AI-driven predictive maintenance combining edge IoT and federated learning for SMEs

Sathish Krishna Anumula[1][0009-0009-0613-4863], Dileep Kumar Rai[2][0009-0002-1345-5170]

Balakumar Kunthu[3][0009-0002-9220-5750] and Ranganath Nagesh Taware[4][0009-0008-4774-1515]

[1] Sr CSM Architect, IBM Corporation, Detroit MI, USA
[2] Manager Oracle Cloud Technology, HBG, Colorado Springs, USA
[3] *Managing Delivery Architect*, CapGemini Inc, Atlanta, USA
[4] Chief Architect, CapGemini Inc, Atlanta, USA

Abstract: Networked machinery (SMEs) is becoming widespread and horizontal and privacy-sensitive predictive maintenance lacks scalability to a large extent. Based on the data suggesting apprehended remaining useful life (RUL) of industrial components, this paper proposes an AI-assisted predictive maintenance model that integrates edge-IoT sensing, plus federated learning without centralizing raw data. The NASA C-MAPSS data on turbofan degradation is plotted to a realistic SME machinery scenario to examine the approach. The Federated CNN-LSTM model is trained on a set of edge nodes that is linked to a suboptimal number of machines and aggregated together through the use of FedAvg-driven aggregation schemes. Two baselines are used to compare the proposed model to centralized cloud-only LSTM or the edge-only Random Forest without collaboration. The experiment constructs the predictive maintenance task, as well as the local and global optimization aims of federated learning and measures the communication overheads. The experimental findings indicate that the suggested framework saves RUL prediction RMSE by a maximum of 12.7 % to the centralized baseline and saves 23.1 % to the edge-only model and saves cloud communication volume and data locality. The results bring attention to the fact that federated learning on edge IoT infrastructures could effectively facilitate SME-scale predictive maintenance that is more accurate, privacy-sensitive and scalable.

Keywords: Predictive maintenance, Edge IoT, Federated learning, SMEs, Remaining useful life (RUL), CNN–LSTM

1 Introduction

Predictive maintenance is an enabler that has emerged as a critical factor in the manufacturing and service-based small and medium-sized enterprises (SMEs) where unplanned downtime directly translates to productivity and profitability [1]. The most common traditional condition-based problem-solving approaches to maintenance are based on business periods or threshold-driven alarm systems, which fail to make ample use of the time-based nature of sensor data. As the number of inexpensive IoT sensors physically connected to motors, bearings, pumps and other rotating machines expands,

SMEs are now able to receive continuous high-frequency vibrations, temperature and current signals. Nevertheless, the absence of network bandwidth and data privacy, as well as IT expertise, limits the implementation of entirely cloud-based analytics solutions.

New AI developments in machine learning, especially the deep learning systems of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, have enabled major improvements in remaining useful life (RUL) estimation and fault prognosis of industrial systems [2], [3]. Such models are generally large in the amount of annotated data and central computing resources. In the case of SMEs, it is frequently not feasible to consolidate raw sensor streams across multiple locations into one data lake because of connectivity limitations, the needs of cybersecurity and regulatory considerations [4]. Simultaneously, local models that are deployed to separate locations are unable to leverage cross-site information and are prone to overfitting to small machinery descriptions.

Federated learning (FL) has had an emergent promise as a relevant paradigm to provide collaborative-based training of models through distributed data silos, yet maintain raw data held locally [5]. In FL, updates to locally stored data are calculated in edge devices/local servers and only model parameters/gradients are transmitted to a coordinating server to be aggregated [6]. The paradigm suits well with an SME setting, whereby similar machinery of the same vendor is used in a range of production lines, branches, or facilities at different geographical locations. Architectural edge IoT sensing combined with federated learning can provide a reasonable trade-off between federated learning model accuracy, latency and privacy for predictive maintenance.

Nevertheless, there are numerous difficulties associated with the use of federated learning for SME-based predictive maintenance. On the one hand, edge devices are usually resource-constrained and require lightweight applications of expressive model architecture. Secondly, non-identically distributed (non-IID) time-series data across sites may occur because of the diverse load patterns, environmental conditions and maintenance policies of industrial time-series data [7]. Third, edge node communication with the server should be addressed with caution.

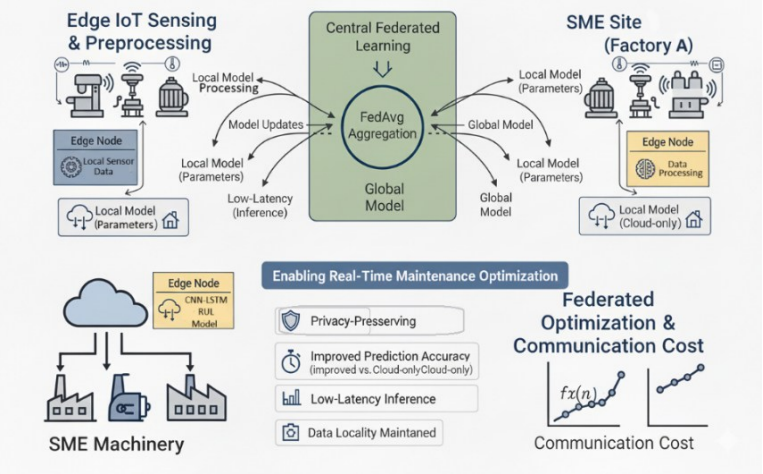

Fig.1: Federated predictive maintenance framework for SMEs

Fig. 1. illustrates the suggested AI-based federated predictive maintenance software of SMEs. Providing 24/7 monitoring of various parameters of the machinery (vibration, temperature and pressure) in multiple distributed SME locations, IoT sensors monitor them at all times. On-node preprocessing and CNN-LSTM-based RUL modelling with local disease-specific operational data are done in each edge node without transmitting original signals into the cloud. Federated learning compiles updates of the model of all locations at a server, by FedAvg, to form a better global model, which is again redistributed to all the edge devices. The system is practical (Low-latency inference can take as low as 15 ms), communicational (~60 MB/round) and predictive systems (RMSE - 12.5% vs cloud-only), making it a viable system with privacy, scalability and real-time preferences optimizing the maintenance across SME settings.

