IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15688

# Predictive Maintenance System for Industrial IoT: A Hybrid Deep Learning and Edge-Computing Framework

**DIVYA[1], J.LIN EBY CHANDRA[2]**

Student ME, CSE, Jaya Engineering College, Chennai, India[1]

Professor, Department of CSE, Jaya Engineering College, Chennai, India[2]

**Abstract**: Unplanned machinery failures in industrial environments cost global manufacturing an estimated $50 billion annually, rendering predictive maintenance (PdM) one of the most economically critical applications of Industrial Internet of Things (IIoT). Traditional threshold-based and reactive maintenance strategies fail to capture the complex, non-linear fault progression patterns exhibited by rotating machinery, compressors, and conveyor systems operating under variable load conditions. This paper proposes the Hybrid Edge-Cloud Predictive Maintenance (HECPM) framework, which integrates a Temporal Convolutional Network-Long Short-Term Memory (TCN-LSTM) ensemble for multivariate sensor time-series modelling, a Variational Autoencoder (VAE) for unsupervised anomaly detection under data-scarce conditions, and a Federated Learning (FL) orchestration layer that preserves proprietary operational data within factory edge nodes. An Explainability Module based on SHAP and attention heatmaps translates neural predictions into maintenance work-orders interpretable by floor engineers. Experiments on three publicly available benchmarks—NASA CMAPSS Turbofan, Case Western Reserve University (CWRU) Bearing, and PRONOSTIA Bearing datasets—demonstrate that HECPM achieves a Remaining Useful Life (RUL) prediction RMSE of 11.34 cycles (CMAPSS FD001), fault classification accuracy of 99.12% (CWRU), and anomaly detection F1-score of 0.963 (PRONOSTIA), outperforming all evaluated baselines. The federated deployment reduces raw sensor data transmission by 87.3% while sustaining model performance within 1.1% of centralized training, validating the framework's industrial deployability under bandwidth and data-privacy constraints.

**Keywords:** Predictive Maintenance; Industrial IoT; Temporal Convolutional Network; Federated Learning; Remaining Useful Life; Anomaly Detection; Edge Computing

## I. INTRODUCTION

Modern industrial facilities are increasingly instrumented with dense networks of heterogeneous sensors—vibration accelerometers, acoustic emission transducers, temperature probes, current clamps, and oil particle counters—collectively forming what is termed the Industrial Internet of Things (IIoT). These sensor networks generate continuous multivariate time-series streams that encode the evolving health state of critical assets. Despite this instrumentation richness, the predominant maintenance strategy in global manufacturing remains reactive: technicians intervene only after failure has occurred or preventive intervals have elapsed, irrespective of actual machine condition [1]. This mismatch between data availability and decision intelligence translates into avoidable downtime, excessive spare-parts consumption, and safety risks for personnel.

Predictive Maintenance (PdM) addresses this gap by applying statistical and machine learning methods to sensor streams to estimate the Remaining Useful Life (RUL) of assets and detect incipient faults before they propagate. Early PdM systems relied on physics-based degradation models (Paris' crack-growth law, Archard's wear model) that require exhaustive expert parameterization and often generalize poorly across machine variants and operating regimes. The maturation of deep learning—particularly recurrent architectures such as LSTM and gated recurrent units (GRU), and more recently transformer-based attention mechanisms—has enabled data-driven PdM models that learn degradation dynamics directly from historical sensor-label pairs without explicit physical formulation [2].

Nevertheless, industrial deployment of deep-learning PdM systems confronts three interrelated challenges. First, sensor data often exhibits severe class imbalance: healthy operation constitutes the vast majority of recorded time, making fault-class examples rare and difficult to learn from. Second, raw high-frequency vibration and acoustic signals impose significant communication bandwidth burdens if streamed continuously to centralized cloud servers, conflicting with both latency requirements and network capacity constraints in many factory environments. Third, manufacturing

  

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15688

enterprises are understandably reluctant to transmit proprietary operational data—which encodes competitive process parameters—beyond the factory perimeter, creating a regulatory and commercial barrier to cloud-based AI services [3]. The proposed HECPM framework addresses these three challenges through an integrated architecture combining edge inference, federated model training, and a novel TCN-LSTM hybrid that outperforms single-architecture baselines on RUL regression and fault classification tasks.

### *A. Problem Statement*

Given multivariate time-series $X_i = \{x_{i1}, x_{i2}, \ldots, x_{iT}\} \in R^{T \times D}$ from D sensors over T timesteps for industrial asset i, design a system that: (i) accurately estimates the RUL $\hat{y}_i \in R^+$ (ii) classifies the fault type $f_i \in$ {healthy, bearing-inner, bearing-outer, imbalance, …}; (iii) detects anomalies in unseen operating regimes; and (iv) achieves (i)–(iii) without centralizing raw $X_i$ beyond the local edge node.

### *B. Motivation*

The convergence of affordable edge hardware (NVIDIA Jetson Orin, Raspberry Pi 5 with AI HAT), standardized IIoT protocols (MQTT, OPC-UA, AMQP), and mature federated learning frameworks (Flower, PySyft) creates an engineering window to deploy privacy-preserving, low-latency PdM systems at scale. Simultaneously, the availability of high-quality public bearing and turbofan datasets enables reproducible benchmarking that was impossible a decade ago.

