Received: 24 May 2026 | Revised: 15 October 2026 | Accepted: 23 October 2026 | Published online: 25 November 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–28
DOI: 10.47852/bonviewJCCE620210516

FFT-Based Spectral–Temporal Deep Learning Framework for Wind Turbine Fault Diagnosis and Predictive Maintenance

Sajib Debnath[1], Md. Uzzal Mia[2,*], Md Abubakkar[3], Arindam Kishor Biswas[4], Md. Sarwar Hosain[2,*] and Ohidujjaman[5]

[1] *O&M Analytics, AES Clean Energy, The AES Corporation, USA*

[2] *Department of Information and Communication Engineering, Pabna University of Science and Technology, Bangladesh*

[3] *Department of Computer Sciences, Midwestern State University, USA*

[4] *Department of Business Analytics, University of the Cumberlands, Williamsburg, USA*

[5] *Department of Computer Science and Engineering, United International University, Bangladesh*

Abstract: Wind turbines operating under harsh dynamic conditions are highly susceptible to mechanical failure, yet traditional time-domain Supervisory Control and Data Acquisition (SCADA) monitoring frequently fails to detect the weak frequency-domain signatures characteristic of incipient faults. This study proposes a novel FFT-based spectral-temporal deep learning framework for wind turbine fault diagnosis and predictive maintenance. The proposed framework extracts a 23-dimensional FFT-derived spectral feature vector and fuses it with 15 time-domain statistical features into a unified 38-dimensional dual-domain representation. A hybrid Transformer–LSTM architecture with multi-head self-attention then simultaneously captures global spectral dependencies and temporal fault-progression patterns across a 640 ms context window. The framework is evaluated on combined SCADA and vibration datasets comprising NREL Gearbox Reliability Collaborative and Case Western Reserve University benchmark data across five health conditions. The proposed model achieves 95.8% accuracy with precision 0.95, recall 0.96, and F1-score 0.955 for binary fault detection and 96.1% overall accuracy with macro-averaged precision, recall, and F1-score of 0.926, 0.964, and 0.944, respectively, for five-class diagnosis. A median fault detection lead time of 18.4 days enables proactive maintenance scheduling. Ablation studies confirm that FFT feature integration, Transformer encoding, and LSTM sequence modeling each contribute incrementally to overall performance. These results demonstrate that the proposed spectral-temporal deep learning framework provides a reliable, interpretable, and computationally efficient solution for intelligent condition monitoring of wind energy systems.

Keywords: wind turbine fault diagnosis, predictive maintenance, FFT, Transformer–LSTM, SCADA

1. Introduction

The global transition toward sustainable and low-carbon energy systems has accelerated the deployment of wind power as a key renewable energy source for electricity generation. Wind power capacity worldwide surpassed the 1,000 GW milestone in 2023 and is projected to double before 2030, driven by accelerating offshore installation programs across Europe, Asia, and North America [1]. Wind turbines are complex rotating machines designed to operate reliably for 20 years or more under conditions that no laboratory test can fully replicate. Offshore and high-altitude onshore turbines are exposed to salt spray, icing, wave-induced loads, and wind shear profiles that accelerate material fatigue throughout the drivetrain [2]. Gearboxes, main bearings, and generator bearings absorb much of this mechanical stress, and their failure patterns reflect this. A systematic review of failure mode distributions in large wind turbine fleets shows that bearings and gearboxes together account for a disproportionate share of both failure events and total downtime, largely because incipient damage in these components progresses slowly and is difficult to detect until it has already caused secondary losses [3]. The financial consequences are significant: offshore repair campaigns require specialist vessels and favorable weather windows; a single gearbox replacement incident routinely costs several

***Corresponding authors:** Md. Uzzal Mia, Department of Information and Communication Engineering, Pabna University of Science and Technology, Bangladesh. Email: uzzal.220605@s.pust.ac.bd. Md. Sarwar Hosain, Department of Information and Communication Engineering, Pabna University of Science and Technology, Bangladesh. Email: sarwar.ice@pust.ac.bd

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

Preproof Version

hundred thousand dollars; and unplanned maintenance may absorb 20–30% of a turbine's lifetime operating expenditure [4].

Most operational turbines are equipped with Supervisory Control and Data Acquisition (SCADA) systems that log temperatures, power output, and low-resolution vibration averages at ten-minute intervals. Although threshold-based alarms built on these signals are effective at detecting faults that have already produced measurable thermal or power anomalies, they are consistently triggered late in the fault-progression cycle [5]. The signatures that distinguish an incipient bearing spall from normal operation are low-amplitude, frequency-specific impulses embedded in broadband vibration noise information that time-averaged SCADA data cannot preserve [6]. Fast Fourier transform (FFT) analysis of high-rate accelerometer data addresses this directly: by decomposing a short vibration window into its spectral components, the FFT renders characteristic defect frequencies derived from bearing geometry and shaft speed directly observable, often weeks before any thermal anomaly appears [7–9]. However, spectral analysis alone is a fixed, knowledge-driven tool that cannot learn from prior failure histories or adapt when a turbine operates outside its design envelope. Deep learning methods overcome this limitation through automatic feature extraction, yet most existing approaches process spectral information implicitly through learned convolutional filters, forgoing the structured frequency-domain knowledge that practitioners have found most diagnostically informative [10, 11]. Scalability to edge hardware poses a further difficulty, as laboratory benchmark studies routinely omit the inference latency and parameter budget constraints that field deployment imposes [12, 13].

To address these limitations, this study proposes a novel FFT-based Spectral-Temporal Deep Learning Framework for wind turbine fault diagnosis and predictive maintenance. The primary objectives of this study are as follows: (1) to develop a structured dual-domain feature representation that explicitly integrates physics-grounded FFT-derived spectral features with time-domain statistical features into a unified input representation; (2) to design a hybrid Transformer–LSTM architecture that simultaneously captures global spectral dependencies and local temporal fault-progression patterns; (3) to evaluate the proposed framework on combined SCADA and vibration datasets across binary and five-class fault diagnosis tasks; and (4) to assess the practical deployment feasibility of the framework in terms of computational efficiency and real-time inference latency.

The significance of this study lies in its principled integration of interpretable signal processing with high-capacity deep sequence learning, bridging a persistent gap between the two paradigms. By explicitly encoding bearing defect frequency amplitudes, harmonic amplitude ratios, and spectral entropy trends as structured input features, the proposed framework preserves the physical interpretability that maintenance engineers require while simultaneously leveraging the adaptive learning capability of deep neural networks. The median fault detection lead time of 18.4 days demonstrated in this study enables proactive maintenance scheduling, substantially reducing unplanned downtime and associated operational costs. Furthermore, with only 1.06 million parameters and a single-sequence inference latency of 5.1 ms, the framework is well-suited for real-time deployment on resource-constrained edge platforms embedded in turbine controllers.

The main contributions of this study are summarized as follows:

1) Structured Physics-Grounded Spectral Input: A 23-dimensional FFT-derived spectral feature vector is constructed from named physical quantities bearing defect frequency amplitudes, harmonic amplitude ratios, spectral entropy, and spectral centroid, providing structured, interpretable input that preserves diagnostic frequency-domain knowledge unlike implicit convolutional spectral learning.
2) Empirically Validated Dual-Domain Fusion: Time-domain (15-D) and FFT spectral (23-D) features are fused into a unified 38-dimensional representation; the ablation study confirms this fusion alone yields a + 3.3 pp accuracy gain (88.2 to 91.5%) on a fixed LSTM backbone, attributable to dual-domain information rather than model capacity.
3) Architecturally Justified Transformer–LSTM Synergy: Multi-head self-attention captures global spectral dependency across the 640 ms context window, while stacked LSTM layers model local fault impulse dynamics, two physically distinct temporal phenomena requiring different inductive biases, contributing + 1.5 pp and + 2.8 pp, respectively, in controlled ablation.
4) Comprehensive Experimental Evaluation: Ablation studies, benchmark comparisons, confusion matrices, ROC analysis, and fault detection lead-time assessment across 12 historical failure events confirm diagnostic reliability across binary and five-class fault diagnosis tasks.
5) Practical Deployment Readiness: With 1.06 M parameters and 5.1 ms GPU inference latency, the framework is edge-compatible; Jetson Xavier NX deployment requires INT8 quantization (≈35 ms estimated E2E), with full edge-platform benchmarking identified as future work.

The novelty of the proposed framework lies not in any individual component but in three integrations not previously combined for wind turbine fault diagnosis. The FFT feature vector provides structured, physics-grounded spectral input, preserving defect frequency interpretability that implicit convolutional learning discards. The Transformer–LSTM architecture is motivated by the dual temporal scale of fault progression: self-attention captures global spectral dependency across the 640 ms context window, while LSTM models local fault impulse dynamics, two phenomena requiring different inductive biases that neither component addresses alone. The ablation study confirms architectural synergy with isolated gains of + 3.3 pp (FFT), + 1.5 pp (Transformer), and + 2.8 pp (LSTM). A parameter-matched comparison further rules out parameter count as the primary driver: the CNN+LSTM baseline (0.70 M) achieves substantially lower accuracy than the proposed model (1.06 M).

The remainder of this paper is organized as follows: Section 2 reviews the related literature. Section 3 presents the mathematical formulations. Section 4 describes the proposed method. Section 5 describes the dataset and experimental setup. Section 6 presents and discusses results. Section 7 concludes the paper.

2. Related Work

Wind turbine fault diagnosis draws on three overlapping bodies of literature: frequency-domain signal processing, classical machine learning, and deep learning. Each has contributed useful tools; however, each also exhibits well-documented limitations that motivated the development of the hybrid framework proposed in this study.

2.1. Signal-processing approaches

FFT-based spectral analysis has underpinned rotating machinery for decades because it maps vibration energy to physically interpretable frequency axes. Recent work has gone beyond

Preprint

simple peak picking: Peeters et al. showed that blind envelope spectrum indicators derived from the FFT can identify fault-related frequency components without requiring the analyst to preselect frequency bands, which matters enormously in field settings where defect frequencies shift with load and speed [8]. Cyclostationary analysis formalizes the statistical periodicity of bearing fault impulses a property the raw FFT does not exploit and has demonstrated reliable detection performance even in the presence of heavy background noise from gear meshing and electrical interference [9].

Nonstationary operation remains a challenging problem. When a turbine's rotor speed changes, fault signatures spread across frequency bins and become difficult to track with a fixed-frequency analysis. Variational mode decomposition addresses this by adaptively separating the vibration record into compact, frequency-localized modes without a predefined filter bank [14]. Combining such decompositions with SCADA operational variables—computing wavelet coherence between accelerometer data and wind speed, for instance—has produced early gearbox anomaly indicators that flag deterioration before any SCADA threshold alarm [15]. Envelope analysis extended to cyclic spectral estimation handles speed variation by targeting second-order cyclostationary patterns that survive rotational speed fluctuation; this approach has shown good sensitivity to outer race defect progression under varying speed test conditions [16]. The shared limitation of these methods is their dependency on experts. Defect frequencies must be estimated from the geometry and shaft speed, frequency bands and mode counts must be selected, and detection thresholds must be calibrated by engineers familiar with the specific machine. Transferring to a different turbine model or unanticipated fault modes requires repeating this engineering effort from scratch.

2.2. Classical machine learning approaches

Supervised learning introduces a degree of automation by fitting decision boundaries from labeled historical data rather than from first principles. Convolutional features extracted from time-frequency image representations have achieved strong diagnostic accuracy on benchmark datasets [17], and systematic benchmark studies on the Case Western Reserve University (CWRU) bearing data have clarified how preprocessing choices, sampling rate, window length, and feature normalization influence classifier performance [10]. Supervised dimensionality reduction using trace ratio linear discriminant analysis has shown that preserving the class-discriminative structure during feature compression produces better fault classification than unsupervised alternatives, such as principal component analysis, particularly when the inter-class separation in the original feature space is moderate [18]. Prognostic and health management (PHM) frameworks that systematically combine multiple classifier outputs have demonstrated that algorithm diversity, rather than any single method, is the main driver of reliable fault detection across various rotating machines [19]. The performance ceiling of classical ML becomes apparent in scenarios where feature engineering assumptions break down. A feature vector designed for one fault mode may perform poorly on another that was not adequately represented during the design. More fundamentally, the temporal dynamics of fault progression, the way a spectral feature grows over hours or days, is discarded when each time window is processed independently.

2.3. Deep learning-based approaches

Multiscale convolutional networks trained directly on vibration waveforms or time-frequency representations eliminate the manual feature engineering step and achieve competitive accuracy on standard fault classification benchmarks [6]. Concerns regarding interpretability, which have historically limited the adoption of deep learning in industrial condition monitoring, have been partially addressed. Li et al. proposed a wavelet kernel network that constrains its convolutional filters to wavelet-like shapes, producing frequency-selective learned representations that can be mapped directly to known fault-frequency bands [11]. Temporal modeling adds another dimension. Hybrid architectures that couple multiscale convolutional layers with LSTM units capture both the spectral character of a fault at a given instant and its evolution over time [20]. When applied to remaining useful life (RUL) estimation, deep temporal models track gradual degradation continuously rather than classifying discrete fault states [21]. Feature fusion methods, for instance, combining convolutional spatial descriptors with enhanced singular value decomposition of multi-sensor data, have extended this capability to heterogeneous data sources [22]. Hybrid architectures that combine wavelet preprocessing with deep convolutional networks have also demonstrated strong performance on standard bearing fault benchmarks [23]. Transformer architectures have pushed performance on long-horizon dependency tasks. The Informer uses sparse self-attention to model hundreds of time steps efficiently, with performance on industrial time series that exceeds the conventional recurrent baselines [24]. A dedicated survey of Transformer variants applied to time-series data confirms that the attention mechanism generalizes well across classification, anomaly detection, and forecasting tasks [25]. A broader review of deep learning methods for temporal sequences reinforces the finding that attention-based and recurrent components work best in combination [26]. Dedicated deep learning studies on wind turbine fault detection using SCADA data have further confirmed that convolutional and recurrent networks outperform threshold-based baselines on field data [13]. Recent advances in multivariate time-series modeling have introduced several foundational architectures that are highly relevant to industrial condition monitoring. Transformer-based patch modeling approaches such as PatchTST have demonstrated that segmenting time series into local semantic patches substantially improves long-range dependency capture in multivariate settings [27]. TimesNet reformulates the one-dimensional time series as two-dimensional temporal variations, enabling convolutional feature extraction across both intra-period and inter-period temporal patterns, with strong generalization across multiple time-series analysis tasks [28]. The iTransformer inverts the traditional Transformer architecture by applying attention across the variate dimension rather than the time dimension, effectively modeling multivariate correlations and achieving state-of-the-art performance on multivariate forecasting benchmarks [29]. Beyond sequential modeling, recent spatiotemporal graph approaches have further expanded the landscape of multivariate time-series learning. Diffusion-based Kriging models with graph-enhanced attention have demonstrated strong performance in capturing complex spatial dependencies and dynamic internode relationships across heterogeneous sensor networks [30], offering a promising direction for multicomponent industrial condition monitoring where physically coupled components such as gearboxes, main bearings, and generator bearings exhibit strong inter-sensor dependencies that sequential models do not explicitly encode.

Pre-proof version

Table 1
Summary of related work in wind turbine and rotating machinery fault diagnosis

Reference	Technique/method	Key contribution	Limitation
Peeters et al. [8]	FFT envelope spectrum indicators	Blind frequency-band selection via spectral sparsity; physically interpretable spectral features	Fixed model; requires known defect frequencies to interpret results
Tang et al. [9]	Cyclostationary analysis	Statistically rigorous detection of periodic fault impulses in heavy noise	Assumes periodic fault signature; sensitive to speed variation
Xiang et al. [15]	Wavelet coherence + SCADA fusion	Multi-resolution monitoring with operational context for wind turbine gearboxes	Expert-designed feature extraction; not end-to-end learnable
Liang et al. [14]	Variational mode decomposition (VMD)	Adaptive frequency-band separation for bearing fault feature extraction	The mode number is a hand-tuned hyperparameter; no temporal learning
Schmidt et al. [16]	Cyclic spectral + envelope analysis	Nonstationary bearing fault detection under varying shaft speed	Manual frequency-band specification; requires prior geometry knowledge
Zhang et al. [17]	CNN with time–frequency image	Strong classification accuracy on benchmark bearing data using learned spatial features	Dependent on feature image quality; limited generalization to unseen operating conditions
Soualhi et al. [19]	PHM ensemble framework	Systematic health management pipeline with diverse classifier combination	Shallow ensemble; limited capacity to capture long-range temporal dependencies
Gu et al. [18]	Trace ratio LDA	Supervised dimensionality reduction preserving class-discriminative spectral structure	Linear projection; may miss nonlinear fault boundaries in feature space
Shao et al. [6]	Multiscale CNN	Hierarchical spatial feature extraction; state-of-the-art on standard fault classification benchmarks	Time-domain input only; no explicit FFT feature integration
Zhang et al. [10]	CNN benchmark study (CWRU)	Comprehensive comparison of preprocessing and classifier choices across multiple DL architectures	Implicit spectral processing; benchmark-specific findings
Li et al. [11]	WaveletKernelNet (interpretable DL)	Physically constrained convolutional filters map to fault-frequency bands	Wavelet basis predefined; adaptivity is limited
Chen et al. [20]	Multiscale CNN + LSTM	Combines spatial pattern matching with LSTM-based temporal memory for bearing fault classification	No structured FFT feature extraction; implicit spectral handling
Li et al. [21]	Deep RUL estimation	Continuous health-state regression from multiscale vibration features	Regression only; no multiclass fault-type identification
Zhou et al. [24]	Informer (sparse Transformer)	Efficient global attention for long industrial time-series sequences	Permutation-invariant architecture; weak local temporal dynamics
Wen et al. [25]	Transformers in time series (survey)	Comprehensive review of attention-based models across classification, detection, and forecasting	Survey paper; not a domain-specific fault diagnosis framework
Lim & Zohren [26]	DL for time-series forecasting (survey)	Systematic assessment of recurrent and attention-based temporal modeling	Forecasting focus; does not target multiclass fault classification
Li et al. [22]	Enhanced SVD feature fusion	Adaptive integration of multi-source features via singular-value decomposition	Implicit feature merging; no structured frequency-domain guidance
Helbing & Ritter [13]	Deep learning for wind turbine fault detection	SCADA-driven fault identification outperforms threshold-based baselines on field data	Black-box model; limited spectral interpretability
This work	FFT + Transformer–LSTM (dual-domain)	Explicit 23-D spectral features fused with 15-D time-domain features, global attention + LSTM temporal modeling, and physically interpretable outputs	Requires further validation on diverse real-world turbine fleets

Pre-proof Version

Beyond the rotating machinery domain, the broader machine learning literature has demonstrated that diverse model families can effectively capture complex nonlinear patterns. Gaussian process regression has been applied with principled uncertainty quantification in industrial and economic settings [31]. Graphical and network-based techniques have proven effective for structured relational modeling [32]. Ensemble and composite methods have demonstrated that combining diverse learners improves generalization [33]. The Transformer–LSTM was preferred over GPR due to its $\mathcal{O}(n^3)$ scaling infeasibility on 760,000 windows and over-ensemble methods due to inference overhead incompatible with the 10 ms edge deployment constraint.

