IOP Conf. Series: Materials Science and Engineering **1344** (2026) 012022 doi:10.1088/1757-899X/1344/1/012022

Federated Learning Framework to support AI-Driven Prescriptive Maintenance in Large-Scale Cryogenic Infrastructures at CERN

Paolo Cacace[1,2], Diogo R. Santos[1] and Luigi Serio[1]

[1]Technology Department, CERN, Meyrin, Switzerland
[2]DIAG Department, Sapienza University, Rome, Italy

E-mail: paolo.cacace@cern.ch

Abstract. This work presents a Federated Learning (FL) framework for anomaly detection and prescriptive maintenance tested on CERN's helium compression systems. To keep its magnets in a superconducting state, the Large Hadron Collider (LHC) relies on an extensive cryogenic system, where predictive maintenance classical approaches are already central in ensuring operational efficiency and reducing infrastructure downtime. Leveraging FL, our solution enables AI-powered predictive maintenance via decentralized model training, using acquisitions from 20 custom motor-compressor systems located in 4 compressor stations around the 27 km LHC accelerator. This approach enables model performance improvement while optimizing energy consumption and maintaining consistent anomaly detection performance. The machine learning model used to detect deviations from nominal behaviour is based on deep autoencoders, an architecture that has proven to be particularly effective for reconstruction tasks. FL is enabled thanks to CERN's FL platform, CAFEIN®, which has already demonstrated its effectiveness in medical applications. Experimental results confirm the efficacy of our framework, trained in federated learning, in detecting anomalies and estimating the RUL (Remaining Useful Life) of the compressors, with performance aligned with prior findings using centralized setup. This approach exemplifies how collaborative AI-driven methodologies can foster innovation, sustainability, and efficiency in industrial settings, accelerating collaboration between research centres, and at the same time complying with data privacy limitations.

1 Introduction

Anomalous events within the technical infrastructure are a major source of downtime in large-scale particle accelerators [1]. Prescriptive maintenance has proven to be an effective instrument that can play a crucial role in monitoring these complex systems, optimizing efficiency, and reducing costs [2]. At CERN, the world's largest cryogenic system plays a crucial role in keeping the LHC magnets in a superconducting state by cooling them to 1.8 K using a complex helium processing cycle [3]. A key component of this system is the helium compression infrastructure, primarily consisting of screw compressors. Monitoring their condition has been essential for the optimal operation of the cryogenic system and the LHC.
Machine learning has already proven to be a powerful and effective tool for consistently diagnosing compressors problems and predicting their Remaining Useful Life (RUL). In our work, vibration data taken from selected measuring points through dedicated equipment, compressor management data and working parameters, have been used to train a model to predict the compressors RUL, demonstrating superior performance when compared with vibration specialists [4].


Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

This paper presents a FL framework for anomaly detection, classification, and RUL estimation, validated on the LHC's helium compression system. CERN offers an ideal testbed thanks to its distributed compressor infrastructure and the availability of CAFEIN® [5], an independently developed FL platform which will be used for model aggregation. Our investigation focuses on 20 motor-compressor units across five stations around the accelerator ring. Each station is divided into low-pressure and high-pressure groups, featuring distinct compressor models, creating a realistic setting for the framework validation. Our approach leverages the analysis of operational parameters and vibration spectrum data acquired on-site, processed by a deep-learning autoencoder to monitor the compressors' health over time. This method allows identifying, classifying, and tracking anomalies while preserving data locality thanks to federated learning. When scaled across various infrastructures, our solution enables energy-efficient [6] and secure model training, enabling seamless collaboration among international organizations and research centres. The tool supports enhanced maintenance strategies and operational efficiency, and it helps reduce downtime. The aim is to extend equipment lifespan, and optimize CAPEX[1] and OPEX[2] costs exemplifying how collaborative AI-driven methodologies can drive innovation, sustainability, and operational efficiency in both industrial and research applications.