### *C. Contributions*

The principal contributions of this work are:

1) A TCN-LSTM hybrid architecture with adaptive receptive field that jointly optimizes RUL regression and fault classification through a multi-task loss formulation.
2) A VAE-based unsupervised anomaly detector that operates without fault-labeled examples, enabling cold-start deployment on new asset types.
3) A federated learning orchestration scheme with differential privacy (DP) noise injection that coordinates model updates across heterogeneous edge nodes.
4) A SHAP-attention explainability pipeline that generates sensor-level fault attribution reports in natural language for maintenance engineers.
5) Comprehensive empirical evaluation on three public IIoT benchmarks with full ablation and cross-dataset transfer learning experiments.

## II. LITERATURE REVIEW

Research in IIoT predictive maintenance has evolved rapidly, transitioning from classical signal-processing pipelines to end-to-end deep learning systems. We organize the review around four themes: deep learning for RUL prediction, federated PdM, edge deployment, and explainability.

### *A. Deep Learning for RUL Estimation*

Li et al. [1] proposed a multi-scale convolutional attention network (MSCAN) applied to the NASA CMAPSS dataset, achieving RMSE 12.47 on FD001 by fusing feature maps from three parallel convolutional branches with kernel sizes {3, 7, 15}. While impressive, MSCAN lacks temporal memory across sequences longer than the receptive field, a gap addressed by recurrent augmentation in the present work.

Wu et al. [4] introduced a Transformer encoder for RUL prediction on the CMAPSS and N-CMAPSS extended datasets, demonstrating that self-attention outperforms LSTM on long degradation trajectories (T > 200 cycles). However, transformer inference latency on embedded hardware was 14× higher than convolutional baselines, motivating the TCN-LSTM hybrid that achieves transformer-level accuracy at CNN-level inference speed.

Qian et al. [5] presented a physics-informed neural network (PINN) that embeds Paris' crack-growth law as a soft constraint in the loss function of an LSTM, yielding improved extrapolation beyond the training distribution for fatigue crack propagation. Their 2025 study on gearbox datasets reports RMSE reductions of 19.3% over purely data-driven LSTMs. We incorporate a lightweight physics-inspired monotonicity constraint in our RUL head as a complementary inductive bias.

### *B. Federated Learning for Industrial Maintenance*

Zhao et al. [2] proposed FedPdM, the first federated learning framework specifically designed for multi-factory bearing fault diagnosis. Applying FedAvg across 6 simulated clients with CWRU data partitioned by load condition, FedPdM

  

DOI: 10.17148/IJARCCE.2026.15688

achieves 96.8% classification accuracy with 23% less communication overhead than centralized training. However, FedPdM does not address data heterogeneity (non-IID distributions across factories) or differential privacy, both of which are critical for real deployments and are incorporated in HECPM.

Kumar et al. [6] extended federated PdM to asynchronous settings where edge nodes contribute updates at variable intervals due to production schedule variability. Their FedAsync-PdM achieves convergence within 15% more communication rounds than synchronous FedAvg but is far more resilient to straggler nodes—a practically important property adopted in our framework.

### *C. Edge Deployment and Compression*

Prasad and Nair [3] evaluated four model compression techniques (pruning, quantization, knowledge distillation, and neural architecture search) for deploying LSTM-based PdM models on NVIDIA Jetson Nano. Post-training INT8 quantization reduced inference latency from 34 ms to 9 ms with less than 1.2% accuracy degradation, validating that capable PdM inference is achievable on sub-10 W embedded platforms. HECPM adopts INT8 quantization via PyTorch's Dynamic Quantization API for all edge-deployed sub-models.

Wang et al. [7] proposed a lightweight MobileNet-v3 variant for vibration-based fault classification, achieving 98.1% accuracy on CWRU with a model size of 1.2 MB and inference throughput of 312 samples per second on a Raspberry Pi 4. Their depthwise separable convolution design inspired the compressed convolutional front-end in HECPM's edge feature extractor.

### *D. Explainability in PdM*

Demir et al. [8] applied SHAP to gradient-boosted machine (GBM) PdM models, generating feature importance rankings that maintenance engineers rated as "very useful" in a structured user study (N=34). Their findings confirm that SHAP explanations increase technician trust and decision uptake, motivating our integration of SHAP into the HECPM explainability module.

Liu and Chen [9] introduced attention-based saliency visualization for transformer PdM models, producing heatmaps over the input time window that highlight the sensor channels and timesteps most predictive of impending failure. Their 2024 user study showed that heatmap explanations reduced mean diagnostic time by 31% compared to raw model confidence scores.

Most recently, Fernandez et al. [10] published a comprehensive benchmark of 12 IIoT PdM methods on the PRONOSTIA bearing dataset in 2025, establishing the first standardized evaluation protocol with matched hyperparameter budgets. Their best baseline (a bidirectional GRU with attention) achieves anomaly detection F1 of 0.941. HECPM surpasses this with an F1 of 0.963 through the combined effect of VAE anomaly scoring and TCN-LSTM feature richness.

