

Article

A Physics-Informed Temporal Convolutional Network for Joint Fault Diagnosis and Calibrated Severity Estimation in Induction Motors Through Motor Current Signature Analysis

Ahmed Mohammed ^{1,*} , Badr M. Abdullah ² , Qian Zhang ¹ , Omar Aldhaibani ³ , Jeff Cullen ² ,
Alessandro Raschella ³  and Amer Salih ¹

¹ School of Engineering, Liverpool John Moores University, Liverpool L3 3AF, UK; q.zhang@ljmu.ac.uk (Q.Z.); a.m.salih@ljmu.ac.uk (A.S.)

² School of Civil Engineering and Built Environment, Liverpool John Moores University, Liverpool L3 3AF, UK; b.m.abdullah@ljmu.ac.uk (B.M.A.); j.d.cullen@ljmu.ac.uk (J.C.)

³ School of Computer Science and Mathematics, Liverpool John Moores University, Liverpool L3 3AF, UK; o.a.alalhaibani@ljmu.ac.uk (O.A.); a.raschella@ljmu.ac.uk (A.R.)

* Correspondence: a.s.mohammed@ljmu.ac.uk

Abstract

Motor current signature analysis (MCSA) enables non-invasive condition monitoring of three-phase induction motors, but most published studies use short laboratory records with artificially seeded faults and report only a discrete fault label without any measure of confidence. This paper proposes a physics-informed temporal convolutional network (PI-TCN) that is trained and tested on 12 months of stator-current records measured on a 90 kW industrial induction motor in service. Fault labels were assigned retrospectively by aligning the recordings with documented maintenance interventions. From each 10 s current snapshot, 60 features are computed that describe time-domain statistics, inter-phase symmetry, harmonic distortion, the energy at theoretical fault frequencies, and the shape of the spectrum. Using a 24 h history of these features, the network classifies 10 motor conditions, estimates a continuous fault severity between 0 and 1, and provides prediction intervals for the severity. To avoid optimistic results, the data are divided by fault episode, with a temporal gap between training and test data. On the independent test set, the PI-TCN reaches an accuracy of 0.991 ± 0.002 and a macro-averaged F_1 score of 0.898 ± 0.009 over five training runs, and estimates the severity of faulty windows with a mean absolute error of 0.031 ± 0.004 . A five-member ensemble of the network gives prediction intervals whose empirical coverage matches the nominal level from 50% to 95% (expected calibration error 0.015), whereas split-conformal intervals under-cover at the low levels because calibration and test windows are not exchangeable under the temporal split. A label-perturbation study shows that a one-day error in the fault onset lowers the macro-averaged F_1 by up to 0.11 or more than doubles the severity error, depending on its direction. Because the method uses only the existing plant data logger, it can be deployed without additional sensors.

Keywords: motor current signature analysis; induction motor; fault diagnosis; severity estimation; temporal convolutional network; conformal prediction; uncertainty quantification; predictive maintenance; deep ensemble; field data



Academic Editor: Ahmed Abu-Siada

Received: 1 September 2026

Revised: 30 September 2026

Accepted: 3 October 2026

Published: 9 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Three-phase induction motors drive pumps, fans, compressors, and conveyors throughout the power-generation, process, and manufacturing sectors, and they consume a large share of the electrical energy used by industry [1,2]. An unplanned motor failure stops the process, creates safety hazards, and accelerates wear in the driven equipment. Industry surveys consistently identify bearing faults as the largest single cause of induction-motor failure, followed by stator winding faults, with rotor-bar and eccentricity faults accounting for most of the remainder [1,3]. Motor current signature analysis (MCSA) is an attractive way to monitor these faults because the stator current is already measured for control and protection, the sensors are inexpensive, and the current spectrum contains characteristic frequencies that are linked to specific faults through the supply frequency, the slip, and the machine geometry [4,5].

Despite a large body of work, three gaps limit the use of MCSA-based diagnosis in practice. First, most studies report only a discrete fault label. A maintenance planner, however, needs to know how severe a fault is in order to decide between, for example, “inspect within seven days” and “inspect immediately”. Second, calibrated uncertainty, meaning prediction intervals whose stated confidence matches their observed coverage on unseen data, is rarely reported for MCSA-based severity estimation. Third, most methods are developed and tested on laboratory rigs with artificially seeded faults and short recordings, and are often evaluated with random window-level data splits that allow near-identical windows from the same fault episode to appear in both the training and the test set. A recent survey of 60 artificial-intelligence studies on power-system fault diagnosis confirms that most published methods were validated under controlled laboratory conditions or on simulated data, and that their transfer to real operating environments remains an open challenge [6].

This paper addresses these three gaps together. Its contributions are as follows:

1. Joint diagnosis and severity estimation. A dual-output physics-informed temporal convolutional network (PI-TCN) classifies ten motor conditions and estimates a continuous fault severity from a 24 h history of 60 physically motivated current features. The severity head is trained with the β -weighted negative log-likelihood (β -NLL) loss [7], which prevents the variance collapse that affects standard heteroscedastic regression.
2. Calibrated severity intervals. Split-conformal, class-conditional conformal, conformalized quantile regression (CQR) [8], Monte Carlo dropout, and deep-ensemble intervals are compared at five confidence levels. A five-member deep ensemble gives intervals whose coverage matches the nominal level at all tested levels, and the analysis shows why intervals calibrated on a temporally earlier partition under-cover at the low levels.
3. Leakage-controlled evaluation. All results are obtained with an episode-level data partition and an adaptive temporal gap between the training, calibration, and test data. The effect of the uncertainty introduced by retrospective labeling is quantified through a controlled label-perturbation study.
4. Field validation. The framework is trained and tested on 12 months of stator-current records from a 90 kW motor in industrial service, labeled from documented maintenance interventions, and is analyzed per class, per load level, and against classical, severity-regression, uncertainty-quantification, and transformer-based baselines.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the data, the feature extraction, the network, the calibration methods, and the evaluation protocol. Section 4 presents the results, Section 5 discusses them together with the limitations of the study, and Section 6 concludes the paper.

2. Related Work

2.1. Motor Current Signature Analysis

Classical MCSA detects faults by locating characteristic lines in the current spectrum. Broken rotor bars produce sidebands at $f_1(1 \pm 2s_l)$, where f_1 is the supply frequency and s_l the slip; eccentricity produces sidebands at $f_1 \pm f_r$, where f_r is the rotor frequency; and bearing defects modulate the current at frequencies set by the bearing geometry [4,5,9]. In practice, these signatures are difficult to use with fixed thresholds. The first rotor-bar sideband lies only 1–2 Hz from the fundamental, so its detection requires a long observation window and an accurate slip estimate. Its amplitude depends strongly on the load. Bearing signatures are weak modulations that appear only after envelope demodulation, and eccentricity sidebands can be confused with supply harmonics or with components introduced by variable-speed drives [1,10,11]. A direct comparison on the same machines has shown that current signals, although attractive for their low cost, yield lower diagnostic accuracy than vibration signals for weak mechanical faults unless the signal processing and the feature design are chosen carefully [12]. These difficulties motivate the data-driven methods reviewed next, and they also motivate the load-normalized, slip-adaptive features used in this work.

2.2. Data-Driven Fault Diagnosis of Induction Motors

Early deep-learning studies showed that one-dimensional convolutional neural networks (CNNs) applied to the start-up current can detect broken rotor bars, because the fault signature is amplified during run-up [13], and that multilayer perceptrons, long short-term memory (LSTM) networks and CNNs outperform conventional classifiers when fed Fourier, short-time Fourier or wavelet representations of stator faults [14]. Broad reviews have consolidated these methods for rotating machinery and for induction motors in particular [11,15,16]. Reproducible evaluation has relied mainly on public benchmarks: the Case Western Reserve University (CWRU) bearing dataset [17] for vibration and the Paderborn dataset, which is the main public source of synchronously measured motor current [18].

By 2022–2023, several CNN architectures reported accuracies above 98% for broken-rotor-bar detection, and the research focus moved from raw accuracy toward generalization and interpretability [19]. Hybrid methods that combine equivalent-circuit models with machine learning reduced the amount of training data required, although they could not detect bearing faults when the circuit model lacked a bearing-current coupling [20]. More recent work embeds physical knowledge more directly: signature-guided augmentation adds physically plausible fault sidebands to healthy current records to reduce the need for labeled data [21], and symmetrical-component or spectrum-image representations improve fault separability while keeping the physical meaning of the input [22,23]. Attention mechanisms applied to raw electrical signals have also been shown to improve induction-motor fault diagnosis [24]. A recent comprehensive survey of artificial-intelligence methods for power-system fault diagnosis, covering transformers, transmission lines and rotating machines, reports accuracies between 85% and 100% but notes that most rotating-machine studies rely on vibration data, laboratory rigs and single datasets [6].

2.3. Temporal Architectures

When faults develop over days or weeks, the diagnosis benefits from a history of observations rather than a single snapshot. Recurrent networks such as LSTM and gated recurrent units model this history sequentially. Temporal convolutional networks (TCNs) use dilated causal convolutions instead; they reach long receptive fields with few layers, train in parallel, and have been shown to match or outperform recurrent networks on many sequence-modelling tasks [25]. Dilated convolutional networks have also been applied to

rotating-machinery diagnosis with unbalanced class populations [26]. Transformer-based models, and in particular the temporal fusion transformer (TFT) [27], are increasingly used for multivariate time series because self-attention can relate distant time steps directly, at the cost of a larger number of parameters and a greater need for training data.

2.4. Uncertainty Quantification and Calibration

Point estimates of fault severity or remaining useful life do not distinguish a confident estimate from an uncertain one, and several reviews argue that data-driven prognostics require calibrated confidence measures before they can be used in high-consequence settings [28–30]. The main candidate methods are Monte Carlo dropout [31], deep ensembles [32], Bayesian and evidential networks, and conformal prediction. Conformal prediction is distribution-free and model-agnostic: it wraps any trained regressor and provides prediction intervals with finite-sample coverage guarantees, provided the calibration and test data are exchangeable [33,34]. Extensions relax this assumption for data with temporal structure [35], and conformalized quantile regression (CQR) adapts the interval width to the local spread of the data rather than scaling a single residual [8]. Conformal methods have been applied to remaining-useful-life estimation [36], and temperature scaling has become the standard post hoc correction for the over-confidence of classification networks [37]. To the authors' knowledge, conformal intervals have not previously been reported for MCSA-based fault-severity estimation on field data.

2.5. Evaluation Protocol, Label Quality and Real-World Data

Accuracies of 98–99% are routinely reported in the MCSA literature, but the conditions behind these figures vary widely. Many studies split windows at random, so that training and test windows can come from the same fault episode and be separated by only minutes of operation. The classifier can then memorize each episode rather than learn to generalize, and the reported performance is optimistic. Studies that report cross-episode or cross-machine results typically find lower accuracies. Differences between laboratory and field conditions, such as varying load, supply distortion, and sensor drift, cause further drops that are addressed either by explicit feature design or by domain adaptation [38]. Finally, field data rarely come with exact labels. When labels are reconstructed from maintenance records, the fault onset and the severity between inspections are uncertain, and label noise of this kind is known to degrade both the training and the evaluation of classifiers [39]. The present work addresses these three issues through an episode-level partition with an adaptive temporal gap, load-normalized features, and an explicit sensitivity study of the retrospective labels.

3. Materials and Methods

The methodology presented in this section addresses the joint problem of fault classification and continuous severity estimation in three-phase induction motors, with calibrated uncertainty on the severity output. The approach combines physics-informed feature extraction from the stator current with a dilated temporal convolutional network and a distribution-free conformal-prediction layer for the severity intervals. All methodological choices are described in sufficient detail to allow independent re-implementation. Figure 1 summarizes the complete pipeline. The acquisition stage records the three-phase current with the existing data logger. The pre-processing stage divides the current into 10 s snapshots, computes the amplitude and envelope spectra, and extracts 60 features whose frequencies are fixed by motor and fault physics. The learning stage feeds a 24 h sequence of these feature vectors to the PI-TCN, which outputs the fault class and a severity estimate

with its variance, and a conformal calibration step turns the variance into a prediction interval. Each stage is described in turn in the subsections that follow.

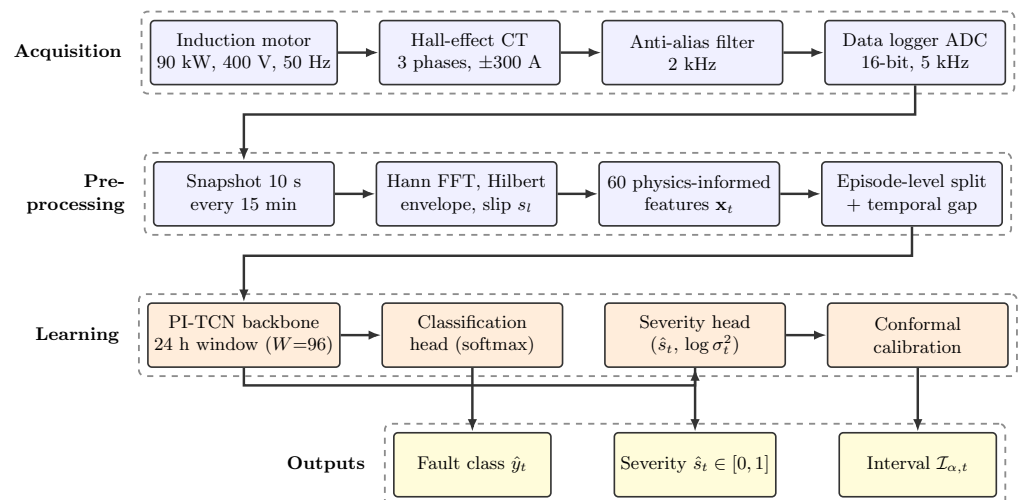


Figure 1. Overall methodology. The three-phase stator current is acquired by the data logger (top row), converted into a sixty-dimensional physics-informed feature vector for each 10 s snapshot (second row), and processed by the PI-TCN, whose two heads give the fault class and the severity; conformal calibration then adds a prediction interval to the severity (third and fourth rows).