2 LHC Helium Compression Infrastructure and Database Overview

Helium Compression Infrastructure Overview: CERN's LHC cryogenic system relies on a critical helium compression infrastructure, composed of eight cryoplants, each featuring an 18 kW system at 4.5 K and a 2.4 kW system at 1.8 K, requiring substantial helium compression capacity. The cryogenic infrastructure includes 68 oil-lubricated screw compressors that supply helium to Cold Boxes at a nominal flow of 1,800 g/s and a pressure level of 18 bars. Our investigation focuses on 20 compressors that supply the 4.5 K cryogenic plants at four LHC points (18, 4, 6, and 8). These compressors are divided into two categories: Boosters (Low Pressure) and High Pressure units, organized over five positions. Booster compressors (positions 1–3) provide medium pressure, while high-stage compressors (positions 6 and 7). During previous LHC runs, these machines accumulated over 100,000 operating hours and experienced various vibration levels, making the development of predictive maintenance models a necessity to ensure operational reliability and minimize unplanned downtime.

Database Organization Overview: For condition monitoring, the compressors undergo monthly manual vibration measurements using tri-axial accelerometers, with subsequent spectral analysis whose results are consolidated into a report detailing faults such as bearing failures, imbalance, and misalignment. In addition, key operating parameters, including compressor load, motor power consumption, and helium pressure and temperature, are continuously logged in CERN's NXCALS Big Data tracking system. Maintenance interventions are recorded via a Computerized Maintenance Management System (CMMS), which is used to schedule and track maintenance activities and log all technical interventions. This fragmented data distribution requires a reorganization that integrates CMMS asset tracking with NXCALS data to ensure the quality and consistency of information available for diagnostics and provides a solid foundation for training machine learning models for predictive maintenance.

Considering this organization, the final step in data preparation is labeling the monthly acquisitions based on expert weekly reports; this gives us a reference to split data from compressors working in nominal conditions and those operating with partial load. All the retrieved information is reorganized into one consistent data structure containing vibration acquisitions and compressor operating parameters, considering the information retrieved from the CMMS. By integrating vibration analysis, operational data, and structured database management, we create a unified and consistent data structure framework for effective machine learning model training and result interpretation.

Data Filtering : Final data filtering was conducted considering compressors operating conditions defined based on sliding valve opening percentage and motor current. A compressor is considered to be under Full Load Condition (FLC) when its sliding valve opening exceeds 95 %, and it is operating at Full Power Condition(FPC) when the normalized Z-score under FLC[3] falls within the range $[-1.5, 1.5]$. When these criteria are met simultaneously, the compressor is considered to be in a nominal operating state, thereby ensuring that only stable, high-load data are used for diagnostic analysis while inconsistent readings from reduced load or power conditions are filtered out.

[1] Capital Expenditures (CAPEX) refer to funds used by an organization to acquire, upgrade, or maintain physical assets.

[2] Operating Expenditures (OPEX) refer to the day-to-day expenses incurred in running a business or system.

[3] The Z-score, also known as the standard score, is calculated using the formula: $Z(t) = \frac{\text{Current}(t) - \mu_{\text{FLC}}}{\sigma_{\text{FLC}}}$, where Current$(t)$ is the data point, and μ_{FLC} and σ_{FL} are the mean and standard deviation of the motor current under FL conditions.

3 Federated Learning

Federated Learning Overview: In recent years, the rise of distributed data sources and the increasingly stringent data protection regulations have intensified interest in (FL) as a viable and privacy-preserving machine learning paradigm. FL, introduced by [7], enables collaborative model training across multiple clients without sharing raw data, making it particularly suitable for domains like medical [8] [9] and critical infrastructure [10] where data privacy and ownership are critical. In this setting, model updates are computed locally and aggregated centrally, using different model combination techniques to iteratively improve a global model.

In our work, we take advantage of CAFEIN®, a FL platform developed at CERN to streamline distributed model training while preserving data privacy. Leveraging a Docker-based design for client nodes, it provides a flexible solution for implementing FL processes in clinical and industrial settings. Built using a client–server architecture with a central Parameter Server (PS), CAFEIN® enables client-server communication using the MQTT[4] protocol, chosen for its reliability and efficiency in FL scenarios. A dedicated control plane manages key FL operations (e.g., orchestrating training rounds, scheduling client participation, and aggregating updates), ensuring the process runs smoothly and securely. Taking advantage of this platform's capabilities, we implemented a FL workflow that involves multiple rounds of communication between a central server and distributed clients. Each client trains locally on its private data, and the aggregated updates at the server side create a new global model at each iteration. By repeating these rounds, the global model refines itself incrementally, combining all the client's local knowledge. The typical steps in such a FL cycle include:

1. The central server initialises the global model and distributes it to a selected subset of clients.
2. Each selected client downloads the global model and performs local updates using its private data. The updates typically involve running one or more steps of a gradient-based optimization algorithm, such as Stochastic Gradient Descent (SGD).
3. Once the local updates are completed, the clients upload their updated model parameters to the central server.
4. The server aggregates the updates from all participating clients using a specific aggregation technique and updates the global model. The updated global model is then distributed back to the clients for the next round of training.