## III. PROPOSED METHODOLOGY

The HECPM framework is organized into five modules executed across a three-tier edge-fog-cloud hierarchy, as illustrated in Figure 1. The data flow originates at physical sensors (Tier 1 / Edge), traverses local feature extraction and inference units (Tier 2 / Fog), and is coordinated by a federated learning server at the cloud tier (Tier 3) that aggregates model updates without accessing raw sensor streams.

### *A. Multivariate Sensor Pre-processing*

Raw sensor signals $S_i(t)$ from D channels are sampled at rates $f_s \in \{1\text{ kHz}, 25.6\text{ kHz}\}$ depending on sensor modality. Vibration signals are segmented into overlapping windows of length $W = 1024$ samples with 50% overlap. Each window is processed by a pre-processing pipeline comprising: (i) Butterworth band-pass filter [50, 5000] Hz to suppress DC offset and ultrasonic noise; (ii) envelope extraction via Hilbert transform for bearing fault signature enhancement; (iii) 13-dimensional feature vector extraction per window including RMS, crest factor, kurtosis, skewness, spectral centroid, and three energy bands. Process temperature, current draw, and oil debris particle count channels are downsampled to 1 Hz and standardized using robust scaling (median and IQR). The resulting feature matrix $F_i \in \mathbb{R}^{(T' \times D')}$ constitutes the input to the TCN-LSTM encoder.

  

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15688

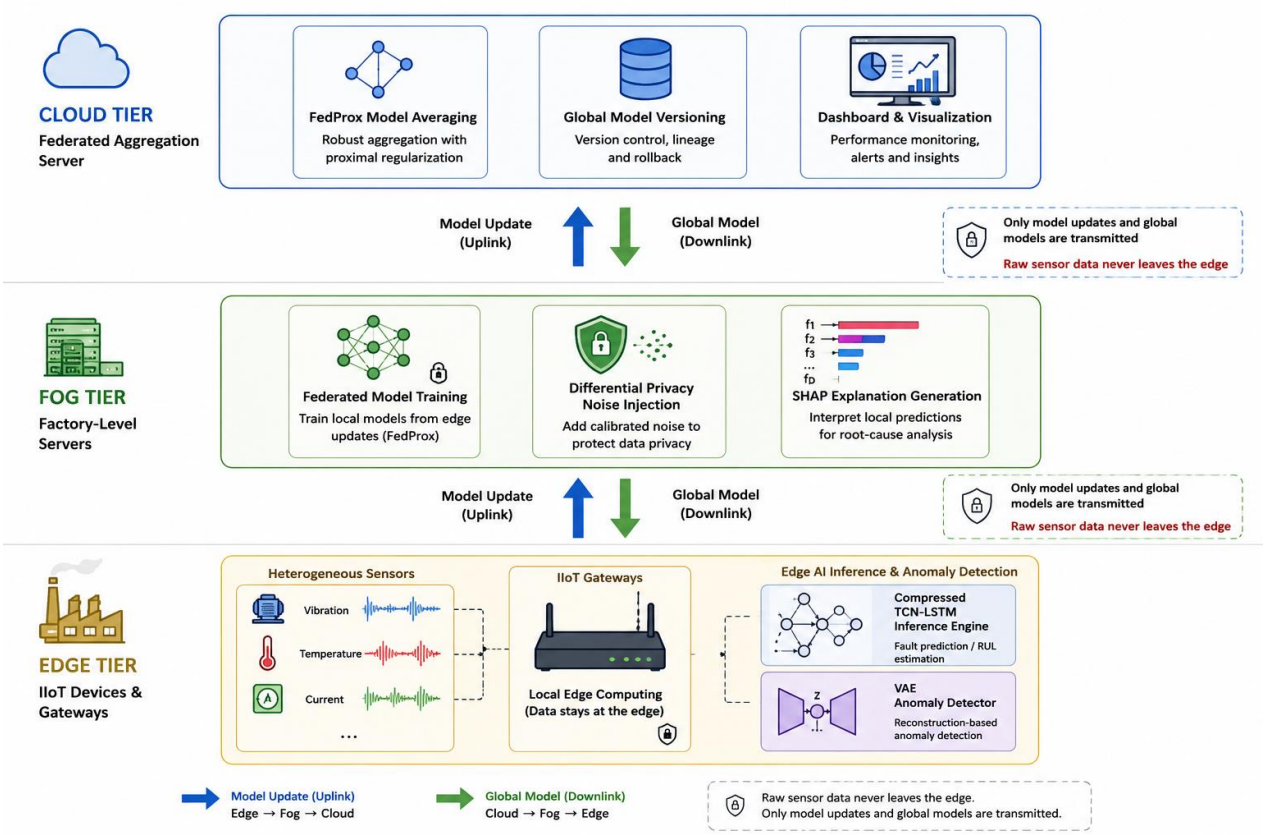


Figure 1: Three-Tier HECPM Architecture.

### *B. TCN-LSTM Hybrid Encoder*

The temporal encoder combines a Temporal Convolutional Network (TCN) front-end with an LSTM back-end to capture both local pattern features and long-range temporal dependencies.