AQ1

2.4. Research gap

Despite the progress reviewed above, the existing approaches exhibit a consistent pattern of limitations, as summarized in Table 2. Signal-processing methods [8, 9, 14–16] provide explicit frequency-domain coverage and physical interpretability but lack any learning component; they cannot adapt to historical failure data. Classical ML methods [10, 17–19] introduce automation but treat spectral information only partially, through handcrafted feature vectors that were not designed to change with data. Deep learning models [6, 10, 11, 20–22, 24, 26] achieve the highest accuracy but operate predominantly in the time domain or treat spectral content implicitly through learned filters, forgoing the structured frequency-domain knowledge that practitioners have found to be the most diagnostically informative.

Four specific gaps motivate the proposed framework. They are as below:

1) Absence of Explicit Dual-Domain Learning: No existing approach simultaneously integrates structured, physics-grounded FFT features with end-to-end deep learning in an interpretable, unified architecture.
2) Separation Between Signal Processing and Deep Learning: Spectral analysis methods and deep learning architectures have been developed largely in parallel, with limited principled integration of the two paradigms.
3) Limited Spectro-Temporal Modeling: Current models do not explicitly capture how fault-related spectral features, defect frequency amplitudes, harmonic amplitude ratios, and spectral entropy evolve across consecutive time windows as faults progress.
4) Reduced Practical Interpretability: Black-box deep learning models remain difficult to validate and audit in field deployments where maintenance engineers need to understand and ultimately trust the model's outputs.

The proposed framework addresses all four gaps by constructing a 38-dimensional dual-domain feature vector from explicit FFT analysis and time-domain statistics, and processing sequences of these vectors through a Transformer–LSTM model, as shown in the final row of Table 2.

3. System Model and Mathematical Formulation

3.1. Signal representation and acquisition

Let the continuous-time vibration signal measured by an accelerometer mounted on a turbine component be denoted as $x(t)$, where $t \in \mathbb{R}$ represents continuous time, and $x : \mathbb{R} \to \mathbb{R}$ represents the instantaneous acceleration. In a digital monitoring system, $x(t)$ is uniformly sampled at a sampling frequency f_s Hz to produce the discrete-time sequence as follows:

$$x[n] = x(nT_s), \quad n = 0, 1, \ldots, N-1 \tag{1}$$

where $T_s = \frac{1}{f_s}$ is the sampling interval, and N is the total number of samples in the analysis window. The sampling frequency must satisfy the Nyquist–Shannon sampling theorem to ensure the perfect reconstruction of the band-limited signal [34, 35]:

$$f_s \geq 2f_{\max} \tag{2}$$

where $f_{\max}$ is the highest frequency component. For gearbox vibration analysis, fault signatures may extend to several kilohertz, necessitating sampling frequencies in the range of 10–50 kHz. The analysis window duration $T_w = \frac{N}{f_s}$ determines the frequency resolution of the subsequent FFT analysis, with longer windows yielding a finer spectral resolution.

3.2. The discrete Fourier transform (DFT)

The discrete Fourier transform (DFT) is a fundamental mathematical operation that projects a finite-duration discrete sequence onto a basis of complex exponentials [34]. Given a real-valued discrete sequence $x[n]$ of length N, the DFT is defined as

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j\frac{2\pi kn}{N}}, \quad k = 0, 1, \ldots, N-1 \tag{3}$$

Table 2
Research gap comparison across wind turbine fault diagnosis approaches

Approach	Time	FrEquation	Deep learning	Temporal modeling	Interpretability
FFT / Spectral [8, 9, 16]	✓	✓	×	×	✓
VMD / Wavelet [14, 15]	✓	✓	×	×	Partial
Feature-Based ML [18, 19]	✓	Partial	×	×	Partial
CNN / LSTM / Transformer [6, 11, 20, 24, 25]	✓	×	✓	Partial	×
Proposed Framework	✓	✓	✓	✓ Transformer–LSTM	✓

Pre-proof version

where $j = \sqrt{-1}$ is the imaginary unit, k is the discrete frequency index, and the complex exponential

$$e^{-j\frac{2\pi kn}{N}} = \cos\left(\frac{2\pi kn}{N}\right) - j\sin\left(\frac{2\pi kn}{N}\right)$$

represents the basis function at the normalized frequency $\omega_k = \frac{2\pi k}{N}$ radians per sample. The corresponding physical frequency associated with index k is expressed as

$$f_k = \frac{kf_s}{N}, \quad k = 0, 1, \ldots, \frac{N}{2} \tag{4}$$

For real-valued signals, the DFT exhibits conjugate symmetry as follows:

$$X[N-k] = X^*[k]$$

This implies that only the first $N/2 + 1$ frequency bins (from DC to the Nyquist frequency $f_s/2$) carry unique information. The inverse DFT (IDFT) recovers the original time-domain sequence [34].

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X[k] \cdot e^{j\frac{2\pi kn}{N}} \tag{5}$$

3.3. The fast Fourier transform algorithm

The direct computation of the DFT requires $O(N^2)$ complex multiplications and additions, which becomes prohibitively expensive for large N. The FFT is an algorithm that exploits the periodicity and symmetry properties of the complex exponential twiddle factors $W_N = e^{-j\frac{2\pi}{N}}$ to recursively decompose the DFT computation, achieving a complexity of $O(N\log_2 N)$ [8, 34]. This represents a reduction by a factor of $\frac{N}{\log_2 N}$ compared to the direct computation; for $N = 1024$ samples, this corresponds to approximately a 100-fold speedup. The Cooley–Tukey radix-2 decimation-in-time (DIT) FFT algorithm 1 decomposes the N-point DFT (with $N = 2^m$ for integer m) into two $N/2$-point DFTs of the even- and odd-indexed subsequences [8, 34]:

$$X[k] = E[k] + W_N^k \cdot O[k], \quad k = 0, 1, \ldots, \frac{N}{2} - 1 \tag{6}$$

$$X\left[k + \frac{N}{2}\right] = E[k] - W_N^k \cdot O[k], \quad k = 0, 1, \ldots, \frac{N}{2} - 1 \tag{7}$$

where $E[k] = \sum_{n=0}^{\frac{N}{2}-1} x[2n] \cdot W_{N/2}^{kn}$ and $O[k] = \sum_{n=0}^{\frac{N}{2}-1} x[2n+1] \cdot W_{N/2}^{kn}$ are the DFTs of the even and odd subsequences, respectively, and $W_N^k = e^{-j\frac{2\pi k}{N}}$ is the twiddle factor. The **butterfly** computation in Equations 6 and 7 is the fundamental processing unit of the FFT, combining two DFT values with a twiddle factor multiplication to produce two output values. This recursive splitting continues for $\log_2 N$ stages, each involving $\frac{N}{2}$ butterfly operations, yielding an overall $O(N\log_2 N)$ complexity.

3.4. Windowing and spectral leakage

The DFT implicitly assumes that a finite signal segment is a single period of an infinite periodic sequence. When the signal contains frequency components that are not integer multiples of the frequency resolution $\Delta f = \frac{f_s}{N}$, the energy from a single spectral peak spreads into adjacent frequency bins, a phenomenon known as *spectral leakage* [34, 36]. To mitigate leakage, the time-domain segment is multiplied by a window function $w[n]$ prior to applying the FFT as follows:

$$x_w[n] = x[n] \cdot w[n], \quad n = 0, 1, \ldots, N-1 \tag{8}$$

The windowed FFT is then expressed as

$$X_w[k] = \sum_{n=0}^{N-1} x[n] \cdot w[n] \cdot e^{-j\frac{2\pi kn}{N}} \tag{9}$$

Commonly used window functions include the Hann window: $w[n] = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right)$ the Hamming window: $w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$ and a flat-top window, which provides excellent amplitude accuracy at the expense of frequency resolution.

The rectangular window ($w[n] = 1, \forall n$) has the finest frequency resolution but the worst leakage characteristics.

In this study, the Hann window was applied to each segment prior to the FFT computation, offering a favorable trade-off between frequency resolution and sidelobe suppression.

3.5. Spectral feature extraction

From the windowed FFT output $X_w[k]$, the following discriminative spectral features are computed:

3.5.1. Spectral magnitude and power spectral density

The spectral magnitude at each frequency bin is

$$|X[k]| = \sqrt{(\Re\{X[k]\})^2 + (\Im\{X[k]\})^2} \tag{10}$$

The one-sided Power Spectral Density (PSD) estimate is

$$P[k] = \frac{2}{N^2}|X[k]|^2, \quad k = 1, \ldots, \frac{N}{2} - 1 \tag{11}$$

with $P[0] = \frac{1}{N^2}|X[0]|^2$ and $P\left[\frac{N}{2}\right] = \frac{1}{N^2}\left|X\left[\frac{N}{2}\right]\right|^2$ for DC and Nyquist components [34, 37].

3.5.2. Spectral energy

Total spectral energy characterizing the overall vibration severity:

$$E_{\text{spec}} = \sum_{k=0}^{\frac{N}{2}} P[k] \tag{12}$$

3.5.3. Spectral centroid

The spectral centroid represents the center of mass of the frequency distribution.

$$f_c = \frac{\sum_{k=0}^{\frac{N}{2}} f_k \cdot P[k]}{\sum_{k=0}^{\frac{N}{2}} P[k]} \tag{13}$$

Pre-proof version

3.5.4. Spectral spread

The spectral spread quantifies the bandwidth of the dominant spectral components as follows:

$$\sigma_f^2 = \frac{\sum_{k=0}^{\frac{N}{2}} (f_k - f_c)^2 \cdot P[k]}{\sum_{k=0}^{\frac{N}{2}} P[k]} \tag{14}$$

3.5.5. Spectral entropy

Spectral entropy measures spectral irregularity and complexity.

$$H_{\text{spec}} = -\sum_{k=0}^{\frac{N}{2}} p[k] \log_2(p[k]) \tag{15}$$

where $p[k] = \frac{P[k]}{\sum_k P[k]}$ is the normalized PSD that acts as a probability distribution over frequencies [35, 37].

3.5.6. Harmonic amplitude ratios

For rotating machinery, fault frequencies often appear as harmonics of the fundamental defect frequency f_d. The harmonic amplitude ratio for the m-th harmonic is expressed as

$$\text{HAR}_m = \frac{|X[k_m]|}{|X[k_1]|}, \quad m = 2, 3, \ldots, M \tag{16}$$

where $k_m = \text{round}\left(\frac{m f_d N}{f_s}\right)$ is the frequency bin index of the m-th harmonic. HAR captures the energy distribution across the harmonic series, which is strengthened in the presence of periodic impulse-type faults.

3.5.7. Peak frequency and amplitude

The peak frequency and corresponding amplitude are defined as

$$f_{\text{peak}} = f_{k^*}, \quad \text{where } k^* = \arg\max_k |X[k]|, k \in [1, N/2] \tag{17}$$

$$A_{\text{peak}} = \frac{|X[k^*]|}{N} \tag{18}$$

3.6. Problem formulation

Let $x[n]$ be a vibration-signal segment of length N collected from a wind turbine sensor. The signal is characterized by a combined time-domain feature vector $\mathbf{F}_t \in \mathbb{R}^{d_t}$ and frequency-domain feature vector $\mathbf{F}_f \in \mathbb{R}^{d_f}$, extracted according to the procedures described above. The fused feature vector is

$$\mathbf{F} = [\mathbf{F}_t; \mathbf{F}_f] \in \mathbb{R}^{d_t + d_f} \tag{19}$$

A sequence of T consecutive feature vectors, $\{\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_T\}$, constitutes the input to the deep learning model. The classification objective is to learn the mapping

$$f_\theta : \{\mathbf{F}_1, \ldots, \mathbf{F}_T\} \to \hat{y} \in \{0, 1, \ldots, C-1\} \tag{20}$$

where θ denotes the set of learnable parameters, C is the number of health condition classes, and $\hat{y}$ is the predicted class label. Class 0 corresponds to normal operation, whereas classes 1 through $C-1$ correspond to distinct fault categories (e.g., gearbox tooth fault, inner race bearing fault, outer race bearing fault, and rotor imbalance) [38, 39].

The optimal parameters were obtained by minimizing the categorical cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{F}, y) \in \mathcal{D}} \sum_{c=0}^{C-1} \log(p_c) \tag{21}$$

where y is the one-hot encoded true label, $p_c = \text{softmax}(f_\theta(\mathbf{F}))_c$ is the predicted probability for class c, and $\mathcal{D}$ denotes the training dataset [40].

4. Proposed Methodology

4.1. Framework overview

The proposed FFT-based Spectral–temporal Deep Learning framework comprises six interconnected stages, forming an end-to-end pipeline from raw sensor data to health condition prediction. Figure 1 illustrates the architecture of the proposed system. The stages are as follows: (1) Multi-source Data Acquisition, (2) Signal Preprocessing, (3) FFT-based Spectral Feature Extraction, (4) Dual-Domain Feature Fusion, (5) Transformer–LSTM Classification Model, and (6) Fault Prediction and Alerting.

4.1.1. Multisource data acquisition

High-frequency vibration data were acquired from triaxial piezoelectric accelerometers mounted on the gearbox, main bearing, and generator bearing housings at a sampling frequency of $f_s = 25,600$ Hz with 16-bit ADC resolution (approximately 96 dB dynamic range). Complementary SCADA operational data were recorded at 1 Hz intervals, including wind speed, active power output, rotational speed (RPM), blade pitch angles, temperatures (generator windings, main bearing, and gearbox oil), and nacelle orientation. AQ2

4.1.2. Signal preprocessing

The preprocessing pipeline consisted of four sequential operations. First, an eighth-order Butterworth lowpass anti-aliasing filter with a cutoff frequency $f_c = \frac{f_s}{2} - 500$ Hz was applied prior to digitization. Subsequent high-pass filtering at Hz removes the DC offset and low-frequency mounting resonances. A comb filter notch at 5 Hz (and harmonics) suppresses power line interference in electrically susceptible measurement channels. Each sensor channel was then standardized to zero mean and unit variance using the training set statistics, preventing channels with larger physical amplitudes from dominating the feature space:

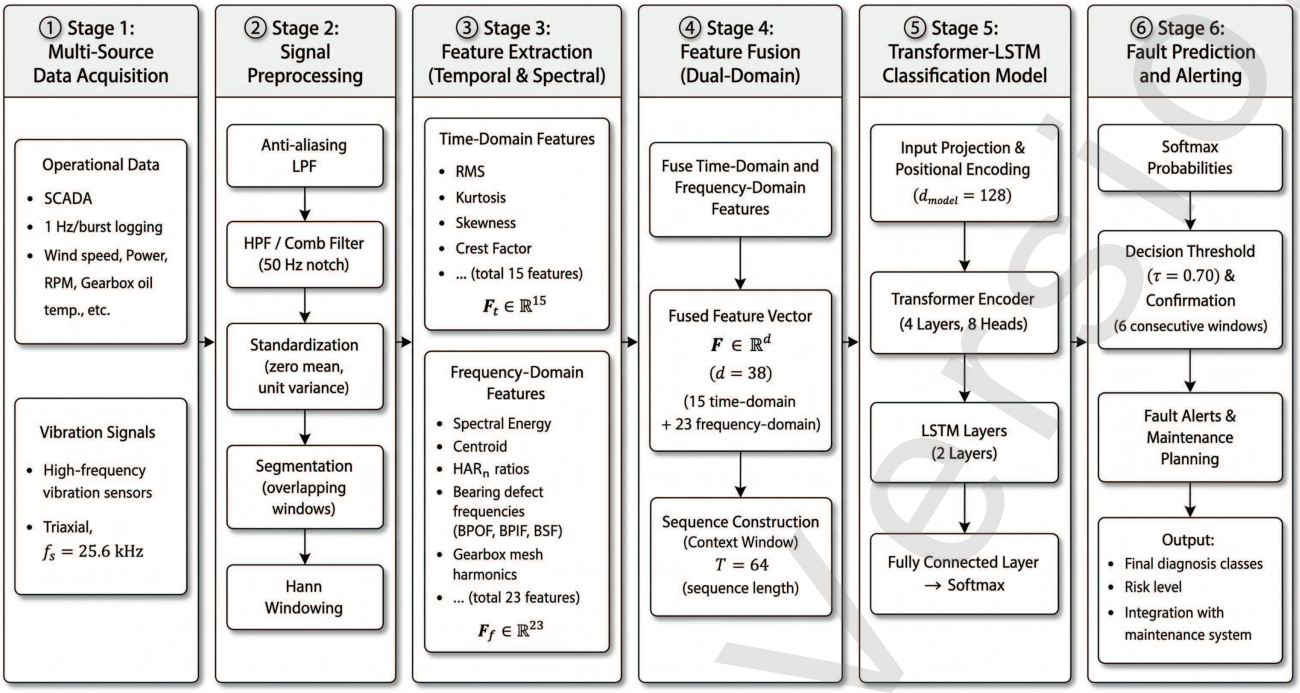
$$\tilde{x}[n] = \frac{x[n] - \mu_x}{\sigma_x} \tag{22}$$

where μ_x and σ_x are the channel mean and standard deviation, respectively, computed over the training partition.