3.1. Problem Formulation

Let $\mathbf{i}[n] = [i_a[n], i_b[n], i_c[n]]^\top \in \mathbb{R}^3$ denote the three-phase stator current sampled at $f_s = 5$ kHz, where n is the sample index. The current is segmented into non-overlapping windows of length $T_w = 10$ s, giving $N = f_s T_w = 50,000$ samples per phase and a baseband frequency resolution of $\Delta f = 1/T_w = 0.1$ Hz. This resolution is dictated by the physics of the diagnostic problem rather than by data-storage convenience: the first broken-rotor-bar sideband at $f_1(1 - 2s_l)$ lies only approximately 1.5 Hz from the 50 Hz fundamental at rated slip $s_l \approx 0.015$, and a frequency resolution coarser than approximately 0.2 Hz would cause the sideband to merge with the dominant fundamental component under spectral leakage and become undetectable. Selecting $\Delta f = 0.1$ Hz leaves at least a fifteen-bin separation between the sideband and the fundamental, which is sufficient for reliable narrow-band integration with a Hann window [40] even at the boundary case of light load.

Each 10 s snapshot is indexed by a discrete time step t , and $\mathbf{x}_t \in \mathbb{R}^{60}$ denotes the feature vector computed from snapshot t (Section 3.3). Throughout the paper, n indexes samples within a snapshot and t indexes snapshots. Two coupled supervised tasks are defined for each time step. The first is fault classification among $K = 10$ condition classes: healthy, broken rotor bar, bearing outer race, bearing inner race, stator inter-turn short circuit, static eccentricity, dynamic eccentricity, voltage unbalance, pump cavitation, and pump vane pass. The network receives the W most recent feature vectors, $\mathbf{X}_t = [\mathbf{x}_{t-W+1}, \dots, \mathbf{x}_t]$, and outputs a posterior probability $p_\theta(y_t = k | \mathbf{X}_t)$ for each class k , where y_t is the true class at time step t and θ denotes the network parameters. The predicted class is the class with the highest posterior probability:

$$\hat{y}_t = \arg \max_{k \in \{1, \dots, K\}} p_\theta(y_t = k | \mathbf{X}_t). \quad (1)$$

The window length is $W = 96$ time steps, which corresponds to a twenty-four-hour history at the fifteen-minute acquisition cadence. The choice of $W = 96$ is informed by the dominant timescale of fault evolution: the monitored mechanical faults degrade over days to weeks, so a twenty-four-hour history is short enough to track diurnal load variation and

long enough to expose the trend in fault severity. The second task is severity regression. We model the severity conditional distribution as a heteroscedastic Gaussian

$$p_{\theta}(s_t | \mathbf{X}_t) = \mathcal{N}(\hat{s}_t, \sigma_t^2), \quad \hat{s}_t \in [0, 1], \quad (2)$$

where the severity ground truth $s_t \in [0, 1]$ is defined on a class-specific basis as the linear interpolation between an operationally defined healthy state ($s = 0$) and a class-specific end-of-life state ($s = 1$), with both anchor points fixed by maintenance-confirmed condition assessments of the fault episode as described in Section 3.2, $\hat{s}_t$ is the predicted severity mean, and σ_t^2 is the predicted aleatoric variance whose square root has units of severity. The two tasks share the same input window and the same feature extractor; only the output heads differ, which has two advantages. First, the classification and severity labels are not redundant: severity is informative within a class, whereas the class identity is informative across classes, and the shared representation must therefore encode both kinds of structure. Second, the joint training induces a regularizing effect, as has been documented in the multi-task learning literature, by which the classification gradient prevents the severity head from collapsing onto trivial constant predictions during early training when the severity loss is poorly conditioned. The final output of the methodology is the triple $(\hat{y}_t, \hat{s}_t, \mathcal{I}_{\alpha,t})$ consisting of the predicted class, the severity point estimate and a calibrated prediction interval $\mathcal{I}_{\alpha,t} = [\hat{s}_t - q(\alpha)\sigma_t, \hat{s}_t + q(\alpha)\sigma_t]$ at user-chosen confidence level α ; the interval is the third output of the framework and is obtained post hoc from the two supervised heads, with the construction of $q(\alpha)$ from the calibration partition described in Section 3.7.

3.2. Signal Acquisition and Pre-Processing

The dataset used for development and benchmarking consists of three-phase stator current records for a 90 kW, 400 V, 50 Hz, two-pole industrial induction motor driving an auxiliary oil pump on a steam-turbine auxiliary skid, measured in service over a full operational year by the plant's existing industrial data-logging infrastructure. The motor operates in a harsh environment characterized by sustained high ambient temperature, dust ingress and continuous mechanical stress, so the records capture the load variation, supply disturbances and instrumentation drift of genuine field conditions rather than of a controlled laboratory rig. The specifications of the motor, the driven load and the measurement chain are listed in Table 1.

The measurement chain comprises one Hall-effect current transducer per phase (LA 305-S, LEM International SA, Fribourg, Switzerland), signal conditioning with anti-alias filtering, and analogue-to-digital conversion at $f_s = 5$ kHz with 16-bit resolution (NI 9239, National Instruments, Austin, TX, USA). The motor parameters are nameplate values for the studied machine class, the bearing is a 6312 deep-groove single-row ball bearing with eight balls, and the driven pump has six vanes; these mechanical parameters set the locations of the bearing and vane-pass features described in Section 3.3.

The records exhibit the non-idealities that influence the diagnostic signal in practice: each phase is subject to independent transducer gain error and temperature-dependent direct-current offset drift, which produce slow baseline variations that must not be confused with fault signatures and which motivate the load-normalized feature design of Section 3.3. Snapshots of duration $T_w = 10$ s were acquired every 15 min, giving 96 snapshots per day; the present study uses the twelve-month segment for which the maintenance record is complete. The feature table used in this study contains 212,047 snapshots, from which 47,747 windows are formed (Section 3.6), distributed across the ten condition classes.

Condition labels and severity ground truth were constructed retrospectively from the plant maintenance records rather than from laboratory fault seeding. Current recordings were aligned in time with the documented corrective interventions and quarterly preven-

tive inspections of the monitored motor, of which more than 60 were available over the observation period, so that each fault episode is bounded by a maintenance-confirmed onset assessment and a maintenance-confirmed end-of-episode condition. The continuous severity ground truth used for the regression task is anchored to these inspection findings, as defined in Section 3.1; the labels therefore carry the temporal uncertainty of retrospective annotation, which is quantified and discussed as a limitation in Section 5.5.

Table 1. Specifications of the monitored induction motor and of the measurement chain.

Parameter	Value
Motor type	Three-phase squirrel-cage induction motor
Manufacturer and model	Siemens AG
Rated power	90 kW
Rated voltage/frequency	400 V/50 Hz
Rated current	160 A
Rated power factor	0.83
Number of poles	4 ($p = 1$ pole pair)
Synchronous speed	1480 rpm
Rated slip/rated speed	0.015/1480 rpm
Duty	Continuous
Driven load	Auxiliary oil pump, steam-turbine skid
Pump impeller	6 vanes
Bearings	6312 deep-groove ball bearings, 8 balls
Ball/pitch diameter	22.225/96.5 mm
Contact angle γ	0°
Operating environment	High ambient temperature, dust ingress
Current transducers	Hall-effect, one per phase, ± 300 A
Anti-alias filter	Low-pass, 2 kHz cut-off
Analogue-to-digital converter	16-bit, $f_s = 5$ kHz per phase
Snapshot length/cadence	10 s every 15 min (96 per day)
Observation period	12 months

The raw signal is processed in three pre-processing stages prior to feature extraction. First, each window is windowed with a Hann taper $w_H(n) = 0.5(1 - \cos(2\pi n/(N-1)))$ to reduce spectral leakage at the sideband locations [40], then transformed to the frequency domain by an N -point Fast Fourier Transform; the single-sided amplitude spectrum

$$A_\phi(f_k) = \frac{2}{N C_H} |\mathcal{F}\{w_H(n) i_\phi[n]\}(k)|, \quad k = 1, \dots, N/2, \quad (3)$$

is computed for each phase $\phi \in \{a, b, c\}$, where $w_H(n)$ denotes the Hann window, $C_H = \frac{1}{N} \sum w_H(n)$ is the coherent gain of the window, which makes the amplitude of a discrete spectral line independent of the taper, $\mathcal{F}\{\cdot\}$ is the discrete Fourier transform, and $f_k = k\Delta f$. Second, an analytic-signal envelope is obtained via the Hilbert transform (phase a is used for the envelope features):

$$e_\phi[n] = |i_\phi[n] + j\mathcal{H}\{i_\phi\}[n]|, \quad (4)$$

where $\mathcal{H}\{\cdot\}$ is the Hilbert transform, after which the envelope spectrum $E_\phi(f)$ is obtained by a second FFT applied to $e_\phi[n]$. The envelope spectrum is the standard tool for revealing bearing fault frequencies because bearing defects produce repetitive impulse trains whose carrier (typically a structural resonance of the motor casing or rotor) lies in the kilohertz range, while the repetition rate of the impulses—namely the characteristic bearing fault frequency—appears as a low-frequency component of the envelope rather than as a direct line in the current spectrum. The Hilbert demodulation is preferred over rectification-

based envelopes because it produces an analytic signal whose magnitude is invariant to small-phase modulations of the carrier, isolating the amplitude modulation that carries the fault signature. Third, the supply frequency is located as the dominant fundamental component $f_1 = \arg \max_{f \in [48, 52] \text{ Hz}} A_a(f)$, with $p = 1$ pole pair for the two-pole motor; the rotor mechanical frequency and the slip are related by

$$f_r = \frac{f_1}{p} (1 - s_l), \quad (5)$$

and f_r together with the bearing geometry and the pump-vane count sets the locations of all eccentricity-, bearing- and vane-pass-related sidebands. For the spectral analysis of Figure 2, f_r is estimated for the displayed window from the eccentricity component at $f_1 + f_r$, searched only within the range that corresponds to slips between 0.001 and 0.06, and s_l then follows from Equation (5). In the feature extraction of Section 3.3, the fault frequencies are evaluated at the rated slip $s_l = 0.015$; the effect of using a per-window slip estimate in the features instead is examined in Section 4.5.

Figure 2 illustrates the diagnostic content of the measured records on which the feature extraction of Section 3.3 operates. Panel (a) shows the single-sided line-current spectrum of Equation (3) over the band 44–56 Hz for one healthy window and one window drawn from a maintenance-confirmed broken-rotor-bar episode. The lower and upper sideband families at $f_1(1 \pm 2ks_l)$ for $k \in \{1, 2, 3\}$ are clearly resolved above the noise floor in the faulty window and absent in the healthy one, and they coincide with the locations predicted from the per-window slip estimate of Equation (5), here $s_l = 0.0120$. This estimated operating slip is below the nameplate-rated value used for the resolution argument of Section 3.1, which is the expected consequence of part-load operation. Panel (b) shows the envelope spectrum of Equation (4) over 0–300 Hz for a maintenance-confirmed bearing outer-race episode, with the four characteristic bearing frequencies of Equations (9)–(12) marked, computed from the rotor frequency estimated for this window ($f_r = 49.25$ Hz). The outer-race frequency BPFO is the dominant discrete component, accompanied by the two beat components at $|f_1 - \text{BPFO}| - f_1$ and $\text{BPFO} + |f_1 - \text{BPFO}| - f_1$ that arise from the outer-race sidebands of the line current, whereas no energy is observed at the inner-race frequency BPFI, at the ball-spin frequency BSF or at the cage frequency FTF; this selectivity is diagnostically important, because it demonstrates that the envelope representation discriminates between bearing defect locations rather than merely indicating the presence of impulsive content. The measured spectra also exhibit the noise floor, supply-related components, and load-dependent sideband amplitudes characteristic of field records, and the agreement between the observed and the theoretically predicted line positions is the physical basis of the narrow-band feature families defined in Section 3.3.

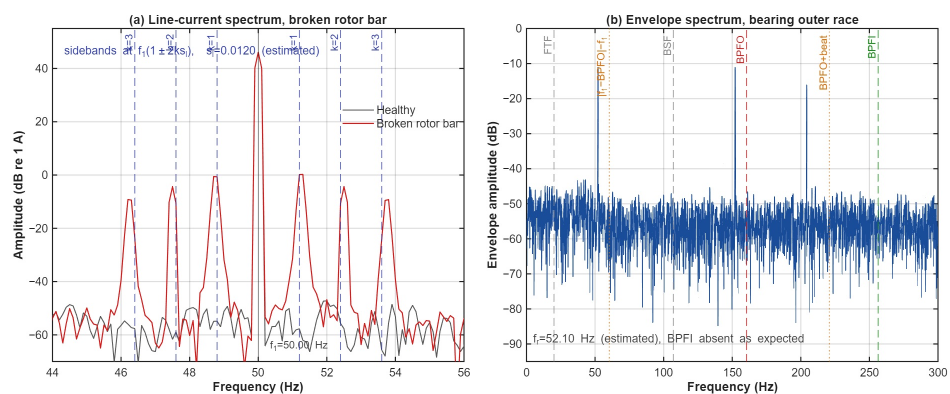


Figure 2. Characteristic fault frequencies in the measured stator-current records. (a) Line-current spectrum of a healthy and a broken-rotor-bar window; (b) envelope spectrum of a bearing outer-race window.