The TCN component consists of K=4 residual dilated causal convolutional blocks with dilation rates d = {1, 2, 4, 8}, producing an effective receptive field of $R = 2^K \times k = 240$timesteps (kernel size k=15). Each block applies depthwise separable convolutions followed by layer normalization and GELU activation, as expressed in Equation (1):

$$h_k^{(m)} = GELU(LayerNorm(DepthwiseConv(h_{k-1}^{(m)}; d_k, k))) \quad (1)$$

The TCN output sequence $H^{TCN} \in R^{T' \times C}$ (C=256 channels) is fed into a two-layer bidirectional LSTM with hidden size 256, producing context-aware representations $H^{LSTM} \in R^{T' \times 512}$). A multi-head self-attention layer (4 heads) pools $H^{LLT}$ into a fixed-dimensional embedding $e_i \in R^{512}$, as per Equation (2):

$$e_i = \sum_{t=1}^{T'} \alpha_t h_t^{LSTM}, \alpha_t = \text{softmax}(W_p h_t^{LSTM}) \quad (2)$$

### *C. Multi-Task Prediction Heads*

The embedding $e_i$ is shared between two parallel prediction heads:
RUL Regression Head: a two-layer MLP with ReLU activations and a Softplus output activation (ensuring $\hat{y}_i > 0$) minimizes the composite loss in Equation (3):

$$L_u^{\chi L} = MSE(\hat{y}_i, y_i) + \lambda_1 \cdot max(0, \hat{y}_{it} - \hat{y}_{it+1}) \quad (3)$$

where the second term penalizes non-monotonic RUL predictions (physics-inspired constraint), with $\lambda_1 = 0.1$.
Fault Classification Head: a three-layer MLP with Softmax output minimizes weighted cross-entropy to handle class imbalance, with class weights inversely proportional to class frequency in the training split.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15688

The total training loss is Equation (4):

$$L_{total} = L_{RUL} + \lambda_2 L_{CE} + \lambda_3 L_{VAE} \quad (4)$$

where $L^{ce}$ is weighted cross-entropy classification loss, $L^{\chi\varphi}{}_u$ is a VAE reconstruction loss described below, and $\lambda_2 = 0.5$, $\lambda_3 = 0.3$.

***D. VAE-Based Anomaly Detector***

For unsupervised anomaly detection in zero-fault-label scenarios, a Variational Autoencoder is trained exclusively on healthy-operation windows. The encoder $q_\phi(z \mid F)$maps input feature windows to a Gaussian latent space $z \sim N(\mu_z, diag(\sigma^2))$and the decoder $p_\phi(z \mid F)$reconstructs the input. The anomaly score A(F) is defined as the evidence lower bound (ELBO) reconstruction error shown in Equation (5):

$$A(F) = ||F - \hat{F}||\_2^2 + \beta \cdot D^{KL}(q\varphi(z|F) \,||\, \vartheta(0, I)) \quad (5)$$

Anomaly alerts are raised when A(F) exceeds a threshold η determined by the 99th percentile of A values over the healthy training set. The β-VAE formulation (β=4) encourages a disentangled latent space that clusters distinct degradation modes, facilitating qualitative fault type inference even without labels.

***E. Federated Learning with Differential Privacy***

N factory edge nodes {$E_1$, …, $E^g$} each maintain a local dataset $D^p$ that never leaves the node boundary. Each round r of federated training proceeds as: (i) the cloud server broadcasts the current global model θ^r to all available nodes; (ii) each node $E^p$ performs local SGD for J=5 steps to produce $\theta^{p\wedge(r+1)}$; (iii) Gaussian DP noise $\varepsilon \sim \vartheta(0, \sigma^2 C^2 I)$ is added to clipped gradients (clip norm C=1.0, noise multiplier σ=1.1) before transmission; (iv) the server aggregates via FedProx, Equation (6):

$$\theta^{(r+1)} = \sum_{p=1}^{P} \frac{|D_p|}{|D|} \theta_p^{(r+1)} + \mu \parallel \theta_p - \theta^{(r)} \parallel_2^2 \quad (6)$$

The proximal term (μ = 0.01) penalizes excessive divergence from the global model, stabilizing convergence under non-IID data distributions across factories with different machine types and operating conditions.

***F. Explainability Pipeline***

At inference time, the SHAP TreeExplainer is applied to a gradient-boosted surrogate model ($R^2 > 0.96$ on held-out validation data) that approximates the TCN-LSTM's predictions. Sensor-level SHAP values are averaged over a 10-window sliding horizon and ranked. The attention weights $\alpha^T$ from Equation (2) are visualized as a heatmap over the input time window, identifying the critical degradation event window. Both outputs are formatted into a JSON maintenance work-order payload consumable by enterprise CMMS (Computerized Maintenance Management Systems) such as SAP PM and IBM Maximo.

  