This decentralized learning approach eliminates the need for data centralization, thereby addressing concerns about data privacy and reducing communication costs. However, challenges remain, particularly when clients have heterogeneous data, or when the number of participating clients fluctuates across rounds. This makes choosing the right model aggregation technique critical for the success of FL.

Federated Averaging (*FedAvg*) Algorithm Overview : Introduced by [7], FedAvg is one of the most widely used aggregation algorithms in Federated Learning. The main idea behind FedAvg is to aggregate the local models, averaging the updates from all participating clients based on their "importance" (i.e., the amount of data they used for training). In CAFEIN®, the FedAvg aggregation algorithm takes inspiration from [7] and works as shown in Algorithm 1. FedAvg works by selecting a subset of clients each round, and each of the selected clients receives the global model and updates it by running a predefined number of local training epochs. At the end of the local training, the models are sent back to the server and aggregated following the steps described in Algorithm 1. This enables FedAvg to combine the knowledge learned by individual clients in a single model, all while keeping the data local and private.

Algorithm 1 FedAvg

Parameters: K clients indexed by k; minibatch size B; learning rate [11] η; clients local epochs E.
Server: Initialize w_0.
for each round $t = 1, 2, \dots$ **do**
  $S_t \leftarrow$ random set of m clients
  for each client $k \in S_t$ **do**
    $w_k^{t+1} \leftarrow \text{ClientUpdate}(k, w_t)$
  $m_t \leftarrow \sum_{k \in S_t} n_k$
  $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_k^{t+1}$
return w_T

ClientUpdate (k, w)**:**
Split local dataset P_k into minibatches of size B
for each local epoch i from 1 to E **do**
  for each minibatch $b \in P_k$ **do**
    $w \leftarrow w - \eta \nabla \ell(w; b)$
return w

[4]MQTT (Message Queuing Telemetry Transport) is a lightweight and robust publish-subscribe network protocol.

4 Data Distribution Overview

After pre-processing, data were categorized based on the technical specifications of the compressors at each station. The stations were then divided into two clusters: one for Low Pressure Compressors (LP) (positions 1 and 2)[5] and another for High Pressure Compressors (HP) (positions 6 and 7).

The training and evaluation sets were then partitioned based on reports from vibration specialists between healthy and unhealthy acquisitions. The training samples, slightly different from what was done in [4], were chosen among the acquisitions from compressors that operated under FLC and FPC with no reading labeled as anomalous during LHC Run 2. Subsequently, for each compressor group, model performance was evaluated on a single validation set that combined all the samples labeled as "good" from all compressors running during LHC Run 3 with samples from Run 2 compressors that experienced either critical or non-critical issues. This organization provides 181 training and 305 validation samples for the LP Compressors and 131 training and 268 validation samples for HP Compressors distributed across the LHC Cryogenic infrastructure as follows:

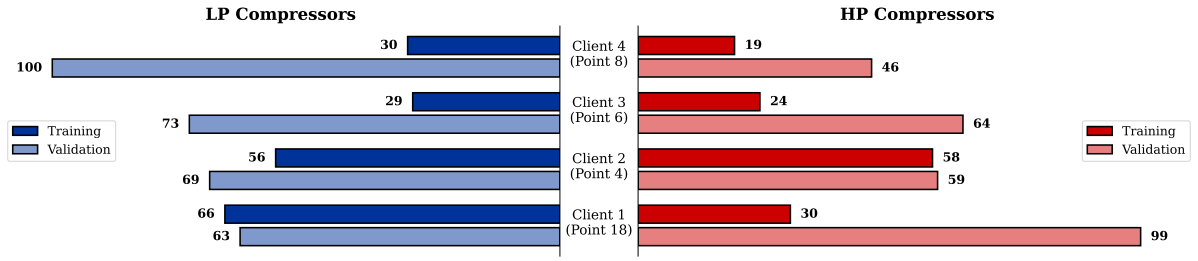

Figure 1. Client Data Distribution

This data distribution enabled effective validation of the FL setup's performance while preserving data privacy by aggregating model updates without sharing data between clients.