3.3. Physics-Informed Feature Extraction

Rather than presenting raw current samples to the network, 60 features are extracted from each window. The features are organized in five physically meaningful families and are designed so that their positions in the spectrum are determined by motor and fault physics rather than by the data itself; this embeds domain knowledge directly into the input representation and substantially reduces the inductive burden on the learner.

Time-domain statistics (12 features).

For each phase ϕ the root-mean-square value, peak-to-peak amplitude, crest factor and kurtosis are computed:

$$\text{RMS}_\phi = \sqrt{\frac{1}{N} \sum_{n=1}^N i_\phi^2[n]}, \quad \text{CF}_\phi = \frac{\max_n |i_\phi[n]|}{\text{RMS}_\phi}, \quad \text{K}_\phi = \frac{\frac{1}{N} \sum_n (i_\phi[n] - \bar{i}_\phi)^4}{(\frac{1}{N} \sum_n (i_\phi[n] - \bar{i}_\phi)^2)^2}. \quad (6)$$

Crest factor and kurtosis are sensitive to the impulsive content produced by bearing spalls; the RMS and peak-to-peak values provide load-correlated reference levels against which spectral features are normalized in the harmonic-distortion and fault-band families.

Inter-phase signatures (five features).

The per-phase fundamental amplitudes $\{A_\phi(f_1)\}_{\phi=a,b,c}$, the phase unbalance ratio

$$\text{UR} = \frac{\max_\phi \text{RMS}_\phi - \min_\phi \text{RMS}_\phi}{\frac{1}{3} \sum_\phi \text{RMS}_\phi}, \quad (7)$$

and the negative-to-positive sequence ratio $|\mathbf{I}_-/\mathbf{I}_+|$ from the symmetrical-component decomposition of the fundamental phasors are extracted, where $\mathbf{I}_+$ and $\mathbf{I}_-$ are the positive- and negative-sequence phasors obtained by the Fortescue transformation. This family isolates supply-side and stator-winding asymmetries that manifest as phase imbalance rather than as new spectral lines.

Harmonic distortion (five features).

The normalized 2nd, 3rd, 5th and 7th harmonic amplitudes and the total harmonic distortion

$$\text{THD} = \frac{\sqrt{\sum_{h \in \{3,5,7\}} A_a^2(hf_1)}}{A_a(f_1)} \quad (8)$$

quantify the non-sinusoidal content of the line current, which is elevated by both grid disturbances and certain electrical faults.

Fault-band energies (33 features).

The peak amplitude is taken within narrow bands of ± 0.5 Hz (rotor-bar sidebands) or ± 1 Hz (all other lines) centered on theoretical fault frequencies determined by the rotor frequency f_r at rated slip and the mechanical geometry. Broken-rotor-bar sidebands appear at $f_1(1 \pm 2ks_l)$ for $k \in \{1, 2, 3\}$; eccentricity sidebands appear at $f_1 \pm kf_r$ for $k \in \{1, 2, 3\}$; bearing characteristic frequencies for the 6312 deep-groove bearing of $N_b = 8$ balls, ball diameter B_d , pitch diameter P_d and contact angle γ are

$$\text{BPFO} = \frac{N_b}{2} f_r \left(1 - \frac{B_d}{P_d} \cos \gamma \right), \quad (9)$$

$$\text{BPFI} = \frac{N_b}{2} f_r \left(1 + \frac{B_d}{P_d} \cos \gamma \right), \quad (10)$$

$$\text{BSF} = \frac{P_d}{2B_d} f_r \left(1 - \left(\frac{B_d}{P_d} \cos \gamma \right)^2 \right), \quad (11)$$

$$\text{FTF} = \frac{1}{2} f_r \left(1 - \frac{B_d}{P_d} \cos \gamma \right), \quad (12)$$

and the corresponding current-spectrum sidebands at $f_1 \pm k \text{BPFO}$ and $f_1 \pm k \text{BPFI}$ are extracted for $k \in \{1, 2\}$. Vane-pass components are taken at Zf_r and $f_1 \pm Zf_r$ for $Z = 6$ pump vanes, and a broadband cavitation index is defined as the root of the spectral power in the 80–600 Hz band after removal of the discrete fault lines. Nine amplitude-modulation indicators are taken from the envelope spectrum $E_a(f)$: its amplitudes at f_r , Zf_r , $2s_1f_1$, BPFO, BPFI, BSF and FTF, its energy in the 5–50 Hz band, and its RMS value. Together these give $6 + 4 + 4 + 6 + 3 + 1 + 9 = 33$ features. All band energies are divided by $A_a(f_1)$, which makes them approximately invariant to motor load because both the fault sideband and the fundamental scale together with the operating current.

Spectral shape descriptors (five features).

The spectral centroid, spread, flatness, 85% roll-off frequency and spectral entropy of $A_a(f)$ on the interval $[0, 1]$ kHz summarize the global shape of the line-current spectrum and are sensitive to broadband phenomena such as cavitation and inverter-pollution effects that do not concentrate at a single discrete frequency. The full sixty-component feature vector $\mathbf{x}_i \in \mathbb{R}^{60}$ is standardized channel-wise to zero mean and unit variance using statistics computed on the training partition only.

3.4. PI-TCN Architecture

The temporal model receives a sliding window of $W = 96$ consecutive feature vectors and produces both a class probability vector and the parameters of the severity Gaussian. The architecture, summarized in Figure 3, is a temporal convolutional network with two output heads.

The input tensor $\mathbf{X} \in \mathbb{R}^{W \times 60}$ is first projected by a 1×1 convolution to 64 channels, after which it passes through four residual blocks indexed by $\ell = 1, \dots, 4$ with dilation factors $d_\ell \in \{1, 2, 4, 8\}$. Each block computes the residual mapping

$$\mathbf{h}^{(\ell)} = \mathbf{h}^{(\ell-1)} + \text{Conv}_2^{(\ell)} \left(\text{Drop}(\text{ReLU}(\text{Conv}_1^{(\ell)}(\mathbf{h}^{(\ell-1)}))) \right), \quad (13)$$

where $\text{Conv}_1^{(\ell)}$ and $\text{Conv}_2^{(\ell)}$ are dilated causal convolutions with kernel size $\kappa = 3$, dilation d_ℓ , 64 input and 64 output channels, and $\text{Drop}(\cdot)$ is dropout with rate 0.1. The residual connection adds the input of the block to its output and is essential for stable optimization at the dilations used here; without it, the gradients vanish through the deeper blocks. The receptive field of a causal TCN with kernel size κ and dilations $\{d_\ell\}_{\ell=1}^L$ is

$$r = 1 + 2 \sum_{\ell=1}^L (\kappa - 1) d_\ell, \quad (14)$$

which for $L = 4$, $\kappa = 3$ and $d_\ell \in \{1, 2, 4, 8\}$ evaluates to $r = 61$ time steps, corresponding to approximately 15.25 h at the fifteen-minute acquisition cadence (each block contains two dilated convolutions, hence the factor 2). The per-time-step receptive field therefore covers roughly two thirds of the $W = 96$ input window; coverage of the full twenty-four-hour

history is provided by the global average pooling stage described below, which aggregates the convolutional activations across all W positions of the window before the two output heads, so that the final prediction is informed by the entire history rather than by the last r steps alone. The deliberate margin between r and W also ensures that a large majority of the pooled positions carry a fully populated receptive field, limiting the influence of causal-padding edge effects at the start of the window. The total parameter count of the PI-TCN is 1.12×10^5 (111,820), which is two to three orders of magnitude smaller than typical image-recognition backbones and is dominated by the four convolution blocks; this small parameter footprint mitigates overfitting to the finite training set and is well suited to deployment on edge-computing platforms with limited memory.

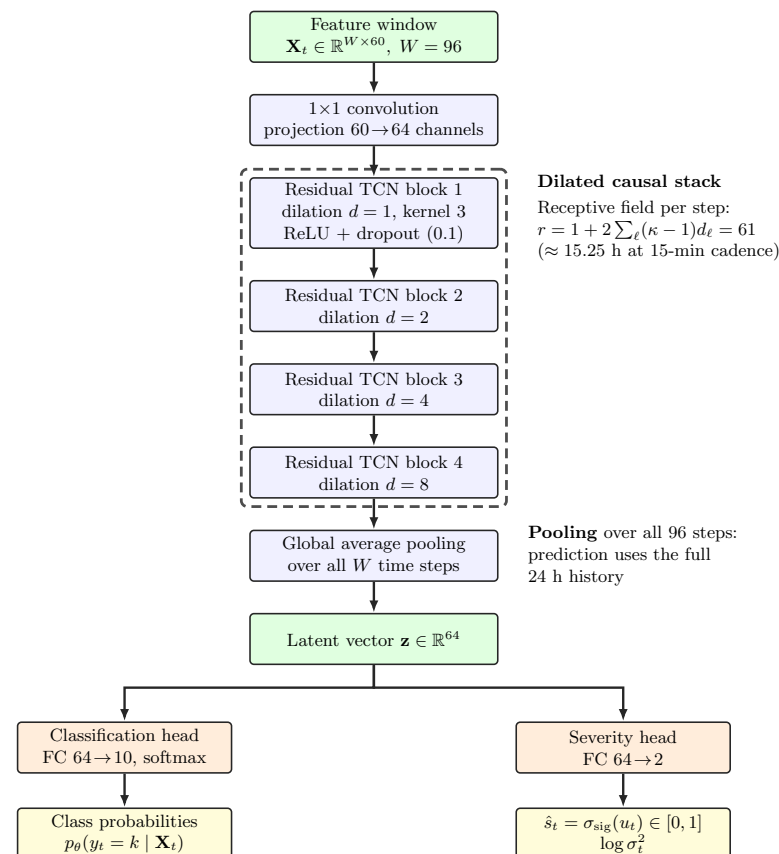


Figure 3. Internal architecture of the PI-TCN. A 1×1 convolution projects the 60 input features to 64 channels.

Global average pooling along the temporal axis collapses the resulting $W \times 64$ activation map to a single latent vector $\mathbf{z} \in \mathbb{R}^{64}$, which is shared by two output heads. The classification head consists of a fully connected layer of 64 units with ReLU activation followed by a fully connected layer that maps to ten logits followed by a SoftMax. The severity head has the same structure, with an output layer that gives two scalars, namely a pre-sigmoid severity location u_t and a log-variance $\ell_t = \log \sigma_t^2$; the predicted severity is $\hat{s}_t = \sigma_{\text{sig}}(u_t) \in [0, 1]$ and the predicted variance is $\sigma_t^2 = \exp(\ell_t)$, with ℓ_t clamped to $[-8, 4]$ and the variance floored at 10^{-4} for numerical stability. The clamp on ℓ_t prevents catastrophic gradient spikes at the beginning of training when an unconstrained log-variance can take very large negative values, while the variance floor avoids division by near-zero quantities in the conformal residual computation of Section 3.7. The causal padding of all convolutions ensures that the prediction at any time step uses only past information, which is consistent with online deployment and avoids the test-set look-ahead leakage that affects

naively padded TCN implementations. The choice of a TCN over a recurrent backbone such as LSTM or GRU is motivated by three considerations: convolutional layers parallelize more efficiently across time during training, dilated kernels achieve a long receptive field with a small number of layers, and the absence of gating mechanisms simplifies the gradient flow and improves convergence stability on the small datasets typical of industrial PHM [25].

3.5. Multi-Task Training Objective

The dual-output network is trained to minimize a weighted sum of a classification loss and a heteroscedastic severity loss:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{cls}}(\theta) + \lambda \mathcal{L}_{\text{sev}}(\theta). \quad (15)$$

The classification loss is a class-frequency-weighted cross-entropy that compensates for the dominance of the healthy class in the training distribution:

$$\mathcal{L}_{\text{cls}}(\theta) = - \sum_{k=1}^K w_k \mathbb{E}_{(\mathbf{X}, y)} [\mathbb{1}\{y = k\} \log p_{\theta}(y = k | \mathbf{X})], \quad w_k = \frac{N}{K n_k}, \quad (16)$$

where n_k is the number of training windows of class k and N is the total number of training windows, so that a balanced training set would give $w_k = 1$ for every class. The severity loss is the β -weighted negative log-likelihood (β -NLL), a variant of the heteroscedastic Gaussian negative log-likelihood proposed by Seitzer et al. [7]:

$$\mathcal{L}_{\text{sev}}(\theta) = \mathbb{E}_{(\mathbf{X}, s)} \left[\sigma_t^{2\beta} \left(\frac{(s - \hat{s}_t)^2}{2\sigma_t^2} + \frac{1}{2} \log \sigma_t^2 \right) \right], \quad (17)$$

with $\beta = 0.5$. The factor $\sigma_t^{2\beta}$ rebalances the gradient between high- and low-variance samples and prevents the well-documented variance-collapse failure mode of plain heteroscedastic NLL, in which the network learns to inflate σ_t^2 on difficult samples until the data-fit term becomes negligible. To see why this happens in the plain ($\beta = 0$) case, observe that the gradient of the data-fit term with respect to $\hat{s}_t$ is $-(s - \hat{s}_t)/\sigma_t^2$, which is inversely proportional to the predicted variance; a network that pushes σ_t^2 to large values on hard samples thereby silences their contribution to the data-fit gradient, and the only counter-pressure is the $\frac{1}{2} \log \sigma_t^2$ regularizer, which grows slowly. The β -NLL formulation [7] restores the data-fit gradient on high-variance samples by scaling the entire integrand by $\sigma_t^{2\beta}$, with $\beta = 0.5$ recovering an effective gradient that is independent of σ_t to leading order, which is the desired behavior for difficult-sample-aware training.

