Uncertainty-Aware Bearing Remaining Useful Life Prediction Based on Conformal Prediction

Shun Wang[1], Yolanda Vidal[1,2], and Francesc Pozo[1,2]

[1] *Department of Mathematics (MAT), Control, Data and Artificial Intelligence (CoDAlab), Barcelona East School of Engineering (EEBE), Universitat Politècnica de Catalunya · BarcelonaTech (UPC), Campus Diagonal-Besòs, Av. Eduard Maristany, 16, 08019 Barcelona, Spain*
shun.wang@upc.edu
yolanda.vidal@upc.edu
francesc.pozo@upc.edu

[2] *Institute of Mathematics (IMTech), Universitat Politècnica de Catalunya · BarcelonaTech (UPC), Pau Gargallo, 14, 08028 Barcelona, Spain*

ABSTRACT

Accurate prediction of the remaining useful life of rolling element bearings is a critical task in prognostics and health management. Although deep learning methods have shown strong predictive capability, purely data-driven approaches still face two important limitations: they may produce physically inconsistent predictions that contradict the irreversible nature of bearing degradation and often fail to provide reliable uncertainty estimates. To address these issues, this paper proposes a physics-informed probabilistic framework to predict the remaining useful life. First, a health index is constructed from logarithmic envelope spectrum features using a variational autoencoder, enabling the extraction of a monotonic degradation indicator without requiring labeled fault data. Second, a Transformer-based predictor is trained with a monotonicity constraint that explicitly enforces the predicted remaining useful life to be non-increasing over time. Third, Monte Carlo dropout is used to quantify epistemic uncertainty, and a post-hoc conformal calibration strategy is applied to construct finite-sample prediction intervals with guaranteed marginal coverage by leveraging historical degradation data. Experiments on the XJTU-SY full-lifecycle bearing dataset show that the proposed framework improves point prediction accuracy relative to controlled feature and model ablations. More importantly, the uncertainty results reveal a substantial mismatch between raw Monte Carlo dropout intervals and the observed prediction errors: the average prediction interval coverage probability increases from 0.4835 before calibration to 0.9445 after conformal calibration. The resulting wider intervals should not be interpreted only as a loss of sharpness, but as a correction of the severe overconfidence of the uncalibrated model under heterogeneous degradation trajectories. Bearings with more irregular or nonstationary degradation behavior require wider calibrated intervals to maintain reliable coverage, indicating that the proposed framework can expose trajectory-dependent prediction difficulty and support risk-aware maintenance decisions.

1. INTRODUCTION

Bearings are critical rotating components in a wide range of mechanical systems, including those in the energy, aerospace, and manufacturing sectors (Mian et al., 2024; S. Wang, Vidal, & Pozo, 2025a). Their progressive degradation can quickly lead to secondary damage, unexpected downtime, safety risks, and considerable economic losses (Zhao, Cai, & Cui, 2024). As a result, remaining useful life (RUL) prediction has become a central task in prognostics and health management, as it enables condition-based maintenance and helps prevent catastrophic failures (Cuesta, Leturiondo, Vidal, & Pozo, 2025). In practice, reliable bearing prognostics requires addressing three closely related objectives: constructing a health index (HI) that captures incipient degradation, learning an RUL prediction model that is both accurate and physically plausible, and quantifying uncertainty so that maintenance decisions can be made with confidence (J. Zhou, Yang, Xiang, & Qin, 2025). Although substantial progress has been made on each of these aspects individually, integrating them into a unified and effective framework remains a challenge.

The construction of a reliable HI is the foundation of data-driven prognostics. Traditional statistical indicators such as

root-mean-square (RMS) and kurtosis are widely used because of their simplicity, but they mainly reflect global energy changes and are often insensitive to early-stage faults (Guo, Li, & Wan, 2023; Zhong, Wang, Guo, Cabrera, & Li, 2020). Signal-processing methods such as envelope analysis can improve the sensitivity to fault-related signatures, but generally require empirical parameter tuning and substantial domain expertise (Hou et al., 2022). To automate feature extraction, learning-based approaches have employed autoencoders and related models (Wei, Wu, & Terpenny, 2021). However, the HIs derived from reconstruction errors often fluctuate and show non-monotonic behavior, which limits their usefulness for degradation tracking. From a physical perspective, bearing faults generate repetitive impulsive excitations that appear in the envelope spectrum (Randall & Antoni, 2011). However, in practice, dominant high-amplitude components in the raw spectrum can mask weak early-stage fault signatures, making it difficult to construct a robust and monotonic HI.

Even when a reliable HI is available, mapping it to an accurate and physically meaningful RUL remains difficult. Recurrent neural networks such as LSTM and GRU can model temporal dependencies, but may still produce non-monotonic and physically implausible predictions when the degradation signal is weak or noisy (Dong et al., 2023; Ni et al., 2024). Transformer-based models have recently attracted attention due to their ability to capture long-range dependencies through self-attention (Cao, Meng, Li, Wu, & Fan, 2024). Nevertheless, when trained in a purely data-driven manner, they do not guarantee that the predicted RUL decreases over time. This is problematic because bearing degradation is inherently irreversible. Without explicit physical constraints, the learned prediction trajectories may contradict the basic degradation process.

In addition to accurate point prediction, practical maintenance decisions also require reliable uncertainty quantification. A common approach is to use Monte Carlo dropout to estimate epistemic uncertainty (Gal & Ghahramani, 2016). However, such methods are based on variational approximation and therefore rely on implicit distributional assumptions, which may lead to poor calibration in practice, especially under distribution shift (Angelopoulos, Bates, Fisch, Lei, & Schuster, 2022). As a result, the resulting prediction intervals can be overconfident and may fail to achieve the desired coverage. Conformal prediction provides a principled way to improve the reliability of such uncertainty estimates. By applying a post-hoc calibration step, it can construct prediction intervals with finite-sample coverage guarantees without requiring strong distributional assumptions (Angelopoulos et al., 2022; X. Zhou, Chen, Gui, & Cheng, 2025). This advantage is particularly important for prognostics, where only a limited number of run-to-failure trajectories are available. In such small-sample scenarios, conformal calibration offers a more robust alternative to uncertainty estimates obtained directly from dropout.