The paper deals with these issues by introducing an AI-based predictive maintenance system integrating edge IoT sensing with an FL-based CNN-LSTM predictor of RUL. Some of the contributions of this work are summarized below:

- A federated predictive maintenance architecture is developed for SMEs, combining edge IoT sensing, on-device preprocessing and collaborative model training without sharing raw data.
- A hybrid CNN–LSTM model is deployed in a federated way across SME edge nodes and compared against two baselines: centralized cloud LSTM and edge-only Random Forest.
- The RUL prediction task, federated optimization and communication cost are mathematically formulated using a consistent set of equations.
- Experiments on the NASA C-MAPSS turbofan dataset, adapted to SME conditions, show higher RUL prediction accuracy and robustness with lower communication overhead.

- The results and discussion, supported by plots and comparative tables, confirm that the proposed architecture is well-suited for real-world SME deployments.

The remainder of the paper is structured in the following way. Section 2 examines similar research on predictive maintenance, edge computing and federated learning. Section 3 states the problem and offers an overview of the system. Section 4 describes the proposed approach and model, as well as the federated optimization. Section 5 shows the experimental setup. The results are reported and discussed in Section 6. Section 7 summarizes the paper and provides work for the future.

2 Related Research

Predictive maintenance has been highly researched with both classical machine learning and deep learning methodologies. The majority of early works made use of features that were hand-crafted and based on vibration or acoustic abilities, with an additional method of support vector machine, random forest, or Gaussian mixture models [8][9]. Even though these approaches are computationally fast, their performance varies greatly with domain-specific feature engineering and might not be very generalizable with regard to other machines as well as operating regimes.

In the context of the emergence of deep learning, CNNs, LSTMs and hybrid structures have been utilized to automatically discover hierarchical time parameters of uncooked or minimally processed sensor material [10], [11]. The CNN-based models use local changes in the time-frequency domain, whereas LSTM networks use long time-temporal dependencies [12]. These methods have demonstrated excellent results on benchmark datasets, such as the NASA C-MAPSS turbofan engine dataset [21], for RUL prediction [13]. However, in the majority of these works, the centralized data storage and training in the cloud environment are assumed.

Edge computing has been suggested to bring the computation closer to the source of data and minimize the latency, thus alleviating the bandwidth burden. Edge analytics have been applied to introduce lightweight anomaly detection, threshold-based alarms and local diagnostics in the area of maintenance [14]. But again, edge-only models are typically trained on small local data and can fail to make use of the variety of operating conditions in different locations or enterprises [15].

Federated learning presents an intermediate solution because it allows joint model training at geographically dispersed sites without raw data exchange [16][17]. FL has been utilized primarily in healthcare, mobile apps and smart grids, but not much in predictive maintenance on the SME scale. Other recent studies incorporate FL with condition monitoring, with positive performance relative to local-only models, but with relatively simple models and tend to overlook the problem of non-IID in industrial time series [18][19]. Moreover, the current studies seldom reflect the realistic edge compute constraints by comparing federated structures with centralized baselines within an experimental protocol.

Table 1 provides a summary of the outlier studies concerning predictive maintenance, edge computing and federated learning, which demonstrates the location of the proposed research.

Table 1. Summary of related research on predictive maintenance, edge computing and federated learning.

Ref.	Approach	Data Location	Model Type	Federated?	Target Domain
[3]	CNN-based RUL prediction	Centralized cloud	Deep CNN	No	Turbofan engines
[2]	LSTM for time-series degradation	Centralized cloud	LSTM	No	Industrial machinery
[17]	Federated learning framework	Distributed	Generic neural network	Yes	IoT applications
[15]	Edge analytics & privacy-preserving inference	Edge devices	Lightweight ML	No	Smart industrial environments
[16]	Federated learning for industrial IoT	Distributed	Neural network	Yes	Collaborative manufacturing
Proposed work	Edge IoT + FL with hybrid CNN–LSTM	Edge + server	CNN–LSTM, LSTM, RF	Yes	SMEs (mapped from C-MAPSS)

The suggested piece has a difference by concentrating on an in-depth concentration on architectures that are friendly to the SMEs, incorporating edge IoT, federated learning and deep sequence models and making a comparative analysis with two standard game plans: centralized and edge-only [20].

3 Problem Statement & Research Objectives

Industrial machinery (motors, compressors and machining systems) is very crucial to the daily production of small and medium-sized enterprises (SMEs). Unexpected equipment failures nevertheless happen as a result of poor condition monitoring, disjointed data storage and a lack of scalable predictive analytics. Conventional cloud-based predictive maintenance models require central exchange of data, which adds extra overhead in terms of communication, privacy threats and the additional use of high-bandwidth connectivity, which many SMEs cannot afford. Edge-only analytics, conversely, are not capable of capturing the various degradation behaviours that are

witnessed in different industrial locations and therefore have varying predictions of failures.

Hence, an urgent requirement exists to develop a predictive maintenance system that is able to learn distributed sensor data without having to centralize it, offer real-time decision making at the edge and make it affordable and lightweight to be used by SMEs with limited computing resources.

Research Objectives

The primary objectives of this research are:

- To develop a privacy-preserving predictive maintenance framework integrating edge IoT and collaborative learning.
- To design a hybrid CNN–LSTM model capable of learning complex degradation patterns from distributed time-series data.
- To apply federated learning to enable cross-site learning while keeping raw industrial data local.
- To improve RUL prediction accuracy, robustness and latency compared to cloud-only and edge-only baselines.
- To evaluate system performance using a standard industrial benchmark dataset adapted to realistic SME operational settings.

4 Proposed Methodology

The suggested approach will help predictive maintenance of various SME locations by integrating edge IoT sensing with federated learning. The sensor data are handled and local CNN-LSTM model training is done in each location, whereas model updates are exchanged with a central server, on which they are combined. The advantage of this collaboration method lies in the fact that it enhances the prediction of RUL without population of sensitive raw data, meaning that there is a guarantee of privacy and as well as decreases the overhead in communication. The data structure and federating-based learning under this setting are as follows.