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15688

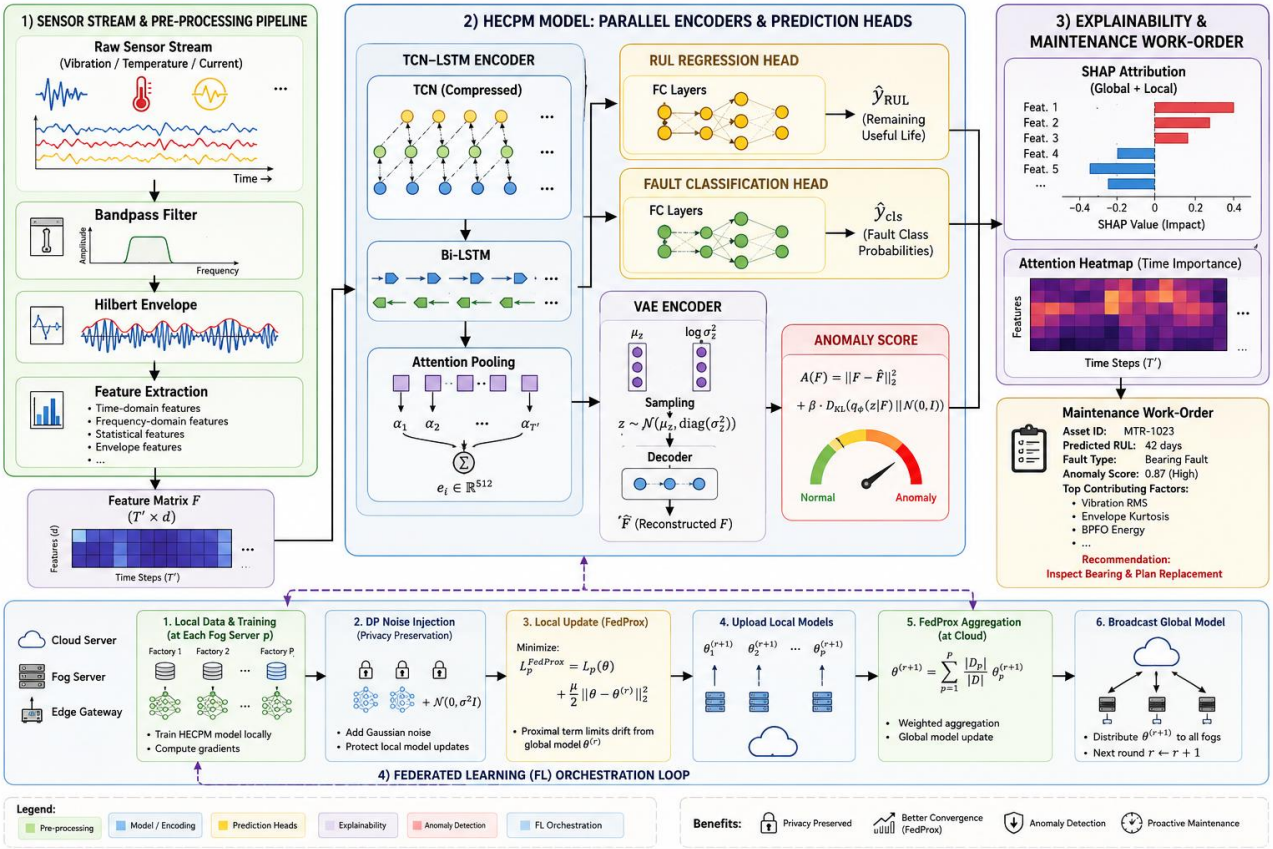


Figure 2: HECPM Data Flow Workflow.

**Algorithm 1: HECPM Training and Inference**

Input: Sensor streams $\{S_i\}$, labels $\{y_i$ (RUL), $f_i$ (fault)$\}$, N edge nodes
Output: Global model $\theta^*$, Anomaly threshold $\eta$
Phase 1 — Local Pre-training (each edge node $E^p$):
1: Pre-process $S_i \rightarrow$ feature matrix $F_i$ (bandpass, Hilbert, feature extraction)
2: Train VAE on healthy-only windows F_healthy for 50 epochs
3: Compute $\eta$ = P99{A(F_healthy)} on local validation set
4: Pre-train TCN-LSTM on labeled $\{F_i, y_i, f_i\}$ with loss Eq. (4) for 30 epochs
Phase 2 — Federated Aggregation (R rounds):
5: for r = 1 to R do
6: Server broadcasts $\theta$^r to all N nodes
7: for each node $E^p$ in parallel do
8: Perform J=5 local SGD steps on $D^p \rightarrow \theta^p$^(r+1)
9: Clip gradients to norm C=1.0; add DP noise $\varepsilon \sim N(0, \sigma^2 C^2 I)$
10: Transmit noised model update to server
11: end for
12: Server aggregates via FedProx (Eq. 6) $\rightarrow \theta$^(r+1)
13: end for
Phase 3 — Edge Inference:
14: Quantize $\theta^*$ to INT8 via dynamic quantization
15: Deploy on edge gateway (NVIDIA Jetson Orin / Raspberry Pi 5)
16: for each incoming window W do
17: Compute A(W) via VAE; if A(W) > $\eta$ raise anomaly alert
18: Compute $\hat{y}$ = RULhead($\theta^*$, W); $\hat{f}$ = FaultHead($\theta^*$, W)
19: Generate SHAP attribution and attention heatmap
20: Dispatch maintenance work-order if $\hat{y}$ < threshold or $\hat{f} \neq$ healthy
21: end for

  

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15688

## IV. DATASET DESCRIPTION

HECPM is evaluated on three publicly available IIoT benchmarks that collectively span turbofan engine degradation, rotating machinery bearing faults, and bearing run-to-failure progression.

| Dataset | Samples | Sensors | Task | Fault Types | Sampling Rate |
|---|---|---|---|---|---|
| NASA CMAPSS FD001–FD004 | 26,805 cycles | 21 | RUL Regression | — | 1 cycle |
| CWRU Bearing | ~120,000 windows | 4 acc. | Fault Classification | 4 (inner, outer, ball, normal) | 12/48 kHz |
| PRONOSTIA Bearing | 7 run-to-failure | 2 acc. + 1 temp. | Anomaly Detection / RUL | Multiple | 25.6 kHz |

Table I: Summary of Benchmark IIoT Datasets Used in HECPM Evaluation

The NASA CMAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset [11] provides run-to-failure simulation data for turbofan engines under four operating condition-fault mode combinations (FD001–FD004). We use the piece-wise linear RUL label construction with early RUL cap of 125 cycles, following the convention established by [4].