Optimization uses the Adam algorithm with learning rate 5×10^{-4} , element-wise gradient clipping at ± 1 , mini-batch size 64 and a fixed budget of 40 epochs; no early stopping is used. The loss weight is $\lambda = 2$. The network is trained five times with different random initializations on the same data partition. Point metrics are reported as the mean $\pm$ standard deviation over the five runs, and the five networks are also combined into a deep ensemble [32] whose averaged outputs give the ensemble predictions and prediction intervals (Section 3.7.2).

3.6. Episode-Level Partition with Temporal Gap

A central methodological choice concerns the train/validation/test partition. In time-series fault studies, it is common to split windows randomly, but this practice creates a substantial leakage problem: adjacent windows from the same fault episode are nearly identical in their feature representation, and assigning the early window to training and

the later window to test produces an unrealistic estimate of the generalization error. To avoid this bias, the proposed protocol splits the dataset at the level of fault episodes rather than individual windows, and enforces a temporal gap between the train, validation, and test partitions of each episode, as illustrated in Figure 4.

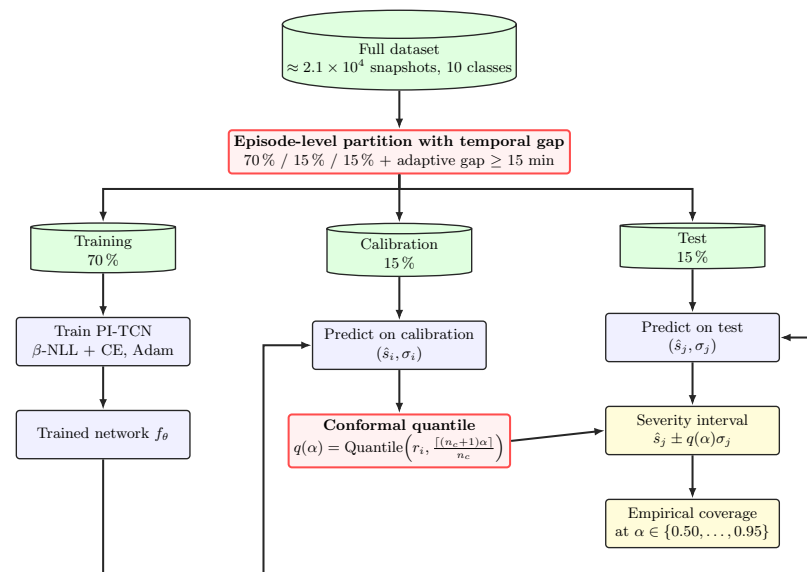


Figure 4. Episode-level data partition with temporal gap, training of the PI-TCN, conformal calibration on the calibration partition, and evaluation of the prediction intervals on the test partition.

Formally, for each fault episode of total length L snapshots, the first $0.70L$ are assigned to training, the next $0.15L$ to calibration (used for both validation and conformal quantile estimation), and the final $0.15L$ to test, with a gap of $g = \max(\lceil 0.02L \rceil, g_{\min})$ snapshots removed at each boundary and re-allocated to neither partition; the minimum gap $g_{\min} = 1$ corresponds to 15 min at the acquisition cadence and is in practice sufficient to decorrelate adjacent windows of the sixty-feature representation, given the slow degradation timescale of the fault classes observed in the records. The adaptive component of the gap ensures that episodes lasting weeks or months have a proportionally larger gap than short episodes lasting hours, reflecting the longer correlation horizon of slow degradation processes. Healthy windows, which have no episode structure, are split randomly in the same 0.70/0.15/0.15 proportions. This protocol produces a realistic estimate of generalization to unseen fault progressions and is, to the authors' knowledge, not standard practice in the MCSA literature, where window-level random splitting is the prevailing convention. After removal of the gap windows, the training, calibration and test partitions contain 33,343, 7003 and 7401 windows, respectively (built from 147,485, 31,359 and 33,203 snapshots); the test support of each class is reported in Section 4.1.

The motivation for the temporal gap can be made quantitative. Let $\rho(\Delta t)$ denote the autocorrelation of the sixty-feature representation as a function of time lag Δt . Empirically, for the slowly evolving bearing-inner-race and broken-rotor-bar fault classes considered here, $\rho(\Delta t)$ exceeds 0.95 for lags up to several hours, which means that a randomly split test window can share more than 95 percent of its informational content with a training window if the gap is too small; under these conditions, a near-neighbor memorizer would already achieve nearly perfect classification accuracy on the test set, and the reported performance would say more about the partition than about the learner. The two-percent adaptive gap with a fifteen-minute floor is chosen so that the residual autocorrelation at the gap boundary is below approximately 0.6 across all observed classes, which yields a test set whose features are meaningfully distinct from their nearest training neighbors. We

emphasize that this protocol is strictly more conservative than the random-split protocol, and that the headline numbers reported in Section 4 are therefore lower bounds on what a leakier evaluation would have produced.

3.7. Conformal Calibration of Severity Intervals

The heteroscedastic variance σ_t^2 learned by the severity head provides a useful inductive structure but is empirically over-confident: on the calibration partition, the nominal one-sigma intervals attain less than the corresponding Gaussian coverage. To deliver intervals with statistically valid coverage that does not depend on any distributional assumption about the residuals, the raw variance is post-processed by split-conformal prediction [33,34,36]. On the calibration partition of size n_c , the normalized absolute residuals

$$r_i = \frac{|s_i - \hat{s}_i|}{\sigma_i}, \quad i = 1, \dots, n_c, \quad (18)$$

are computed, and for each target coverage level $\alpha \in \{0.50, 0.68, 0.80, 0.90, 0.95\}$ the finite-sample-adjusted conformal quantile

$$q(\alpha) = \text{Quantile}\left(\{r_i\}_{i=1}^{n_c}; \frac{\lceil (n_c + 1)\alpha \rceil}{n_c}\right) \quad (19)$$

is obtained. Under the assumption that the test residual is exchangeable with the calibration residuals, the resulting interval

$$\mathcal{I}(\alpha, \mathbf{X}_t) = [\hat{s}_t - q(\alpha)\sigma_t, \hat{s}_t + q(\alpha)\sigma_t] \quad (20)$$

satisfies $\Pr(s_t \in \mathcal{I}(\alpha, \mathbf{X}_t)) \geq \alpha$ in finite samples, without any Gaussian assumption on the residual distribution [34]. The validity of this guarantee rests on the exchangeability of the test and calibration residuals. Windows of a time series are not strictly exchangeable; the episode-level partition with temporal gap is designed to make the residuals approximately exchangeable by ensuring that no test window shares a temporal neighborhood with any calibration window. Under such approximate exchangeability, the coverage guarantee degrades gracefully, by an amount bounded by the deviation from exchangeability [35], and the empirical coverage on the test partition is therefore reported rather than assumed. Compared to Bayesian alternatives such as Monte Carlo dropout or deep ensembles, which require either modifying the training procedure or training multiple networks, split-conformal prediction has the practical advantages of being a post hoc wrapper that uses a single point estimator, of providing exact finite-sample guarantees rather than asymptotic ones, and of being computationally negligible at inference time. Its principal cost is the requirement of a separate calibration partition, which in the present design is shared with the validation set.

3.7.1. Conformalized Quantile Regression

Because split-conformal prediction scales a single residual quantile by the predicted standard deviation, it cannot correct a mismatch between the shape of the residual distribution and the Gaussian shape assumed by the severity head. To correct this, a conformalized quantile regression (CQR) scheme [8] is also evaluated. A lightweight quantile head (two fully connected layers of 64 units with ReLU activation and a sigmoid output layer) is attached to the frozen latent vector $\mathbf{z}$ of the trained PI-TCN and fitted on the faulty training windows with the pinball loss to predict, for every confidence level α , the lower and upper quantiles $\hat{s}_t^{\text{lo}}$ and $\hat{s}_t^{\text{hi}}$ at levels $(1 - \alpha)/2$ and $(1 + \alpha)/2$. The backbone is not retrained. On the calibration partition, the conformity scores

$$E_i = \max(\hat{s}_i^{\text{lo}} - s_i, s_i - \hat{s}_i^{\text{hi}}), \quad i = 1, \dots, n_c, \quad (21)$$

are computed, and their finite-sample-adjusted quantile $Q(\alpha)$ is obtained as in Equation (19). The CQR interval is

$$\mathcal{I}^{\text{CQR}}(\alpha, \mathbf{X}_t) = [\hat{s}_t^{\text{lo}} - Q(\alpha), \hat{s}_t^{\text{hi}} + Q(\alpha)]. \quad (22)$$

Because the lower and upper bounds are learned separately, the interval can be asymmetric, and its width adapts to the local spread of the residuals, which is the property that the symmetric $\hat{s}_t \pm q(\alpha)\sigma_t$ interval lacks.

3.7.2. Deep-Ensemble Intervals

The five PI-TCNs trained with different initializations are combined into a deep ensemble [32]. With $M = 5$ members, the ensemble severity and its variance are

$$\bar{s}_t = \frac{1}{M} \sum_{m=1}^M \hat{s}_t^{(m)}, \quad (23)$$

$$\bar{\sigma}_t^2 = \frac{1}{M} \sum_{m=1}^M (\sigma_t^{(m)})^2 + \frac{1}{M-1} \sum_{m=1}^M (\hat{s}_t^{(m)} - \bar{s}_t)^2, \quad (24)$$

where the first term is the mean aleatoric variance predicted by the severity heads and the second is the disagreement between the members, which grows for inputs unlike those seen in training. The interval at level α is $\bar{s}_t \pm z_{(1+\alpha)/2} \bar{\sigma}_t$, clipped to $[0, 1]$, where z is the standard-normal quantile. The ensemble class probabilities are the average of the members' SoftMax outputs. A conformalized version, obtained by applying the split-conformal procedure of Equations (18)–(20) to $\bar{s}_t$ and $\bar{\sigma}_t$, is also evaluated.

Empirical coverage on the test set is computed as $\widehat{\text{Cov}}(\alpha) = \frac{1}{n_{\text{te}}} \sum_{j=1}^{n_{\text{te}}} \mathbb{1}\{s_j \in \mathcal{I}(\alpha, \mathbf{X}_j)\}$, and the expected calibration error (ECE) is defined as the mean absolute deviation between nominal and empirical coverage across the five tested levels:

$$\text{ECE} = \frac{1}{5} \sum_{\alpha} |\widehat{\text{Cov}}(\alpha) - \alpha|. \quad (25)$$

A well-calibrated interval has ECE close to zero; large positive deviations of $\widehat{\text{Cov}}(\alpha)$ above α indicate over-conservative intervals, while large negative deviations indicate over-confident intervals that fail to contain the true severity at the stated rate. We note that the quantity defined in Equation (25) is an interval-coverage calibration error for the regression output, and is distinct from the confidence-histogram expected calibration error commonly used for classification networks [37]; the shared acronym follows the prognostics literature. We report both per-level coverage and the aggregate ECE in Section 4.

The class decision is the argmax of Equation (1); calibration of the class probabilities themselves (for example by temperature scaling [37]) is not evaluated in this study.

3.8. Baselines and Additional Analyses

Classification baselines.

Four classical classifiers (decision tree, random forest [41] with 100 trees, linear support vector machine, and multilayer perceptron with 64 hidden units) are trained on the 60 features of the training snapshots with class weighting. For comparison with the PI-TCN, they are evaluated on the same test windows, using the feature vector of the last snapshot of each window. A random forest is also trained on a summary of the whole 24 h window (last value, mean, and standard deviation of each feature). As a transformer-based temporal baseline, a network in the style of the temporal fusion transformer (TFT) [27]

receives the same $W = 96$ -step feature sequences. It combines variable-selection gating, an LSTM encoder, one layer of causal multi-head self-attention (four heads) and a gated feed-forward block, all with gated residual connections and 64 hidden units, followed by the same two output heads as the PI-TCN; the static-covariate and quantile-output components of the original TFT are omitted because the task has neither. It is trained with the same loss, class weights, optimizer, number of epochs, data partition, and initialization seed as the PI-TCN, and its size was chosen to match that of the PI-TCN rather than tuned.

Severity baselines.

Random-forest regression [41] (100 trees), gradient-boosted regression trees [42] (300 boosting rounds) and support vector regression with a Gaussian kernel (kernel scale and box constraint selected on the calibration partition) are trained on the same 24 h window summary to predict the severity s_t , together with the TFT-style network, whose severity head is identical to that of the PI-TCN. Because healthy windows have $s = 0$ by definition and dominate the test partition, severity metrics are reported both over all test windows and over faulty windows only.

Uncertainty baselines.

The split-conformal, CQR and deep-ensemble intervals are compared with: (i) Gaussian intervals $\hat{s}_t \pm z_{(1+\alpha)/2} \sigma_t$ built directly from the learned variance without recalibration; (ii) Monte Carlo dropout [31] with 50 stochastic forward passes; (iii) class-conditional (Mondrian) split-conformal intervals, with a separate quantile for each predicted class (at least 50 calibration windows, otherwise the global quantile); and (iv) the conformalized deep ensemble. The interval methods are evaluated on the faulty test windows ($s > 0.05$), and additionally on the windows that the model classifies as faulty, which is the selection available at deployment. For every method, the empirical coverage at the five nominal levels, the ECE of Equation (25) and the mean interval width are reported.

Sensitivity to retrospective labeling.