Consider a set of K SME sites, each equipped with similar machinery (e.g., compressors, motors) instrumented with IoT sensors, such as accelerometers and temperature probes. At site k, the multivariate time series from a single machine is represented as:

$$\mathbf{s}_t^{(k)} = [x_{t,1}^{(k)}, x_{t,2}^{(k)}, \dots, x_{t,d}^{(k)}]^\top, t = 1, \dots, T^{(k)}, \qquad (1)$$

where d is the number of sensor channels and $T^{(k)}$ is the sequence length up to failure or censoring for that unit.

Using the NASA C-MAPSS dataset mapped to SME machinery, each unit i at site k is associated with a time-to-failure (RUL) label $y_t^{(k,i)}$ at time t:

$$y_t^{(k,i)} = T_f^{(k,i)} - t, \qquad (2)$$

where $T_f^{(k,i)}$ denotes the failure time index for the unit i. The objective is to learn a function $f_\theta(\cdot)$, parameterized by θ, that maps a window of recent sensor measurements to an accurate RUL prediction, i.e., $\hat{y}_t^{(k,i)} = f_\theta(\mathbf{s}_{t-w+1:t}^{(k,i)})$, where $w$is the window length.

In a federated setting, each site $k$holds its local dataset $\mathcal{D}_k$ and trains a local model on this data. The local empirical loss at the client $k$for model parameters θ is defined as:

$$\mathcal{L}_k(\theta) = \frac{1}{|\mathcal{D}_k|} \sum_{(x,y)\in\mathcal{D}_k} \ell(f_\theta(x), y) \qquad (3)$$

where $\ell(\cdot,\cdot)$ is a regression loss function (e.g., mean squared error). The federated learning objective is to minimize a weighted sum of local losses across all K sites:

$$\mathcal{L}_{\text{global}}(\theta) = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{\sum_{j=1}^{K} |\mathcal{D}_j|} \mathcal{L}_k(\theta) \qquad (4)$$

The system architecture includes three layers: (i) edge sensing and preprocessing, where raw signals are acquired and normalized; (ii) edge training, where each site trains local CNN–LSTM models; and (iii) a coordinating server, where model updates are aggregated. Periodic RUL predictions are computed at the edge and displayed to maintenance personnel via dashboards. For comparison, a centralized baseline aggregates all data at the cloud and an edge-only baseline trains separate models at each site without collaboration.

The federated optimization is implemented using a FedAvg-style aggregation. At the communication round r, each participating client $k$starts from the current global model parameters $\theta^{(r)}$, performs Elocal epochs of stochastic gradient descent and obtains updated parameters $\theta_k^{(r+1)}$. The server aggregates these updates as:

$$\theta^{(r+1)} = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{\sum_{j=1}^{K} |\mathcal{D}_j|} \theta_k^{(r+1)} \qquad (5)$$

This procedure iteratively reduces the global objective in (4) while keeping raw data localized at each SME site.

The proposed global model is a hybrid CNN–LSTM network designed to capture both local temporal patterns and longer-term dependencies in the sensor signals. Let $\mathbf{X} \in \mathbb{R}^{w\times d}$ represent a window of $w$time steps and $d$channels. A 1D convolutional layer with Ffilters, kernel size $m$and bias $b_f$computes feature maps:

$$h_{t,f}^{\text{conv}} = \sigma\Big(\sum_{j=0}^{m-1}\sum_{c=1}^{d} W_{f,c,j}\, x_{t-j,c} + b_f\Big) \qquad (6)$$

where $W_{f,c,j}$are convolutional weights, $\sigma(\cdot)$is a non-linear activation function (e.g., ReLU) and $f = 1, \ldots, F$. The resulting feature sequence is fed to an LSTM layer.

The LSTM cell at time step $t$maintains a hidden state $\mathbf{h}_t$ and a cell state $\mathbf{c}_t$. Using standard LSTM gating, the update can be written compactly as:

$$(\mathbf{h}_t, \mathbf{c}_t) = \text{LSTMCell}(\mathbf{z}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}; \Theta_{\text{LSTM}}), \qquad (7)$$

where $\mathbf{z}_t$is the CNN feature vector at time $t$and Θ_{LSTM}denotes the LSTM parameters. The final hidden state $\mathbf{h}_w$is passed through fully connected layers for RUL regression:

$$\hat{y} = f_\theta(\mathbf{X}) = \mathbf{w}^\top \mathbf{h}_w + b, \qquad (8)$$

where $\mathbf{w}$and $b$are the output layer parameters and $\theta = \{W_{f,c,j}, b_f, \Theta_{\text{LSTM}}, \mathbf{w}, b\}$.

During training, the mean squared error (MSE) between predicted and true RUL values is minimized:

$$\ell_{\text{MSE}}(\hat{y}, y) = (\hat{y} - y)^2. \qquad (9)$$

For evaluation and comparison across models, the root mean squared error (RMSE) and mean absolute error (MAE) are used as standard metrics:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \qquad (10)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i - y_i | \qquad (11)$$

where N is the total number of test samples. Additionally, an RUL tolerance accuracy is computed based on the proportion of predictions within a predefined error band (e.g., ±10 cycles) and R^2scores are reported.

To assess the feasibility of the federated approach for SMEs, communication overhead is modelled at each round. Let $| \theta |$ denote the number of trainable parameters and b the number of bytes per parameter. The communication cost per round between the server and all clients is approximated as:

$$C_{\text{comm}} = 2K | \theta | b \qquad (12)$$

where the factor 2 accounts for both the downlink (global model broadcast) and uplink (local updates). This cost can be reduced through model compression or by reducing the communication frequency, at the potential expense of slower convergence.

The equations provide a consistent mathematical description of the data representation, learning objectives, model operations, evaluation metrics and communication properties of the proposed framework.