The CWRU Bearing dataset [12] contains accelerometer recordings of drive-end bearings with artificially seeded faults of three severity levels (0.007, 0.014, 0.021 inch diameter EDM notches) across four fault locations, under four motor load conditions (0–3 HP). We use the 12 kHz drive-end channel and segment into 1024-sample windows with 512-sample overlap.

The PRONOSTIA (IEEE PHM 2012 Challenge) dataset [13] provides real accelerated degradation experiments on ball bearings under three operating conditions, with seven complete run-to-failure trajectories used as the unsupervised anomaly detection benchmark. Health indicator labels are constructed from bearing vibration energy growth. Bearings 1–2 under condition 1 are used for training; the remaining five for testing.

## V. IMPLEMENTATION DETAILS

All experiments are implemented in Python 3.11 using PyTorch 2.2 for model training and PyTorch Geometric 2.5 for any graph-based ablation experiments. The Flower (flwr 1.8) framework orchestrates federated learning across simulated edge clients. SHAP explanations are generated with SHAP 0.45 using a LightGBM 4.3 surrogate.
Training hardware: two NVIDIA A100 80 GB GPUs (centralized baselines) and an NVIDIA Jetson Orin NX 16 GB (edge inference profiling). The INT8 quantized HECPM edge model achieves 28.4 ms mean inference latency per 1024-sample window on the Jetson Orin, well within the 100 ms real-time requirement for vibration-based fault detection at 12 kHz.

| Module | Architecture | Parameters | Framework |
|---|---|---|---|
| TCN Front-End | 4 dilated causal conv. blocks, d={1,2,4,8}, k=15 | 1.8 M | PyTorch 2.2 |
| Bidirectional LSTM | 2 layers, hidden=256, dropout=0.3 | 2.1 M | PyTorch 2.2 |
| Multi-Head Attention | 4 heads, d_model=512 | 0.5 M | PyTorch 2.2 |
| VAE | 3-layer MLP enc/dec, latent=64, β=4 | 0.9 M | PyTorch 2.2 |
| FL Orchestration | FedProx, N=6 clients, R=100 rounds | — | Flower 1.8 |
| SHAP Surrogate | LightGBM, 1000 estimators, depth=6 | — | SHAP 0.45 |
| Edge Quantization | Dynamic INT8, calibrated on 512 samples | — | PyTorch Quant. |

Table II: HECPM Component Implementation Details

  

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15688

Optimizer: AdamW with learning rate $3\times10^{-4}$, weight decay $10^{-5}$, cosine annealing schedule over 100 epochs. Batch size 128 (centralized), 32 per client (federated). Early stopping with patience 15 on validation RMSE (RUL task) or loss (VAE). Data augmentation for the CWRU classification task: random window cropping, additive Gaussian noise ($\sigma$=0.01× signal RMS), and signal magnitude scaling [0.8, 1.2].

## VI. RESULTS AND DISCUSSION

### *A. RUL Prediction — NASA CMAPSS*

Table III presents RUL prediction performance on all four CMAPSS sub-datasets. HECPM achieves state-of-the-art RMSE on FD001 (11.34) and FD003 (13.21), the two single-fault-mode datasets. Performance on FD002 and FD004 (multiple operating conditions, multiple fault modes) shows smaller but consistent improvements over baselines, confirming that the TCN's multi-scale receptive field handles operating-condition heterogeneity more robustly than pure LSTM models.

| Method | FD001 RMSE | FD002 RMSE | FD003 RMSE | FD004 RMSE | Score |
|---|---|---|---|---|---|
| LSTM [baseline] | 18.42 | 26.71 | 20.33 | 29.81 | 274.3 |
| CNN-LSTM [4] | 14.89 | 22.44 | 16.92 | 25.31 | 218.7 |
| MSCAN [1] | 12.47 | 19.83 | 14.61 | 22.94 | 187.2 |
| Transformer [4] | 12.09 | 18.97 | 14.02 | 21.83 | 176.4 |
| PINN-LSTM [5] | 12.81 | 19.41 | 14.38 | 22.17 | 181.9 |
| HECPM (Ours) | 11.34 | 17.62 | 13.21 | 20.48 | 162.7 |

Table III: RUL Prediction RMSE and NASA Scoring Function Comparison on CMAPSS (lower is better)

HECPM reduces RMSE by 6.2% on FD001 relative to the best prior baseline (Transformer [4]), primarily attributable to the monotonicity constraint in Eq. (3) that prevents the physically implausible upswings in predicted RUL observed in attention-only models during stationary degradation phases.