Motivated by the above challenges, this paper proposes a unified physics-informed uncertainty-aware framework for bearing remaining useful life prediction. Rather than introducing a new standalone learning architecture, the contribution of this work lies in linking three requirements that are often treated separately in existing studies: degradation-sensitive health representation, physically consistent RUL evolution, and statistically calibrated uncertainty estimation. First, a physically interpretable health index is constructed from the logarithmic envelope spectrum using a variational autoencoder, enabling weak degradation signatures to be captured without labeled fault data. Second, a Transformer encoder with a differentiable monotonicity constraint is employed to enforce non-increasing RUL trajectories and reduce physically implausible predictions caused by local fluctuations in the health indicator. Third, Monte Carlo dropout is combined with conformal calibration to convert overconfident predictive intervals into statistically reliable bounds with finite-sample marginal coverage under the exchangeability assumption. Finally, beyond reporting interval metrics, this study analyzes how calibrated uncertainty varies across heterogeneous degradation trajectories, clarifying that wider intervals may reflect genuinely more difficult and less representative degradation behavior rather than merely poor model performance. The proposed framework is validated on the XJTU-SY full-lifecycle bearing dataset (B. Wang, Lei, Li, & Li, 2018).

The remainder of this paper is organized as follows. Section 2 describes the proposed methodology. Section 3 presents the experimental results. Finally, Section 4 concludes the paper.

2. PROPOSED METHODOLOGY

The proposed framework is designed to satisfy two practical requirements of bearing prognostics: first, the health indicator should be physically interpretable and robust to non-stationary noise; second, the RUL predictor should be accurate, physically consistent, and able to provide confidence-aware outputs. To this end, the method consists of two stages. The first stage constructs a degradation indicator using the SimUFD framework (S. Wang, Vidal, & Pozo, 2025b). The second stage learns a constrained sequence-to-RUL mapping with a physics-informed Transformer, followed by conformal calibration to obtain statistically valid prediction intervals. The general pipeline is illustrated in Figure 1.

2.1. Health Index Construction

Rolling-element bearing faults typically produce impulsive excitations at fault-characteristic frequencies (FCFs) and their harmonics, which appear as modulation components in the envelope spectrum (Randall & Antoni, 2011). In

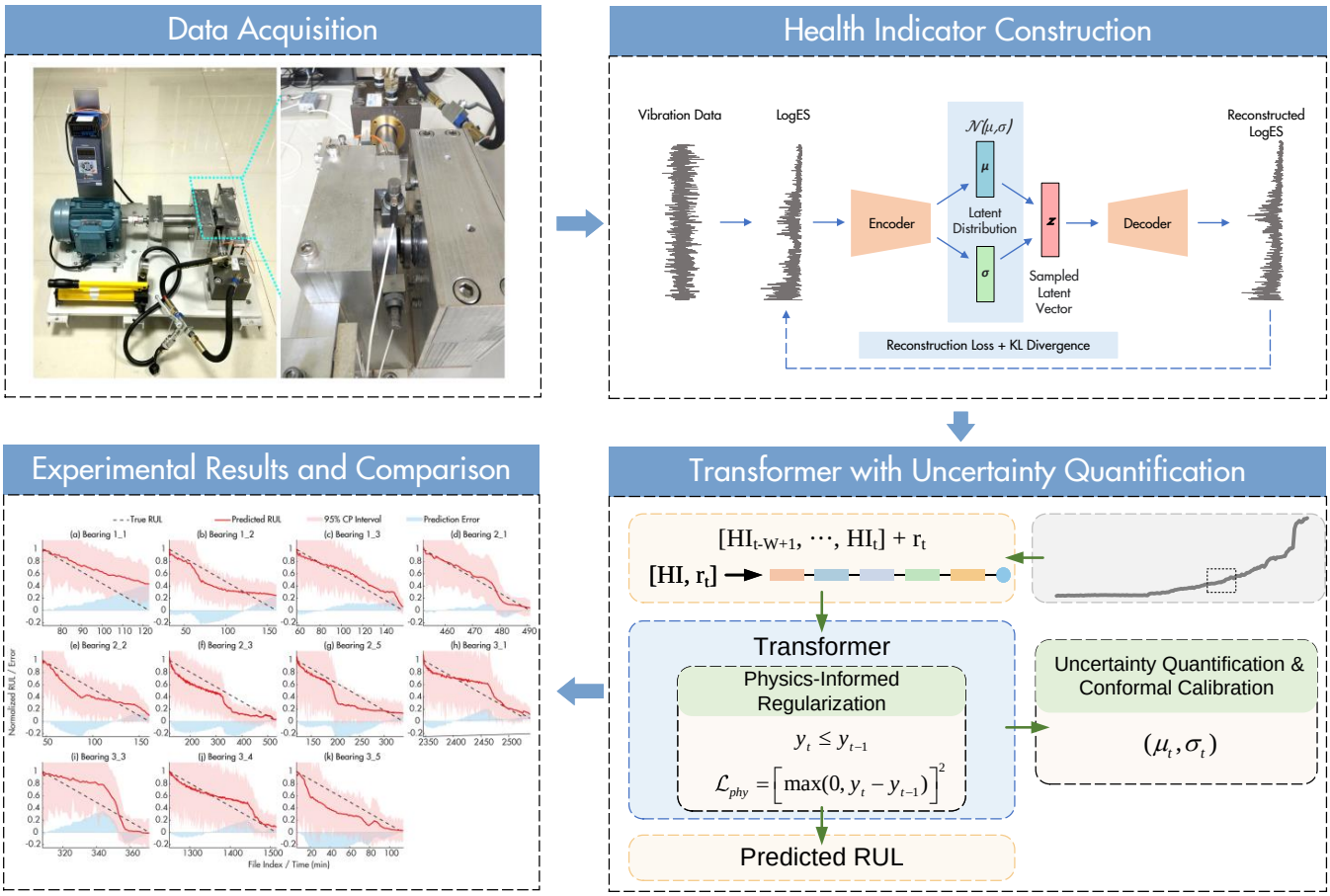

Figure 1. Overview of the proposed framework, from vibration data acquisition and VAE-based health indicator construction to RUL prediction using a physics-informed Transformer and conformal calibration.

this work, the HI is constructed using the SimUFD framework (S. Wang et al., 2025b), which operates directly in the envelope-spectrum domain and therefore preserves physical interpretability.