The normalized continuous signal was subsequently segmented into overlapping windows of length $N = 1024$ samples with a hop size of $H = 256$ samples (75% overlap), yielding a feature vector approximately every 10ms at 25, 600 Hz. The overlap ensures that transient fault impulses near the window boundaries are captured in at least one full window. Finally, each segment is multiplied element-wise by the Hann window

Proof Version

Figure 1
System architecture of the proposed FFT-based spectral–temporal deep learning framework



(Equation 8) prior to the FFT computation to suppress spectral leakage, with the window pre-computed once and applied via vectorized multiplication. For each window, one complete set of time-domain and frequency-domain features was computed. The relationship betwee/n the analysis window, hop size, and temporal context is as follows: Each analysis window of $N = 1024$ samples at $f_s = 25,600$ Hz spans approximately 40 ms, producing one feature vector every $H = 256$ samples (10 ms hop). A sequence of $T = 64$ consecutive feature vectors therefore covers $64 \times 10\,\text{ms} = 640\,\text{ms}$ of temporal context sufficient to capture multiple fault-frequency cycles at typical gearbox operating speeds above 200 RPM and at least one full shaft revolution at speeds as low as 100 RPM.

4.1.3. FFT-based spectral feature extraction

Figure 2 illustrates the STFT spectrogram comparison between normal and faulty vibrations, clearly demonstrating the periodic impulse structure at the ball pass outer frequency (BPOF), which becomes visible in the frequency domain during fault conditions.

The STFT spectrogram in Figure 2(b) shows the outer race bearing fault condition at BPOF = 287 Hz.

Figure 3 presents a comparison of the normalized spectral features across Normal, Incipient Fault, and Developed Fault conditions, highlighting the strong discriminative capability of the frequency-domain features. In particular, the BPOF amplitude exhibits an approximately 18.5× increase from normal to developed fault conditions (corresponding to a normalized bar value of $\approx$ 0.054 for normal operation in Figure 3, displayed as 0.05 after rounding), reflecting the emergence of periodic impulsive behavior associated with the bearing defects. Additionally, the harmonic amplitude ratios (HAR_2, HAR_3, HAR_4) show progressive amplification with fault severity, indicating the strengthening of harmonic structures.

Spectral entropy exhibits a non-monotonic progression across fault severity levels; it initially increases from the normal condition to the incipient fault stage, reflecting the introduction of additional broadband spectral components associated with early-stage mechanical looseness and distributed micro-contact impacts. As the fault developed further into the developed-fault condition, the spectral entropy decreased as the dominant periodic impulse structure consolidated, and the spectrum became increasingly regular. This non-monotonic behavior is physically consistent with the competing effects of spectral broadening during fault initiation and harmonic peak consolidation during advanced fault progression and is captured by the proposed FFT feature pipeline as a discriminative indicator of the fault stage. Other features, including spectral energy, centroid, and spread, also increase with fault severity, further confirming the effectiveness of FFT-based spectral features as robust and interpretable indicators of fault diagnosis.

To provide rigorous feature-level interpretability beyond attention weight visualization, SHapley Additive exPlanations (SHAP) attribution analysis was applied post hoc to all 38 input features across all five fault classes; results are reported in Section 6.12 and confirm that the framework assigns highest attribution to the physically correct fault-frequency features for each fault type.

The structured 23-dimensional FFT feature vector bridges expert spectral knowledge with end-to-end deep learning rather than replacing one with the other. Learned frequency embeddings and wavelet packet coefficients were considered but rejected: the former discards physical interpretability required for field deployment, while the latter reintroduces basis-selection hyperparameters that recreate the expert dependency the framework otherwise avoids. Cyclostationary indicators were excluded due to their requirement for shaft-speed synchronization unavailable in the SCADA pipeline. A feature sensitivity analysis (Table 21, Section 6.11) confirms model robustness to partial feature loss under field conditions.

Preprint Version

Figure 2
STFT spectrograms: (a) normal operation; (b) outer race bearing fault

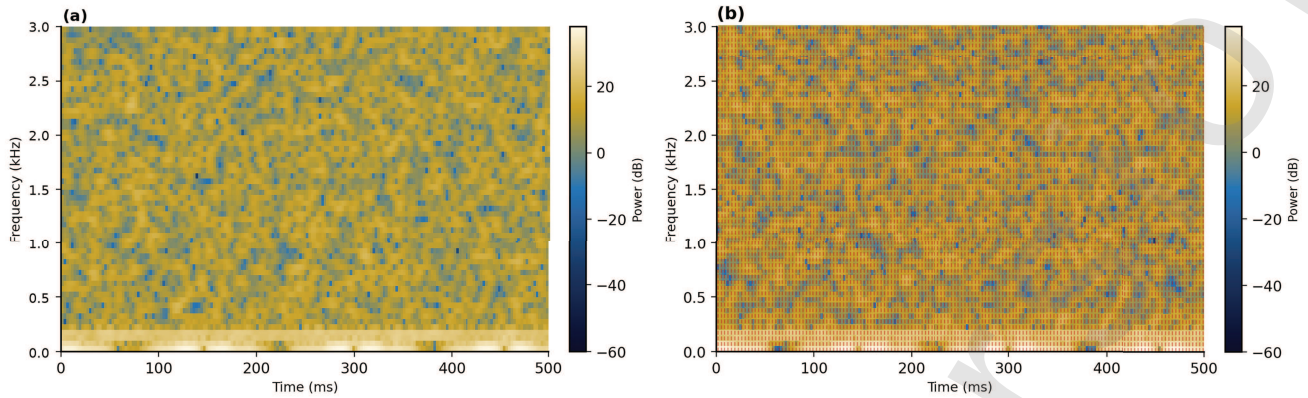
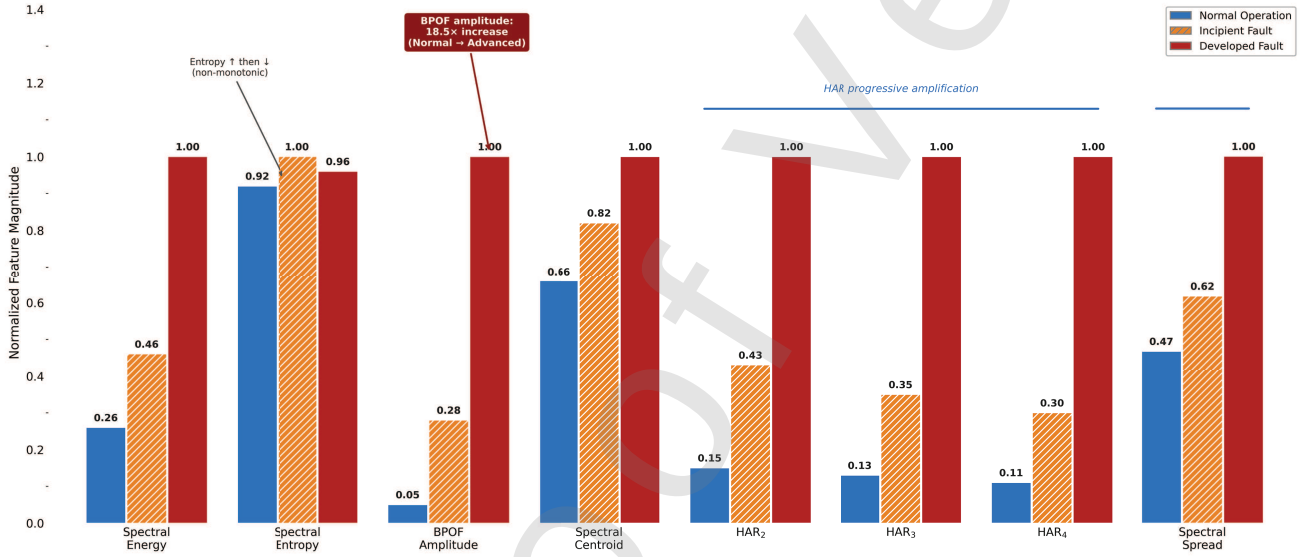


Figure 3
Normalized spectral feature comparison across Normal, Incipient Fault, and Developed Fault conditions



4.1.4. Dual-domain feature fusion

In parallel with the spectral feature extraction pipeline, a 15-dimensional time-domain feature vector $F_t \in \mathbb{R}^{15}$ was extracted from each raw signal segment, comprising RMS amplitude, peak amplitude, peak-to-peak amplitude, crest factor, kurtosis, skewness, variance, mean absolute value, shape factor, impulse factor, margin factor, signal energy entropy, zero-crossing rate, and two turbine operational-state features (normalized power, normalized RPM).

The time- and frequency-domain vectors were concatenated to form a 38-dimensional fused feature vector per window:

$$F = \left[F_t \in \mathbb{R}^{15}, F_f \in \mathbb{R}^{23}\right] \in \mathbb{R}^{38} \tag{23}$$

Although the time-domain and frequency-domain features originate from distinct mathematical perspectives, both characterize the same underlying physical vibration signal through complementary lenses, making them physically commensurate descriptors of the same fault state. Critically, the direct concatenation is followed by a fully learnable linear projection (Equation 24, $W_{\text{proj}} \in \mathbb{R}^{38 \times 128}$, which learns to re-weight and recombine all 38 input dimensions jointly during training, effectively discovering the optimal cross-domain manifold mapping in a data-driven manner. The ablation study (Table 14) empirically validates this choice: the dual-domain fusion configuration achieves 91.5% accuracy, consistently outperforming any single-domain alternative, and the full proposed model achieves 95.8%. Dual-stream cross-domain encoding is identified as a promising direction for future work.

A sequence of $T = 64$ consecutive fused feature vectors is assembled using the sliding window mechanism, forming the input tensor $\mathcal{X} \in \mathbb{R}^{T \times d}$ (with $d = d_t + d_f = 15 + 23 = 38$, consistent with Equation 19), which is fed to the Transformer–LSTM model. The 64-step sequence spans a temporal context of approximately 640 ms at the 10 ms feature extraction rate, which is sufficient to capture at least one full shaft revolution for gearbox speeds as low as 100 RPM and multiple fault-frequency cycles at typical operating speeds (above $\sim$ 200 RPM for most onshore turbines).

It should be noted that the current feature fusion treats each sensor channel independently: the 38-dimensional feature vector aggregates spectral and time-domain features from all channels but does not explicitly encode interchannel dependencies arising

Preproof Version

from mechanical coupling between the gearbox, main bearing, and generator bearing. The Transformer–LSTM architecture implicitly learns soft interchannel relationships through joint attention over the fused feature sequence, partially compensating for this independence assumption. Explicit cross-channel features including cross-power spectral density, coherence functions at fault frequencies, and time-lagged cross-correlation amplitudes represent a lightweight extension that could be appended to the existing feature vector without architectural modification and are identified as a near-term improvement direction in Section 7.2.

4.1.5. Transformer–LSTM architecture

The raw 38-dimensional feature vector at each time step is linearly projected to the transformer model dimension $d_{\text{model}} = 128$ via a learned linear transformation:

$$Z^{(0)} = FW_{\text{proj}} + b_{\text{proj}}, \quad W_{\text{proj}} \in \mathbb{R}^{38 \times 128} \tag{24}$$

Positional encodings are added to $Z^{(0)}$ to inject temporal ordering information because the self-attention mechanism is inherently permutation-invariant. Sinusoidal positional encodings [24] are used as follows:

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \tag{25}$$

$$\text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \tag{26}$$

The projected and positionally encoded sequences were processed using a stack of $L = 4$ transformer encoder layers. Each encoder layer applies multi-head self-attention (MHSA).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{27}$$

$$\text{head}_i = \text{Attention}(ZW_i^Q,\ ZW_i^K,\ ZW_i^V) \tag{28}$$

$$\text{MHSA}(Z) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\, W^O \tag{29}$$

where $H = 8$ attention heads are used, $d_k = \frac{d_{\text{model}}}{H} = 16$ is the key/query dimension per head, and $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are the projection matrices for the ith head. The scaled dot-product scores $\frac{QK^{\top}}{\sqrt{d_k}}$ prevent excessively small gradients in the softmax function for large d_k. The 8-head configuration allows the model to simultaneously attend to different aspects of the feature sequence.

A position-wise feed-forward network (FFN) is then applied as follows:

$$\text{FFN}(Z) = \text{ReLU}(ZW_1 + b_1)\, W_2 + b_2 \tag{30}$$

where $W_1 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$ and $W_2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$, with inner dimension $d_{\text{ff}} = 512$. Layer normalization and residual connections are applied to both sublayers:

$$Z' = \text{LayerNorm}(Z + \text{MHSA}(Z)) \tag{31}$$

$$Z'' = \text{LayerNorm}(Z' + \text{FFN}(Z')) \tag{32}$$

Dropout with probability $p = 0.1$ was applied within the attention and FFN sublayers for regularization.

The output of the final transformer encoder layer $Z^{(L)} \in \mathbb{R}^{T \times d_{\text{model}}}$ is passed sequentially to a two-layer LSTM network. Each LSTM cell maintains a hidden state $h_t \in \mathbb{R}^{d_h}$ and cell state $c_t \in \mathbb{R}^{d_h}$ governed by:

$$f_t = \sigma\left(W_f[h_{t-1}, z_t] + b_f\right) \tag{33}$$

$$i_t = \sigma(W_i[h_{t-1}, z_t] + b_i) \tag{34}$$

$$g_t = \tanh\left(W_g[h_{t-1}, z_t] + b_g\right) \tag{35}$$

$$o_t = \sigma(W_o[h_{t-1}, z_t] + b_o) \tag{36}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{37}$$

Table 3
Spectral feature vector components extracted from the FFT output for each signal window

No.	Feature	Formula/method	Physical significance
1	Spectral Energy (E_{spec})	Equation (12)	Overall vibration severity
2	Spectral Centroid (f_c)	Equation (13)	Dominant frequency shift
3	Spectral Spread (σ_f^2)	Equation (14)	Bandwidth of spectral activity
4	Spectral Entropy (H_{spec})	Equation (15)	Signal complexity / regularity
5–6	Peak Frequency & Amplitude	Equation (17)–(18)	Strongest resonance detection
7–9	HAR_2, HAR_3, HAR_4	Equation (16)	Harmonic progression of fault-frequencies
10–12	Sub-band energies (3 bands)	$\sum P[k]$ per band	Zonal spectral energy distribution
13–15	BPOF / BPIF / BSF amplitudes	$\lvert X[k_{\text{fault}}]\rvert$	Bearing defect frequency magnitudes
16–18	Gear mesh frEquation amplitudes (3×)	$\lvert X[k_{\text{mesh}}]\rvert$	Gearbox tooth fault indicators
19–20	Fundamental + 1× sidebands	Spectral analysis	Modulation by shaft unbalance
21–23	RMS, Kurtosis, Crest factor (on PSD)	Statistical moments on $P[k]$	Distribution shape of PSD

Pre-proof version

$$h_t = o_t \odot \tanh(c_t) \tag{38}$$

where f_t, i_t, and o_t are the forget, input, and output gate activations, respectively; g_t is the candidate cell update; σ denotes the sigmoid function; and $\odot$ denotes the Hadamard (element-wise) product. The LSTM hidden dimension was $d_h = 128$ per layer. Two LSTM layers with dropout ($p = 0.2$) between the layers were stacked to increase the representational depth.

The final LSTM hidden state h_T (output at the last time step) is passed through a fully connected classification head as follows:

$$\hat{y} = \text{softmax}\,(W_{\text{cls}}\, h_T + b_{\text{cls}}), \quad W_{\text{cls}} \in \mathbb{R}^{d_h \times C} \tag{39}$$

The predicted class is $\hat{y} = \arg\max_c \hat{y}_c$. For binary fault detection, $C = 2$; for multiclass diagnosis, $C = 5$ covering normal operation, gearbox tooth fault, inner race bearing fault, outer race bearing fault, and rotor imbalance.

Implementation of Transformer–LSTM Architecture: To enhance reproducibility and provide practical insight into the proposed model, a concise implementation of the Transformer–LSTM architecture is presented below. As shown in Listing 1, the proposed model integrates transformer-based feature encoding with LSTM-based temporal modeling. The code reflects the core processing pipeline, including linear feature projection, multi-head self-attention-based transformer encoding, sequential temporal modeling using LSTM layers, and final classification. This implementation aligns with the mathematical formulation described in Equations (24)–(39), noting that the linear projection (Equation 24) and sinusoidal positional encoding (Eqs. 25–26) are incorporated in the full implementation but omitted from this simplified listing for brevity and demonstrate how spectral–temporal features are effectively learned for fault diagnosis.

Listing 1
Transformer–LSTM fault diagnosis implementation in PyTorch

```
import torch
import torch.nn as nn
class TransformerLSTM(nn.Module):
       def __init__(self, d_in=38,
d_model=128, nhead=8, n_cls=5):
           super().__init__()
           self.m = nn.Sequential(
                nn.Linear(d_in, d_model),
                nn.TransformerEncoder(
                    nn.TransfEncoderLayer(
                        d_model, nhead,
                        dim_feedforward=512,
                        dropout=0.1,
                        batch_first=True
                    ), 4
                )
           )
           self.lstm = nn.LSTM(d_model, 128,
2, batch_first=True)
           self.fc = nn.Linear(128, n_cls)

       def forward(self, x):
           x = self.m(x)
           _, (h, _) = self.lstm(x)
           return self.fc(h[-1])
```

Implementation Details and Reproducibility Clarification: The PyTorch implementation shown in Listing 1 provides a simplified representation of the proposed Transformer–LSTM architecture for clarity and readability. For conciseness, several components described in the mathematical formulation Equations (24–39) are not explicitly included in the code snippet.