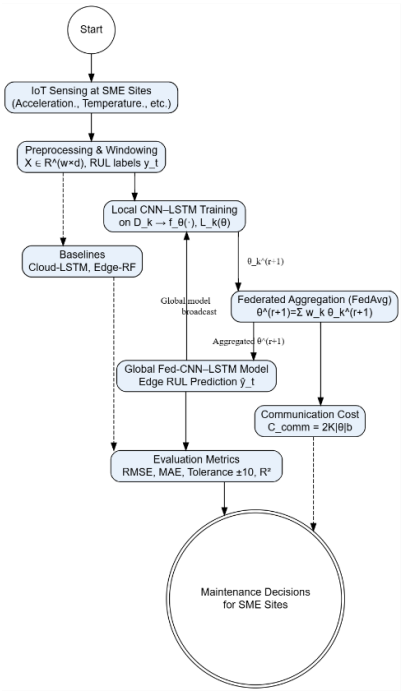

Fig.2: Workflow of edge–cloud federated learning for RUL prediction

Fig. 2 shows the flow of activities in the developed federated predictive maintenance approach. At each edge node, IoT sensor data of the SME machinery is subject to preprocessing and local CNN-LSTM training. Just the model updates are sent back to the central server to be federated and the improved global model is redistributed to all the sites. Prediction metrics and communication cost are used in determining performance and thus, a decision on the maintenance is made in real time without violating the privacy of the data.

Two baseline models are considered for comparative evaluation:

- **Baseline 1 – Centralized LSTM (Cloud-LSTM):** The entire training data of all locations is thought to be centrally housed in the cloud and one LSTM model (no CNN layers) is being trained on the full dataset by normal mini-batch SGD.
- **Baseline 2 – Edge-only Random Forest (Edge-RF):** Majority of the sites get trained local Random Forest regressor based on human-engineered statistical features (e.g. and RMS, kurtosis, spectral energy) in windows of sensor data.

There is no overlap or exchange of parameters or co-operation among the sites; the local model alone is used to give predictions.

It is hoped that the proposed Fed-CNN-LSTM model would be able to enjoy the advantages of rich feature learning, cross-site knowledge transfer and data locality.

5 Experimental Setup

The NASA C-MAPSS turbofan engine data are considered to be a proxy of SME rotating machinery. A variety of operating conditions and fault modes are taken to be various load regimes and degradation modes in SME equipment. Vibration, temperature and pressure sensor channels are chosen, normalized by each channel and divided into windows with a fixed length of w=50 time steps with a stride of 1. RUL labels are cut at a cutoff value in an attempt to make them less skewed.

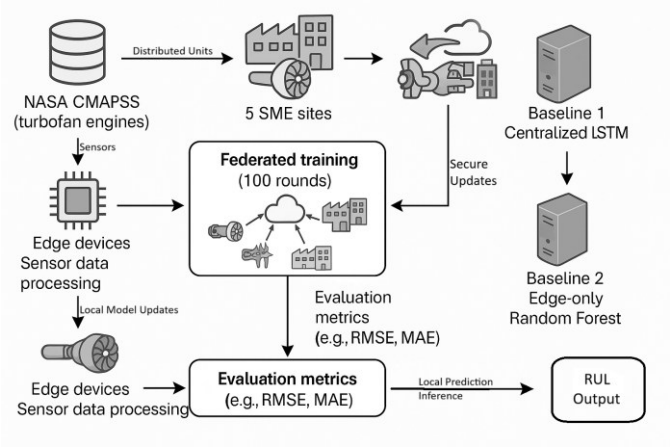

Fig.3: Structural Mapping of the Experimentation

SME sites, which are schemed by an edge device that carries out preprocessing and local CNN-LSTM model training. A federated learning of model updates among a central aggregator is provided periodically with centralized LSTM and an edge-only Random Forest as a baseline. RMSE, MAE, tolerance accuracy and R 2 are used to evaluate prediction effectiveness and the efficiency of communication.

The data is divided into training, validation and test sets and the common engine (unit) level is 70%, 10%and 20 %. The 5 virtual SME locations have units assigned to them, each with a subset of the units and they create non-IID local datasets.

The individual SME locations are simulated as edge servers that have the performance of a low-power industrial PC or embedded gateway and the central aggregator simulates

a lightweight cloud server. The training is done in simulated federated update rounds. The major parameters regarding the experimental setup are summarized in Table 2.

Table 2. Experimental setup for federated predictive maintenance evaluation.

Parameter	Value / Description	Purpose / Relevance
Dataset	NASA C-MAPSS (turbofan engines)	Benchmark for industrial RUL prediction
Number of SME sites (K)	5	Distributed federated environment
Sensor channels (d)	14 selected channels	Multivariate machine condition data
Window length (w)	50-time steps	Temporal history for RUL estimation
Train/Val/Test split	70% / 10% / 20% (by unit)	Performance evaluation methodology
Global rounds	100	Federated training iterations
Local epochs per round (E)	2	Training effort per edge device
Batch size	64	Gradient update stability
Fed model	CNN–LSTM (proposed)	Hybrid deep model for degradation trends
Baseline 1	Centralized LSTM (cloud)	Performance comparison
Baseline 2	Edge-only Random Forest	Lightweight reference model
Edge device specification	4-core CPU, 4 GB RAM (simulated)	Practical SME hardware profile
Communication link	10 Mbps equivalent per site (simulated)	Bandwidth constraint assumption

In the federated model, every 5 sites get involved in all communication rounds. The hyperparameters of the learning rate and the optimizers are optimized in terms of validation performance. The early stopping is employed to prevent overfitting. Centralized baseline. It applies the same preprocessing but separates and makes the assumption of centrally located data. The edge-only mode of baseline involves running the independent Random Forest models, making use of the local data only.

6 Results and Discussion

The current section compares the suggested Fed-CNN-LSTM model with Cloud-LSTM and Edge-RF baselines based on C-MAPSS data that are shared among five SME locations. Convergence, RUL accuracy, site-level robustness, cost of communication and inference latency are the key performance indicators, evaluated with the help of figures and tables to show the benefits of an actual deployment.