2.1.1. Logarithmic Envelope Spectrum

Given a vibration signal $s(t)$, the analytic signal is first obtained using the Hilbert transform $\mathcal{H}\{\cdot\}$, and the envelope is computed as $\mathrm{En}(t) = \sqrt{s(t)^2 + \mathcal{H}\{s(t)\}^2}$. The corresponding single-sided envelope spectrum is then defined as $\mathrm{ES}(f) = |\mathcal{F}\{\mathrm{En}(t)\}|$, where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform. Although the envelope spectrum reveals FCF-related peaks, a small number of dominant high-amplitude components can mask weak early fault signatures and make the derived HI unstable. To alleviate this effect, a logarithmic transformation is applied:

$$\mathrm{LogES}(f) = \log\big(1 + \mathrm{ES}(f)\big), \tag{1}$$

where the constant 1 is added to ensure numerical stability near zero. Previous studies have shown that this transformation offers two important advantages (S. Wang et al., 2025b). First, it compresses the spectral dynamic range, making weak fault-related components more visible relative to dominant peaks. Second, it introduces an amplitude-adaptive normalization effect when deviations from a reference spectrum are measured, thereby reducing sensitivity to large-amplitude fluctuations and improving the stability of the degradation indicator.

2.1.2. VAE-Based Health Index

The extracted LogES features are fed to a variational autoencoder (VAE) trained exclusively on healthy-state data. The encoder $q_\phi(\mathbf{z}|\mathbf{x})$ maps each LogES frame $\mathbf{x}$ to a latent Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$, from which a latent code $\mathbf{z}$ is sampled via the reparameterization trick. The decoder $p_\theta(\mathbf{x}|\mathbf{z})$ reconstructs the original LogES from $\mathbf{z}$. Training minimizes the evidence lower bound:

$$\mathcal{L}_{\mathrm{VAE}} = \underbrace{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}_{\mathcal{L}_{\mathrm{rec}}} + D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\big). \tag{2}$$

During online monitoring, the HI at time step t is defined as the mean squared reconstruction error over all frequency bins

N_f:

$$\mathrm{HI}(t) = \frac{1}{N_f} \sum_{i=1}^{N_f} \big[\mathrm{LogES}_i(t) - \hat{\mathrm{LogES}}_i(t)\big]^2. \tag{3}$$

As the bearing degrades, spectral shifts between the current state and the learned healthy distribution increase the reconstruction error, causing $\mathrm{HI}(t)$ to rise.

A fault alarm point (FAP) is detected using the three-sigma rule applied to the HI computed over the initial N_h healthy files:

$$\tau = \mu_h + 3\sigma_h, \quad t_{\mathrm{FAP}} = \min\big\{t \mid \mathrm{HI}(t+k) > \tau,\ k = 0, 1, 2\big\}, \tag{4}$$

where a persistence window of three consecutive exceedances suppresses transient false alarms. The normalized RUL label is then defined as a piecewise-linear function that equals 1 at t_{FAP} and decreases linearly to 0 at the end of life T:

$$y_t = \begin{cases} 1, & t \le t_{\mathrm{FAP}}, \\ 1 - \dfrac{t - t_{\mathrm{FAP}}}{T - t_{\mathrm{FAP}}}, & t > t_{\mathrm{FAP}}. \end{cases} \tag{5}$$

2.2. Physics-Informed Transformer

Given the HI sequence and normalized RUL labels produced in the previous stage, the second stage trains a Transformer-based predictor to map degradation trajectories to RUL estimates. A physics-informed penalty is incorporated to enforce the monotonic decay property inherent to irreversible bearing degradation, preventing physically implausible predictions during inference.

2.2.1. Input Tokenization and Temporal Encoding

At each time step t, the input to the predictor is a fixed-length sliding window containing the most recent L HI observations:

$$\mathbf{x}_t = \big[x_{t-L+1}, \ldots, x_t\big]^\top \in \mathbb{R}^L, \tag{6}$$

with left-padding applied when $t < L - 1$. Instead of using only a generic positional index, a degradation-aware relative-time scalar is introduced to encode the physical progression since the FAP:

$$r_t = \frac{\log\big(1 + \max(t - t_{\mathrm{FAP}},\, 0)\big)}{\log(1 + T_{\mathrm{ref}})}, \tag{7}$$

where $T_{\mathrm{ref}} = 500$ is a fixed normalization horizon. This logarithmic mapping emphasizes differences in the early degradation stage while compressing large time indices, thus providing more informative temporal context to the self-attention mechanism.

2.2.2. Transformer Predictor

The sequence $\tilde{\mathbf{H}}_t$ is processed by an N_l-layer Transformer encoder. Each layer consists of multi-head self-attention and a position-wise feed-forward network, together with residual connections and layer normalization. The contextualized CLS token $\mathbf{U}_{t,0} \in \mathbb{R}^d$ is then passed through an MLP regression head to produce the predicted RUL mean:

$$\hat{\mu}_t = g_\mu(\mathbf{U}_{t,0}). \tag{8}$$

Because bearing degradation is driven by irreversible damage accumulation, the predicted RUL should satisfy the monotonic decay constraint $d\hat{y}/dt \le 0$. In discrete form, this physical prior can be expressed as $\hat{\mu}_t - \hat{\mu}_{t-1} \le 0$. Following the physics-informed learning paradigm, this constraint is incorporated as a differentiable penalty:

$$\mathcal{L}_{\mathrm{phy}} = \frac{1}{B} \sum_{b=1}^{B} \big[\max\big(0,\ \hat{\mu}_t^{(b)} - \hat{\mu}_{t-1}^{(b)}\big)\big]^2, \tag{9}$$

where B denotes the batch size and paired inputs $(\mathbf{x}_{t-1}, \mathbf{x}_t)$ are used during training. The ReLU operator ensures that only violations of the monotonicity constraint are penalized, allowing the model to remain flexible enough to capture different degradation rates across bearings.

The supervised training objective combines the mean squared error (MSE) loss with the physics-based penalty:

$$\mathcal{L}_{\mathrm{sup}} = \frac{1}{|\Omega|} \sum_{t \in \Omega} (y_t - \hat{\mu}_t)^2, \tag{10}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \lambda\, \mathcal{L}_{\mathrm{phy}}, \tag{11}$$

where $\lambda > 0$ controls the strength of the physical constraint and Ω denotes the supervised training set.