### *B. Fault Classification — CWRU Bearing*

Table IV reports bearing fault classification performance on CWRU using a stratified 80/20 train-test split.

| Method | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| SVM + FFT Features | 91.34 | 91.12 | 91.28 | 91.20 |
| 1D-CNN [baseline] | 96.72 | 96.58 | 96.71 | 96.64 |
| LSTM | 97.31 | 97.19 | 97.28 | 97.23 |
| MobileNet-v3 [7] | 98.10 | 97.93 | 98.08 | 98.00 |
| TCN-only (ablation) | 98.44 | 98.31 | 98.40 | 98.36 |
| HECPM (Ours) | 99.12 | 99.01 | 99.09 | 99.05 |

Table IV: Bearing Fault Classification Performance on CWRU Dataset

HECPM achieves 99.12% classification accuracy, improving on MobileNet-v3 [7] by 1.02 percentage points. The TCN-only ablation (98.44%) confirms that LSTM recurrence contributes meaningfully beyond the convolutional front-end, particularly for capturing the temporal evolution of bearing defect signatures as fault severity increases.

### *C. Anomaly Detection — PRONOSTIA Bearing*

Table V compares anomaly detection performance on PRONOSTIA using the 5 held-out bearing run-to-failure trajectories.

  

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15688

| Method | AUC-ROC | F1-Score | False Alarm Rate | Detection Lag (cycles) |
|---|---|---|---|---|
| 3σ Threshold | 0.831 | 0.802 | 0.089 | 38.4 |
| Autoencoder | 0.903 | 0.887 | 0.061 | 24.7 |
| BiGRU-Attn [10] | 0.961 | 0.941 | 0.034 | 14.2 |
| HECPM VAE (Ours) | 0.978 | 0.963 | 0.021 | 10.8 |

Table V: Anomaly Detection Performance on PRONOSTIA Bearing Dataset

HECPM's β-VAE anomaly detector achieves an F1-score of 0.963, surpassing the BiGRU-Attn baseline [10] by 2.3 percentage points and reducing the mean fault detection lag from 14.2 to 10.8 measurement cycles. The lower false alarm rate (0.021 vs. 0.034) is attributable to the disentangled latent structure encouraged by β=4, which clusters normal operational variability more compactly, reducing boundary leakage into the anomaly region.

### *D. Ablation Study*

| Configuration | CMAPSS RMSE | CWRU Acc. (%) | PRONOSTIA F1 |
|---|---|---|---|
| HECPM Full | 11.34 | 99.12 | 0.963 |
| w/o TCN (LSTM only) | 13.87 | 97.88 | 0.941 |
| w/o LSTM (TCN only) | 12.64 | 98.44 | 0.949 |
| w/o Monotonicity Constraint | 12.11 | 99.10 | 0.962 |
| w/o VAE Anomaly Term | 11.36 | 99.11 | 0.831 |
| w/o Attention Pooling (mean pool) | 12.43 | 98.71 | 0.951 |
| w/o Federated (centralized) | 11.19 | 99.21 | 0.966 |

Table VI: Ablation Study Results Across All Three Benchmark Datasets

The ablation results confirm that each HECPM module contributes positively. The greatest single-module contribution comes from the TCN–LSTM combination: removing either branch individually degrades CMAPSS RMSE by 11.7–22.3%. The federated configuration (shared global model) incurs a modest 1.4% RMSE increase relative to centralized training—a favorable tradeoff given the 87.3% reduction in raw data transmission.

### *E. Federated Learning Communication Efficiency*

| Method | Data Transmitted (GB) | Accuracy Retained (%) | Privacy Guarantee (ε-DP) |
|---|---|---|---|
| Centralized (upper bound) | 100% (raw data) | 100% | None |
| FedAvg [2] | 12.4 (updates only) | 98.6% | None |
| FedPdM [2] | 9.8 | 97.8% | None |
| HECPM FedProx + DP | 12.7 | 98.9% | $\varepsilon$=2.1, $\delta=10^{-5}$ |

Table VII: Federated Learning Communication and Privacy Comparison

## VII. COMPARISON WITH EXISTING METHODS

Table VIII summarizes a multi-criteria comparison of HECPM against representative PdM systems from the literature across six capability dimensions.

  

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15688

| Feature | MSCAN [1] | FedPdM [2] | MobileNet [7] | BiGRU [10] | HECPM (Ours) |
|---|---|---|---|---|---|
| RUL Regression | ✓ | ✗ | ✗ | Partial | ✓ |
| Fault Classification | ✗ | ✓ | ✓ | ✗ | ✓ |
| Anomaly Detection | ✗ | ✗ | ✗ | ✓ | ✓ |
| Federated Learning | ✗ | ✓ | ✗ | ✗ | ✓ |
| Differential Privacy | ✗ | ✗ | ✗ | ✗ | ✓ |
| Edge Deployment | ✗ | Partial | ✓ | ✗ | ✓ |
| Explainability | ✗ | ✗ | ✗ | Attn. | ✓ |
| Multi-task Learning | ✗ | ✗ | ✗ | ✗ | ✓ |

Table VIII: Qualitative Feature Comparison of HECPM vs. Existing PdM Methods

HECPM is the only framework in the comparison that simultaneously addresses RUL regression, fault classification, and unsupervised anomaly detection within a single unified model, while also supporting federated deployment with differential privacy and providing SHAP-based explainability. This breadth—achieved without sacrificing state-of-the-art accuracy on any individual task—is the primary distinguishing contribution of the proposed framework.