The data distribution was further examined using Principal Component Analysis (PCA) [12], a widely used dimensionality reduction technique that transforms correlated input variables into a set of orthogonal principal components. By retaining only the most informative components, PCA reduces the size of the feature space while preserving the underlying structure of the data, allowing for both data distribution, visualization and analysis. In this study, a three-dimensional PCA projection with a total retained variance of 68.8 % was computed by fitting the transformation on a set contained data used for training both LP and HP compressor prediction models. Validation samples distribution has then been visualized in two separate scatter plots, one for LP compressors samples and one for HP compressors samples. Points were further separated by LHC station (P18, P4, P6 and P8). For each station cluster, we overlaid a confidence ellipsoid 95 % derived from a multivariate normal fit, providing a compact visual summary of the spread and overlap of the cluster.

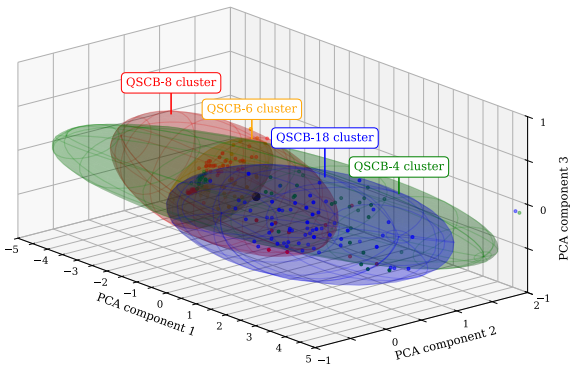

Figure 2. Data Distribution LP Compressors

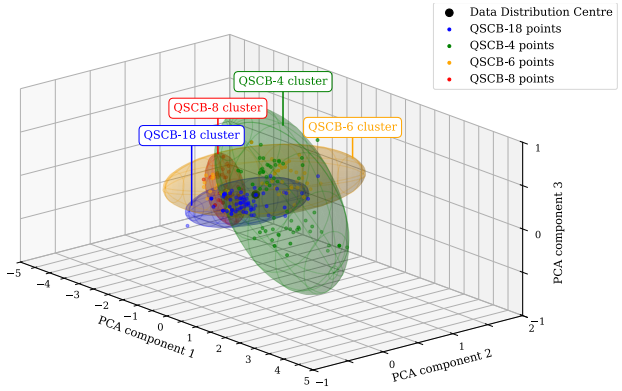

Figure 3. Data Distribution HP Compressors

[5] Compressors at position 3 were excluded from this analysis due to their fault-free performance, lacking significant data to conduct a consistent analysis.

The distributions presented in Fig.2 and Fig.3, suggest that the validation data from LP compressors appear significantly more dispersed compared to the HP compressors, suggesting a higher degree of variability or noise. This dispersion reflects greater operational heterogeneity and less stable vibration patterns in LP compressors samples, potentially making them more challenging to model consistently. In contrast, the HP validation data clusters more tightly, indicating more homogeneous behaviour and potentially higher model results reliability for this class.

5 Federated Anomaly Detection Infrastructure

To assess the benefits and trade-offs of collaborative training, we evaluated model performance under three distinct data-partitioning setups:

- **Centralized Learning:** All vibration spectra of HP and LP compressors are aggregated on a central server. A model is trained on each compressor group dataset; the result serves as an upper-bound reference under ideal data-availability conditions.
- **Isolated Learning:** Each compression station trains two independent autoencoders, one on its HP data and one on its LP data, without any parameter sharing. This approach quantifies the performance degradation that arises when information exchange between stations is prohibited.
- **Federated Learning:** Similar to the isolated learning setup, each station maintains a separate model for HP and LP compressors groups that are trained and evaluated locally on its data. Subsequently, to balance accuracy, privacy and increase model training efficiency, at each local epoch, the corresponding model weights (HP with HP, LP with LP) are sent to a central server and aggregated following the FedAvg algorithm. The resulting global HP and LP models are then redistributed to all stations for the next round of local training.

The comparison of the results obtained between those setups provides a uniform basis for quantifying and understanding how full data centralization, complete decentralization, and federated collaboration influence the algorithm's training dynamics and overall performance.

