neural Network Architectures," *IEEE Access*, vol. 13, pp. 44445–44465, 2025, https://doi.org/10.1109/access.2025.3548693.

[30] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," *2008 Eighth IEEE International Conference on Data Mining*, Dec. 2008, https://doi.org/10.1109/icdm.2008.17.