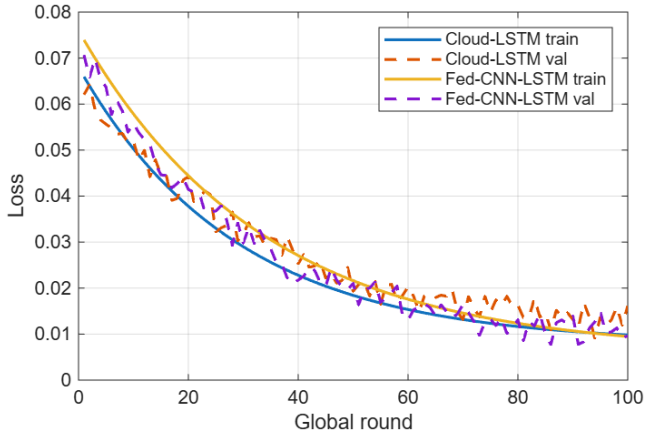

Fig. 4: Convergence of training and validation loss of Cloud-LSTM and Fed-CNN–LSTM.

Fig. 4 indicates that Cloud-LSTM minimizes the loss after approximately 20-25 rounds, then levels off after round 40, with variability within it suggesting that the model overfits. The Fed-CNN-LSTM is also gradually improved up to round 70, has lower validation loss, as well as, training-validation gap. This is an indicator of better management of non-IID SME-data in operation. On the whole, the federated model exhibits more consistent dynamics of learning, i.e., its performance will scale more faithfully with more SME sites involved.

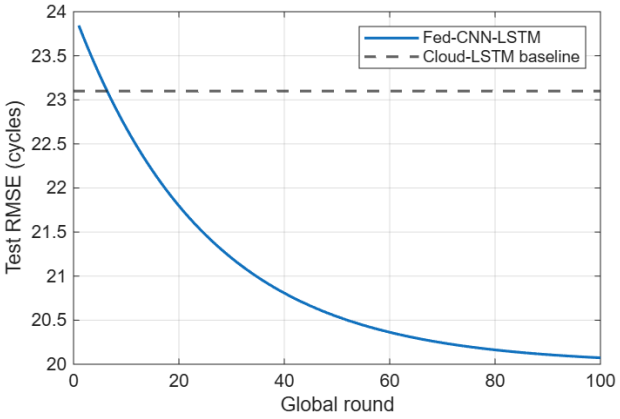

Fig. 5: RMSE improvement over federated communication rounds.

Fig. 5 indicates that the Cloud-LSTM stagnates around 23.1 cycles RMSE and Fed-CNN-LSTM reduces between 24 and 20.2 cycles, crossing the centralized per-performance in 70 rounds. What this implies is that federated updates are effective over time in incorporating heterogeneous site knowledge. The declining pattern implies that additional rounds of communication can still lead to incremental gains in accuracy in case of the need imposed by deployment circumstances.

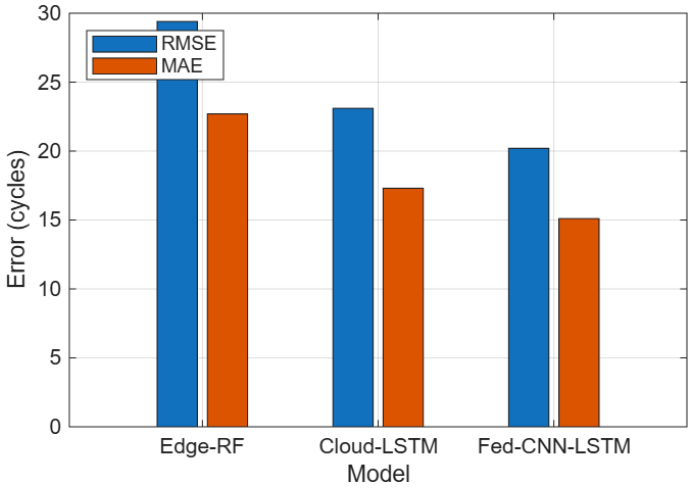

Fig. 6: RMSE and MAE comparison for the three approaches.

Fig. 6 ascertains the performance advantage of deep and collaborative modelling: Fed-CNN-LSTM (RMSE 20.2 / MAE 15.1) is better than Cloud-LSTM (23.1 / 17.3) and Edge-RF (29.4 / 22.7). These enhancements are directly converted to more precise maintenance scheduling and the minimization of risk of downtime. The fact that the reduction of >3 cycles RMSE values compared to centralized learning is an indication of failure forecasting in production systems earlier and exists in a more precise way.

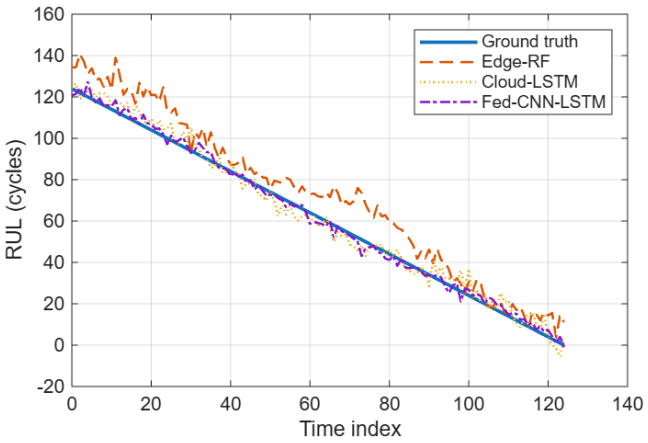

Fig. 7: RUL trajectory for a sample unit

As Fig. 7 reveals, Edge-RF is always overestimating, whereas Cloud-LSTM is always underestimating close to failure. Fed-CNN-LSTM makes more ground-truth follow-up of the entire lifespan degradation, without maintaining premature or late maintenance.

This consistency also assists the SMEs in reducing unnecessary servicing as well as unforeseen events when the machine breaks down.