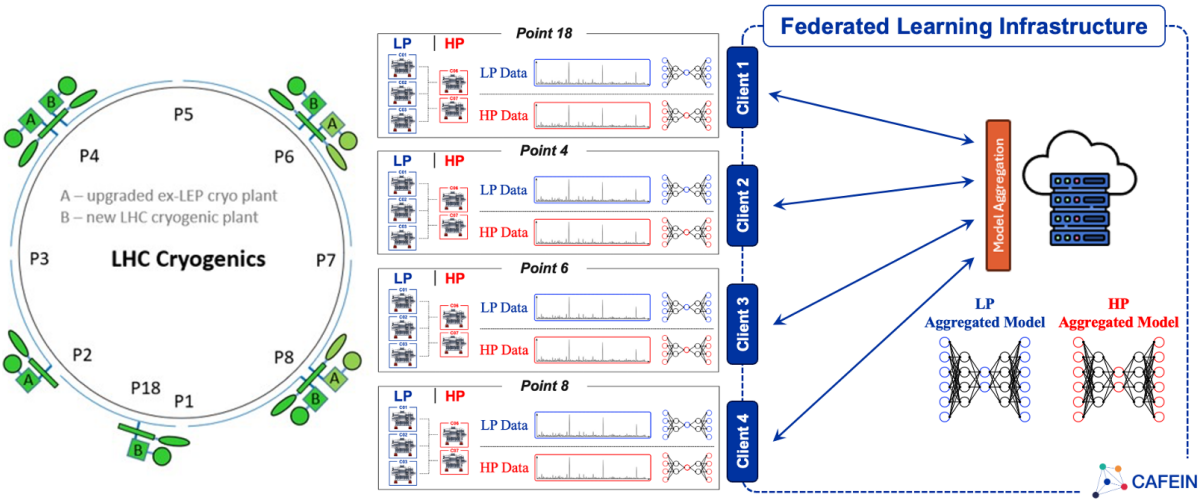

Figure 4. Federated Learning infrastructure

All experiments employ the same fully-connected autoencoder, an architecture ideally suited for anomaly detection and classification tasks. It consists of two symmetric components: an encoder and a decoder. The encoder compresses the input vector $\mathbf{x}$ into a lower-dimensional latent representation defined by $z = \mathsf{enc}(\mathbf{x}) = \sigma\big(W_e\,\mathbf{x} + b_e\big)$. The decoder then reconstructs the input from z as $\hat{\mathbf{x}} = \mathsf{dec}(z) = \sigma\big(W_d\,z + b_d\big)$. During training, the parameters W_e, b_e, W_d, b_d are adjusted to minimize the Mean Absolute Error (MAE) loss[6]: $\mathcal{L} = \sum_i \big\|\mathbf{x}_i - \hat{\mathbf{x}}_i\big\|$. By training exclusively on spectra from compressors operating

[6]The MAE reconstruction error, defined as $\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|$, measures the average absolute discrepancy between each original value x_i and its reconstruction $\hat{x}_i$.

under healthy conditions, the autoencoder learns the nominal data distribution, and at inference time, any input that deviates from this learned distribution yields a MAE reconstruction error, enabling the detection of anomalies as outliers. Specific frequencies related to known faults are also taken into account to label the faults by comparing the original signal and the reconstructed one, frequency by frequency.

The autoencoder is trained using identical hyper-parameters across centralized, isolated, and federated settings: a batch size of 8, the Adam optimizer with a learning rate of 1×10^{-3}, and in the FL setup, one local epoch per aggregation round using FedAvg. The performance and the learning process of all the setups have then been compared and evaluated using the validation loss.

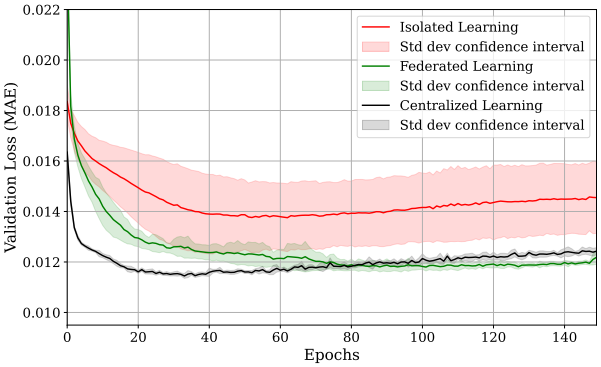

Figure 5. Model Convergence with different setups (LP compressors)