In the full implementation used for the experimental evaluation, the following components were incorporated:

1) Positional Encoding: Sinusoidal positional encodings are added to the input feature sequence prior to Transformer encoding to preserve temporal order.
2) Dropout Regularization: Dropout is applied within Transformer encoder layers (attention and feed-forward sublayers, $p = 0.1$) and between stacked LSTM layers ($p = 0.2$) to mitigate overfitting.
3) Feed-Forward Dimension: In the full implementation, dim_feedforward = 512 is explicitly set in each TransformerEncoderLayer to match equation 30, overriding PyTorch's default of 2048.
4) Layer Normalization: Residual connections with layer normalization are employed in each Transformer encoder block, consistent with standard Transformer design.
5) Training Pipeline: The complete training procedure includes mini-batch processing, cross-entropy loss optimization, gradient clipping, and adaptive learning rate scheduling (ReduceLROnPlateau), as detailed in Algorithm 1.

The code in Listing 1 is therefore intended to illustrate the high-level architecture, whereas the full implementation adheres strictly to the detailed formulation and training pipeline described in this study. This ensures consistency between the theoretical framework and reported experimental results.

Table 4
Layer-by-layer architecture summary of the proposed Transformer–LSTM model

Layer	Configuration	Output Dimension
Input Feature Sequence	$T = 64$, $d = 38$	$\mathbb{R}^{64\times38}$
Linear Projection	$38 \rightarrow 128$	$\mathbb{R}^{64\times128}$
Positional Encoding	Sinusoidal	$\mathbb{R}^{64\times128}$
Transformer Encoder ×4	$H = 8$, $d_{ff} = 512$	$\mathbb{R}^{64\times128}$
LSTM Layer 1	$d_h = 128$, dropout 0.2	$\mathbb{R}^{64\times128}$
LSTM Layer 2	$d_h = 128$, dropout 0.2	$\mathbb{R}^{128}$
Fully Connected + Softmax	$128 \rightarrow C$	$\mathbb{R}^{C}$

4.1.6. Fault prediction and alerting

The softmax output $\hat{y}$ from the classification head yields a probability vector across $C = 5$ health classes. A fault alert is triggered when the predicted fault-class probability exceeds a decision threshold $\tau = 0.70$ for at least six consecutive inference windows (6 windows× $\sim$ 10 ms/window $=\sim$ 60 ms), suppressing transient false positives caused by nonrecurring impulse noise. For the five-class model, P_{fault} is defined as $1 - P_{\text{Normal}}$ (i.e., the cumulative probability assigned to all fault classes), and

Preproof Version

the predicted fault type is $\arg\max_{c\in\{1,\ldots,4\}} \hat{y}_c$; see Section 4.2 for the extended 6-hour criterion used in lead time evaluation. The predicted class is mapped to a severity tier: *Advisory* ($0.70 \leq P_{\text{fault}} < 0.85$), *Warning* ($0.85 \leq P_{\text{fault}} < 0.95$), or *Critical* ($P_{\text{fault}} \geq 0.95$) and transmitted via SCADA integration to the maintenance management system. A rolling alert log records the fault class, probability score, timestamp, and turbine ID, enabling trend tracking and RUL estimation. This three-tier thresholding strategy substantially reduces false alarm rates compared to single-threshold classification while maintaining a high binary fault detection recall, as validated on the held-out test set.

4.2. Lead time definition

To quantitatively evaluate the predictive capability of the proposed framework, the fault detection lead time is defined as the temporal difference between the actual fault occurrence and the first reliable model detection, given by $\text{Lead Time} = T_{\text{failure}} - T_{\text{first detection}}$. Here, T_{failure} denotes the timestamp of the confirmed fault event obtained from maintenance logs or failure records, and $T_{\text{first detection}}$ represents the earliest time at which the predicted fault probability exceeds the decision threshold ($\tau = 0.70$) for at least six consecutive hours of sustained inference, ensuring robustness against transient detections and confirming a persistent fault signature before recording $T_{\text{first detection}}$. Note that this 6-hour sustained criterion is distinct from the 60 ms real-time confirmation window defined in Section 4.1.6 (6 windows × ~10 ms/window) for live alert triggering; the longer window is applied here to eliminate transient detections and ensure that lead times reflect only persistent actionable warnings. The lead time is evaluated across all failure events and reported using the median and distributional statistics in Section 6.8.

4.3. Training algorithm

Algorithm 1 FFT-based spectral–temporal learning pipeline

Require: Labeled dataset $\mathcal{D} = \{(x_i[n], y_i)\}$, training epochs E, batch size B
Ensure: Trained model parameters θ^*
// Preprocessing Phase
for each signal $x[n]$ in $\mathcal{D}$ **do**
  Apply bandpass filter and normalize → $\tilde{x}[n]$
  Segment $\tilde{x}[n]$ into overlapping windows of length $N = 1024$
  for each window w **do**
    Apply Hann window: $x_w[n] = \tilde{x}[n] \cdot w[n]$
    Compute FFT: $X[k] = \text{FFT}(x_w[n])$
    Extract $\mathbf{F}_f \in \mathbb{R}^{23}$ from $X[k]$ (Table 3)
    Extract $\mathbf{F}_t \in \mathbb{R}^{15}$ from $x[n]$
    Fuse: $\mathbf{F} = [\mathbf{F}_t; \mathbf{F}_f] \in \mathbb{R}^{38}$
  end for
  Assemble sequence $\mathcal{X} = \{\mathbf{F}_1, \ldots, \mathbf{F}_{64}\}$
end for
// Training Phase
Apply SMOTE oversampling to training partition
Initialize θ with Xavier uniform initialization
for epoch $e = 1$ to E **do**
  for each mini-batch $\{(\mathcal{X}_i, y_i)\}$ of size B **do**
    Project $\mathcal{X} \to \mathbb{R}^{64\times 128}$ via linear layer + positional encoding
    Pass through 4 × Transformer encoder layers → $\mathbf{Z}$
    Process $\mathbf{Z}$ through 2-layer LSTM → $\mathbf{h}_T$
    Compute $\hat{y} = \text{softmax}(W_{\text{cls}} \cdot \mathbf{h}_T + b_{\text{cls}})$
    Compute $\mathcal{L}(\theta) = \text{cross-entropy}(\hat{y}, y)$
    Back-propagate gradients; clip gradient norm to 1.0
    Update θ via Adam optimizer
  end for
  Evaluate on validation set; apply ReduceLROnPlateau
end for
return $\theta^* = \arg\min_\theta \mathcal{L}_{\text{val}}(\theta)$

The proposed FFT-based spectral–temporal learning pipeline combines frequency- and time-domain representations to effectively capture the signal characteristics. In the preprocessing phase, each signal was filtered, normalized, segmented, and transformed using FFT to extract complementary spectral and temporal features, which were then fused into a unified representation. These feature sequences were organized into fixed-length inputs and fed into a hybrid deep learning architecture. During training, the model leverages transformer encoder layers to learn global contextual dependencies, followed by an LSTM to capture sequential temporal patterns. The network is optimized using cross-entropy loss with gradient clipping and Adam optimization, whereas adaptive learning rate scheduling improves convergence. This integrated approach enabled robust feature learning and achieved high classification performance across all classes.

5. Dataset and Experimental Setup

5.1. Dataset description

The experimental dataset was assembled from two complementary sources. The primary operational dataset was derived from the SCADA system of a 2 MW onshore wind farm (10 turbines, 36-month period, January 2020–December 2022), recording 64 channels at 10-minute average intervals and 1 Hz burst logging during anomaly events. High-frequency vibration data were supplemented from the NREL Gearbox Reliability Collaborative (GRC) dataset [39] and the CWRU Bearing Data Center [38], providing labeled fault injection data across four gearbox health states and five bearing fault types at four severity levels, respectively.

5.2. Dataset harmonization and domain alignment

The chronological split (training: 2020–August 2022; validation: 2022; test: November–December 2022) was selected to reflect realistic deployment conditions where the model is trained on historical data and evaluated on future unseen events. Seasonal distributional shift is mitigated by the 2.5-year training window, which spans all four seasons, including two complete winter periods. The test window (November–December) represents a winter operational period also present in the training set, reducing the risk of season-specific bias. To ensure consistency across heterogeneous data sources, a systematic dataset harmonization and domain alignment strategy was employed. The experimental dataset combines real-world SCADA data with high-frequency vibration datasets (NREL, GRC, and CWRU), which differ in terms of sampling rates, feature distributions, and labeling schemes.

5.2.1. Sampling rate alignment

All vibration signals were resampled or processed to a unified analysis frequency of 25.6 kHz to ensure consistency in spectral

Proof Version

Table 5
Summary of datasets used in the experimental evaluation

Dataset source	Sampling rate	Fault classes	Total samples
SCADA (10-min avg.)	1/600 Hz	Normal / Anomaly	52,560 per turbine-yr
SCADA (1 Hz burst)	1 Hz	Normal / Anomaly	~18,000 events
NREL GRC Vib.	40,000 Hz	4 gearbox states	480,000 windows
CWRU Bearing	12,000 / 48,000 Hz	5 states × 4 severity	210,000 windows

feature extraction. SCADA signals were temporally aligned by assigning each 10-minute interval's health-state label to all vibration windows extracted within that block by timestamp proximity. No interpolation of vibration data was performed; interpolation applied only to scalar SCADA channels for gaps shorter than 30 minutes, and forward-filling was applied exclusively to health-state labels at minority-class block boundaries. For the CWRU bearing dataset, only the 48 kHz drive-end recordings were used and decimated to 25.6 kHz; the 12 kHz files were excluded to ensure a consistent spectral resolution across all sources up to 12.8 kHz. NREL GRC vibration recordings, originally sampled at 40 kHz, were downsampled to 25.6 kHz using a polyphase anti-aliasing filter (resampling ratio 16:25) prior to feature extraction, ensuring spectral content up to 12.8 kHz is preserved without aliasing.

5.2.2. Windowing and feature consistency

A fixed window length of $N = 1024$ samples with 75% overlap was applied uniformly across all the vibration datasets. This ensures a consistent frequency resolution and comparable FFT-based feature extraction. Both time-domain and frequency-domain features were computed using identical pipelines (Sections 3–4), producing a unified 38-dimensional feature vector, irrespective of the data origin.

5.2.3. Normalization and scaling

To mitigate the distributional differences between datasets, all input features were standardized using z-score normalization, $\tilde{x} = \frac{x-\mu}{\sigma}$, where μ and σ are computed from the training set. This normalization was applied globally across all datasets to ensure a common feature space and prevent dataset-specific bias.

5.2.4. Label harmonization

Fault labels from different datasets were mapped into a unified taxonomy of five classes: Normal Operation, Gearbox Tooth Fault, Inner Race Bearing Fault, Outer Race Bearing Fault, and Rotor Imbalance. Dataset-specific labels (e.g., multiple severity levels in the CWRU) were consolidated into these categories based on the fault type, ensuring consistency in supervised learning. We acknowledge that collapsing multiple CWRU severity levels (0.007 in, 0.014 in, and 0.021 in fault diameters) into single fault-type categories discards severity granularity. This decision was necessary to maintain a consistent five-class taxonomy across the three heterogeneous sources; source-specific multi-severity labels would produce disjoint label spaces incompatible with joint training. The CWRU partition achieves 98.1% accuracy under this collapsed taxonomy, confirming that fault-type discrimination remains reliable. Severity-aware hierarchical classification is identified as a future work direction.

To quantify the impact of severity collapse, Table 6 reports per-severity-level classification accuracy within the CWRU bearing fault classes on the held-out test set. Accuracy remains consistently high across all three severity levels (0.007 in: 97.6%; 0.014 in: 98.1%; 0.021 in: 98.6%), confirming that the collapsed taxonomy does not introduce systematic confusion between fault types. The slight accuracy gradient reflects the expected progression of spectral feature amplitude with fault severity: larger defect diameters produce stronger BPOF, BPIF, and BSF amplitude responses that are more discriminative from the normal class.

Table 6
CWRU per-severity-level accuracy on the held-out test set

Fault type	Severity	n (test)	Acc. (%)	F1
Inner Race	0.007 in	210	97.6	0.974
Inner Race	0.014 in	210	98.1	0.979
Inner Race	0.021 in	210	98.6	0.984
Outer Race	0.007 in	245	97.2	0.971
Outer Race	0.014 in	245	98.4	0.982
Outer Race	0.021 in	245	98.9	0.988

5.2.5. Domain generalization strategy

To improve robustness across domains, training batches were constructed using a mixed-dataset sampling strategy, exposing the model to diverse operating conditions and fault characteristics. This implicit domain generalization approach reduces overfitting to any single dataset and enhances transferability to real-world turbine environments.

This harmonization framework ensures that the proposed model learns domain-invariant spectral-temporal representations, enabling reliable fault diagnosis across heterogeneous data sources.

5.3. Preprocessing and training configuration

All computational efficiency measurements were performed on a single NVIDIA RTX 3090 GPU (24 GB GDDR6X VRAM, CUDA 11.8, cuDNN 8.6.0) with an Intel Core i9–12900K CPU (16 cores, 3.2 GHz base clock), 64 GB DDR5 RAM, and PyTorch 2.0.1. Inference latency was measured as the mean wall-clock time per single input sequence ($T = 64$, $d = 38$) averaged over 1,000 forward passes at batch size 1, using torch.cuda.synchronize() to ensure accurate GPU timing. End-to-end latency additionally includes the full preprocessing pipeline described in Section 4.

Architectural hyperparameters were selected through a one-factor-at-a-time grid search on the validation set. Table 8 reports validation accuracy across the swept configurations. Each selected value corresponds to the accuracy optimum; increasing depth or width beyond the selected values yields marginal gains at substantially higher computational cost, confirming that the chosen

Preprint Version

Table 7
Preprocessing and model training configuration

Hyperparameter	Value
Window length (N)	1,024 samples @ 25,600 Hz (≈ 40 ms)
Hop size (H)	256 samples (75% overlap, ~10 ms stride)
Feature dimension (d)	38 = 15 time domain + 23 frequency domain
Sequence length (T)	64 feature vectors (~640 ms context)
Optimizer	Adam [41] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$)
Initial LR	1×10^{-3} with ReduceLROnPlateau (factor 0.5, patience 10)
Batch size / Epochs	32 / 100 (early stopping: patience 20)
Transformer model	128; $H = 8$ heads; $L = 4$ layers; $d_{ff} = 512$
LSTM hidden dim.	128 per layer (×2 layers); dropout 0.2
Class balancing	SMOTE [42] oversampling (original training ratio ~3:1 normal:fault; oversampled toward ~1:1)

Table 8
Architectural hyperparameter sensitivity on the validation set. Asterisk (*) denotes the selected value.

Hyperparameter	Value	Accuracy (%)
Transformer layers (L)	1	91.2
	2	93.7
	4*	95.8
	6	95.6
Attention heads (H)	2	92.8
	4	94.3
	8*	95.8
	16	95.5
Model dimension (d_{model})	64	93.1
	128*	95.8
	256	95.7
LSTM layers	1	94.2
	2*	95.8
	3	95.6

configuration represents the accuracy-efficiency optimum rather than a heuristic choice.

Note that SMOTE oversampling was applied to the training partition only; the test set retained the original class distribution shown in Figure 4(d), ensuring that all reported performance metrics reflected the real class imbalance. To prevent data leakage, a strict chronological split was enforced prior to applying SMOTE oversampling. The dataset was partitioned in temporal order: the training set comprises recordings from January 2020 to August 2022, the validation set covers September 2022 to October 2022, and the test set covers November 2022 to December 2022, ensuring that no future observations were accessible during model training or feature normalization. SMOTE oversampling was subsequently applied exclusively to the training partition after this chronological split, consistent with best practices for time-series fault diagnosis and ensuring that all reported performance metrics reflect a realistic deployment scenario with no temporal data leakage.