2.3. Uncertainty Quantification and Conformal Calibration

Reliable remaining useful life prediction requires not only accurate point estimates, but also well-calibrated uncertainty bounds to support risk-aware maintenance decisions. In the proposed framework, predictive uncertainty is modeled in two stages: epistemic uncertainty estimation using Monte Carlo dropout, followed by statistical calibration using conformal prediction.

2.3.1. Epistemic Uncertainty via Monte Carlo Dropout

Epistemic uncertainty, arising from limited training data and model ambiguity, is approximated using Monte Carlo dropout (Gal & Ghahramani, 2016). During inference, dropout layers remain active and S stochastic forward passes are performed for each input sample. The predictive mean

and epistemic variance are estimated as follows:

$$\bar{\mu}_t = \frac{1}{S}\sum_{s=1}^{S}\hat{\mu}_t^{(s)}, \qquad \hat{\sigma}^2_{\mathrm{epi}}(t) = \frac{1}{S-1}\sum_{s=1}^{S}\left(\hat{\mu}_t^{(s)} - \bar{\mu}_t\right)^2. \tag{12}$$

Although Monte Carlo dropout provides a practical approximation of model uncertainty, such estimates are typically not well calibrated and may lead to overconfident prediction intervals, particularly under distribution shift or small-sample conditions. To improve the statistical reliability of the uncertainty bounds, a post-hoc conformal calibration procedure is therefore introduced.

2.3.2. Conformal Calibration

Although Monte Carlo dropout provides a sample-dependent estimate of predictive uncertainty, the resulting intervals are not guaranteed to be well calibrated. To improve their reliability, a conformal calibration step is applied. Specifically, a conformal prediction strategy is adopted, in which the prediction error is scaled by the corresponding Monte Carlo uncertainty estimate.

Given a calibration dataset, the nonconformity score for sample i is defined as

$$s_i = \frac{|y_i - \bar{\mu}_i|}{\hat{\sigma}_{\mathrm{epi},i} + \varepsilon}, \tag{13}$$

where $\bar{\mu}_i$ is the predictive mean obtained by Monte Carlo dropout, $\hat{\sigma}_{\mathrm{epi},i}$ is the corresponding predictive standard deviation, and ε is a small positive constant introduced for numerical stability.

The scores are aggregated into a calibration set $\mathcal{S} = \{s_i\}_{i=1}^{|\mathcal{S}|}$, and the conformal quantile is computed as

$$\hat{q} = \mathrm{Quantile}\left(\mathcal{S}, \frac{\lceil (|\mathcal{S}|+1)(1-\alpha)\rceil}{|\mathcal{S}|}\right). \tag{14}$$

For a test sample at time step t, the calibrated prediction interval is then constructed as

$$\mathcal{C}_\alpha(\mathbf{x}_t) = \left[\bar{\mu}_t - \hat{q}\hat{\sigma}_{\mathrm{epi}}(t),\ \bar{\mu}_t + \hat{q}\hat{\sigma}_{\mathrm{epi}}(t)\right]. \tag{15}$$

In this way, the interval width is adaptively adjusted according to the predictive uncertainty estimated by Monte Carlo dropout. Samples with larger uncertainty are assigned wider intervals, whereas samples with smaller uncertainty receive tighter bounds.

3. EXPERIMENTS

This section evaluates the proposed physics-informed probabilistic framework on a publicly available run-to-failure bearing dataset.

3.1. Dataset and Experimental Setup

The proposed framework is evaluated on the XJTU-SY run-to-failure bearing dataset (B. Wang et al., 2018). This dataset contains 15 bearings operating under three different conditions, defined by different combinations of rotational speed and radial load. The vibration signals are sampled at 25.6 kHz and recorded at 1-minute intervals. The test rig and representative failure modes are shown in Figure 2.

Following common practice, bearings with highly irregular degradation behavior and without a clear degradation trend (e.g., Bearing 3-2) are excluded from the evaluation. As a result, experiments are conducted on the 11 bearings listed in Table 1. A leave-one-bearing-out protocol is adopted, where one bearing is used for testing and all remaining non-test bearings are used to construct the training, validation, and conformal calibration sets. It should be noted that the bearings in this dataset exhibit heterogeneous degradation trajectories across operating conditions and failure modes. Therefore, the leave-one-bearing-out setting is not intended to create an idealized identically distributed scenario, but rather to evaluate whether the proposed framework can provide physically plausible predictions and meaningful uncertainty information when the test trajectory differs from the historical trajectories available for training and calibration. Although several prior studies have reported RUL results on the XJTU-SY dataset, direct cross-paper numerical comparison is often confounded by differences in selected bearings, feature representations, degradation-point definitions, and validation protocols. For this reason, the present study emphasizes controlled within-study comparisons and the interpretation of calibrated uncertainty rather than cross-study ranking.

The health indicator is generated using the procedure described in Section 2.1, and the fault alarm point is detected using the three-sigma rule. For the Transformer predictor, the input sequence length is set to $L = 8$. The model is trained for 100 epochs using the AdamW optimizer with a cosine annealing schedule. The weight of the monotonicity constraint is set to $\lambda = 1.0$, and the dropout rate is set to $p = 0.05$. For uncertainty estimation, Monte Carlo dropout is performed with $S = 50$ stochastic forward passes. The conformal miscoverage level is set to $\alpha = 0.05$, corresponding to a nominal 95% prediction interval.