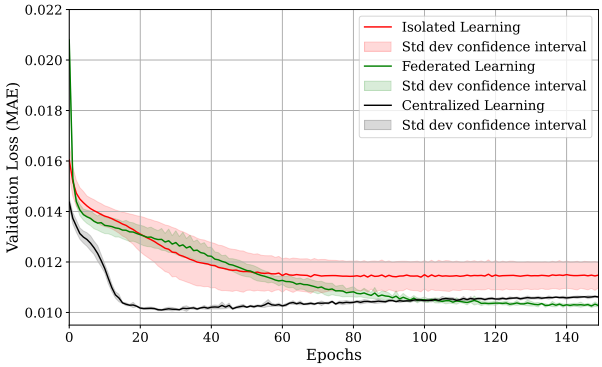

Figure 6. Model Convergence with different setups (HP compressors)

As illustrated in Fig. 5 and Fig. 6, the FL configurations exhibit slower convergence compared to the centralized setup. However, they achieve comparable performance relative to the centralized approach, with the best validation MAE being only 3.5 % and 2.0 % higher for the LP and HP Compressors, respectively. The isolated learning approach, exhibits higher convergence error compared to the other 2 setups, lacking generalization capabilities due to the limited amount of data used during training, thus confirming the benefits of collaborative training in federated setups.

Table 1. Best Validation MAE Comparison

Setup	***MAE (LP)***	***MAE (HP)***
Centralized Learning	*0.0114 ± 0.0001*	*0.0101 ± 0.00005*
Isolated Learning	*0.0138 ± 0.0013*	*0.0114 ± 0.0006*
Federated Learning (FedAvg)	*0.0118 ± 0.0002*	*0.0103 ± 0.0001*

The model trained in a federated setup was benchmarked against its centralized counterpart using standard machine learning metrics [13] to ensure its capability to accurately predict the RUL of helium compressors remains intact. RUL estimation is carried out through a two-stage hybrid approach: first, the autoencoder processes vibration acquisitions and computes the reconstruction error (MAE) for each sample; then, a similarity-based degradation model produces a refined RUL prediction based on historical fault data. This second stage analize data sample by sample and for each motor-compressor unit leverages Z-score reconstruction error[7] and emphasize bearing-related frequencies to estimate equipment condition. The Z-Score computed dynamically by simulating the data stream into the model requires at least six consecutive readings per motor-compressor unit to ensure results stability. Each sample is then assigned a health state (Normal, Warning, or critical), and the final RUL is estimated by analyzing the most recent Z-score error value, the presence of bearing fault-related frequencies and the time spent in each state. To effectively and objectively evaluate model outputs, predictions were compared with ground-truth labels calculated based on historical compressors shutdowns due to technical problems. For the compressors that experienced a critical issue, acquisitions within one month of shutdown were labeled as critical; those

[7] Z-score rescales a value x as $z = \frac{x-\mu}{\sigma}$, where $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the sample mean and $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$ is the corresponding standard deviation. Those metrics are computed for every new sample in the system, with z expresses how many standard deviations a new measurement deviates from nominal behaviour.

from one to thirteen months before critical conditions were labeled as Warning; and the acquisitions taken before, along with those from compressors without critical conditions, were labeled as Normal.

Table 2. RUL estimation performance comparison Type-B Compressors (C01-C02)

Metric	*Precision*		*Recall*		*F1-score*		*Support*
Setup	*Centralized*	*Federated*	*Centralized*	*Federated*	*Centralized*	*Federated*	
Normal	0.94	0.94	0.97	0.96	0.96	0.95	201
Warning	0.74	0.71	0.59	0.59	0.65	0.64	29
Critical	1.00	1.00	1.00	1.00	1.00	1.00	3
Centralized Model Aggregated Results				***Federated Model Aggregated Results***			
Metric	*Precision*	*Recall*	*F1-score*	*Metric*	*Precision*	*Recall*	*F1-score*
Macro avg	0.89	0.85	0.87	*Macro avg*	0.88	0.85	0.86
Weighted avg	0.92	0.92	0.92	*Weighted avg*	0.91	0.92	0.92
Accuracy	0.92			*Accuracy*	0.92		

Table 3. RUL estimation performance comparison Type-H Compressors (C06-C07)