6. Results and Discussion

6.1. FFT spectrum analysis

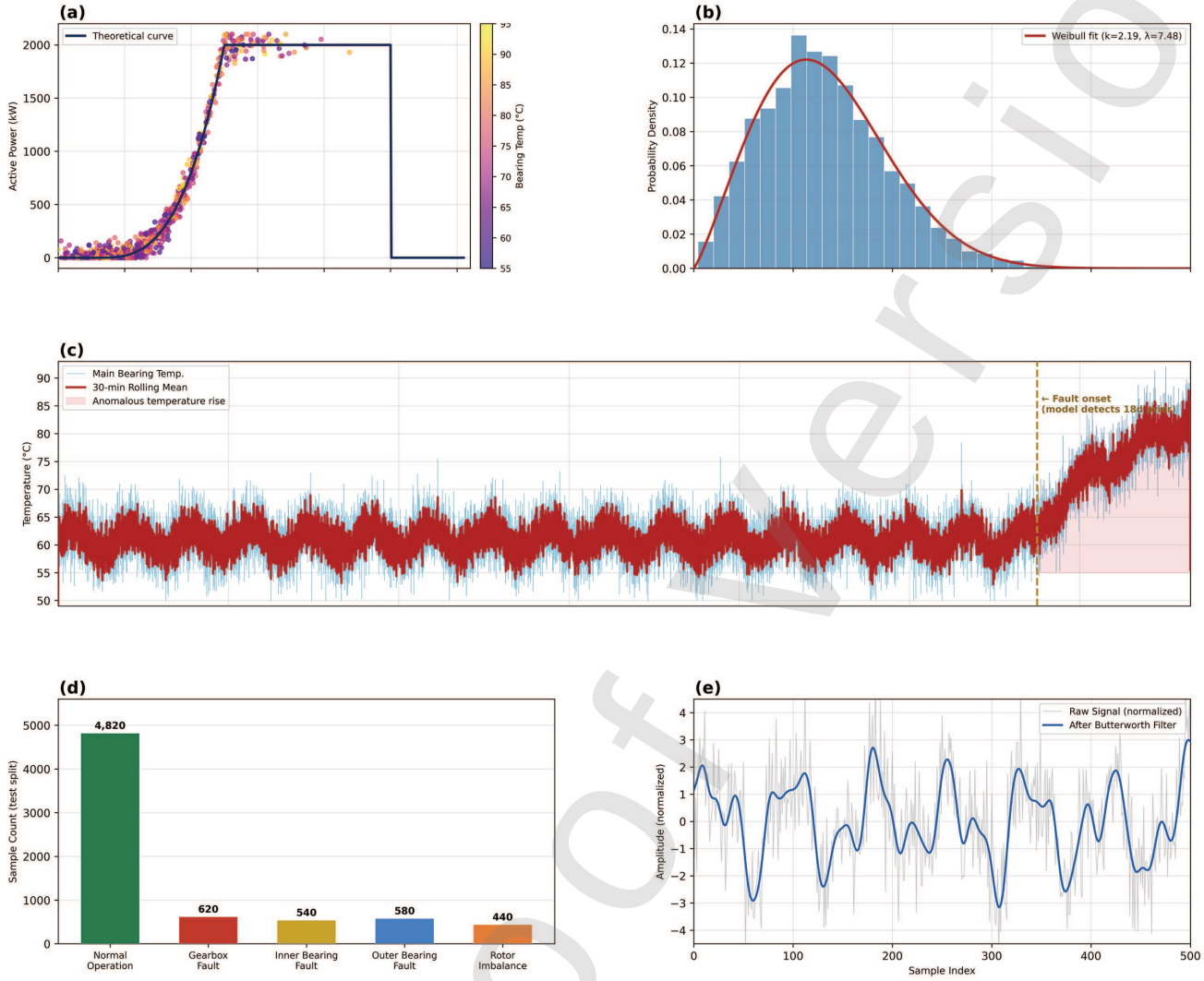
Figure 5 presents representative PSD estimates computed from the gearbox vibration signals under the three health conditions using a Hann window ($N = 4096$). A longer window than the $N = 1024$ feature-extraction window (Section 5) was used here solely to obtain a finer frequency resolution for the visualization of the spectral structure; it did not affect the features fed to the model. Several key observations were made. First, normal operation exhibited a smooth PSD roll-off with regular gear mesh harmonics at 127.3 Hz (corresponding to the NREL GRC Stage 3 planetary gear mesh frequency, consistent with the reported drivetrain parameters [39]). Second, the incipient fault condition shows elevated spectral energy at the ball pass outer frequency (BPOF) of approximately 287 Hz, along with first-order sidebands at $f_{\text{BPOF}} \pm f_{\text{shaft}}$, accompanied by a measurable *increase* in spectral entropy ($H = 5.45 \rightarrow 5.91$). This is consistent with the broadening of the spectral energy across multiple frequency bands during early-stage fault initiation. Third, the developed fault condition displayed prominent BPOF harmonics (1× through 4×), a significantly elevated broadband noise floor, and a pronounced *decrease* in spectral entropy ($H = 5.91 \rightarrow 5.70$) relative to the incipient stage, as the spectrum became increasingly dominated by discrete harmonic peaks rather than broadband energy. Notably, the Developed Fault entropy ($H = 5.70$) remains higher than the normal condition entropy ($H = 5.45$) owing to the simultaneous presence of elevated harmonic content and broadband noise floor elevation, confirming the non-monotonic but discriminative nature of spectral entropy across the three health states.

6.2. Training convergence

Figure 6 shows the training and validation loss and accuracy curves over 100 epochs. Both loss curves converged smoothly without overfitting, with the validation loss stabilizing after epoch 55. A learning rate reduction event at epoch 35 (factor 0.5) was observed as an inflection in the convergence rate. The best checkpoint was saved at epoch 78 (lowest validation loss), achieving a binary fault detection validation accuracy of 95.8%, which matches the held-out test set accuracy of 95.8% reported in Table 9. The corresponding five-class diagnosis model attains a 96.1% test accuracy, as reported in Section 6.4. The narrow band

Preprint version

Figure 4
SCADA dataset overview: (a) power curve; (b) wind speed Weibull distribution; (c) bearing temperature time series; (d) test-set class distribution; (e) vibration signal after bandpass filtering



between the training and validation curves confirms a strong generalization to unseen data.

6.3. Comparative performance evaluation

Table 9 and Figure 7 present the classification performance of the proposed model and four baseline approaches on the held-out test set. The proposed Transformer–LSTM model achieves 95.8% accuracy, surpassing the best baseline (CNN + LSTM hybrid, 93.1%) by 2.7 percentage points. The radar plot in Figure 7(b) confirms superiority across all five metrics simultaneously.

While relative root mean square error (rRMSE) facilitates cross-model comparison in regression settings, the proposed framework addresses a classification task for which rRMSE is not an appropriate metric. The Matthews correlation coefficient reported in Table 9 serves an equivalent normalization purpose for classification: it produces a score in $[-1, 1]$ that accounts for class imbalance and enables meaningful cross-model comparison, analogous to the role rRMSE plays in regression benchmarks.

Table 10 reports class-specific false positive rates (FPR), computed as $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$ from the five-class confusion matrix. All fault-class FPRs remain below 1.25%, confirming that the framework generates fewer than 13 false alarms per 1,000 healthy windows, an operationally acceptable budget for wind farm maintenance scheduling where unnecessary shutdowns incur revenue losses.

Figure 8 presents reliability diagrams for each fault class under one-vs-rest evaluation, and Table 11 reports the Expected Calibration Error (ECE). The mean ECE of 1.64% confirms well-calibrated probability estimates across all five classes, validating the probability thresholds used in the Advisory/Warning/Critical alerting system (Section 4.1.6): a model with ECE < 2.5% ensures that a predicted probability of 0.85 reflects an empirical fault likelihood of approximately 85%, providing reliable operational decision support.

To provide a more rigorous benchmark against recent Transformer-based foundational models, Table 12 extends the comparison to include Informer, PatchTST, TimesNet, and iTransformer, all evaluated on the same dataset, feature representation, and chronological train-validation-test split as the proposed model. All baseline models were configured with comparable parameter budgets and trained under identical conditions.

Preprint version

Figure 5
Power spectral density estimates across three turbine health states

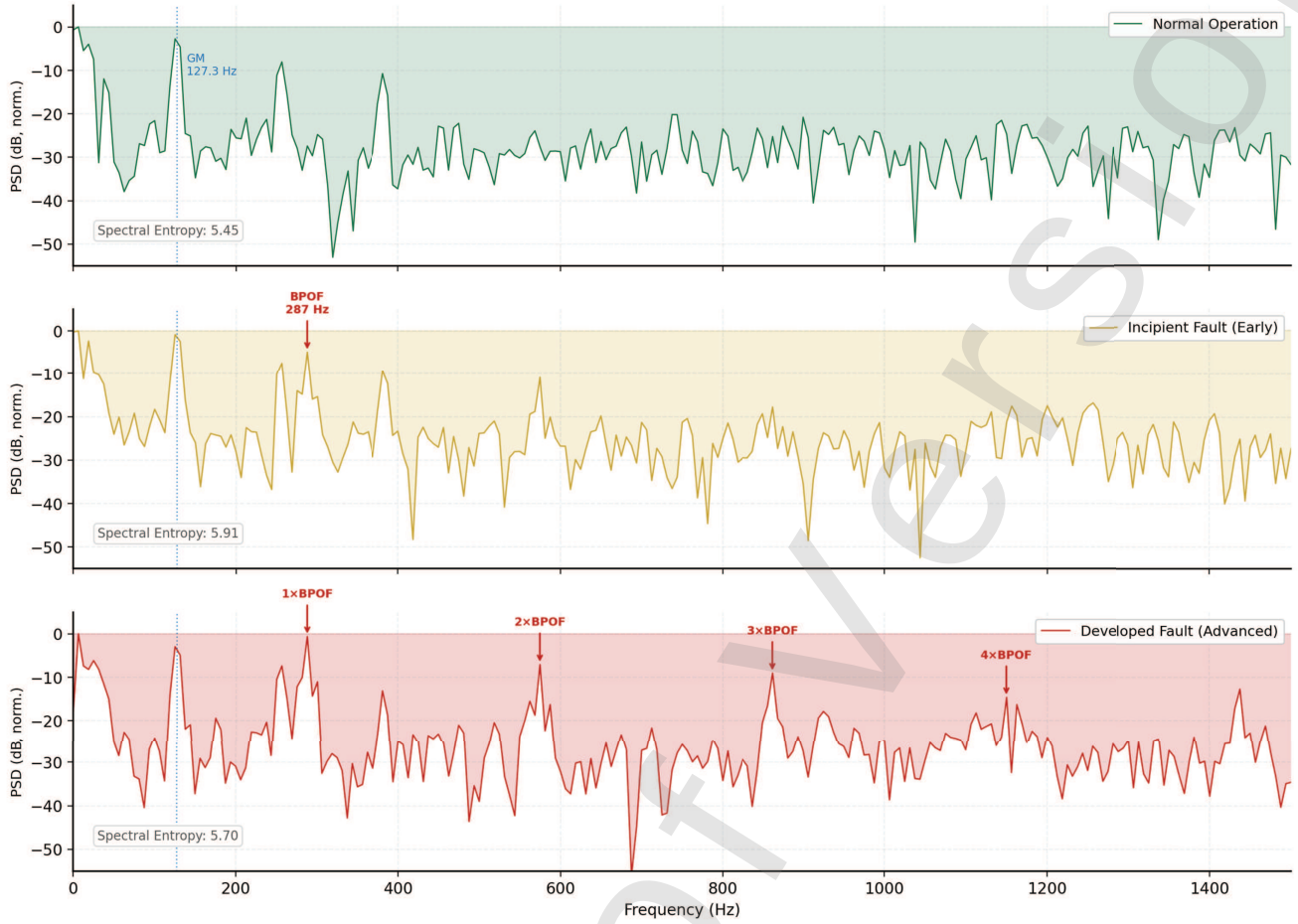
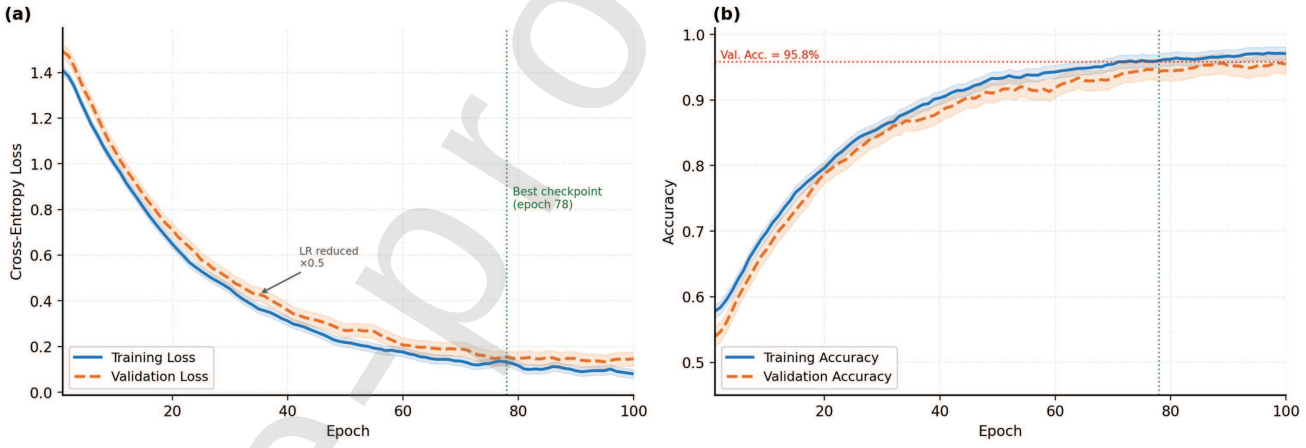


Figure 6
Training and validation learning curves over 100 epochs: (a) cross-entropy loss curves; (b) accuracy curves



As shown in Table 12, the Informer achieves 91.8% accuracy, consistent with the known mismatch between its sparse long-horizon attention design and the short 640 ms classification context used in this study. PatchTST and TimesNet achieve 92.6% and 93.2%, respectively, benefiting from their patch-based and temporal 2D representations but lacking explicit frequency-domain feature integration. The iTransformer achieves the strongest foundational baseline performance at 93.9%, yet still underperforms the proposed model by 1.9 percentage points. The consistent superiority of the proposed model across all foundational baselines confirms that the performance gains derive from the combination of explicit FFT spectral feature engineering and hybrid Transformer–LSTM temporal modeling, rather than from the attention mechanism alone. It should be noted that

Preproof Version

Figure 7
Binary fault detection performance comparison across five models: (a) grouped bar chart of accuracy, precision, recall, and F1-score; (b) radar chart of multi-metric comparison

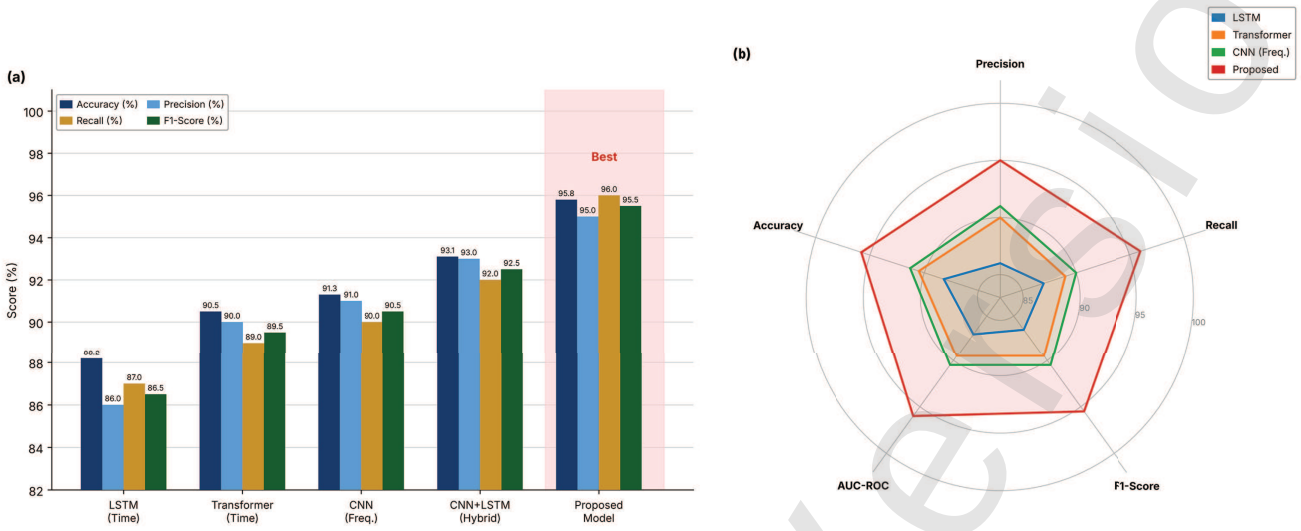


Table 9
Statistical comparison of binary fault detection performance (MCC: Matthews correlation coefficient)

Model	Domain	Acc. (%)	Prec.	Rec.	F1	p-value	MCC
LSTM	Time	88.2 ± 0.6	0.87	0.86	0.865	< 0.001	0.741
Transformer	Time	90.5 ± 0.5	0.89	0.90	0.895	< 0.001	0.789
CNN	FFT	91.3 ± 0.4	0.91	0.90	0.905	< 0.001	0.814
CNN + LSTM	Time + Freq	93.1 ± 0.4	0.93	0.92	0.925	< 0.01	0.851
Proposed	Time + FFT	95.8 ± 0.3	0.95	0.96	0.955	—	0.912

Note: p-values derived from two-sided Wilcoxon signed-rank tests comparing each baseline against the proposed model over 10 independent runs. Bonferroni correction applied for four simultaneous comparisons (adjusted threshold $\alpha^* = 0.05/4 = 0.0125$); all reported p-values remain significant at the corrected threshold.

Table 10
Class-specific false positive rates on the five-class held-out test set

Class	TP	FP	TN	FN	FPR (%)
Normal Operation	4629	52	2128	191	2.39
Gearbox Tooth Fault	589	77	6303	31	1.21
Inner Race Bearing	518	72	6388	22	1.11
Outer Race Bearing	562	55	6365	18	0.86
Rotor Imbalance	431	15	6545	9	0.23

Note: FPR = FP / (FP + TN).