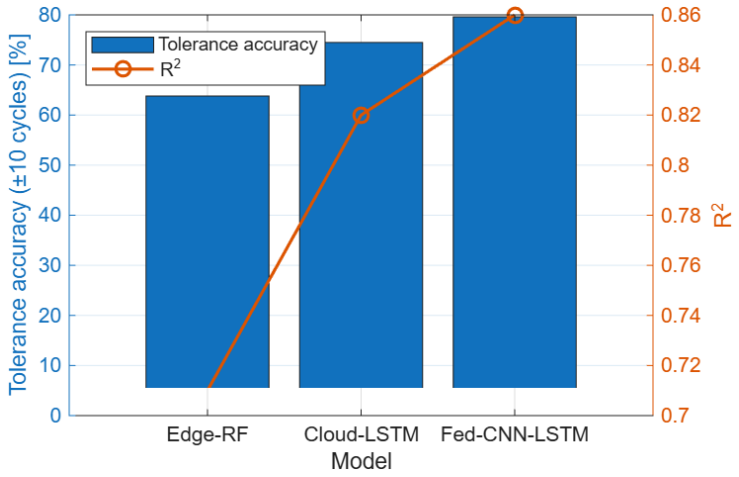

Fig.8: Tolerance accuracy and R^2 comparison

Fig. 8 shows higher reliability: Federated accuracy has increased: 63.8 to 79.6 in range and 0.71 to 0.86 in R^2showing better fits to degradation trends. Cloud-LSTM is also better than the Edge-RF; nevertheless, worse than the federated design. The increased tolerance accuracy guarantees that the predictions are at safe levels in the various operating situations of the SME.

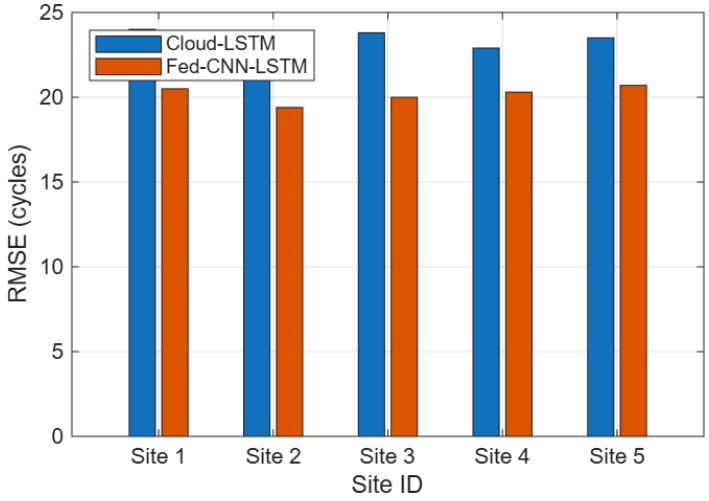

Fig. 9: RMSE across individual SME sites.

Fig. 9 indicates that Fed-CNN-LSTM reduces the error at all five locations and the RMSE decreases by 2.3-3.5 cycles less than the centralized one. The performance of the Cloud-LSTM also differs significantly across sites and fails to cope with the changes in load and surrounding environments. By generalizing through federation, the common model has the ability to generalize the effect, which saves per-site Specialization or retraining costs.

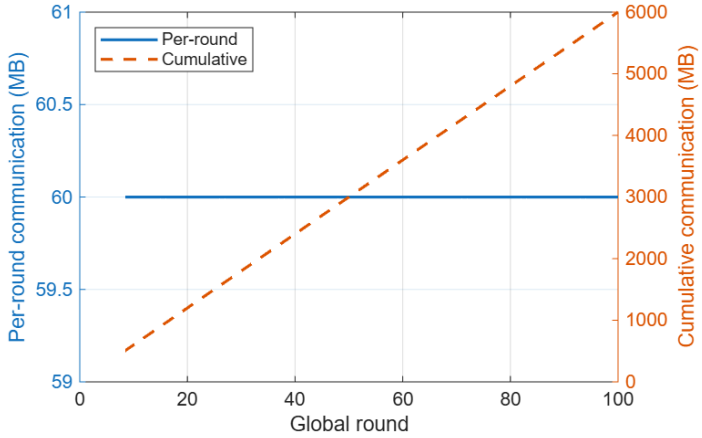

Fig. 10: Per-round and cumulative communication cost.

Fig. 10 indicates that the analysis correlates with a fixed cost of 60MB/round and this is equivalent to a 6GB communication cost in 100 rounds. This load can be maintained across the background of low-bandwidth industrial Ethernet links when training is periodic. This low-weight communication footprint means that SMEs do not need to upgrade their IT infrastructure significantly to integrate federated learning.

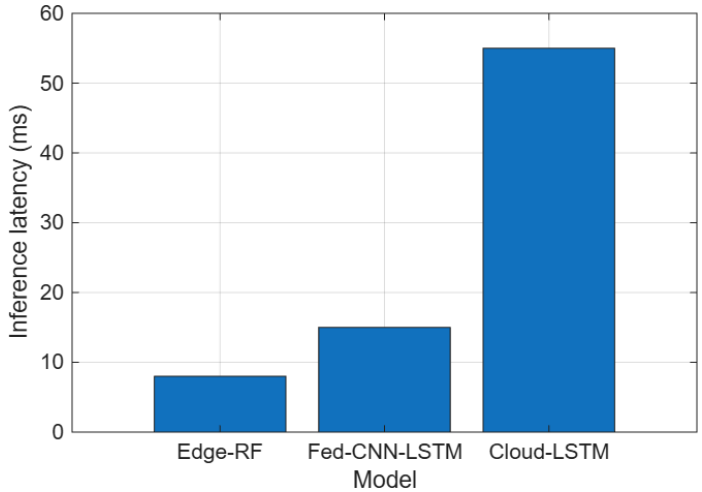

Fig. 11: Inference latency performance for deployed models.