Metric	*Precision*		*Recall*		*F1-score*		*Support*
Setup	*Centralized*	*Federated*	*Centralized*	*Federated*	*Centralized*	*Federated*	
Normal	0.95	0.95	0.97	0.93	0.96	0.96	183
Warning	0.71	0.71	0.52	0.52	0.60	0.60	23
Critical	0.75	0.75	1.00	1.00	0.86	0.86	6
Centralized Model Aggregated Results				***Federated Model Aggregated Results***			
Metric	*Precision*	*Recall*	*F1-score*	*Metric*	*Precision*	*Recall*	*F1-score*
Macro avg	0.80	0.83	0.81	*Macro avg*	0.80	0.83	0.81
Weighted avg	0.92	0.93	0.92	*Weighted avg*	0.92	0.93	0.92
Accuracy	0.93			*Accuracy*	0.93		

The results show that the model trained with FL performs comparably to the centralized model. For the booster stations, performance is slightly lower, likely due to greater data heterogeneity, as also indicated by the variability of the isolated learning results. Overall, the findings support the effectiveness of FL in predicting the RUL of cryogenic compressors, indicating that it can serve as an effective, energy-efficient and privacy preserving alternative to centralized approaches, even in scenarios characterized by data heterogeneity and sample distributional shifts across clients.

6 Conclusions

This work presents a FL framework for anomaly detection and RUL estimation of CERN's cryogenic compressor infrastructure. Building on previous successes with centralized machine learning, the proposed solution leverages FL using CAFEIN® FL platform to enable energy efficient model training across distributed compressor stations without data sharing. Results show that federated models achieve performance comparable to centralized ones while preserving data privacy and improving generalization compared to isolated setups. The lightweight model, made to fit edge devices, can be used in a real-time monitoring setup to ideally craft ready to use devices for AI-driven prescriptive maintenance.

By leveraging efficient, privacy-preserving collaboration and significantly reducing computational and operational costs, this system presents a scalable and adaptable solution for predictive maintenance in large-scale industrial environments. Our approach enables model training without the need to exchange data, thereby promoting cross organizational collaborations. This decentralized infrastructure not only enhances the robustness of predictive models but also enables the integration of diverse data inputs, increasing the capabilities of the model and leading to tangible improvements such as reduced downtime, extended equipment lifespans, and minimized environmental impact. Ultimately, our solution, powered by CAFEIN® FL exemplifies how collaborative AI-driven strategies can drive innovation, sustainability, and economic efficiency across the industrial and research landscapes.

Acknowledgements
This work is based on vibration data collected over the years by the CERN cryogenics maintenance team (TE-CRG-ML), whose expertise and continuous efforts in the field have been instrumental to this paper.

References

[1] Luigi Serio et al. A smart framework for the availability and reliability assessment and management of accelerators technical facilities. *Journal of Physics: Conference Series*, 1067(7):072029, 2018.

[2] Kalaiarasan Sekar et al. Role of machine learning approaches in remaining useful prediction: A review. In *Intelligent Computing and Innovation on Data Science*, pages 361–370, Singapore, 2021. Springer Nature Singapore.

[3] Luigi Serio et al. CERN experience and strategy for the maintenance of cryogenic plants and distribution systems. *IOP Conference Series: Materials Science and Engineering*, 101:012140, 2015.

[4] Paolo Cacace et al. Machine learning framework for anomaly detection and maintenance optimization in large-scale cryogenic systems. *IOP Conference Series: Materials Science and Engineering*, 1327(1):012030, 2025.

[5] CAFEIN. https://cafein.web.cern.ch/. Accessed: 02-05-2025.

[6] Luca Barbieri et al. A close look at the communication efficiency and the energy footprints of robust federated learning in industrial IoT. *IEEE Internet of Things Journal*, 12(11):15130–15150, 2025.

[7] H. Brendan McMahan et al. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 1273–1282. PMLR, 2017.

[8] Diogo Reis Santos et al. Feasibility analysis of federated neural networks for explainable detection of atrial fibrillation. In *2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, 2024.

[9] Andrea Protani et al. Federated GNNs for EEG-based stroke assessment. *arXiv preprint arXiv:2411.02286*, 2024.

[10] Ilias Siniosoglou et al. Federated learning models in decentralized critical infrastructure. In *Shaping the Future of IoT with Edge Intelligence*, chapter 5. River Publishers, 2023.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[12] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

[13] Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14, 03 2024.