3.2. Health Indicator Evaluation

Figure 3 presents the normalized health indicator trajectories for all 15 bearings. Across different operating conditions, a consistent pattern can be observed: the health indicator remains close to zero and relatively stable during the healthy stage, and then increases as degradation progresses. This behavior provides a clear degradation trend for the downstream RUL predictor. Under Condition 1, the degradation evolution is relatively gradual, whereas under Condition 3 the transi-

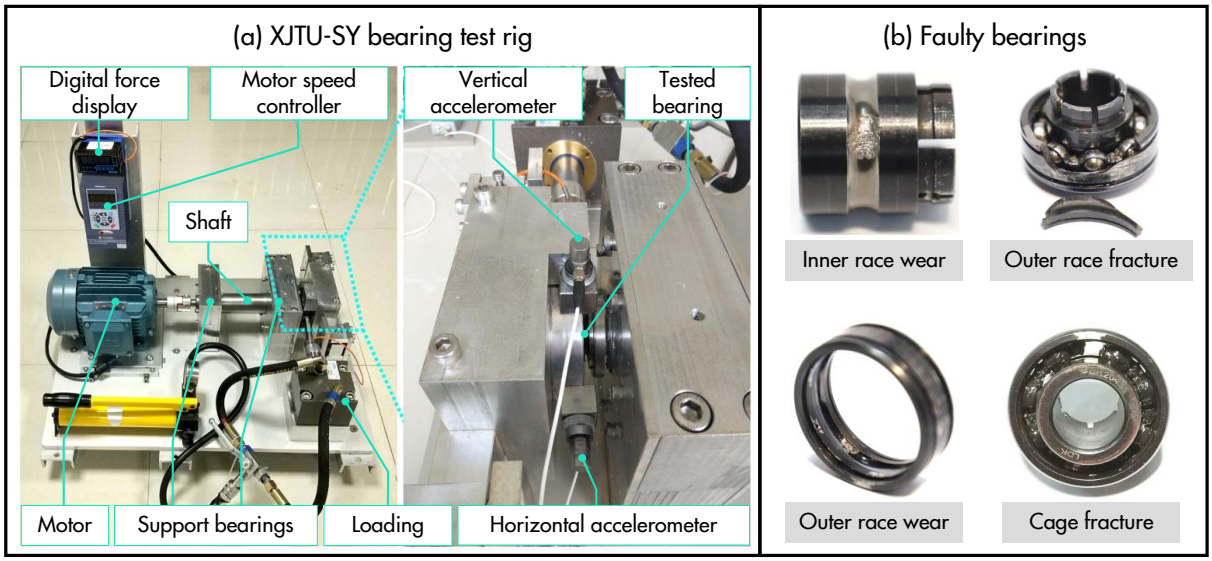

Figure 2. XJTU-SY accelerated life test platform and representative bearing failure modes. (a) Bearing test rig configuration. (b) Examples of failed bearings illustrating typical fault modes.

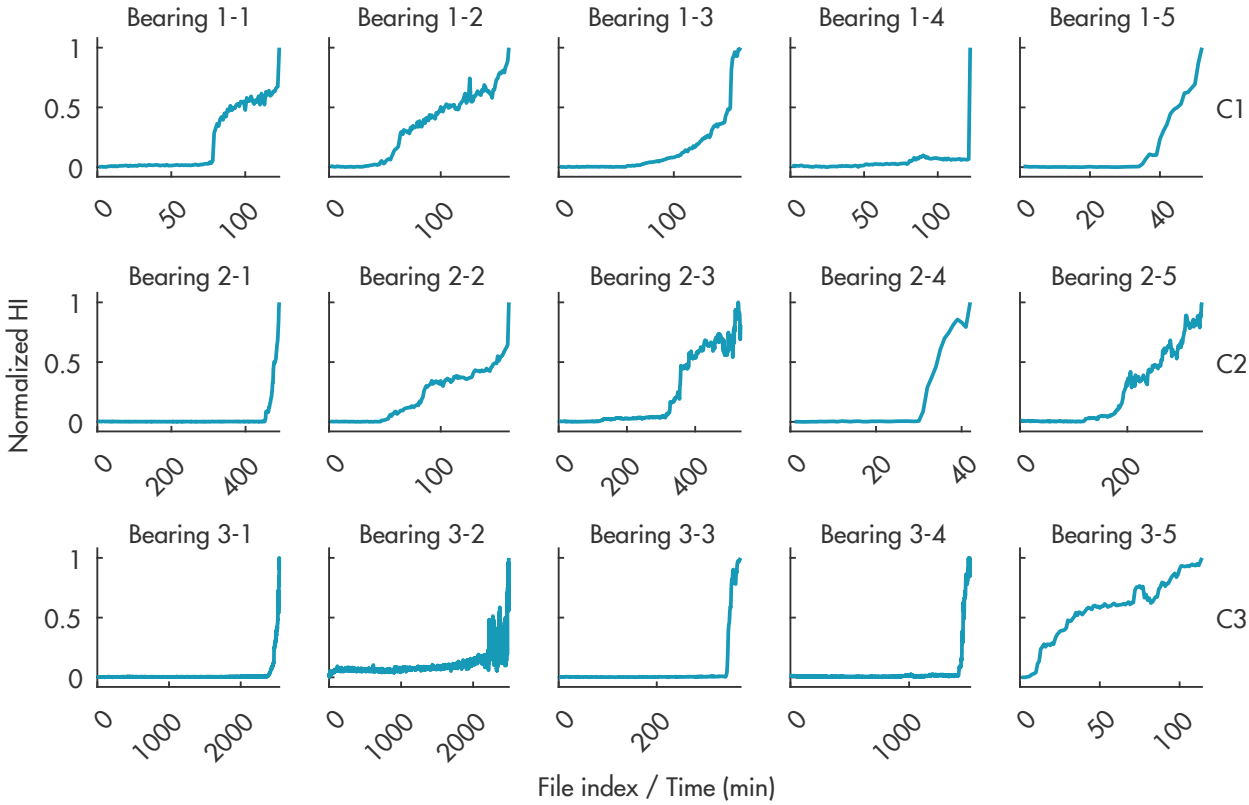

Figure 3. Normalized health indicators constructed by the proposed framework for all 15 XJTU-SY bearings under three operating conditions. Each curve represents the VAE reconstruction error computed from the logarithmic envelope spectrum, followed by min-max normalization within each run. Rows correspond to Condition 1 (2100 r/min, 12 kN), Condition 2 (2250 r/min, 11 kN), and Condition 3 (2400 r/min, 10 kN), while columns correspond to Bearings 1 to 5 within each condition. The condition label (C1 to C3) is shown on the right side of each row.