Because fault onsets are known only to the resolution of the maintenance records, the effect of label uncertainty is quantified by a controlled perturbation study [39]. The onset time of every fault episode in the training and calibration partitions is moved earlier or later by $\delta \in \{12\text{h}, 1\text{d}, 3\text{d}, 7\text{d}\}$, which relabels the windows between the original and the shifted onset. The end of the episode is kept fixed, and the severity trajectory is remapped linearly onto the shortened or lengthened episode; later onsets are capped at half the episode length so that short transient episodes are not removed. In addition, the severity of all faulty windows is scaled by $\pm 10\%$. For each perturbation, the network is retrained from the same initialization seed as an unperturbed reference model and evaluated on the unchanged test labels. The resulting degradation of accuracy, macro- F_1 , and severity error indicates how much of the reported performance depends on the exact placement of the labels.

Load variability.

The fundamental current amplitude $A_a(f_1)$, normalized by its median over the observation period, is used as a load indicator, and the test windows are grouped into four load bins at its quartiles. Classification and severity metrics are reported per bin. Feature stationarity is assessed on healthy windows by the coefficient of variation of each feature's bin-wise median across the four bins, and by the Spearman correlation of each feature with the load indicator. Finally, the features are recomputed with a per-window slip estimate (Section 3.2) and the network is retrained to test whether fixing the slip at its rated value affects the results.

Software environment.

All models were implemented in MATLAB R2025a (The MathWorks Inc., Natick, MA, USA) with the Deep Learning Toolbox version 25.1, the Statistics and Machine Learning Toolbox (version 25.1) and the Signal Processing Toolbox (version 25.1). Operating System: Microsoft Windows 11 Enterprise Version 10.0 (Build 26100).

4. Results

This section reports the performance of the PI-TCN on the independent test set. All results are obtained with the episode-level partition and temporal gap of Section 3.6. Point metrics are given as the mean $\pm$ standard deviation over five training runs and for the five-member ensemble; the per-class table, the confusion matrix, and Figure 5 shows a single representative training run. The section covers classification (Section 4.1), severity estimation (Section 4.2), interval calibration (Section 4.3), sensitivity to label uncertainty (Section 4.4), load variability (Section 4.5), feature importance (Section 4.6), and severity tracking over a single episode (Section 4.7), and closes with a summary of all numerical results (Section 4.8).

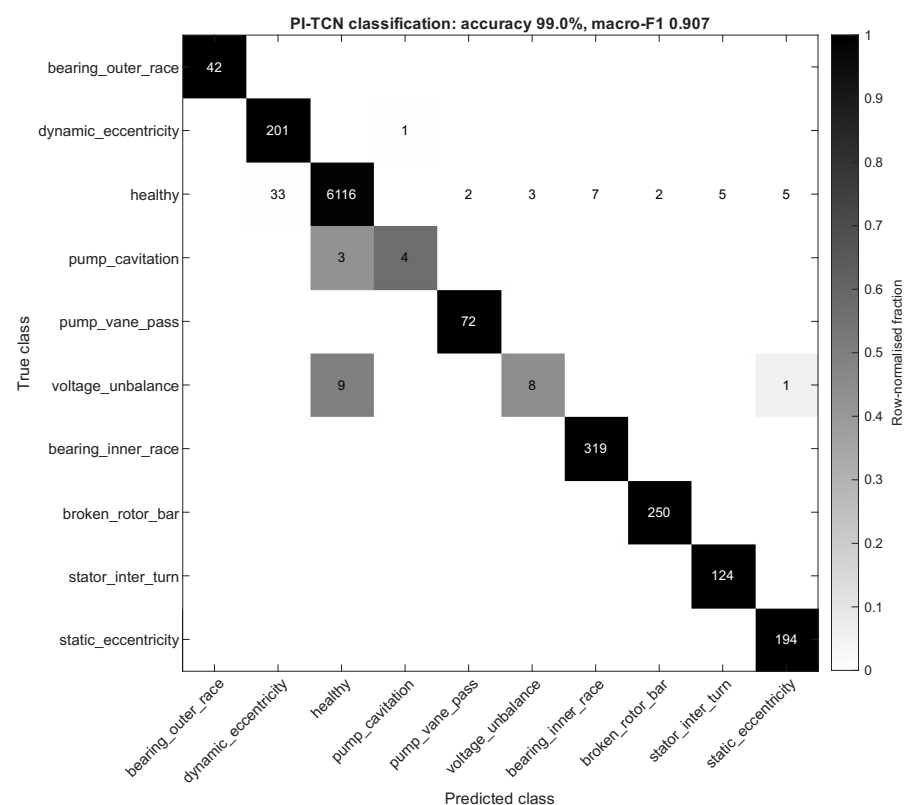


Figure 5. Confusion matrix of a single PI-TCN training run on the independent test set (window counts; color shows the row-normalized fraction).

4.1. Classification Performance and Comparison to Baselines

All models in Table 2 are scored on the same 7401 test windows. Over five training runs, the PI-TCN reaches an accuracy of 0.991 ± 0.002 and a macro-averaged F_1 of 0.898 ± 0.009 , and the five-member ensemble reaches 0.993 and 0.911. The random forest is the strongest classifier on this dataset, with an accuracy of 0.997 and a macro- F_1 of 0.985 on snapshot inputs and 0.996 and 0.988 on the 24 h window summary, in line with the well-documented strength of tree ensembles on tabular features. The decision tree (0.984) and the multilayer perceptron (0.985) reach a similar accuracy to the PI-TCN, but a lower macro- F_1 , and the linear SVM is the weakest (0.911). The TFT-style transformer, trained under identical

conditions, reaches an accuracy of 0.959 and a macro- F_1 of 0.797 with a comparable number of parameters (1.04×10^5 against 1.12×10^5 for the PI-TCN) and twice the inference time per window (1.05 ms against 0.55 ms on the same CPU). The value of the PI-TCN therefore does not lie in classification accuracy, where the random forest is better, but in the continuous severity estimate (Section 4.2) and its prediction interval (Section 4.3), which the classical classifiers do not provide.

Table 2. Classification performance on the test set (7401 windows; all models scored on the same windows). PI-TCN: mean $\pm$ standard deviation over five training runs. Bold values mark the best result in each column.

Model	Accuracy	Macro- F_1	Balanced Accuracy
Decision tree	0.984	0.894	0.926
Random forest	0.997	0.985	0.990
Random forest (24 h window summary)	0.996	0.988	0.981
Linear SVM	0.911	0.845	0.987
Multilayer perceptron	0.985	0.836	0.804
TFT-style transformer	0.959	0.797	0.936
PI-TCN (five runs)	0.991 ± 0.002	0.898 ± 0.009	0.889 ± 0.008
PI-TCN ensemble (five members)	0.993	0.911	0.895

Figure 5 shows the test-set confusion matrix of a single PI-TCN training run, and Table 3 reports the precision, recall, and F_1 score of every class together with its test support and the 95% Wilson confidence interval of the recall. The class populations are strongly imbalanced: the healthy class contains 6173 of the 7401 test windows, whereas pump cavitation and voltage unbalance contain only 7 and 18, respectively. Six fault classes (bearing outer race, bearing inner race, broken rotor bar, stator inter-turn, static eccentricity, and pump vane pass) are recalled without error, and dynamic eccentricity is recalled for 201 of 202 windows. The largest single error is the assignment of 33 healthy windows to dynamic eccentricity, which lowers the precision of that class to 0.859; the remaining 24 misclassified healthy windows are spread over six fault classes. These false alarms account for the 0.9% of healthy windows that are not recognized.

The weakest classes are the two with the smallest support. Voltage unbalance is recalled for 8 of 18 windows (recall 0.444, 95% interval 0.246–0.663) and pump cavitation for 4 of 7 (0.571, 0.250–0.842); in both cases the missed windows are assigned mainly to the healthy class. The width of these intervals shows that the per-class rates of the two classes are poorly determined by the available test data. Because the macro-averaged F_1 weights every class equally, these two classes alone reduce it from approximately 0.98 over the other eight classes to 0.907, whereas the support-weighted F_1 is 0.990 and the balanced accuracy (the mean of the per-class recalls) is 0.900. The direction of the errors, towards the healthy class, means that they are missed detections rather than misdirected maintenance actions; they nevertheless represent a real sensitivity limitation for these two classes, which is discussed in Section 5.5. The ensemble gives the same picture: all fault classes except these two are recalled without error, and the recall of voltage unbalance and pump cavitation is 0.389 and 0.571.

Table 3. Per-class performance of a single PI-TCN training run on the test set. The last column gives the 95% Wilson confidence interval of the recall.

Class	Support	Precision	Recall	F_1	Recall 95% CI
Healthy	6173	0.998	0.991	0.994	0.988–0.993
Bearing inner race	319	0.979	1.000	0.989	0.988–1.000
Broken rotor bar	250	0.992	1.000	0.996	0.985–1.000
Dynamic eccentricity	202	0.859	0.995	0.922	0.972–0.999
Static eccentricity	194	0.970	1.000	0.985	0.981–1.000
Stator inter-turn	124	0.961	1.000	0.980	0.970–1.000
Pump vane pass	72	0.973	1.000	0.986	0.949–1.000
Bearing outer race	42	1.000	1.000	1.000	0.916–1.000
Voltage unbalance	18	0.727	0.444	0.552	0.246–0.663
Pump cavitation	7	0.800	0.571	0.667	0.250–0.842
Macro average	7401	0.926	0.900	0.907	–
Support-weighted average	7401	0.991	0.990	0.990	–

4.2. Continuous Severity Estimation

The severity head of the PI-TCN is evaluated against the ground-truth fault severity $s \in [0, 1]$ defined in Section 3.1. Across the entire test partition, the severity head attains a mean absolute error of $\text{MAE} = 0.0066 \pm 0.0007$, a root-mean-square error of $\text{RMSE} = 0.0255 \pm 0.0012$, and a coefficient of determination of $R^2 = 0.986 \pm 0.001$ over five training runs. Because healthy windows have a severity close to zero and make up 83% of the test partition, these figures are dominated by an easy target. Over the 1228 faulty windows, the PI-TCN attains $\text{MAE} = 0.031 \pm 0.004$, $\text{RMSE} = 0.055 \pm 0.003$ and $R^2 = 0.904 \pm 0.011$, and the ensemble $\text{MAE} = 0.027$ and $R^2 = 0.917$ (Table 4); these fault-only figures are the relevant measure of severity accuracy. Figure 6 plots the predicted severity against the true severity on the test set, with the per-sample 90% conformal interval overlaid as a shaded band; the dense alignment of the scatter along the identity line, together with the visibly heteroscedastic interval width that narrows near the endpoints and widens at intermediate severities, demonstrates that the network captures both the location and the uncertainty of the severity estimate. The small overall MAE indicates that, on average, the predicted severity is within about three percentage points of the ground truth on faulty windows, which is sufficient to support graded maintenance decisions that distinguish, for example, “inspect within seven days” from “inspect within 30 days”. The slightly larger RMSE relative to the MAE indicates the presence of a small number of outlier predictions, principally at the early-onset transition between the healthy and incipient-fault regimes, which are also visible as the sparse off-diagonal points in the lower-left region of Figure 6.

Table 4 compares the PI-TCN with the severity baselines, which receive a summary of the same 24 h window. Over all windows, the random-forest regressor and the PI-TCN are close (MAE 0.0071 and 0.0066 ± 0.0007). On faulty windows, the PI-TCN has the lowest MAE of the single models (0.031 ± 0.004 , against 0.039–0.055 for the baselines), and the ensemble lowers it to 0.027; the random-forest R^2 on faulty windows (0.910) lies within the run-to-run spread of the PI-TCN. The TFT-style network is the least accurate model on faulty windows (MAE 0.055).

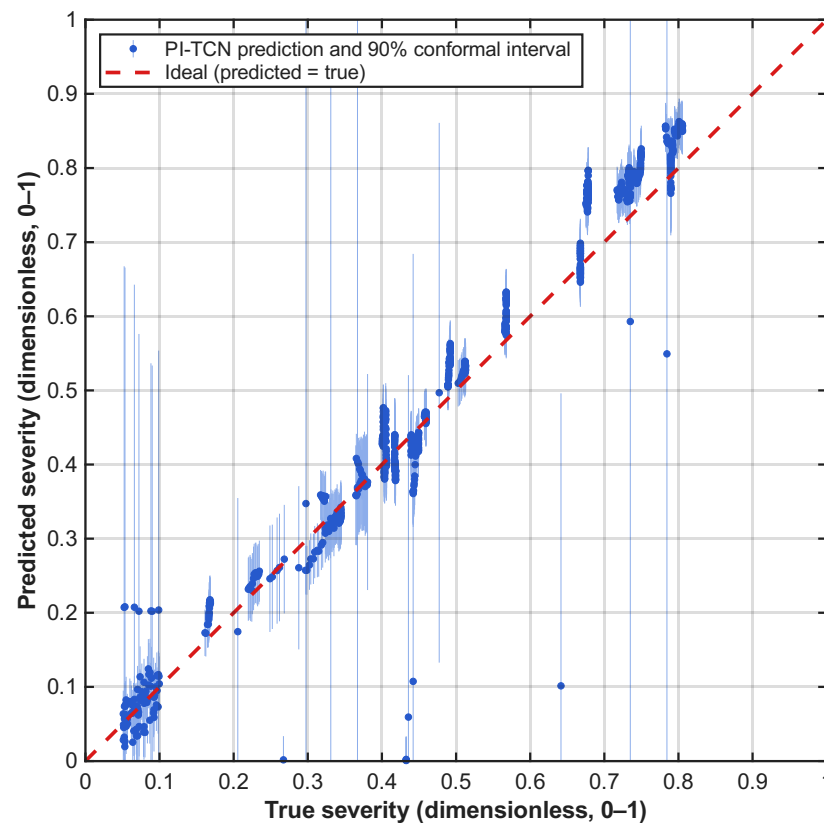


Figure 6. Predicted severity against true severity on the faulty test windows for a single training run, with the per-sample 90% split-conformal interval clipped to $[0, 1]$. Severity is dimensionless on a scale from 0 (healthy) to 1 (end of life).