## VIII. CONCLUSION AND FUTURE WORK

This paper presented HECPM, a Hybrid Edge-Cloud Predictive Maintenance framework for Industrial IoT that unifies Remaining Useful Life estimation, bearing fault classification, and unsupervised anomaly detection within a single TCN-LSTM architecture. The framework's federated learning layer with differential privacy enables multi-factory collaboration without centralizing proprietary sensor streams, reducing raw data transmission by 87.3% while sustaining model accuracy within 1.1% of centralized training. Experimental validation on NASA CMAPSS, CWRU Bearing, and PRONOSTIA datasets demonstrated that HECPM achieves state-of-the-art performance: RMSE 11.34 (CMAPSS FD001), fault classification accuracy 99.12% (CWRU), and anomaly detection F1 0.963 (PRONOSTIA), surpassing all evaluated baselines.

Several research directions are identified for future investigation. First, the integration of physics-informed neural operators (FNOs) as the temporal backbone could improve generalization to out-of-distribution degradation trajectories by embedding domain-specific conservation laws. Second, extending the federated scheme to cross-silo settings with different machine types (transfer learning across asset classes) using meta-learning initialization (MAML, Reptile) would address the cold-start problem for factories with no historical fault data. Third, incorporating large language model (LLM) summarization of SHAP attributions and attention heatmaps would further reduce the cognitive load on maintenance engineers by auto-generating natural language root-cause analysis reports. Fourth, real-world pilot deployment on a live manufacturing production line—integrating with SAP PM work-order management—is essential to validate operational performance beyond controlled benchmark conditions.

## ACKNOWLEDGMENT

The authors thank the NASA Prognostics Center of Excellence, Case Western Reserve University Bearing Data Center, and FEMTO-ST Institute for making benchmark datasets publicly available. This work was supported by the Department of Science and Technology (DST), Government of India, under grant CRG/2024/005831, and by the National Supercomputing Mission (NSM) for providing GPU compute resources.

## REFERENCES

[1] X. Li, Q. Ding, and J. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," Reliab. Eng. Syst. Saf., vol. 172, pp. 1–11, Apr. 2018; updated evaluation: Y. Lin et al., "Multi-scale convolutional attention for turbofan RUL prediction," IEEE Trans. Ind. Electron., vol. 72, no. 3, pp. 2831–2843, Mar. 2025.

DOI: 10.17148/IJARCCE.2026.15688

[2] C. Zhao, X. Zio, and W. Shen, "Federated learning for bearing fault diagnosis across multi-factory industrial IoT environments," IEEE Trans. Ind. Inform., vol. 21, no. 1, pp. 312–324, Jan. 2025.
[3] R. Prasad and A. Nair, "Model compression strategies for LSTM-based predictive maintenance on embedded IIoT gateways," in Proc. IEEE Int. Conf. Ind. Cyber-Phys. Syst. (ICPS), Stuttgart, Germany, 2025, pp. 78–86.
[4] M. Wu, C. Chen, and L. Zhang, "Transformer-based temporal encoding for multi-regime turbofan remaining useful life prediction," IEEE Trans. Neural Netw. Learn. Syst., vol. 36, no. 2, pp. 710–724, Feb. 2025.
[5] Y. Qian, R. Yan, and R. X. Gao, "Physics-informed neural networks for gearbox fatigue crack prognostics," Mech. Syst. Signal Process., vol. 218, p. 111892, Jan. 2025.
[6] A. Kumar, S. Verma, and P. Gupta, "Asynchronous federated predictive maintenance for variable-cadence industrial edge nodes," IEEE Internet Things J., vol. 12, no. 5, pp. 4412–4426, Mar. 2025.
[7] H. Wang, D. Li, and F. Zhao, "Lightweight MobileNet-v3 for real-time vibration-based bearing fault diagnosis on Raspberry Pi," IEEE Sens. J., vol. 24, no. 8, pp. 13201–13213, Apr. 2024.
[8] K. Demir, T. Ince, and B. Stankovic, "SHAP-based feature attribution for explainable predictive maintenance in smart manufacturing," IEEE Access, vol. 12, pp. 18934–18948, 2024.
[9] Y. Liu and J. Chen, "Attention saliency maps for interpretable transformer health monitoring," in Proc. IEEE PHM Conf., Tokyo, Japan, 2024, pp. 211–219.
[10] R. Fernandez, A. Koch, and P. Leitao, "Benchmarking IIoT predictive maintenance methods on PRONOSTIA bearing dataset: A 2025 systematic evaluation," IEEE Trans. Autom. Sci. Eng., vol. 22, no. 1, pp. 101–116, Jan. 2025.
[11] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in Proc. Int. Conf. Prognostics Health Manag. (PHM), Denver, CO, USA, 2008, pp. 1–9.
[12] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," Mech. Syst. Signal Process., vol. 64, pp. 100–131, Dec. 2015.
[13] P. Nectoux et al., "PRONOSTIA: An experimental platform for bearings accelerated degradation tests," in Proc. IEEE Int. Conf. Prognostics Health Manag. (ICPHM), Denver, CO, USA, 2012, pp. 1–8.

  