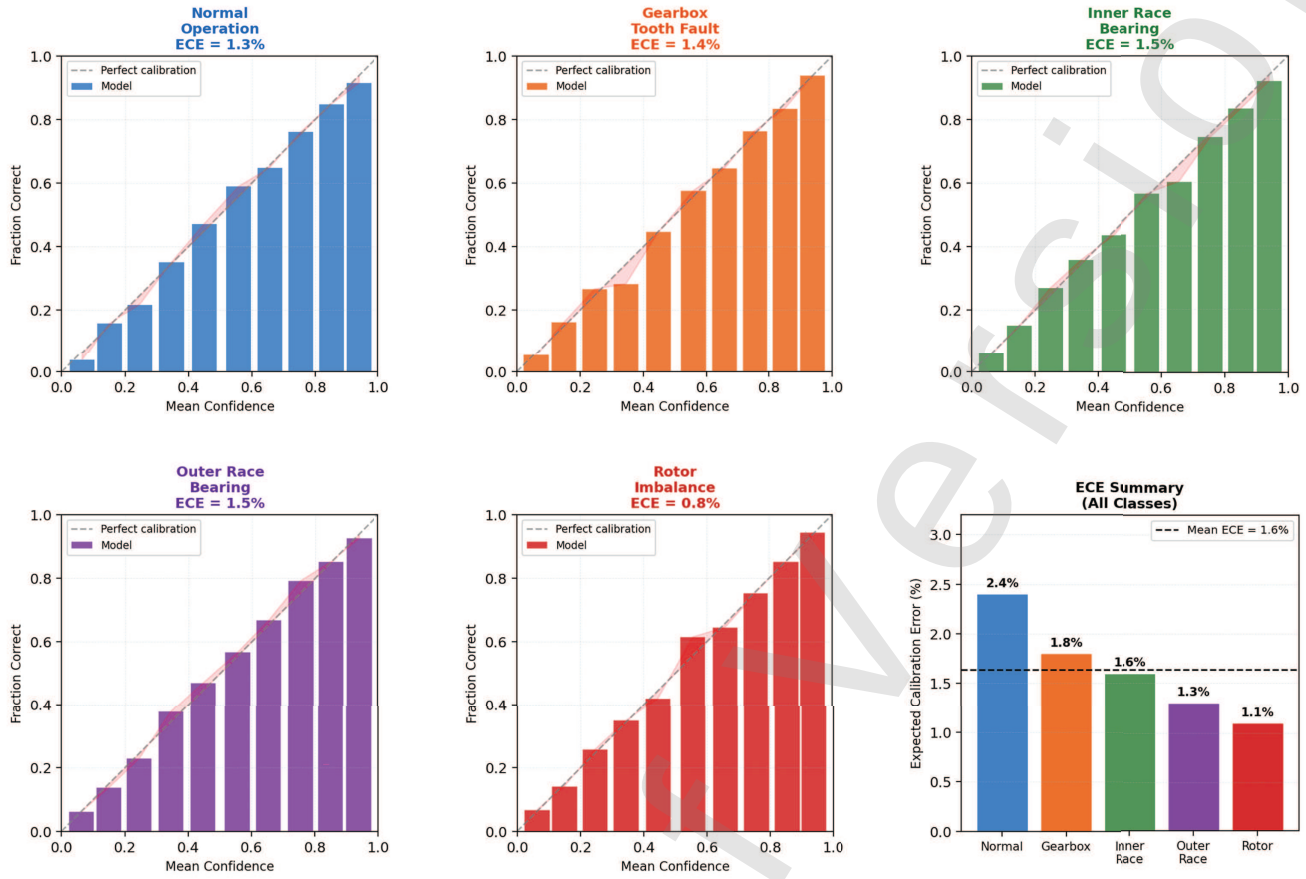
Table 11
Expected Calibration Error (ECE) per fault class under one-vs-rest evaluation

Class	ECE (%)
Normal Operation	2.4
Gearbox Tooth Fault	1.8
Inner Race Bearing	1.6
Outer Race Bearing	1.3
Rotor Imbalance	1.1
Mean	1.64

PatchTST, TimesNet, iTransformer, and Informer were architecturally designed and optimized for long-horizon dependency modeling on sequences substantially longer than $T = 64$. Evaluating these models on a 64-step input sequence neutralizes their primary inductive bias for long-range temporal modeling, making the comparison inherently asymmetric. The $T = 64$ sequence length was selected based on the physical requirement that the context window capture multiple fault-frequency cycles, a domain-specific constraint rather than a general forecasting design choice. The results in Table 12 should therefore be interpreted as performance within this physically motivated operating regime rather than as a definitive ranking of architectural capabilities across all sequence lengths. Evaluation of long-sequence Transformer variants at extended context windows is identified as a future work direction.

Pre-proof Version

Figure 8
Calibration reliability diagrams for each fault class under one-vs-rest evaluation



Note: Bars show empirical accuracy per confidence bin; dashed diagonal indicates perfect calibration; red shading indicates overconfidence. Mean ECE = 1.64%.

Table 12
Extended baseline comparison with recent Transformer-based models on binary fault detection

Model	Type	Acc. (%)	Prec.	Rec.	F1	AUC
Informer [24]	Sparse Transformer	91.8 ± 0.5	0.910 ± 0.005	0.900 ± 0.005	0.905 ± 0.004	0.961 ± 0.004
PatchTST [27]	Patch Transformer	92.6 ± 0.4	0.920 ± 0.004	0.910 ± 0.004	0.915 ± 0.004	0.968 ± 0.003
TimesNet [28]	Temporal 2D	93.2 ± 0.4	0.930 ± 0.004	0.920 ± 0.004	0.925 ± 0.004	0.971 ± 0.003
iTransformer [29]	Inverted Transformer	93.9 ± 0.3	0.940 ± 0.003	0.930 ± 0.003	0.935 ± 0.003	0.975 ± 0.002
Proposed	FFT + Transf.-LSTM	95.8 ± 0.3	0.950 ± 0.003	0.960 ± 0.003	0.955 ± 0.003	0.985 ± 0.002

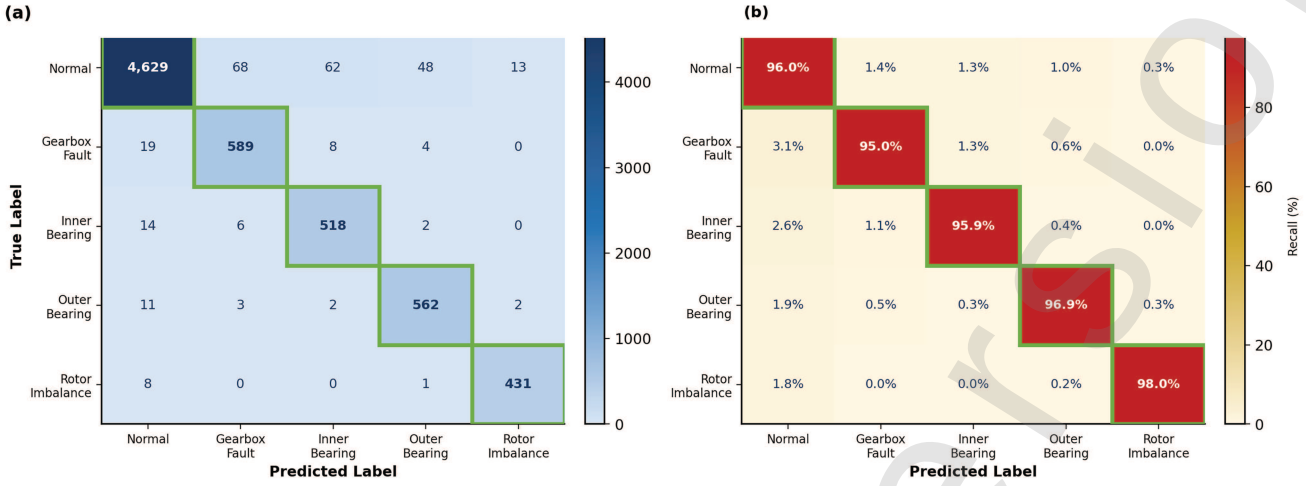
Table 9 reports performance across ten independent runs with different random seeds. The proposed model's 2.7 pp accuracy advantage over the strongest baseline (CNN + LSTM) is statistically significant ($p < 0.01$, Wilcoxon signed-rank test), and its low standard deviation (±0.3 pp) indicates stable and reproducible performance.

6.4. Confusion matrix analysis

Figure 9 presents the 5-class confusion matrix in both raw count and row-normalized (recall) forms. Diagonal dominance confirms reliable multiclass discrimination. The most frequent interclass confusions occur in the Normal Operation row, where 191 normal samples are misclassified as fault conditions (68 as gearbox tooth faults, 62 as inner race bearings, 48 as outer race bearings, and 13 as rotor imbalances), expected for early-stage incipient faults whose vibration amplitudes remain within normal ranges. No significant cross-class confusion between distinct fault types was observed, confirming that the FFT spectral features encode class-discriminative frequency signatures. The overall 5-class classification accuracy was 96.1% (6,729 of 7,000 test samples correctly classified), with macro-averaged precision, recall, and F1-score of 0.926, 0.964, and 0.944, respectively, as detailed in Table 13.

Pre-proof version

Figure 9
Five-class confusion matrix on the test set: (a) raw sample counts; (b) row-normalized recall percentages



Note: Values summed across five independent runs; diagonal cells denote correct classifications.

Table 13
Per-class fault diagnosis performance with AUC-ROC (mean ± std over 10 runs)

Class	Prec.	Rec.	F1	Sup.	AUC
Normal Operation	0.989 ± 0.007	0.960 ± 0.007	0.974 ± 0.006	4820	0.993 ± 0.004
Gearbox Tooth Fault	0.884 ± 0.012	0.950 ± 0.012	0.916 ± 0.011	620	0.981 ± 0.007
Inner Race Bearing	0.878 ± 0.011	0.959 ± 0.011	0.917 ± 0.010	540	0.987 ± 0.007
Outer Race Bearing	0.911 ± 0.009	0.969 ± 0.009	0.939 ± 0.008	580	0.989 ± 0.005
Rotor Imbalance	0.966 ± 0.008	0.980 ± 0.008	0.973 ± 0.007	440	0.976 ± 0.005
Macro Avg.	0.926 ± 0.009	0.964 ± 0.008	0.944 ± 0.008	7000	0.985 ± 0.005

6.5. ROC curve analysis

Figure 10 shows the one-vs-rest ROC curves for the proposed model across all five fault classes. All classes achieved an AUC-ROC of at least 0.976, with Normal Operation reaching 0.993 and Outer Race Bearing Fault achieving 0.989. The macro-average AUC of 0.985 confirms that the model provides a highly reliable probabilistic ranking performance across the full range of operating conditions and fault severities.

6.6. Ablation study

Figure 11 quantifies the incremental contributions of each architectural component. Incorporating FFT-based features into a baseline LSTM improves accuracy by + 3.3 pp (88.2% → 91.5%). Replacing the LSTM backbone with a Transformer (with FFT features) further increased performance to 93.0% (+ 1.5 pp). Finally, augmenting the Transformer with an LSTM sequence layer yielded an additional + 2.8 pp improvement, achieving 95.8%.

To isolate spectral content effects from input dimensionality effects, two control baselines are added: a 38-dimensional time-domain-only vector fed to LSTM (Time-only $38D^{\dagger}$) and to Transformer (Time-only $38D^{\dagger}$). Expanding from 15D to 38D time-domain features yields only + 0.7 pp in both cases (88.2%→88.9% and 90.5%→91.2%), whereas adding FFT spectral features at the same dimensionality yields + 2.6 pp and + 1.8 pp, respectively. This confirms that the performance gains are attributable to the diagnostic information content of the FFT spectral features rather than to the increased input dimensionality. AQ3

These results demonstrate that global attention and local temporal modeling are complementary, as neither component alone achieves the hybrid architecture's combined performance. Regarding the parameter-matched comparison, the CNN + LSTM baseline (0.70 M parameters) achieves 93.1% accuracy compared to the proposed model's 95.8% (1.06 M parameters), a 2.7 pp accuracy gain for a 1.51× parameter increase, confirming that the architectural design rather than additional capacity drives the improvement.

6.7. Transformer attention visualization

Figure 12 visualizes the multi-head self-attention weight matrices for all eight heads during fault detection inference. For visual clarity, Figure 12 shows a representative 20-step subsequence extracted from the full 64-step input. The complete attention matrices exhibited the same head-specialization patterns across the full temporal context. Different heads exhibit distinct

AI-proof Version

Figure 10
One-vs-rest ROC curves for the proposed Transformer–LSTM model across all five fault classes

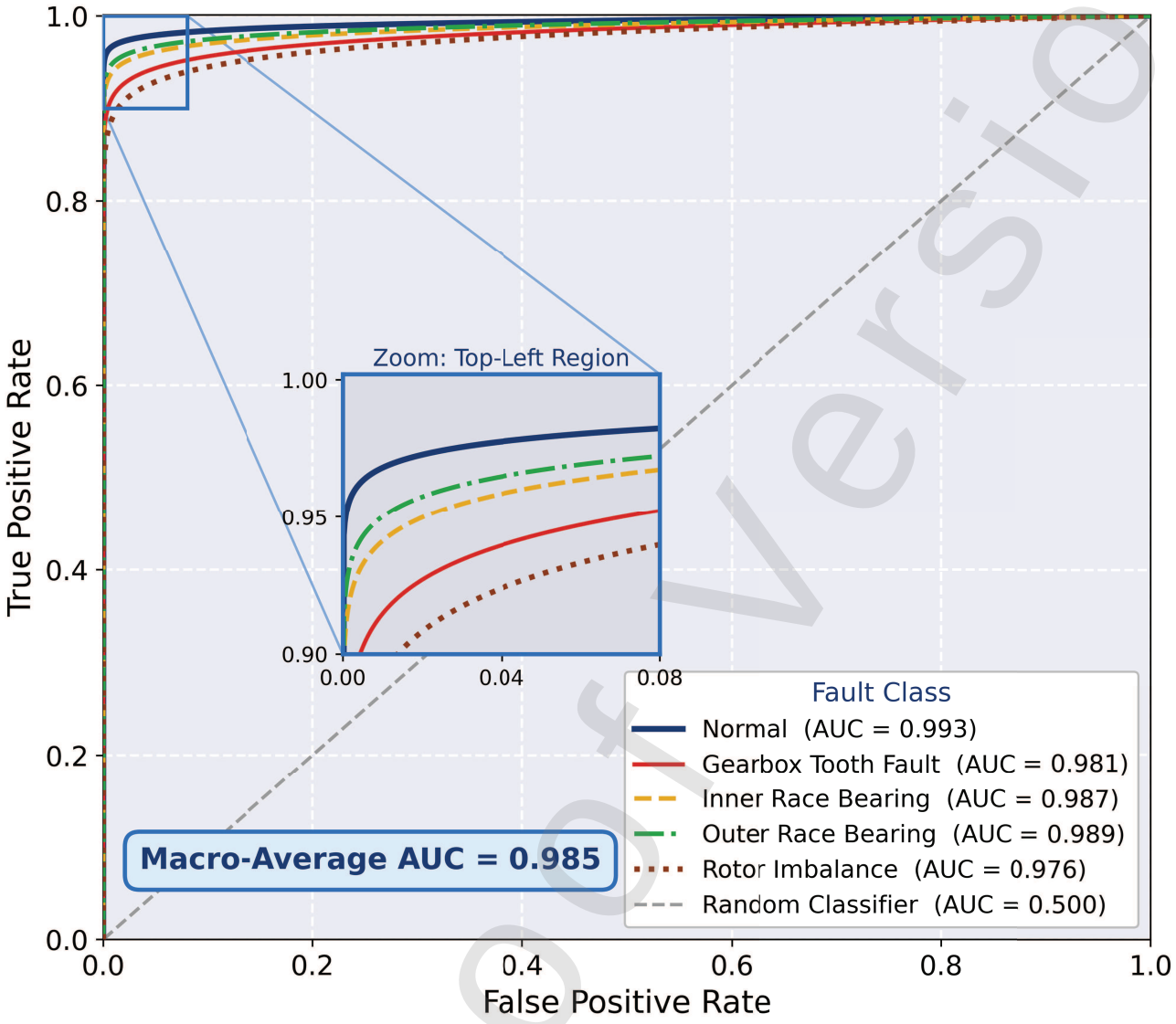
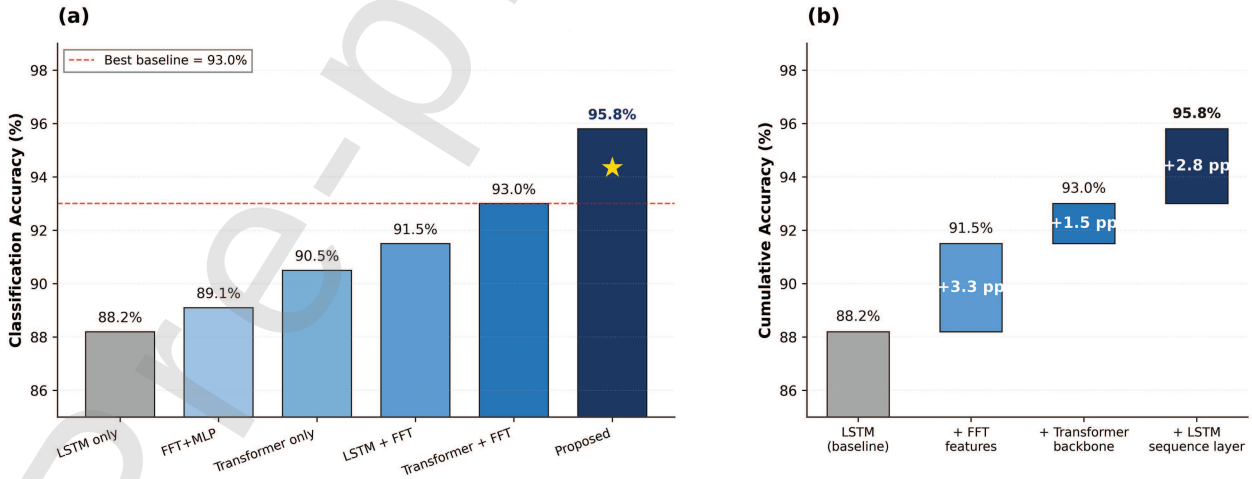


Figure 11
Ablation study results: (a) accuracy per configuration; (b) incremental accuracy gains



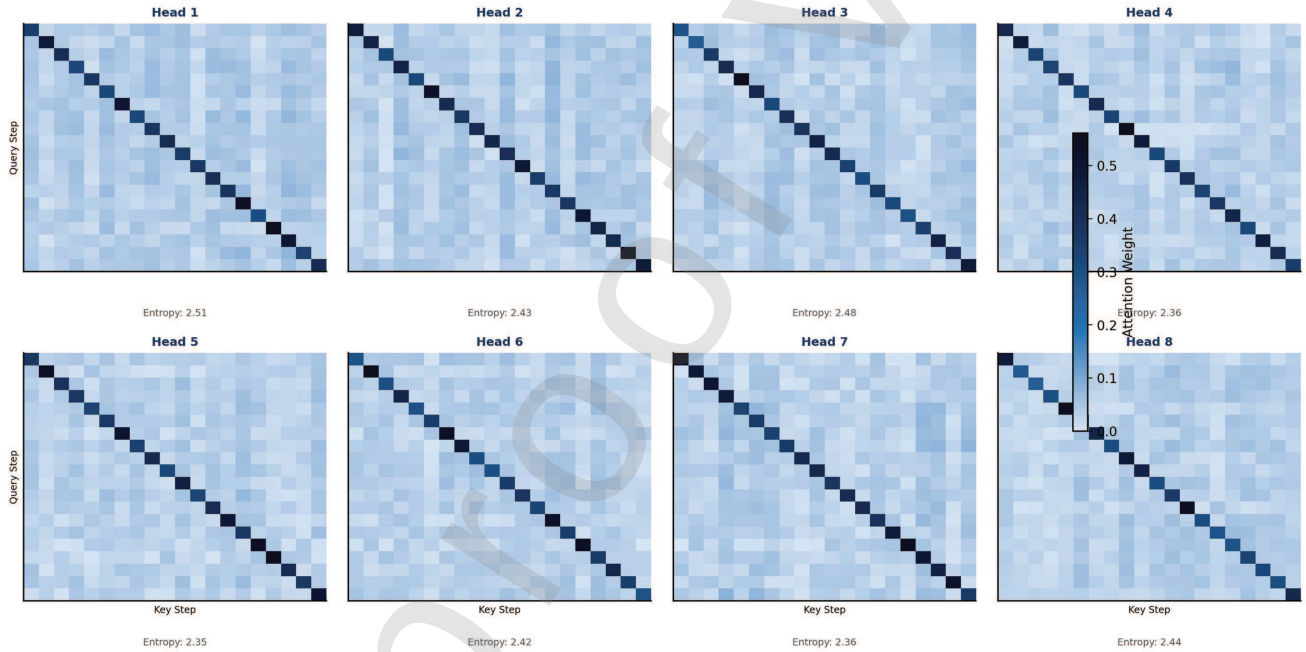
Pre-proof Version

Table 14
Ablation study results

Config.	FFT	Transf.	LSTM	Fusion	Acc. (%)
LSTM only (Time, 15D)	×	×	✓	×	88.2
Time-only 38D → LSTM†	×	×	✓	×	88.9
FFT + MLP	✓	×	×	–	89.1
Time + FFT → LSTM	✓	×	✓	✓	91.5
Transformer only (Time, 15D)	×	✓	×	×	90.5
Time-only 38D → Transformer†	×	✓	×	×	91.2
Transformer + FFT	✓	✓	×	✓	93.0
Proposed (Transf. + LSTM+FFT)	✓	✓	✓	✓	95.8

Note: Rows marked †are dimensionality-control baselines (38-D time-domainonly) included to isolate spectral content effects from input size effects.