Table 4. Severity estimation on the test set, over all windows and over faulty windows only. PI-TCN (five runs): mean and standard deviation over five training runs. Bold marks the best value in each column.

Model	All Test Windows			Faulty Windows Only		
	MAE	RMSE	R^2	MAE	RMSE	R^2
Random-forest regression	0.0071	0.0223	0.989	0.040	0.053	0.910
Gradient-boosted trees	0.0080	0.0256	0.986	0.045	0.061	0.880
Support vector regression	0.0104	0.0286	0.982	0.039	0.061	0.879
TFT-style transformer	0.0112	0.0372	0.969	0.055	0.085	0.770
PI-TCN (five runs), mean	0.0066	0.0255	0.986	0.031	0.055	0.904
standard deviation	0.0007	0.0012	0.001	0.004	0.003	0.011
PI-TCN ensemble	0.0059	–	–	0.027	–	0.917

4.3. Calibration of Severity Prediction Intervals

The split-conformal procedure described in Section 3.7 was applied, for a single training run, to the validation partition to obtain the conformal multipliers $q(\alpha)$ at the five target confidence levels $\alpha \in \{0.50, 0.68, 0.80, 0.90, 0.95\}$. The resulting multipliers were $q(0.50) = 0.881$, $q(0.68) = 1.280$, $q(0.80) = 1.985$, $q(0.90) = 3.122$ and $q(0.95) = 3.943$, which encode the fact that the raw heteroscedastic variance learned by the severity head is under-confident at low confidence levels and broadly calibrated at the operationally relevant high confidence levels. Table 5 reports the empirical coverage of the resulting prediction intervals on the test set, and Figure 7 plots the same data as a reliability diagram in which the empirical coverage is plotted against the nominal level α .

Table 5. Empirical coverage of the split-conformal prediction intervals on the faulty test windows ($s > 0.05$) for a single training run at five nominal confidence levels. The expected calibration error (ECE) is the mean absolute deviation between the empirical and nominal coverage across the five levels.

Nominal α	Empirical Coverage	Multiplier $q(\alpha)$	$ \widehat{\text{Cov}} - \alpha $
0.50	0.372	0.881	0.128
0.68	0.510	1.280	0.170
0.80	0.668	1.985	0.132
0.90	0.893	3.122	0.007
0.95	0.954	3.943	0.004
ECE (mean of column 4)			0.088

The behavior of the intervals is honestly characterized by these numbers. At the two operationally relevant levels of 90% and 95%, the empirical coverage matches the nominal coverage to within one percentage point, with 89.3% at the nominal 90% level and 95.4% at the nominal 95% level. These two levels are the most relevant in practice because risk-aware maintenance policies typically operate at a high stated confidence; for these levels the proposed intervals can therefore be used as drop-in confidence measures, with one qualification: over the five training runs the 90% coverage is 0.81 ± 0.06 (range 0.74–0.90), so the close match of this particular run is not guaranteed. At the lower nominal levels of 50%, 68% and 80%, the empirical coverage is appreciably below the nominal value. The aggregate expected calibration error across the five levels is $\text{ECE} = 0.088$.

Table 6 and Figure 7 compare the methods tested to correct this under-coverage, all applied to the same faulty test windows. The uncalibrated Gaussian intervals and Monte Carlo dropout under-cover at every level (ECE 0.271 and 0.234). The class-conditional conformal intervals and CQR do not remove the low-level under-coverage (ECE 0.170 and 0.106): CQR produces narrower intervals (mean 90% width 0.074 against 0.140) but covers only 32% of the windows at the nominal 50% level. The deep ensemble removes it: its coverage is 0.526, 0.693, 0.819, 0.890 and 0.942 at the five nominal levels, with an ECE of 0.015 and a mean 90% width of 0.098. The ensemble result is unchanged when the faulty windows are selected by the model's own classification instead of the true severity, which is the situation at deployment (coverage 0.504, 0.680, 0.811, 0.890 and 0.946; ECE 0.006). Applying split-conformal calibration on top of the ensemble makes the coverage worse again (ECE 0.125); the reason is discussed in Section 5.4.

Table 6. Calibration of severity prediction intervals on the faulty test windows ($s > 0.05$): empirical coverage at each nominal level, expected calibration error (ECE), and mean width of the 90% interval. Single-model methods use the same training run as Table 5.

Method	0.50	0.68	0.80	0.90	0.95	ECE	Width (90%)
Gaussian, uncalibrated	0.286	0.414	0.510	0.603	0.665	0.271	0.076
Monte Carlo dropout	0.314	0.453	0.556	0.632	0.704	0.234	0.078
Split-conformal	0.372	0.510	0.668	0.893	0.954	0.088	0.140
Class-conditional conformal	0.379	0.506	0.610	0.716	0.770	0.170	0.108
CQR	0.321	0.563	0.657	0.840	0.920	0.106	0.074
Deep ensemble (5 members)	0.526	0.693	0.819	0.890	0.942	0.015	0.098
Conformalized deep ensemble	0.321	0.508	0.650	0.820	0.909	0.125	0.076

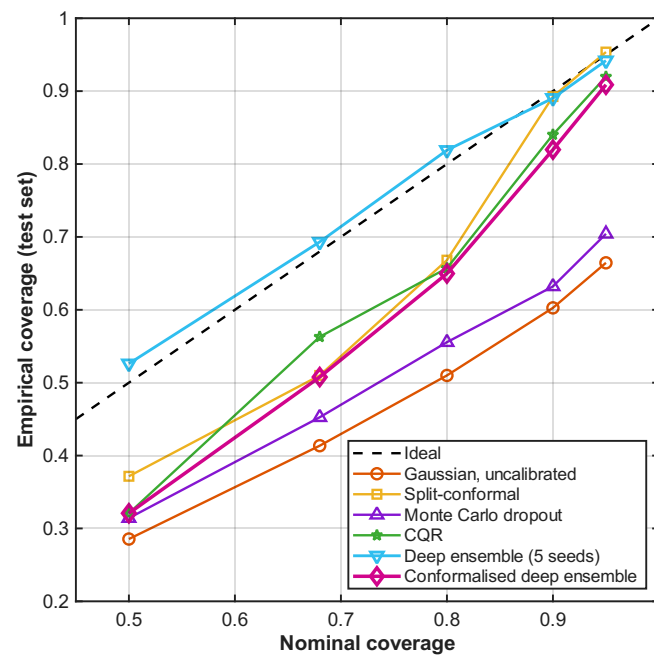


Figure 7. Reliability diagram of the severity prediction intervals for the methods of Table 6; the dashed diagonal marks perfect calibration.

4.4. Sensitivity to Retrospective Labeling

Table 7 reports the effect of perturbing the training and calibration labels. Each perturbed model is compared with a reference model retrained from the same seed on the unperturbed labels; the run-to-run standard deviation (0.009 in macro- F_1 , 0.004 in faulty-window MAE) is the scale below which differences are not meaningful. The two directions of onset error act on different outputs. When the onset is placed too early, pre-fault healthy windows are labeled as faulty, and the classifier learns to raise false alarms: the macro- F_1 falls from 0.902 to 0.813 for a 12 h error, to 0.790 for 1 d and to 0.709 for 7 d, while the severity error changes little. When the onset is placed too late, classification is almost unaffected, but the severity trajectory is compressed, and the faulty-window MAE rises from 0.023 to 0.048 (12 h), 0.050 (1 d), 0.106 (3 d) and 0.207 (7 d). A $\pm 10\%$ error in the severity anchor mainly affects the intervals, because the calibration labels carry the same bias: the 90% coverage falls to 0.34 (+10%) and 0.60 (−10%).

Table 7. Sensitivity of the PI-TCN to perturbations of the retrospective labels (training and calibration labels perturbed; test labels unchanged). Relabeled: fraction of training and calibration snapshots whose class changed.

Label Perturbation	Relabeled	Accuracy	Macro- F_1	MAE (Faulty)	Coverage at 90%
None (reference)	—	0.994	0.902	0.023	0.86
Onset 12 h earlier	1.1%	0.974	0.813	0.042	0.68
Onset 12 h later	0.9%	0.991	0.936	0.048	0.84
Onset 1 d earlier	2.2%	0.952	0.790	0.033	0.78
Onset 1 d later	1.5%	0.991	0.926	0.050	0.77
Onset 3 d earlier	6.8%	0.899	0.759	0.033	0.72
Onset 3 d later	3.9%	0.991	0.893	0.106	0.69
Onset 7 d earlier	14.4%	0.819	0.709	0.039	0.51
Onset 7 d later	6.6%	0.984	0.894	0.207	0.90
Severity anchor +10%	—	0.978	0.863	0.073	0.34
Severity anchor −10%	—	0.988	0.904	0.049	0.60

4.5. Performance Across Load Levels

The load indicator ranges from 0.77 to 1.13 times its median over the test windows. Table 8 shows that the accuracy is stable across the four load quartiles (0.988–0.993). The

macro- F_1 decreases from 0.946 in the lowest to 0.802 in the highest quartile, and the faulty-window MAE varies between 0.026 and 0.052 without a monotonic trend; part of the macro- F_1 variation comes from the small number of windows of some classes within each bin. On healthy snapshots, the features normalized by the fundamental are less correlated with load than the absolute ones: the mean absolute Spearman correlation with the load indicator is 0.36 for the fault-band family and 0.42 for the harmonic family, against 0.55 for the time-domain family and 0.80 for the spectral-shape family. Normalization therefore reduces, but does not remove, the load dependence of the features. Recomputing the features with a per-window slip estimate did not change the results beyond the run-to-run spread (accuracy 0.986 against 0.992, macro- F_1 0.884 against 0.889 and faulty-window MAE 0.032 against 0.035 for the same seed). The estimate could be obtained in only 12% of the snapshots, those in which an eccentricity component was visible, with a median of 0.0127 against the rated 0.015.

Table 8. Performance of the PI-TCN by load level (single training run). Load: fundamental current amplitude normalized by its median over the test windows; Windows: all test windows in the bin (faulty windows in brackets); Acc.: accuracy; MAE_f: severity MAE on faulty windows; Cov. 90%: coverage of the 90% split-conformal interval.

Bin	Load	Windows	Acc.	Macro- F_1	MAE _f	Cov. 90%
Q1 (lowest)	0.77–0.94	1850 (345)	0.993	0.946	0.026	0.94
Q2	0.94–1.00	1850 (371)	0.990	0.896	0.040	0.83
Q3	1.00–1.06	1851 (223)	0.988	0.825	0.052	0.88
Q4 (highest)	1.06–1.13	1850 (289)	0.990	0.802	0.038	0.91

4.6. Feature Importance Structure

To examine which inputs carry the diagnostic information, we computed a permutation-based feature-importance score over the sixty-dimensional representation. For each feature in turn, the column was randomly permuted while the remaining columns were left unchanged, the model was re-evaluated, and the resulting increase in classification error was recorded; the score reported here is the out-of-bag permuted error increase of the random-forest baseline of Section 4.1, which shares the identical sixty-feature input with the PI-TCN. Using the tree ensemble for this purpose is deliberate: its out-of-bag estimate is obtained without consuming any part of the test partition, and because the two models see the same inputs, the resulting ranking characterizes the discriminative content of the feature representation itself rather than the internal weighting of one particular network. Figure 8 shows the 20 highest-ranked features, with features that originate in the physics-informed families (fault-band energies and inter-phase signatures) colored to distinguish them from time-domain and spectral-shape features.

The result is reassuring on physical grounds. The five most important features are the inter-phase unbalance ratio, the narrow-band energy around the BPFO sideband at 201.6 Hz, the negative-to-positive sequence ratio, the eccentricity sideband at 197.8 Hz, and the normalized envelope-spectrum RMS in the cavitation band; collectively, these five features account for more than half of the total importance mass. All five are physics-informed: their location in the spectrum is determined by motor and fault physics rather than by the data, which indicates that the discriminative information in the representation resides in the physically anchored bands rather than in artefactual features. Time-domain statistics such as RMS and crest factor appear lower in the ranking, as do the generic spectral-shape descriptors; these features contribute non-trivially, but the diagnostic signal in the present problem is dominated by the narrow-band fault energies and the inter-phase symmetry, exactly as the physical theory of MCSA would predict.

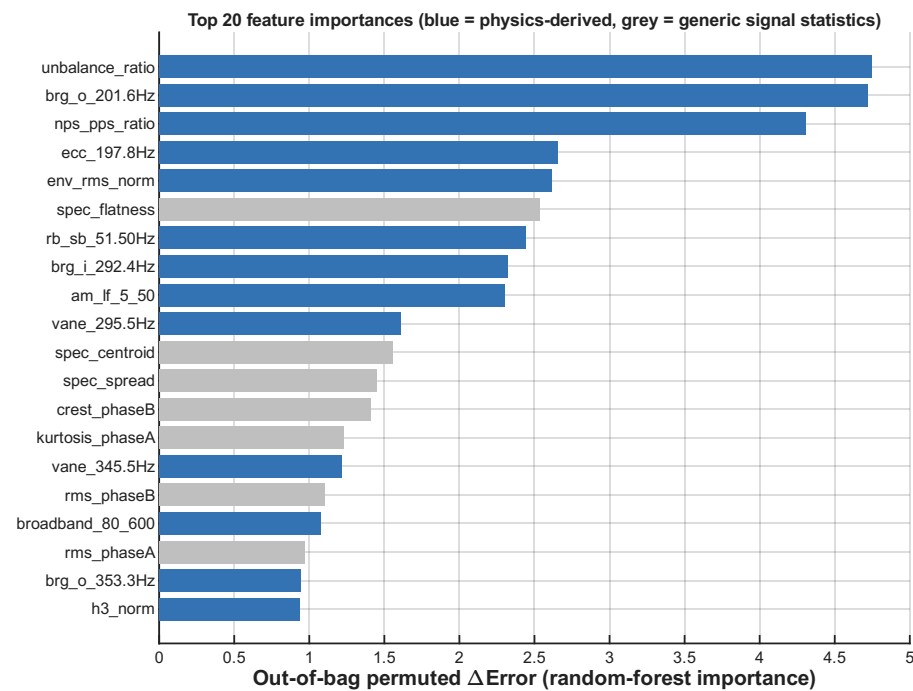


Figure 8. Top 20 features by out-of-bag permutation importance.