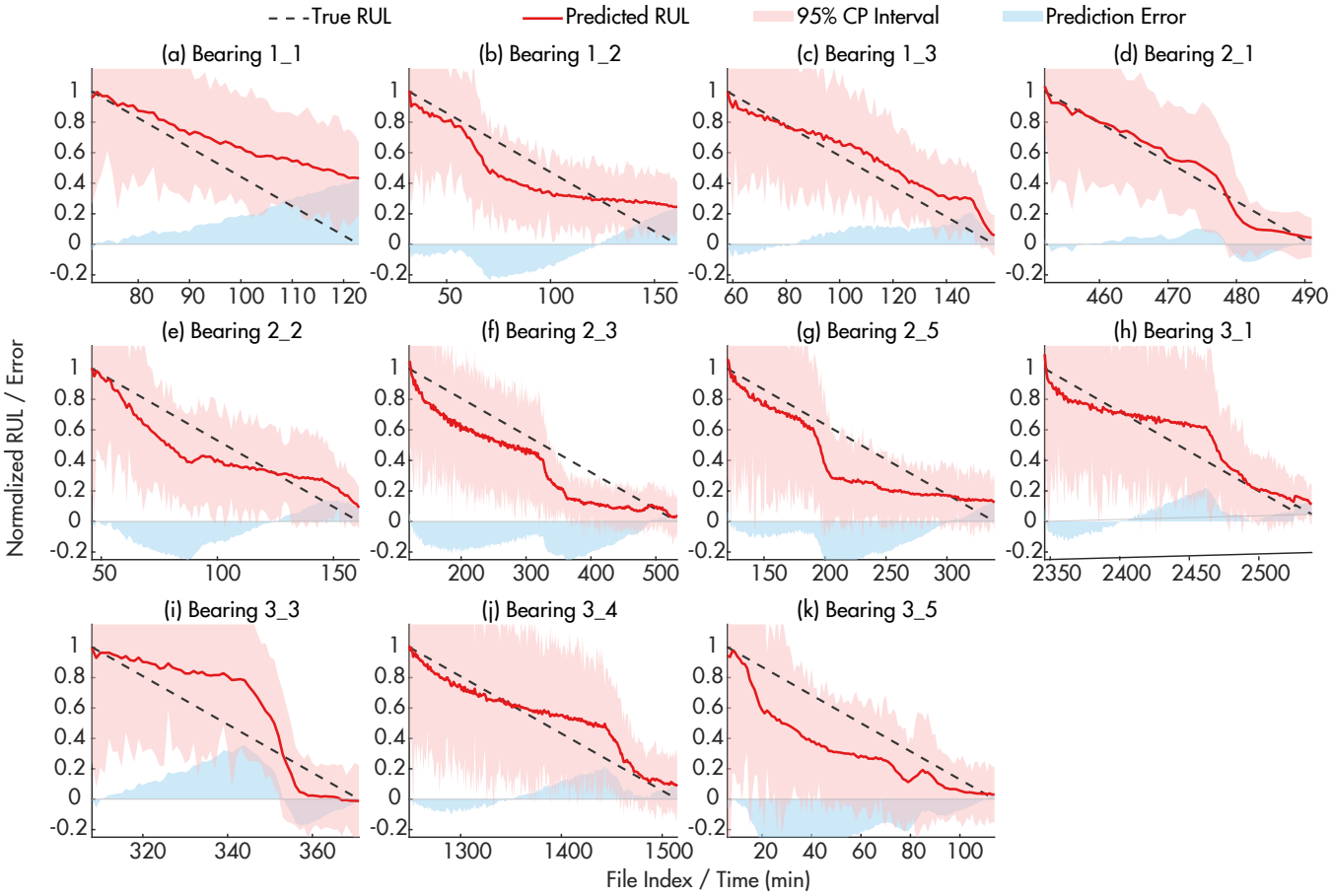

Figure 4. RUL prediction results and calibrated 95% prediction intervals for the test bearings in the XJTU-SY dataset. The red line denotes the predicted mean, the dashed black line denotes the true RUL, the shaded red region indicates the calibrated interval, and the shaded blue region at the bottom represents the prediction error.

tion is often sharper because of the higher rotational speed. Overall, the results indicate that the proposed logarithmic-envelope-spectrum-based health indicator provides a clear degradation trend and is robust to noise across different failure modes

3.3. RUL Prediction Results

Point prediction performance is evaluated using root mean squared error (RMSE) and mean absolute error (MAE) on the normalized RUL scale. Interval quality is evaluated using the prediction interval coverage probability (PICP) and the mean prediction interval width (MPIW):

$$\mathrm{PICP} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[y_i \in \hat{C}(x_i)\right], \tag{16}$$

$$\mathrm{MPIW} = \frac{1}{N}\sum_{i=1}^{N} (\hat{u}_i - \hat{l}_i), \tag{17}$$

where $\hat{u}_i$ and $\hat{l}_i$ denote the upper and lower bounds of the prediction interval, respectively. A desirable probabilistic model should achieve a coverage rate close to or above the nominal target while keeping the interval width reasonably small.

Figure 4 illustrates representative prediction trajectories and calibrated uncertainty bounds for several test bearings. Overall, the proposed framework produces smooth and physically consistent RUL decay trends that closely follow the ground truth. The calibrated prediction intervals successfully capture most of the true trajectories, indicating reliable uncertainty quantification across different degradation patterns.

A comprehensive quantitative comparison is presented in Table 1. To ensure a fair evaluation, the same physics-informed Transformer predictor is adopted for both RMS and SDAE-based features. In addition, an ablation variant using the proposed LogES-VAE feature without the monotonicity constraint is included to isolate the contribution of the physics-informed modeling strategy.

The results show that the proposed configuration, namely

Table 1. Comparison of RUL prediction performance (RMSE / MAE) under different feature representations and physics constraints.