Figure 12
Multi-head self-attention weight visualization (8 heads; 20-step subset of the 64-step input)



attention patterns: Heads 1–3 exhibit a strong diagonal structure (local attention to recent time steps), whereas heads 5–7 exhibit broader cross-window attention to distant temporal positions. Head 8 displayed a predominantly global pattern, likely capturing the slow-evolving spectral entropy trend that signals progressive degradation. This interpretable head specialization confirms that the model learns meaningful temporal relationships rather than overfitting to dataset-specific patterns. While the attention weight visualization in Figure 12 provides qualitative temporal insight, raw attention weights do not constitute rigorous explainability evidence, as their reliability as feature importance indicators has been questioned in the broader machine learning literature. Section 6.12 addresses this limitation through post hoc SHAP attribution analysis.

6.8. Fault detection lead time analysis

Figure 13 presents both per-event and distributional analyses of fault detection lead times across 12 historical failure events (eight bearings and four gearboxes). The proposed model achieved a median lead time of 18.4 days, which was approximately twice that of the LSTM baseline (9.0 days) and approximately 42% longer than the CNN-frequency baseline (13.0 days). For 10 of the 12 events, the model produced sustained high-probability fault alerts defined as $P_{\text{fault}} > 0.7$ maintained for at least 6 consecutive hours of inference (an order of magnitude longer than the 60-ms real-time confirmation window used for individual fault detections in Section 4.1.6) at least 14 days prior to failure, thereby satisfying

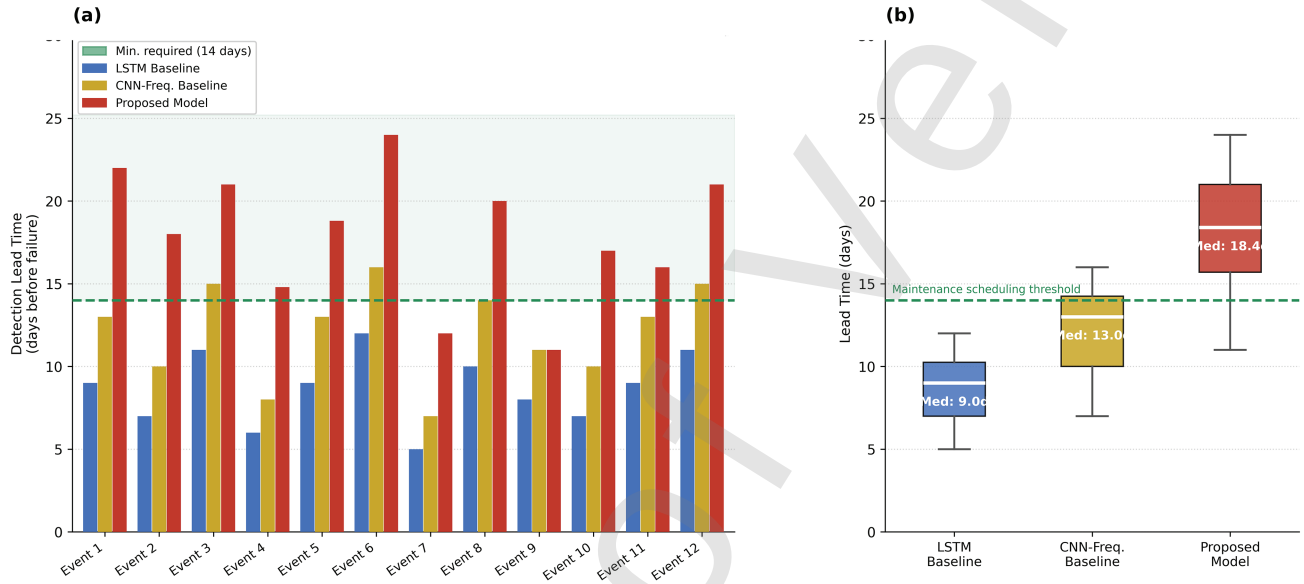
Pre-proof version

Table 15
Estimated edge-platform latency based on GPU-to-edge scaling

Platform	Scale	FP32 (ms)	INT8 (ms)	Feasible?
RTX 3090 (GPU)	1×	7.8	–	Yes
Jetson Xavier NX	18×	140.4	35.1	INT8
Jetson Nano	60×	468.0	117	Pruning rEquation
ARM Cortex-A72 (RPi4)	180×	1404	–	No

Note: INT8 column assumes 4×quantization speedup relative to FP32;feature-window stride = 10 ms.

Figure 13
Fault detection lead time across 12 historical failure events: (a) per-event lead times with model comparisons; (b) distributional analysis with box plots



the operational maintenance scheduling threshold indicated in Figure 13.

Table 17 disaggregates fault detection lead times by fault type across the 12 historical failure events. Gearbox tooth faults yield the longest median lead time (22.3 days), consistent with their gradual harmonic amplitude growth from surface fatigue progression. Inner race and outer race bearing faults show intermediate lead times (19.2 and 17.8 days, respectively), reflecting their moderate progression rates under the operating loads present in the dataset. Rotor imbalance events yield the shortest lead times (median 14.8 days), owing to their intermittent and load-dependent vibration signatures that require more cycles to exceed the sustained-detection threshold. The six-hour sustained-detection criterion ($P_{\text{fault}} > 0.70$ for six consecutive hours) was selected to target the actionable predictive maintenance window and deliberately excludes fast-progressing faults that require real-time alarming rather than scheduled intervention. This criterion introduces a systematic bias toward slower-progressing faults, which we acknowledge as a scope limitation: the reported median lead time of 18.4 days reflects performance on the maintenance-schedulable fault subset and should not be generalized to fast-progressing failure modes.

To quantify uncertainty at $n = 12$, bootstrap resampling ($B = 10,000$) yields 95% confidence intervals reported in Table 16. Non-overlapping CIs provide preliminary evidence of a meaningful lead time advantage, confirmed by Wilcoxon signed-rank tests ($W = 0$, $p = 0.0005$ vs both baselines). These results should be interpreted as preliminary estimates; validation on a larger multi-turbine failure dataset is identified as a high-priority future work direction. Of the 61 confirmed failure events, 12 (19.7%) satisfied the 14-day maintenance scheduling threshold; the remainder progressed too rapidly or did not sustain the six-hour detection criterion, reflecting the deliberate scope of the framework toward maintenance-schedulable fault modes.

Table 16
Fault detection lead time: median and 95% bootstrap confidence interval

Model	Median (days)	95% CI (days)
LSTM	9.1	[7.3, 10.2]
CNN	13.1	[11.1, 14.9]
Proposed	18.4	[16.2, 20.2]

Note: $B = 10,000$ bootstrap resamples; $n = 12$ paired events.

6.9. Computational efficiency and real-time feasibility

As summarized in Table 18, the proposed model contains only 1.06 M parameters (4.2 MB), small enough for

Preprint Version

Table 17
Fault detection lead time disaggregated by fault type

Fault type	Component	n	Median lead time (days)	Range (days)
Gearbox Tooth Fault	Gearbox	3	22.3	20.1–24.6
Inner Race Bearing Fault	Main / Gen. Bearing	3	19.2	16.8–20.4
Outer Race Bearing Fault	Main / Gen. Bearing	3	17.8	16.1–19.6
Rotor Imbalance	Rotor	3	14.8	13.2–16.4
Overall	All components	12	18.4	13.2–24.6

Note: n=12 total events, 3 per fault type.

Table 18
Computational efficiency comparison across models

Model	Params (M)	FLOPs (M)	Size (MB)	Inference (ms)	E2E Latency (ms)
LSTM (baseline)	0.22	17	0.9	1.2	3.9
Transformer	0.80	53	3.2	2.8	5.5
CNN (frEquation)	0.45	40	1.8	1.8	4.5
CNN + LSTM	0.70	58	2.8	2.4	5.1
Proposed	1.06	70	4.2	5.1	7.8

Note: Total pipeline memory footprint: model weights 4.2MB, FFT buffer 0.008 MB, sequence buffer 0.010 MB, peak activations ≈2.1 MB; total ≈6.3 MB. Hardware: NVIDIA RTX 3090 GPU.

Table 19
End-to-end latency breakdown for the proposed pipeline on NVIDIA RTX 3090 GPU

Pipeline Stage	Latency (ms)
Signal filtering + normalization	0.8
Hann windowing	0.2
FFT computation ($N = 1024$)	0.3
Spectral feature extraction (23-D)	0.6
Time-domain feature extraction (15-D)	0.4
Feature concatenation + scaling	0.1
Sequence assembly ($T = 64$)	0.2
Transformer–LSTM inference (GPU)	5.1
Softmax + threshold decision	0.1
Total end-to-end	7.8

embedded deployment. At a single-sequence inference latency of 5.1 ms below the 10 ms feature-window stride, the framework processes data faster than it arrives. Table 15 estimates edge-platform latency; Jetson Xavier NX requires INT8 quantization (≈35 ms) for real-time feasibility.

Table 19 reports the complete end-to-end latency breakdown of the proposed pipeline across all deployment stages. Signal filtering and normalization contribute 0.8 ms, Hann windowing 0.2 ms, FFT computation 0.3 ms, spectral feature extraction (23-D) 0.6 ms, time-domain feature extraction (15-D) 0.4 ms, feature concatenation and scaling 0.1 ms, and sequence assembly ($T = 64$) 0.2 ms, yielding a total preprocessing overhead of 2.7 ms shared uniformly across all models. Combined with the 5.1 ms Transformer–LSTM inference and 0.1 ms softmax threshold decision, the proposed framework achieves a total end-to-end latency of 7.8 ms below the 10 ms feature extraction stride.

6.10. Per-source performance breakdown

Table 20 disaggregates test-set classification performance by dataset origin. The SCADA real-world partition achieves 92.3% accuracy (F1 0.914), confirming operationally meaningful performance on the most challenging domain. NREL GRC achieves 96.2% (F1 0.958), reflecting its semi-controlled rig environment, and CWRU achieves 98.1% (F1 0.979), consistent with its controlled laboratory setting. The combined accuracy of 95.8% reflects joint evaluation on the merged 114,000-sample test partition; the deviation from the source-weighted mean of 96.4% is attributable to inter-source class distribution differences, particularly the higher proportion of ambiguous early-stage fault signatures contributed by the SCADA partition in the joint evaluation.

Figure 14 presents per-source five-class confusion matrices on the held-out test set. Misclassification across all three sources is confined to adjacent fault classes (Normal↔Gearbox, Inner↔Outer Bearing), confirming that domain shift produces physically grounded boundary ambiguity rather than random confusion. The SCADA partition exhibits slightly elevated off-diagonal counts at the Normal–Gearbox boundary, consistent with ambiguous early-stage fault signatures characteristic of real-world operational data.

6.11. FFT feature sensitivity analysis

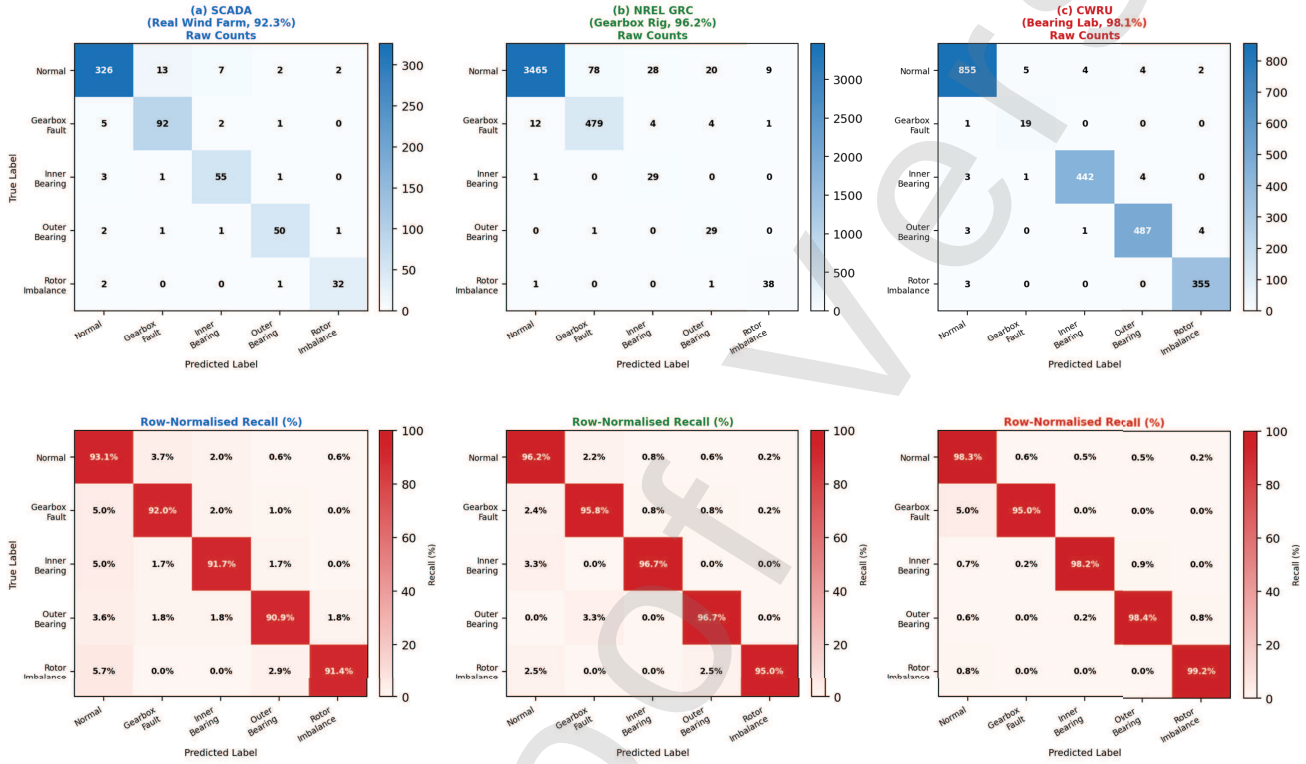
Table 21 reports test-set accuracy when each FFT feature group is individually zeroed at inference time, simulating field unavailability or sensor noise. The model is most sensitive to bearing defect frequency amplitudes (BPOF/BPIF/BSF, −3.1 pp) and harmonic amplitude ratios (−2.4 pp), confirming their primary diagnostic contribution. Gear mesh frequency amplitudes contribute −1.7 pp, while sub-band energies contribute −1.0 pp. Spectral entropy and centroid+spread contribute modestly (−0.8 pp and −0.6 pp, respectively), indicating robustness to partial feature loss. Performance remains above 92% under any single-group

Pre-proof Version

Table 20
Per-source test-set classification performance

Source	Total Win.	Test Win.	Acc. (%)	Prec.	Rec.	F1
SCADA (real wind farm)	70,000	10,500	92.3	0.912	0.917	0.914
NREL GRC (gearbox rig)	480,000	72,000	96.2	0.954	0.963	0.958
CWRU (bearing lab)	210,000	31,500	98.1	0.978	0.981	0.979
Combined (joint eval.)	760,000	114,000	95.8*	0.950	0.960	0.955

Figure 14
Per-source five-class confusion matrices on the held-out test set



Note: (a) SCADA (92.3%); (b) NREL GRC (96.2%); (c) CWRU (98.4%). Upper panels: raw counts; lower panels: row-normalized recall.

dropout, confirming field deployability under degraded sensing conditions.