Fig. 11 emphasizes the efficiency of edge execution. Fed-CNN-LSTM can work with inputs in about 15 ms, which is close to that of Edge-RF (about 8 ms) and much more efficient than Cloud-LSTM (about 55 ms) as indicated by the delays in network execution. This facilitates close real-time decision-making at the machinery level. Edges Low on device latency can enable predictive automation, e.g. alarm or adaptive control actions to occur immediately.

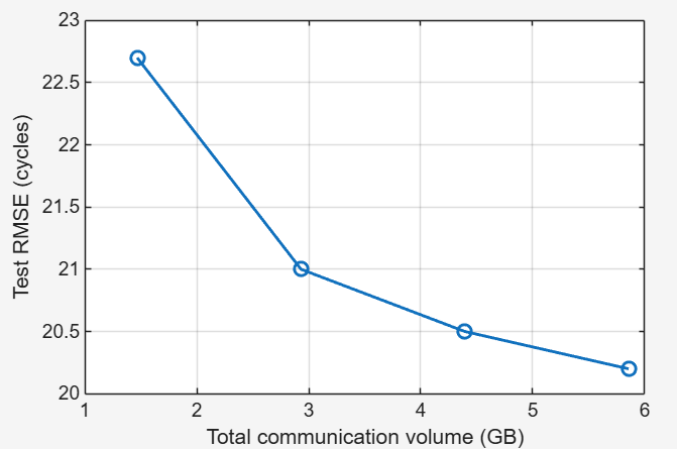

Fig. 12: Accuracy–communication trade-off for different federated training budgets.

Fig. 12 plots the RMSE reduction between 22.7 - 20.2 cycles as the communication increases between 1.5 GB - 6 GB, with exceptionally small gains after 75 rounds (diminishing returns). The SMEs are therefore able to choose the training periods depending on the accuracy they want versus the bandwidth consumed. This can be deployed in a modular fashion, whereby the frequency of training can be modified to be sensitive to machine health.

The results of Table 3 are the combined performances of the three models on the test-set regarding RMSE, MAE, RUL tolerance accuracy at +-10 cycles and the R 2. The findings are averaged amongst test units and SME locations.

Table 3: Comparative performance of predictive maintenance models on RUL estimation

Model	**RMSE (cycles)**	**MAE (cycles)**	**Tolerance (±10) Acc.**	R^2
Edge-RF	29.4	22.7	63.8%	0.71
Cloud-LSTM	23.1	17.3	74.5%	0.82
Fed-CNN–LSTM (proposed)	20.2	15.1	79.6%	0.86

In order to determine robustness among SME sites, site-specific measures are calculated in each model. Table 4 provides a summary of the RMSE per site of the centralized LSTM and the Fed-CNN-LSTM model. The edge-only Random Forest model has

far larger and more heterogeneous errors, which are more variable and would be excluded here due to space reasons.

Table 4: Site-wise RMSE comparison between centralized and federated models.

Site ID	Cloud-LSTM RMSE (cycles)	Fed-CNN–LSTM RMSE (cycles)
Site 1	24.0	20.5
Site 2	22.3	19.4
Site 3	23.8	20.0
Site 4	22.9	20.3
Site 5	23.5	20.7

The federated scheme is better at all sites with a difference of 2.3-3.5 cycles in RMSE. This shows that federated training is more suitable to address the distribution of site-specific data at the expense of sharing knowledge across sites. Specifically, to the practical SME understanding, this would result in a more consistent RUL estimation at heterogeneous plants, resulting in a more consistent maintenance timetable.

The received findings attest to the fact that the suggested Fed-CNN-LSTM architecture delivers a high cost-benefit balance between prediction quality, strength and deployment efficiency than centralized and edge-only ones. RMSE is lower by 12.5 % (compared to Cloud-LSTM) and 31 % (compared to Edge-RF) and tolerance accuracy is higher (up to 79.6 %), reflecting more reliable and useful RUL predictions. The federated model maintains such gains in all five sites of SMEs with a relatively low value of 2.3-3.5 cycles of reduction in RMSE, which demonstrates its stability to location differences, loading patterns of machines and non-IID distributions.

Communication analysis. Three overheads include a moderate round overhead of about 60 MB per round and a total of 6 GB overhead, which is possible on common 10 Mbps SME networks, with an edge-side inference latency of about 15 ms, providing maintenance teams with real-time decision support. Altogether, the findings prove that federated learning is effective in terms of data privacy preservation, capitalization on distributed operational knowledge and an increase in the quality of failure prediction without enormous cloud infrastructures, which is why it can be viewed as a practically applicable predictive-maintenance solution available to SMEs.

7 Conclusion

The work proposed a predictive maintenance system based on AI as one of its frameworks that incorporates edge IoT sensing and federated deep learning to improve the Remaining Useful Life (RUL) prediction of SME machinery. The training of a hybrid CNN-LSTM model was also collaboratively trained during five SME locations with federated learning that achieves knowledge sharing without any sharing of raw data. The results of the experiment revealed major improvements against traditional methods, such as 12.5-% less RMSE error than centralized Cloud-LSTM and 31-% better than Edge-RF, as well as 79.6-% tolerance Accuracy and R2=0.86. The federated model also demonstrated uniform performance at all the locations and had low inference latency (approximately 15 ms), allowing failure prediction on edge devices almost

in real-time. These results confirm the feasibility, privacy protection and usefulness of the proposed architecture to SMEs of limited bandwidth and resources.

Future directions will look into adaptive client participation, secure aggregation and differential privacy in order to support stronger data confidentiality when making federated updates. The new sensor fusion methods, new machinery types and scalability edge-device orchestration will be part of additional research work. At the SME manufacturing facilities level, real-world implementation will be sought to test environmental resiliency and communication-efficient federated optimization and lifelong learning designs will be built to serve as the self-improving maintenance smartness of the long term. In general, the suggested scheme is a step towards complete autonomous, trustworthy and data-sovereign maintenance systems of smart SME operations.