Bearing	RMS + PI-Transformer		SDAE + PI-Transformer		LogES-VAE + Transformer		LogES-VAE + PI-Transformer	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Bearing 1_1	0.2180	0.1781	0.2283	0.1909	0.2121	0.1756	0.2246	0.1849
Bearing 1_2	0.1889	0.1521	0.1372	0.1150	0.1432	0.1159	0.1368	0.1188
Bearing 1_3	0.1737	0.1497	0.2191	0.1849	0.1204	0.1064	0.1003	0.0873
Bearing 2_1	0.1712	0.1469	0.0507	0.0352	0.0750	0.0697	0.0589	0.0489
Bearing 2_2	0.1529	0.1325	0.1310	0.1041	0.1246	0.1117	0.1305	0.1110
Bearing 2_3	0.1121	0.0968	0.1727	0.1532	0.1537	0.1387	0.1534	0.1386
Bearing 2_5	0.2268	0.1971	0.0935	0.0818	0.1416	0.1210	0.1513	0.1250
Bearing 3_1	0.1572	0.1350	0.1076	0.0929	0.0798	0.0707	0.0949	0.0780
Bearing 3_3	0.2230	0.1990	0.1929	0.1648	0.1665	0.1365	0.1822	0.1517
Bearing 3_4	0.0878	0.0771	0.0705	0.0591	0.1124	0.0934	0.0926	0.0759
Bearing 3_5	0.2566	0.2087	0.2773	0.2234	0.2419	0.2006	0.2005	0.1736
Average	0.1789	0.1521	0.1528	0.1278	0.1428	0.1218	0.1387	0.1176

LogES-VAE combined with the physics-informed Transformer, achieves the best overall performance, yielding the lowest average RMSE (0.1387) and MAE (0.1176) among all compared settings. Compared with traditional RMS-based inputs, the learned spectral representation provides a clearer degradation signal that improves prediction stability, particularly in the early degradation stage.

Furthermore, comparing the two LogES-VAE variants highlights the effectiveness of the monotonicity constraint. Removing physics regularization increases the average RMSE from 0.1387 to 0.1428, suggesting that purely data-driven training is more susceptible to local fluctuations in the health indicator. By explicitly enforcing irreversible RUL evolution, the proposed constraint improves both prediction robustness and physical consistency.

Overall, these results demonstrate that the integration of physically interpretable feature construction and physics-informed temporal modeling leads to more accurate and reliable bearing RUL prediction.

3.4. Uncertainty Calibration

The reliability of predictive uncertainty is evaluated using interval coverage metrics reported in Table 2. When uncertainty bounds are constructed directly from the MC Dropout standard deviation using a fixed Gaussian scaling factor ($z = 1.96$), severe miscalibration is observed. The average raw PICP is only 0.4835, indicating that the variational approximation substantially underestimates prediction uncertainty and produces overly narrow intervals. After applying conformal calibration, interval reliability improves significantly. By calibrating prediction errors using pooled historical degradation data, the average PICP increases to 0.9445, and the majority of test bearings achieve coverage close to or exceeding the nominal 95% target. This improvement is accompanied by a wider average interval width (MPIW increases from 0.2649 to 0.9238), reflecting the inherent trade-off between interval sharpness and statistical validity.

Table 2. Interval uncertainty metrics before and after conformal calibration across all test bearings. PICP: prediction interval coverage probability (target ≥ 0.95); MPIW: mean prediction interval width. Bold indicates post-calibration PICP meeting the nominal 95% target.

Bearing	$\hat{q}$	Raw ($z = 1.96$)		Conformal (95%)	
		PICP	MPIW	PICP	MPIW
Bearing 1_1	5.6221	0.4151	0.2756	0.8113	0.7905
Bearing 1_2	5.7423	0.4154	0.1962	0.9308	0.5749
Bearing 1_3	5.9091	0.5248	0.2282	**0.9703**	0.6880
Bearing 2_1	5.6938	0.8500	0.2248	**1.0000**	0.6531
Bearing 2_2	6.5158	0.4138	0.2045	**0.9914**	0.6798
Bearing 2_3	6.4720	0.2681	0.1714	0.7778	0.5660
Bearing 2_5	7.8286	0.4682	0.1751	0.9227	0.6996
Bearing 3_1	7.5784	0.6891	0.2155	**0.9948**	0.8333
Bearing 3_3	6.8350	0.3750	0.2649	**1.0000**	0.9238
Bearing 3_4	7.5162	0.7154	0.2315	**1.0000**	0.8877
Bearing 3_5	10.3310	0.1835	0.1551	**0.9908**	0.8173
Mean	6.9312	0.4835	0.2130	0.9445	0.7376

Despite the overall improvement, the uncertainty behavior is not uniform across bearings. Bearing 2_3 and Bearing 1_1 achieve post-calibration PICPs of 0.7778 and 0.8113, respectively, both below the nominal 95% target. In contrast, Bearings 2_1, 3_3, and 3_4 reach nearly perfect coverage, although with wider calibrated intervals. This variability is informative rather than incidental. The conformal intervals are calibrated from historical degradation trajectories, and their width reflects both the magnitude of the raw predictive uncertainty and the mismatch between the current trajectory and the calibration distribution. When a test bearing exhibits abrupt, irregular, or non-stationary degradation behavior that is weakly represented in the pooled calibration set, the associated nonconformity scores become larger and wider intervals are required to preserve reliability.

These results clarify the meaning of uncertainty in the proposed framework. A narrow interval is not necessarily preferable if it fails to contain the true RUL trajectory, as observed for the raw Monte Carlo dropout intervals. Conversely, a wider calibrated interval may be appropriate when the degradation process is less regular or less similar to previously observed trajectories. In this sense, calibrated interval width acts as an indicator of trajectory difficulty: smoother and more representative degradation patterns tend to admit tighter intervals, whereas atypical trajectories require larger uncertainty bounds. Therefore, the proposed model is uncertainty-aware not because it always generates narrow intervals, but because it corrects overconfident estimates and reveals when the prediction task itself is less certain. This observation also suggests that future trajectory-aware or locally weighted conformal calibration may improve the balance between coverage and sharpness under heterogeneous operating conditions.

4. CONCLUSION

This paper presented a unified physics-informed uncertainty-aware framework for bearing remaining useful life prediction. A logarithmic-envelope-spectrum-based health indicator is constructed using unsupervised variational learning to capture weak degradation signatures, and a Transformer predictor with a differentiable monotonicity constraint is used to enforce physically consistent RUL evolution. Predictive uncertainty is first estimated using Monte Carlo dropout and then calibrated through conformal prediction to obtain statistically reliable prediction intervals.

The experimental results on the XJTU-SY run-to-failure dataset show that the proposed framework improves point prediction performance relative to controlled feature and model ablations. More importantly, the uncertainty analysis demonstrates that raw Monte Carlo dropout intervals are severely overconfident, while conformal calibration substantially improves coverage reliability. The accompanying increase in interval width is not merely a degradation in sharpness, but reflects the heterogeneity of degradation trajectories and the difficulty of making reliable predictions for irregular or weakly represented cases. These findings highlight that useful prognostic uncertainty should not be judged only by how narrow the interval is, but by whether it faithfully indicates when predictions are more or less trustworthy.