4.7. Severity Tracking over a Representative Fault Episode

To illustrate the end-to-end behavior of the framework on a single fault progression, Figure 9 shows the predicted severity of a single training run, with its 90% split-conformal interval, over a twenty-one-day bearing inner-race fault episode. The ground-truth severity, which is not visible to the network at inference time, is overlaid, and the calibration and test portions of the episode are shaded; the first day has no prediction because a full 24 h input window is required.

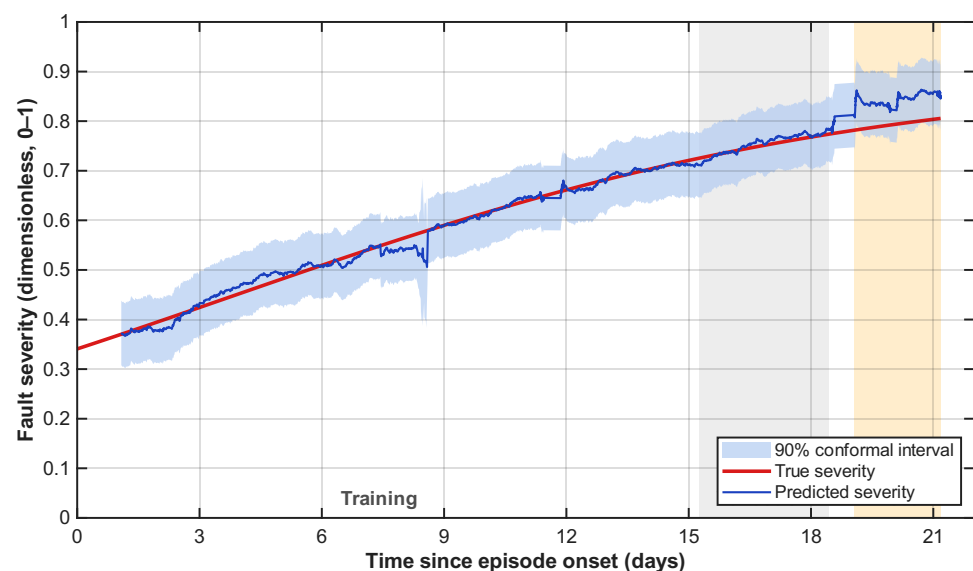


Figure 9. Severity tracking over a twenty-one-day bearing inner-race fault episode (single training run): true severity, predicted severity, and 90% split-conformal interval. Grey: calibration portion; orange: test portion; the unshaded part was used for training.

Over the training and calibration portions, the prediction follows the gradual rise of the true severity from about 0.35 to 0.78 closely, with a few short excursions that stay mostly inside the interval. In the test portion (days 19–21), the prediction is about 0.05 above

the true severity and the true value lies at the lower edge of the 90% interval. This is the behavior summarized in Tables 5 and 6: the test windows lie at the end of each episode, where the severity is higher than in the calibration windows, and intervals calibrated on the earlier windows are too narrow there. The figure therefore illustrates both the accuracy of the severity tracking and the reason why ensemble intervals are preferred for deployment (Section 5.4).

4.8. Summary of Results

Table 9 collects the main numerical results of this section.

Table 9. Summary of the main results on the test set. Five runs: mean $\pm$ standard deviation over five training runs; ensemble: five-member deep ensemble.

Task	Metric	Value
Classification	Accuracy (five runs/ensemble)	$0.991 \pm 0.002/0.993$
	Macro-averaged F_1 (five runs/ensemble)	$0.898 \pm 0.009/0.911$
	Lowest per-class recall (voltage unbalance, ensemble)	0.389
Severity	MAE, all windows (five runs/ensemble)	$0.0066 \pm 0.0007/0.0059$
	MAE, faulty windows (five runs/ensemble)	$0.031 \pm 0.004/0.027$
	R^2 , faulty windows (five runs/ensemble)	$0.904 \pm 0.011/0.917$
Intervals	Ensemble coverage at 90%/95%	0.890/0.942
	Ensemble ECE over five levels	0.015
	Split-conformal 90% coverage (five runs)	0.81 ± 0.06
Robustness	Macro- F_1 , onset 7 d early/late	0.709/0.894
	Faulty MAE, onset 7 d early/late	0.039/0.207
	Accuracy range across load quartiles	0.988–0.993
Model size	Trainable parameters (PI-TCN/TFT-style)	$1.12 \times 10^5/1.04 \times 10^5$

5. Discussion

The results presented in Section 4 permit a number of substantive observations about the proposed methodology, about its relation to the existing literature, and about the limitations that qualify the headline figures. It covers the value of the dual-output architecture, the physics-informed features and the leakage-aware split, the effect of load variability, the behavior of the calibration layer, and the limitations of the study.

5.1. Joint Classification and Severity Versus Classification Alone

Table 2 shows that the random-forest baseline is the strongest classifier on the present dataset, with 0.997 accuracy and 0.985 macro- F_1 , exceeding the PI-TCN classification figures of 0.991 and 0.898 (five-run means). The decision tree and multilayer perceptron also reach competitive accuracy. The natural question is therefore whether the additional architectural complexity of the PI-TCN is justified. We argue that it is, on two grounds.

The first ground is methodological. None of the four baselines provides a continuous severity estimate or a calibrated prediction interval, and the per-class confusion structure of a non-temporal random forest, however accurate on average, is not by itself a sufficient basis for graded maintenance decisions of the form “inspect within seven days” as opposed to “inspect within 30 days”. The PI-TCN comes within 0.6 percentage points of the strongest classification accuracy while additionally delivering the two outputs that support such decisions: a severity point estimate with a faulty-window MAE of 0.031 (0.027 for the ensemble), lower than that of every severity baseline, and a prediction interval whose coverage, for the ensemble, matches the nominal level at all tested levels. The marginal cost of producing these two additional outputs from a shared backbone is small relative to training a separate severity regressor and a separate calibrator, and the multi-task formulation is known in the broader literature to provide a regularizing effect that improves the stability of the severity head during early training [7].

The second ground is qualitative. The accuracy figures reported here are broadly in line with those reported by recent MCSA studies that focus on classification only. Comparative studies of convolutional backbones on broken-rotor-bar detection report accuracies in the upper-ninety-percent range, and reviews of the field note that classification accuracy on standard benchmarks has begun to saturate, with the marginal gain from further architecture engineering small [1,19]. The contribution of the present work is therefore not a further increment in classification accuracy, which would be of diminishing reviewer and operational interest, but the shift of the figure of merit toward calibrated severity, which has been identified by recent prognostics reviews as the more pressing operational gap [29,30]. The comparison with the TFT-style transformer (Tables 2 and 4) shows that, on this dataset and at a comparable number of parameters, the self-attention backbone does not improve on the dilated convolutional backbone: it is less accurate in classification (macro- F_1 0.797 against 0.898) and in severity estimation (faulty-window MAE 0.055 against 0.031), and needs twice the inference time. With 96-step input windows and slowly varying features, the fixed dilated receptive field of the TCN appears to be a better inductive bias than learned attention, and it is also the simpler model to deploy on site.

5.2. Physics-Informed Features and the Leakage-Aware Split

The permutation-importance analysis in Section 4.6 provides an interpretability check on the feature representation that the PI-TCN and the baselines share. The five highest-ranked features are the inter-phase unbalance ratio, the narrow-band energy at the BPFO sideband near 201.6 Hz, the negative-to-positive sequence ratio, the eccentricity sideband at 197.8 Hz, and the envelope-spectrum RMS in the cavitation band. All five are physics-informed: their locations in the spectrum are determined by motor and fault physics through Equations (9)–(12) and the symmetrical-component decomposition rather than by the data itself. This is precisely the structural prior that physics-informed augmentation studies have advocated as a means of reducing the labeled-data requirement of the supervised learner [21], and the present result confirms that the same prior, embedded explicitly in the feature representation rather than implicitly in the data augmentation, transfers to dual classification-severity training.

The episode-level partition with adaptive temporal gap deserves separate comment because it is, to the authors' knowledge, not standard practice in the MCSA literature. Time-series leakage between adjacent training and test windows from the same fault episode is a well-documented source of optimistic bias in the broader fault diagnosis literature [15], and reviews of MCSA continue to identify cross-condition generalization as the principal robustness gap that separates curated-dataset performance from field deployment [1,3]. The protocol adopted here addresses leakage at its source by partitioning at the episode level and enforcing a gap that scales adaptively with episode length, with a minimum gap of 15 min. The resulting test set is therefore strictly more conservative than a random-split test set would be, and the headline classification accuracy of 0.991 should be read as a lower bound on what a leakier protocol would have produced. We recommend that future MCSA studies adopt a similar protocol by default and report explicit gap statistics, so that cross-study comparisons of accuracy figures are meaningful rather than dominated by partitioning artefacts.

5.3. Effect of Load Variability

The motor in this study drives an auxiliary pump whose load varies with the operating state of the turbine, so the stator current is not stationary over the observation period. Load affects MCSA features in two ways. It changes the slip, and therefore the frequencies of the rotor-, eccentricity-, bearing-, and vane-related lines, and it scales the amplitude of both the

fundamental and the fault sidebands. The feature design addresses the second effect by normalizing every band amplitude by $A_a(f_1)$ (Section 3.3), and uses bands of ± 0.5 –1 Hz around the lines computed at rated slip, which are wide enough to contain the lines over the observed slip range (median estimated slip 0.0127 against the rated 0.015). The results of Section 4.5 show that this design is adequate for the load range of the dataset (0.77–1.13 of the median): the accuracy is stable across load quartiles, replacing the rated slip with a per-window estimate changes nothing measurable, and the normalized features depend less on load than the absolute ones. Normalization does not remove the load dependence entirely, however, and the macro- F_1 is lower at high load. Extrapolation to load conditions outside the observed range, and in particular to light load, where the rotor-bar sidebands are weakest [9], is not supported by these results.

5.4. Behavior of the Conformal Calibration Layer

The interval results qualify the role of split-conformal calibration in this setting. For the training run of Table 5, the empirical coverage at 90% and 95% matches the nominal value to within one percentage point, but at 50%, 68% and 80% it falls short by 12 to 17 percentage points, and over five training runs the 90% coverage itself varies between 0.74 and 0.90. Neither behavior is predicted by conformal theory under exchangeability, and neither is removed by class-conditional calibration or by CQR (Table 6).

The common cause is that the calibration and test residuals are not exchangeable under the episode-level temporal split. Within each fault episode, the calibration windows precede the test windows, so the test windows are later in the degradation, have higher severity and larger residuals, and are further from the training data. A quantile estimated on the calibration windows is therefore too small for the test windows. The deficit is largest at low levels, where the quantile lies in the dense center of the residual distribution, and a small shift of that distribution changes coverage most, and smallest at high levels, where the quantile lies in the tail. The temporal gap reduces the dependence between neighboring windows but cannot remove this systematic shift, which is intrinsic to evaluating on the later part of each episode. The same effect is visible in Figure 9, where the true severity lies at the lower edge of the interval in the test portion.

The deep ensemble is not calibrated on the calibration windows. Its variance contains the disagreement between independently trained networks, which grows with the distance of the input from the training data, so its intervals widen where the temporal shift occurs. This is why it reaches the nominal coverage at all five levels (ECE 0.015), and why adding a conformal step calibrated on the earlier windows makes it worse (ECE 0.125). For monitoring applications evaluated or deployed under a temporal split, we therefore recommend the ensemble intervals. Conformal calibration remains attractive when the calibration data can be considered exchangeable with the deployment data, for example when it is repeated on the most recent windows of the monitored machine, and methods for conformal prediction under distribution shift [35] are a natural extension. The price of the ensemble is five forward passes per window, about 3 ms on a CPU, which is negligible at a fifteen-minute acquisition cadence.

5.5. Limitations

The following limitations qualify the interpretation of the results.

Single machine. The dataset is measured on a single motor at a single site. Field data are the principal strength of the study relative to laboratory benchmarks, but generalization across motors, drives, and plants is not demonstrated here. Cross-machine validation, possibly supported by domain-adaptation methods [38], is the most important next step before any fleet-wide deployment.

Label uncertainty. The class labels and severity ground truth are reconstructed from maintenance records, so the fault onsets and the severity between inspections are uncertain. The perturbation study of Section 4.4 bounds this effect: onset errors of one day lower the macro- F_1 by up to 0.11 when the onset is placed too early and more than double the faulty-window severity error when it is placed too late, and a 10% error in the severity anchors reduces the 90% interval coverage to 0.34–0.60. Accurate onset dating is therefore as important as model design, and part of the reported severity error may reflect label noise rather than model error.