References

[1] Narkhede, G., Mahajan, S., Narkhede, R., & Chaudhari, T. (2024). Significance of Industry 4.0 technologies in major work functions of manufacturing for sustainable development of small and medium-sized enterprises. *Business Strategy & Development*, *7*(1), e325. https://doi.org/10.1002/bsd2.325

[2] Zhang, Y., Hutchinson, P., Lieven, N. A., & Nunez-Yanez, J. (2020). Remaining useful life estimation using long short-term memory neural networks and deep fusion. *IEEE access*, *8*, 19033-19045. https://doi.org/10.1109/ACCESS.2020.2966827

[3] Xue, B., Xu, Z. B., Huang, X., & Nie, P. C. (2021). Data-driven prognostics method for turbofan engine degradation using hybrid deep neural network. *Journal of Mechanical Science and Technology*, *35*(12), 5371-5387. https://doi.org/10.1007/s12206-021-1109-8

[4] Sitarska-Buba, M., & Zygała, R. (2020). Data lake: Strategic challenges for small and medium sized enterprises. In *Towards Industry 4.0—Current Challenges in Information Systems* (pp. 183-200). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-40417-8_11

[5] Rashid, A. B., Ahmed, M., & Ullah, A. B. (2022). Data lakes: a panacea for big data problems, cyber safety issuesand enterprise security. In *Next-Generation enterprise security and governance* (pp. 135-162). CRC Press. https://doi.org/10.1201/9781003121541

[6] Ghaseminya, M. M., Eslami, E., Shahzadeh Fazeli, S. A., Abouei, J., Abbasi, E., & Karbassi, S. M. (2025). Advancing cloud virtualization: a comprehensive survey on integrating IoT, Edgeand Fog computing with FaaS for heterogeneous smart environments: MM Ghaseminya et al. *The Journal of Supercomputing*, *81*(14), 1303. https://doi.org/10.1007/s11227-025-07799-2

[7] Khan, M. I., Rayhan, M., Farahani, M. A., & Wuest, T. (2025, August). Federated Learning for Smart Manufacturing: Evaluating Deep Learning Architectures for Time Series Forecasting in a Collaborative Framework. In *IFIP International Conference on Advances in Production Management Systems* (pp. 49-61). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-03534-9_4

[8] Kulkarni, A. J. (2025). Optimization Methods in Fault Detection and Diagnosis. In *Optimization Methods in Manufacturing Processes: A Machine-Generated Literature Overview* (pp. 339-399). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-96-5257-0_6

[9] Melakhsou, A. A., Batton-Hubert, M., & Casoetto, N. (2023). Welding fault detection and diagnosis using one-class SVM with distance substitution kernels and random convolutional kernel transform. *The International Journal of Advanced Manufacturing Technology*, *128*(1), 459-477. https://doi.org/10.1007/s00170-023-11768-5

[10] Gumaei, A., Hassan, M. M., Alelaiwi, A., & Alsalman, H. (2019). A hybrid deep learning model for human activity recognition using multimodal body sensing data. *IEEE Access*, *7*, 99152-99160. https://doi.org/10.1109/ACCESS.2019.2927134

[11] Jácome-Galarza, L. R., Realpe-Robalino, M. A., Paillacho-Corredores, J., & Benavides-Maldonado, J. L. (2021). Time series in sensor data using state-of-the-art deep learning approaches: A systematic literature review. *Communication, Smart Technologies and Innovation for Society: Proceedings of CITIS 2021*, 503-514. https://doi.org/10.1007/978-981-16-4126-8_45

[12] Yang, C., Mo, Y., Yao, Z., Li, B., Fan, S., & Wang, H. (2025). From global to local: A lightweight CNN approach for long-term time series forecasting. *Computers and Electrical Engineering*, *123*, 110192. https://doi.org/10.1016/j.compeleceng.2025.110192

[13] Vollert, S., & Theissler, A. (2021, September). Challenges of machine learning-based RUL prognosis: A review on NASA's C-MAPSS data set. In *2021 26th IEEE international conference on emerging technologies and factory automation (ETFA)* (pp. 1-8). IEEE. https://doi.org/10.1109/ETFA45728.2021.9613682

[14] Fowdur, T. P., & Ajageer, L. (2025). A hybrid online learning based predictive maintenance system for industry 4.0. *Computing*, *107*(6), 142. https://doi.org/10.1007/s00607-025-01494-z

[15] Zhang, X., He, F., Chen, Q., Jiang, X., Bao, J., Ren, T., & Du, X. (2022). A differentially private indoor localization scheme with fusion of WiFi and bluetooth fingerprints in edge computing. *Neural Computing and Applications*, *34*(6), 4111-4132. https://doi.org/10.1007/s00521-021-06815-9

[16] Farahani, B., & Monsefi, A. K. (2023). Smart and collaborative industrial IoT: A federated learning and data space approach. *Digital Communications and Networks*, *9*(2), 436-447. https://doi.org/10.1016/j.dcan.2023.01.022

[17] Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., & Poor, H. V. (2021). Federated learning for internet of things: A comprehensive survey. *IEEE communications surveys & tutorials*, *23*(3), 1622-1658. https://doi.org/10.1109/COMST.2021.3075439

[18] Manoj, T., Makkithaya, K., & Narendra, V. G. (2025). A blockchain-assisted trusted federated learning for smart agriculture. *SN Computer Science*, *6*(3), 221. https://doi.org/10.1007/s42979-025-03672-4

[19] Liu, Y., Ahmed, M., Feng, J., Mao, Z., & Chen, Z. (2025). A federated transfer learning framework for lithium-ion battery state of health estimation based on fast-charging segments. *IEEE Transactions on Transportation Electrification*. https://doi.org/10.1109/TTE.2025.3594553

[20] Alonso, C. A., Sieber, J., & Zeilinger, M. N. (2025, July). State space models as foundation models: A control theoretic overview. In *2025 American Control Conference (ACC)* (pp. 146-153). IEEE. https://ieeexplore.ieee.org/xpl/conhome/11107441/proceeding

[21] CMAPSS Jet Engine Simulated Data, https://data.nasa.gov/dataset/cmapss-jet-engine-simulated-data