Future work will investigate time-varying operating conditions and cross-domain scenarios. In addition, adaptive conformal prediction strategies that account for trajectory similarity or degradation stage may enable tighter and more informative prediction intervals while preserving reliability in practical industrial applications.

ACKNOWLEDGMENT

This work is partially funded by grant PID2021-122132OB-C21 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", by the "European Union".

REFERENCES

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2022). Conformal risk control. *arXiv preprint arXiv:2208.02814*.

Cao, W., Meng, Z., Li, J., Wu, J., & Fan, F. (2024). A remaining useful life prediction method for rolling bearing based on tcn transformer. *IEEE Transactions on Instrumentation and Measurement*.

Cuesta, J., Leturiondo, U., Vidal, Y., & Pozo, F. (2025). A review of prognostics and health management techniques in wind energy. *Reliability Engineering & System Safety*, *260*, 111004.

Dong, S., Xiao, J., Hu, X., Fang, N., Liu, L., & Yao, J. (2023). Deep transfer learning based on bi-lstm and attention for remaining useful life prediction of rolling bearing. *Reliability Engineering & System Safety*, *230*, 108914.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Guo, W., Li, X., & Wan, X. (2023). A novel approach to bearing prognostics based on impulse-driven measures, improved morphological filter and practical health indicator construction. *Reliability Engineering & System Safety*, *238*, 109451.

Hou, B., Wang, D., Chen, Y., Wang, H., Peng, Z., & Tsui, K.-L. (2022). Interpretable online updated weights: Optimized square envelope spectrum for machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, *169*, 108779.

Mian, Z., Deng, X., Dong, X., Tian, Y., Cao, T., Chen, K., & Al Jaber, T. (2024). A literature review of fault diagnosis based on ensemble learning. *Engineering Applications of Artificial Intelligence*, *127*, 107357.

Ni, Q., Ji, J., Feng, K., Zhang, Y., Lin, D., & Zheng, J. (2024). Data-driven bearing health management using a novel multi-scale fused feature and gated recurrent unit. *Reliability Engineering & System Safety*, *242*, 109753.

Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—a tutorial. *Mechanical systems and signal processing*, *25*(2), 485–520.

Wang, B., Lei, Y., Li, N., & Li, N. (2018). A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability*, *69*(1), 401–412.

Wang, S., Vidal, Y., & Pozo, F. (2025a). Recent advances in wind turbine condition monitoring using scada data: A

state-of-the-art review. *Reliability Engineering & System Safety*, 111838.

Wang, S., Vidal, Y., & Pozo, F. (2025b). An unsupervised approach to early fault detection and performance degradation assessment in bearings. *Advanced Engineering Informatics*, *68*, 103620.

Wei, Y., Wu, D., & Terpenny, J. (2021). Learning the health index of complex systems using dynamic conditional variational autoencoders. *Reliability Engineering & System Safety*, *216*, 108004.

Zhao, D., Cai, W., & Cui, L. (2024). Adaptive thresholding and coordinate attention-based tree-inspired network for aero-engine bearing health monitoring under strong noise. *Advanced Engineering Informatics*, *61*, 102559.

Zhong, J., Wang, D., Guo, J., Cabrera, D., & Li, C. (2020). Theoretical investigations on kurtosis and entropy and their improvements for system health monitoring. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–10.

Zhou, J., Yang, J., Xiang, S., & Qin, Y. (2025). Remaining useful life prediction methodologies with health indicator dependence for rotating machinery: A comprehensive review. *IEEE Transactions on Instrumentation and Measurement*.

Zhou, X., Chen, B., Gui, Y., & Cheng, L. (2025). Conformal prediction: A data perspective. *ACM computing surveys*, *58*(2), 1–37.

BIOGRAPHIES

Shun Wang received the master degree in Aerospace Science and Technology in 2023, following a B.Eng. in Aircraft Control and Information Engineering in 2020, both from Northwestern Polytechnical University, Xi'an, China. He is currently a PhD student and pre-doctoral researcher at Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. His research interests include signal processing, fault diagnosis, condition monitoring, and entropy.

Yolanda Vidal holds a degree in Mathematics (1999) and a PhD in Applied Mathematics (2005) from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. As a Full Professor at UPC and an IEEE Senior Member, she actively engages in multidisciplinary research. Her areas of expertise include condition monitoring, structural health monitoring, fault diagnosis and prognosis, predictive maintenance, machine/deep learning, mathematical modeling, and the application of these disciplines in wind turbine technologies. She serves on the Editorial Board of several international journals, including Engineering Applications of AI (Elsevier), Wind Energy (Wiley), Wind Energy Science (Copernicus), Journal of Vibration and Control (SAGE), Mathematics, Sensors, Energies, Frontiers in Built Environment, and Frontiers in Energy Research. Her prolific contributions are evidenced by more than 80 high-impact journal articles, 23 competitive R+D+I projects, 22 book chapters, 11 books, 3 supervised PhD theses, 1 invention patent, a collaboration contract with an industry partner, and over 135 conference papers.

Francesc Pozo obtained his degree in mathematics from the University of Barcelona in 2000, and his PhD in applied mathematics from the Universitat Politècnica de Catalunya (UPC) in 2005. Since 2000, he has been with the Department of Mathematics at UPC, where he is now a Full Professor and the coordinator of the Control, Data, and Artificial Intelligence research group. His expertise encompasses condition monitoring, control systems, data-driven modeling, identification, and structural health monitoring, with a special focus on wind turbines. He is an Editorial Board Member for several international journals, such as Structural Control and Health Monitoring, International Journal of Distributed Sensor Networks, Mathematical Problems in Engineering, Mathematics, Sensors, Algorithms, Journal of Vibration and Control, Frontiers in Built Environment, Frontiers in Energy Research, and Energies. His contributions to his field include over 70 high-impact journal articles, participation in 23 competitive R&D&I projects, authorship of 34 book chapters and 12 books, mentorship of 6 PhD candidates, the filing of 1 invention patent, a collaboration contract with an industry partner, and more than 130 conference papers.