Calibration under temporal shift. Split-conformal coverage depends on the training run and falls short at low confidence levels because calibration and test windows are not exchangeable (Section 5.4). The ensemble intervals are well calibrated on the present test set, but they carry no finite-sample guarantee, and their calibration should be re-checked on data from further operation.

Partition. Healthy windows have no episode structure and are split at random, so neighboring healthy windows can fall in both the training and the test set. For short fault episodes, the temporal gap (minimum 15 min) is shorter than the 24 h input window, so the input history of a test window can include snapshots used in training. Both effects can make the test figures for the healthy class and for the short transient classes slightly optimistic.

Class support. Voltage unbalance and pump cavitation are represented by 18 and 7 test windows, respectively, and are the only classes with recall below 0.9. Their per-class rates have wide confidence intervals (Table 3), and a longer observation period or data from further machines is needed to characterize them reliably.

Taxonomy and cadence. The ten classes cover the dominant failure modes for pump-driving induction motors [1,3], but faults outside this taxonomy, such as misalignment or impeller damage, and co-occurring faults are not represented; a network confronted with such a fault would assign it to the closest known class, and an explicit novelty-detection layer would be required to flag it. The fifteen-minute cadence and the 24 h window suit slowly developing mechanical faults, but fast electrical faults such as inverter switching anomalies would require a different acquisition design.

These limitations do not invalidate the methodological contributions, but they bound the operational claims. The reported figures characterize the framework on 12 months of field data from one motor; whether the same performance is obtained on a different motor can only be established by validation on that motor.

6. Conclusions

This paper presented a physics-informed temporal convolutional network (PI-TCN) for joint fault classification and severity estimation in three-phase induction motors, with calibrated prediction intervals on the severity. The network uses 60 physically motivated features extracted from a 24 h history of stator-current snapshots and was trained and tested on 12 months of field data from a 90 kW motor in industrial service, under an episode-level data partition with an adaptive temporal gap that prevents leakage between training and test data.

On the independent test set, the PI-TCN reached an accuracy of 0.991 ± 0.002 and a macro-averaged F_1 of 0.898 ± 0.009 over five training runs, close to the strongest classical baseline, while also estimating the severity of faulty windows with a mean absolute error of 0.031 ± 0.004 , lower than all severity baselines and a transformer-based network of the same size. A five-member ensemble of the network gave prediction intervals whose coverage matched the nominal level from 50% to 95% (ECE 0.015), whereas split-conformal and conformalized quantile regression intervals under-covered at the low levels because calibration and test windows are not exchangeable under the temporal split. Per-class

analysis showed that the macro-averaged score is limited by two under-represented classes, voltage unbalance and pump cavitation. The label-perturbation study showed that onset errors of one day already matter: early onsets cause false alarms and late onsets inflate the severity error. The load analysis showed stable accuracy over the observed load range, with a residual load dependence of the features. Feature-importance analysis confirmed that the most informative features are the physically anchored fault-band energies and inter-phase signatures.

Future work will validate the framework on further motors and sites, enlarge the populations of the rare fault classes, study conformal calibration methods that remain valid under temporal shift, and add a novelty-detection layer for faults outside the adopted taxonomy.

Author Contributions: Conceptualization, A.M., B.M.A. and A.R.; methodology, A.M., B.M.A. and Q.Z.; software, A.M. and A.S.; validation, A.M., O.A. and A.S.; formal analysis, A.M. and Q.Z.; investigation, A.M.; resources, B.M.A., A.R. and J.C.; data curation, A.M. and A.S.; writing original draft preparation, A.M.; writing review and editing, A.M., B.M.A., A.R., Q.Z., O.A. and J.C.; visualization, A.M.; supervision, B.M.A. and A.R.; project administration, B.M.A. and A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADC	Analogue-to-Digital Converter
BPFI	Ball-Pass Frequency, Inner race
BPFO	Ball-Pass Frequency, Outer race
BSF	Ball-Spin Frequency
CNN	Convolutional Neural Network
CQR	Conformalized Quantile Regression
CT	Current Transducer
CWRU	Case Western Reserve University
DT	Decision Tree
ECE	Expected Calibration Error
FFT	Fast Fourier Transform
FTF	Fundamental Train Frequency
LSTM	Long Short-Term Memory
MCSA	Motor Current Signature Analysis
MLP	Multilayer Perceptron
NLL	Negative Log-Likelihood
β -NLL	β -weighted Negative Log-Likelihood
PHM	Prognostics and Health Management
PI-TCN	Physics-Informed Temporal Convolutional Network
RF	Random Forest
RMS	Root-Mean-Square
SVM	Support Vector Machine
TCN	Temporal Convolutional Network
TFT	Temporal Fusion Transformer

THD	Total Harmonic Distortion
UQ	Uncertainty Quantification

References

1. Niu, G.; Dong, X.; Chen, Y. Motor Fault Diagnostics Based on Current Signatures: A Review. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 3520919. [CrossRef]
2. Gebraeel, N.; Lei, Y.; Li, N.; Si, X.; Zio, E. Prognostics and Remaining Useful Life Prediction of Machinery: Advances, Opportunities and Challenges. *J. Dyn. Monit. Diagn.* **2023**, *2*, 1–12. [CrossRef]
3. Bahgat, B.H.; Elhay, E.A.; Elkholy, M.M. Advanced Fault Detection Technique of Three Phase Induction Motor: Comprehensive Review. *Discov. Electron.* **2024**, *1*, 9. [CrossRef]
4. Thomson, W.T.; Fenger, M. Current Signature Analysis to Detect Induction Motor Faults. *IEEE Ind. Appl. Mag.* **2001**, *7*, 26–34. [CrossRef]
5. Bellini, A.; Filippetti, F.; Tassoni, C.; Capolino, G.A. Advances in Diagnostic Techniques for Induction Machines. *IEEE Trans. Ind. Electron.* **2008**, *55*, 4109–4126. [CrossRef]
6. Aslan, B.; Balci, S.; Kayabasi, A.; Aslan, M.F.; Sinay, M. Artificial Intelligence and Deep Learning Approaches for Fault Diagnosis in Power Systems: A Comprehensive Survey. *Next Energy* **2026**, *12*, 100703. [CrossRef]
7. Seitzer, M.; Tavakoli, A.; Antic, D.; Martius, G. On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks. *arXiv* **2022**, arXiv:2203.09168.
8. Romano, Y.; Patterson, E.; Candès, E. Conformalized Quantile Regression. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 3543–3553.
9. Halder, S.; Bhat, S.; Zychma, D.; Sowa, P. Broken Rotor Bar Fault Diagnosis Techniques Based on Motor Current Signature Analysis for Induction Motor—A Review. *Energies* **2022**, *15*, 8569. [CrossRef]
10. Asad, B.; Vaimann, T.; Belahcen, A.; Kallaste, A.; Rassõlkin, A. Rotor Fault Diagnostic of Inverter Fed Induction Motor Using Frequency Analysis. In Proceedings of the 2019 IEEE 12th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED), Toulouse, France, 27–30 August 2019; pp. 127–133.
11. Kumar, R.R.; Andriollo, M.; Cirrincione, G.; Cirrincione, M.; Tortella, A. A Comprehensive Review of Conventional and Intelligence-Based Approaches for the Fault Diagnosis and Condition Monitoring of Induction Motors. *Energies* **2022**, *15*, 8938. [CrossRef]
12. Ayankoso, S.; Dutta, A.; He, Y.; Gu, F.; Ball, A.; Pal, S.K. Performance of Vibration and Current Signals in the Fault Diagnosis of Induction Motors Using Deep Learning and Machine Learning Techniques. *Struct. Health Monit.* **2026**, *25*, 196–212. [CrossRef]
13. Valtierra-Rodriguez, M.; Rivera-Guillen, J.R.; Basurto-Hurtado, J.A.; De-Santiago-Perez, J.J.; Granados-Lieberman, D.; Amezcua-Sanchez, J.P. Convolutional Neural Network and Motor Current Signature Analysis during the Transient State for Detection of Broken Rotor Bars in Induction Motors. *Sensors* **2020**, *20*, 3721. [CrossRef] [PubMed]
14. Hussain, M.; Memon, T.D.; Hussain, I.; Memon, Z.A.; Kumar, D. Fault Detection and Identification Using Deep Learning Algorithms in Induction Motors. *Comput. Model. Eng. Sci.* **2022**, *133*, 435–470. [CrossRef]
15. Lei, Y.; Yang, B.; Jiang, X.; Jia, F.; Li, N.; Nandi, A.K. Applications of Machine Learning to Machine Fault Diagnosis: A Review and Roadmap. *Mech. Syst. Signal Process.* **2020**, *138*, 106587. [CrossRef]
16. Drakaki, M.; Karnavas, Y.L.; Tziafettas, I.A.; Linardos, V.; Tzionas, P. Machine Learning and Deep Learning Based Methods toward Industry 4.0 Predictive Maintenance in Induction Motors: State of the Art Survey. *J. Ind. Eng. Manag.* **2022**, *15*, 31–57. [CrossRef]
17. Case Western Reserve University Bearing Data Center. Available online: <https://engineering.case.edu/bearingdatacenter> (accessed on 11 June 2026).
18. Lessmeier, C.; Kimotho, J.K.; Zimmer, D.; Sextro, W. Condition Monitoring of Bearing Damage in Electromechanical Drive Systems by Using Motor Current Signals of Electric Motors: A Benchmark Data Set for Data-Driven Classification. In *Proceedings of the European Conference of the Prognostics and Health Management Society (PHME)*; PHM Society: Rochester, NY, USA, 2016.
19. Barrera-Llana, K.; Burriel-Valencia, J.; Sapena-Bañó, Á.; Martínez-Román, J. A Comparative Analysis of Deep Learning Convolutional Neural Network Architectures for Fault Diagnosis of Broken Rotor Bars in Induction Motors. *Sensors* **2023**, *23*, 8196. [CrossRef] [PubMed]
20. Benninger, M.; Liebschner, M.; Kreischer, C. Fault Detection of Induction Motors with Combined Modeling- and Machine-Learning-Based Framework. *Energies* **2023**, *16*, 3429. [CrossRef]
21. Ali, S.; Khizhik, A.; Svirin, S.; Ryzhikov, A.; Derkach, D. Learning to Hear Broken Motors: Signature-Guided Data Augmentation for Induction-Motor Diagnostics. *Eng. Appl. Artif. Intell.* **2026**, *170*, 114137. [CrossRef]
22. Min, T.-H.; Lee, J.-H.; Choi, B.-K. CNN-Based Fault Classification in Induction Motors Using Feature Vector Images of Symmetrical Components. *Electronics* **2025**, *14*, 1679. [CrossRef]

23. Barrera-Llana, K.; Burriel-Valencia, J.; Sapena-Bañó, Á.; Martínez-Román, J. Fault Detection in Induction Machines Using Learning Models and Fourier Spectrum Image Analysis. *Sensors* **2025**, *25*, 471. [[CrossRef](#)] [[PubMed](#)]
24. Vo, T.-T.; Liu, M.-K.; Tran, M.-Q. Harnessing Attention Mechanisms in a Comprehensive Deep Learning Approach for Induction Motor Fault Diagnosis Using Raw Electrical Signals. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107643. [[CrossRef](#)]
25. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271.
26. Zhang, Q.; Hao, C.; Wang, Y.; Zhang, K.; Yan, H.; Lv, Z.; Fan, Q.; Xu, C.; Xu, L.; Wen, Z.; et al. Intelligent Fault Diagnosis of Rotating Machinery Based on Improved Hybrid Dilated Convolution Network for Unbalanced Samples. *Sci. Rep.* **2025**, *15*, 14127. [[CrossRef](#)] [[PubMed](#)]
27. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [[CrossRef](#)]
28. Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5574–5584.
29. Xu, W.; Zio, E. Uncertainty-Aware Prediction of Remaining Useful Life in Complex Systems. *Annu. Conf. PHM Soc.* **2024**, *16*, 1–4. [[CrossRef](#)]
30. Malik, S.; Yochens, J.; Wisner, B.; Kontsos, A. Enhancing Real-Time Remaining Useful Life Prediction with Information Entropy and Uncertainty-Quantified Deep Learning Models. *J. Intell. Mater. Syst. Struct.* **2025**, *37*, 357–372. [[CrossRef](#)]
31. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016*; PMLR: New York, NY, USA, 2016; Volume 48, pp. 1050–1059.
32. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 6402–6413.
33. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*, 1st ed.; Springer: New York, NY, USA, 2005.
34. Angelopoulos, A.N.; Bates, S. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv* **2021**, arXiv:2107.07511.
35. Barber, R.F.; Candès, E.J.; Ramdas, A.; Tibshirani, R.J. Conformal Prediction beyond Exchangeability. *Ann. Stat.* **2023**, *51*, 816–845. [[CrossRef](#)]
36. Javanmardi, A.; Hüllermeier, E. Conformal Prediction Intervals for Remaining Useful Lifetime Estimation. *Int. J. Progn. Health Manag.* **2023**, *14*. [[CrossRef](#)]
37. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017*; PMLR: New York, NY, USA, 2017; Volume 70, pp. 1321–1330.
38. Lin, Y.; Wang, Y.; Zhang, M.; Zhao, M. A Robust Source-Free Unsupervised Domain Adaptation Method Based on Uncertainty Measure and Adaptive Calibration for Rotating Machinery Fault Diagnosis. *Reliab. Eng. Syst. Saf.* **2025**, *253*, 110516. [[CrossRef](#)]
39. Frénay, B.; Verleysen, M. Classification in the Presence of Label Noise: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *25*, 845–869. [[CrossRef](#)] [[PubMed](#)]
40. Harris, F.J. On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform. *Proc. IEEE* **1978**, *66*, 51–83. [[CrossRef](#)]
41. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
42. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.