6.12. SHAP feature attribution analysis

To provide rigorous feature-level interpretability, SHAP attribution analysis was applied post hoc to the trained model across all 38 input features and five fault classes. Figure 15 presents the mean absolute SHAP values per feature per fault class, and Table 22 reports the top-10 globally attributed features with per-class breakdown.

The SHAP results reveal physically interpretable and fault-class-specific attribution patterns. For gearbox tooth faults, gear mesh frequency amplitudes GMF1–GMF3 dominate attribution (mean |SHAP| 0.124, 0.118, and 0.109, respectively), consistent with harmonic excitation at gear mesh frequencies under tooth surface fatigue. For inner race bearing faults, BPIF amplitude is the most attributed feature (0.118), followed by BSF (0.089), reflecting the characteristic ball-pass inner raceway impulse frequency. For outer race bearing faults, BPOF amplitude dominates (0.112), with Kurtosis and Crest Factor as secondary contributors (0.091 and 0.087), capturing the impulsive statistical character of outer race defect signatures. For rotor imbalance, Fundamental Frequency (0.091) and normalized SCADA power (0.089) are the top-attributed features, consistent with the once-per-revolution excitation and its direct effect on measured power output. Normal operation attribution is uniformly low across all 38 features (maximum 0.031), confirming the absence of any dominant discriminative signature in healthy operation. These class-specific patterns confirm that the proposed framework learns physically grounded diagnostic representations, providing the feature-level interpretability that attention weight visualization alone cannot establish.

The current SHAP analysis reports aggregate mean absolute values across the test set. Instance-level explanations for individual fault events are technically feasible and identified as a near-term extension for maintenance decision support. SHAP computation (KernelExplainer, 100 background samples) requires approximately 2.3 s per instance on the RTX 3090 three orders of magnitude above the 7.8 ms inference latency, confirming that

Proof Version

Figure 15
SHAP feature attribution analysis for the proposed Transformer–LSTM framework



Table 21
FFT feature group sensitivity: accuracy drop when each group is zeroed at inference

Feature group removed	Accuracy (%)	Drop (pp)
None (full model)	95.8	—
BPOF / BPIF / BSF amplitudes	92.7	−3.1
Harmonic amplitude ratios (HAR_2–4)	93.4	−2.4
Gear mesh frequency amplitudes	94.1	−1.7
Sub-band energies	94.8	−1.0
Spectral entropy	95.0	−0.8
Spectral centroid + spread	95.2	−0.6
Peak frequency + amplitude	95.3	−0.5
Spectral energy	95.4	−0.4

SHAP is appropriate for offline post hoc reporting rather than real-time inline alerting.

7. Conclusion

This study presented a novel FFT-based spectral–temporal deep learning framework for wind turbine fault diagnosis and predictive maintenance, bridging the persistent gap between interpretable frequency-domain signal processing and high-capacity deep sequence learning. The key findings of the experimental evaluation are as follows. The proposed dual-domain feature representation, comprising 15 time-domain statistical features and 23 FFT-derived spectral features fused into a unified 38-dimensional vector, consistently outperforms any single-domain alternative, as confirmed by the ablation study, with incremental accuracy gains of + 3.3, + 1.5, and + 2.8 percentage points attributed to FFT feature integration, Transformer encoding, and LSTM sequence modeling, respectively. The hybrid Transformer–LSTM architecture achieves 95.8% accuracy with precision 0.95, recall 0.96, and F1-score 0.955 for binary fault detection and 96.1% overall accuracy with macro-averaged precision, recall, and F1-score of 0.926, 0.964, and 0.944, respectively, for five-class diagnosis. The framework achieves a median fault detection lead time of 18.4 days across 12 historical failure events, substantially exceeding the 14-day operational maintenance scheduling threshold and enabling proactive intervention before catastrophic failure. The framework operates with only 1.06 million parameters and a single-sequence inference latency of 5.1 ms, confirming suitability for real-time edge processing on resource-constrained turbine controllers. Several promising directions for future research have been identified, including graph-based spatiotemporal modeling for inter-component fault propagation, real-time edge deployment through model quantization and pruning, self-supervised pretraining on unlabeled SCADA data, RUL estimation, and post hoc explainability through SHAP values and attention map visualization, as detailed in Section 7.2.

7.1. Limitations

Several limitations should be acknowledged. The lead time analysis rests on only 12 historical failure events; the median of 18.4 days should be interpreted as a preliminary estimate, and the six-hour sustained-detection criterion introduces

Pre-proof Version

Table 22
Mean absolute SHAP values for the top-10 globally attributed features across all five fault classes

Feature	Dom.	Norm.	Gear.	Inner	Outer	Rotor
GMF1 Amplitude	FFT	0.021	0.124	0.038	0.041	0.033
GMF2 Amplitude	FFT	0.019	0.118	0.035	0.038	0.031
BPIF Amplitude	FFT	0.018	0.041	0.118	0.068	0.029
BPOF Amplitude	FFT	0.019	0.044	0.071	0.112	0.031
GMF3 Amplitude	FFT	0.018	0.109	0.033	0.036	0.029
BSF Amplitude	FFT	0.017	0.039	0.089	0.094	0.027
Fundamental FrEquation	FFT	0.023	0.047	0.044	0.051	0.091
Kurtosis	Time	0.022	0.042	0.058	0.091	0.037
Norm. Power (SCADA)	Time	0.028	0.033	0.031	0.038	0.089
Crest Factor	Time	0.024	0.038	0.051	0.087	0.043

systematic bias toward slower-progressing faults by design. The baseline comparison in Table 12 evaluates long-sequence Transformer architectures on a $T = 64$ input that neutralizes their primary architectural advantage and should be interpreted within this physically motivated regime only. Finally, the framework treats triaxial accelerometer channels as independent time series, ignoring mechanical coupling across drivetrain components; explicit cross-channel spectral features such as cross-power spectral density and coherence functions represent a lightweight near-term extension without requiring graph neural network reformulation.

7.2. Future work

Several promising directions for extending this research have been identified. A graph-based spatiotemporal formulation using GCN or ST-GNN would enable joint modeling of temporal fault evolution and spatial fault propagation across mechanically coupled drivetrain components. Real-time edge deployment can be pursued through quantization, pruning, and NAS for execution on embedded platforms such as Raspberry Pi 4 and NVIDIA Jetson. The framework can be extended to a fleet-level GNN formulation for cross-turbine anomaly correlation and predictive maintenance optimization. Self-supervised pre-training on unlabeled SCADA data via masked autoencoders or contrastive learning can reduce labeled data requirements and improve generalization. The methodology can further be extended to RUL estimation for probabilistic failure time prediction. Near-term improvements include augmenting the feature vector with 6–9 pairwise cross-channel spectral features, cross-power spectral density, coherence functions, and time-lagged cross-correlations to capture inter-component fault propagation without architectural modification. Future work will also evaluate long-sequence Transformer variants at extended context windows and validate lead time estimates on larger multi-turbine failure datasets.

Funding Support

This research was funded by the **Institute for Advanced Research** Publication Grant of United International University, Ref. No.: IAR-2026-Pub-127.

Ethical Statement

This study does not contain any studies with human subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest related to this work.

Data Availability Statement

The NREL Gearbox Reliability Collaborative (GRC) dataset used in this study is publicly available at https://doi.org/10.2172/1048981. The CWRU Bearing Data Center dataset is accessible at https://engineering.case.edu/bearingdatacenter/download-data-file. The SCADA operational data used in this study are proprietary and cannot be shared publicly because of confidentiality agreements. The feature extraction code and model implementation, including the complete source code for the FFT pipeline, Transformer–LSTM model, and training scripts, are publicly available at https://github.com/uzzal2200/FFT-Based-Spectral-Temporal-Deep-Learning.git. To enable partial reproduction without proprietary SCADA data, results on the publicly available NREL GRC and CWRU datasets in isolation are reported in Table 20 (NREL GRC: 96.2%, CWRU: 98.1%). A dedicated training script (scripts/train_public_only.py) is provided in the GitHub repository to reproduce these per-source results using only public data.

Author Contribution Statement

Sajib Debnath: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Supervision. **Md. Uzzal Mia:** Software, Validation, Data curation, Formal analysis, Visualization, Writing – original draft. **Md Abubakkar:** Investigation, Resources, Writing – review & editing. **Arindam Kishor Biswas:** Validation, Writing – review & editing, Project administration. **Md. Sarwar Hosain:** Supervision, Conceptual guidance, Writing – review & editing. **Ohidujjaman:** Supervision, Research direction, Writing – review & editing.

Pre-Proof Version

References

[1] Veers, P., Bottasso, C. L., Manuel, L., Naughton, J., Pao, L., Paquette, J., . . . , & Rinker, J. (2023). Grand challenges in the design, manufacture, and operation of future wind turbine systems. *Wind Energy Science*, *8*(7), 1071–1131. https://doi.org/10.5194/wes-8-1071-2023

[2] Ren, Z., Verma, A. S., Li, Y., Teuwen, J. J. E., & Jiang, Z. (2021). Offshore wind turbine operations and maintenance: A state-of-the-art review. *Renewable and Sustainable Energy Reviews*, *144*, 110886. https://doi.org/10.1016/j.rser.2021.110886

[3] Liu, Z., & Zhang, L. (2020). A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings. *Measurement*, *149*, 107002. https://doi.org/10.1016/j.measurement.2019.107002

[4] Zhang, J., Dong, Y., Frangopol, D. M., Zhu, S., & Yang, H. (2026). Synergistic operation and maintenance enabling lifecycle-aware opportunistic management of offshore wind energy. *Applied Energy*, *408*, 127424. https://doi.org/10.1016/j.apenergy.2026.127424

[5] Zhu, Z., Lei, Y., Qi, G., Chai, Y., Mazur, N., An, Y., & Huang, X. (2023). A review of the application of deep learning in intelligent fault diagnosis of rotating machinery. *Measurement*, *206*, 112346. https://doi.org/10.1016/j.measurement.2022.112346

[6] Zhuang, J., Jia, M., & Zhao, X. (2022). An adversarial transfer network with supervised metric for remaining useful life prediction of rolling bearing under multiple working conditions. *Reliability Engineering & System Safety*, *225*, 108599. https://doi.org/10.1016/j.ress.2022.108599

[7] Chen, G., Yuan, J., Zhang, Y., Zhu, H., Huang, R., Wang, F., & Li, W. (2024). Enhancing reliability through interpretability: A comprehensive survey of interpretable intelligent fault diagnosis in rotating machinery. *IEEE Access*, *12*, 103348–103379. https://doi.org/10.1109/ACCESS.2024.3430010

[8] Peeters, C., Antoni, J., & Helsen, J. (2022). Blind filters based on envelope spectrum sparsity indicators for rotating machine fault diagnosis. *Mechanical Systems and Signal Processing*, *138*, 106556. https://doi.org/10.1016/j.ymssp.2022.109021

[9] Tang, S., Yuan, S., & Zhu, Y. (2020). Cyclostationary analysis towards fault diagnosis of rotating machinery. *Processes*, *8*(10), 1217. https://doi.org/10.3390/pr8101217

[10] Zhang, X., Zhao, B., & Lin, Y. (2021). Machine learning based bearing fault diagnosis using the Case Western Reserve University data: A review. *IEEE Access*, *9*, 155598–155608. https://doi.org/10.1109/ACCESS.2021.3128669

[11] Li, T., Zhao, Z., Sun, C., Cheng, L., Chen, X., Yan, R., & Gao, R. X. (2022). WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *52*(4), 2302–2312. https://doi.org/10.1109/TSMC.2020.3048950

[12] Clark, I., Chun, S., O'Sullivan, K., & Pierse, N. (2021). Energy poverty among tertiary students in Aotearoa New Zealand. *Energies*, *15*(1), 76. https://doi.org/10.3390/en15010076

[13] Helbing, G., & Ritter, M. (2018). Deep learning for fault detection in wind turbines. *Renewable and Sustainable Energy Reviews*, *98*, 189–198. https://doi.org/10.1016/j.rser.2018.09.012

[14] Liang, T., Lu, H., & Sun, H. (2021). Application of parameter optimized variational mode decomposition method in fault feature extraction of rolling bearing. *Entropy*, *23*(5), 520. https://doi.org/10.3390/e23050520

[15] Huang, B., Wang, Y., Yang, C., Li, Y., & Sun, Q. (2021). A neurodynamic-based distributed energy management approach for integrated local energy systems. *International Journal of Electrical Power & Energy Systems*, *128*, 106737. https://doi.org/10.1016/j.ijepes.2020.106737

[16] Huang, K., Yuen, K.-V., & Wang, L. (2022). Real-time simultaneous input-state-parameter estimation with modulated colored noise excitation. *Mechanical Systems and Signal Processing*, *165*, 108378. https://doi.org/10.1016/j.ymssp.2021.108378

[17] Zhang, Y., Xing, K., Bai, R., Sun, D., & Meng, Z. (2021). An enhanced convolutional neural network for bearing fault diagnosis based on time–frequency image. *Measurement*, *157*, 107667. https://doi.org/10.1016/j.measurement.2020.107667

[18] Wang, X., Si, S., & Li, Y. (2022). Variational embedding multiscale diversity entropy for fault diagnosis of large-scale machinery. *IEEE Transactions on Industrial Electronics*, *69*(3), 3109–3119. https://doi.org/10.1109/TIE.2021.3063979

[19] Ai, L., Cheng, H., Wang, X., Chen, C., Liu, D., Zheng, X., & Wang, Y. (2022). Approximate nearest neighbor search using enhanced accumulative quantization. *Electronics*, *11*(14), 2236. https://doi.org/10.3390/electronics11142236

[20] Chen, X., Zhang, B., & Gao, D. (2021). Bearing fault diagnosis based on multi-scale CNN and LSTM model. *Journal of Intelligent Manufacturing*, *32*(4), 971–987. https://doi.org/10.1007/s10845-020-01600-2

[21] Li, X., Zhang, W., & Ding, Q. (2019). Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. *Reliability Engineering & System Safety*, *182*, 208–218. https://doi.org/10.1016/j.ress.2018.11.011

[22] Li, H., Liu, T., Wu, X., & Chen, Q. (2021). A bearing fault diagnosis method based on enhanced singular value decomposition. *IEEE Transactions on Industrial Informatics*, *17*(5), 3220–3230. https://doi.org/10.1109/TII.2020.3001376

[23] Xu, Y., Li, Z., Wang, S., Li, W., Sarkodie-Gyan, T., & Feng, S. (2021). A hybrid deep-learning model for fault diagnosis of rolling bearings. *Measurement*, *169*, 108502. https://doi.org/10.1016/j.measurement.2020.108502

[24] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325

[25] Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023). Transformers in time series: A survey. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 6778–6786. https://doi.org/10.24963/ijcai.2023/759

[26] Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200209. https://doi.org/10.1098/rsta.2020.0209

[27] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. In *Proceedings of the International Conference on Learning Representations (ICLR 2023)*. https://doi.org/10.48550/arXiv.2211.14730

[28] Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2023). TimesNet: Temporal 2D-variation modeling for

PREPRINT VERSION

general time series analysis. In *Proceedings of the International Conference on Learning Representations (ICLR 2023)*. https://doi.org/10.48550/arXiv.2210.02186

[29] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted transformers are effective for time series forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR 2024)*. https://doi.org/10.48550/arXiv.2310.06625

[30] Zhang, M., Yang, G., Zhu, Z., Jin, G., Wang, M., & Bai, X. (2026). Diffusion-based Kriging model with graph-enhanced attention. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*, 499–510. https://doi.org/10.1145/3774904.3792104

[31] Jin, B., & Xu, X. (2024). Office real estate price index forecasts through Gaussian process regressions for ten major Chinese cities. *Advances in Computational Intelligence*, *4*(2), 8. https://doi.org/10.1007/s43674-024-00075-5

[32] Jin, B., & Xu, X. (2026). Inter-city housing price causality in Fujian: Contemporaneous analysis via vector error-correction and directed acyclic graph models. *Quality and Quantity*, *60*, 13637–13664. https://doi.org/10.1007/s11135-026-02740-y

[33] Xu, X. (2020). Corn cash price forecasting. *American Journal of Agricultural Economics*, *102*(4), 1297–1320. https://doi.org/10.1002/ajae.12041

[34] Yu, H., Liu, Y., Deng, J., Zhang, S., & Chen, W. (2022). A collaborative excitation method for piezoelectric stick-slip actuator to eliminate rollback and generate precise smooth motion. *Mechanical Systems and Signal Processing*, *170*, 108815. https://doi.org/10.1016/j.ymssp.2022.108815

[35] Chen, W., Li, J., Wang, Q., & Han, K. (2021). Fault feature extraction and diagnosis of rolling bearings based on wavelet thresholding denoising with CEEMDAN energy entropy and PSO-LSSVM. *Measurement*, *172*, 108901. https://doi.org/10.1016/j.measurement.2020.108901

[36] Chen, G., Liu, S., Wei, X., Fang, J., & Zhang, Q. (2022). Incipient fault detection for rolling element bearing based on wavelet packet transform and improved frequency band entropy criterion. *Sensors*, *22*(3), 1008. https://doi.org/10.3390/s22031008

[37] Zhang, M., Liang, K., Miao, Y., Lin, J., & Ding, C. (2022). Application of improved double-dictionary K-SVD for compound-fault diagnosis of rolling element bearings. *Measurement*, *187*, 110168. https://doi.org/10.1016/j.measurement.2021.110168

[38] Raj, K. K., Kumar, S., Kumar, R. R., & Andriollo, M. (2024). Enhanced fault detection in bearings using machine learning and raw accelerometer data: A case study using the Case Western Reserve University dataset. *Information*, *15*(5), 259. https://doi.org/10.3390/info15050259

[39] Kovanič, Ľ., Blistan, P., Štroner, M., Urban, R., & Blistanova, M. (2021). Suitability of aerial photogrammetry for dump documentation and volume determination in large areas. *Applied Sciences*, *11*(14), 6564. https://doi.org/10.3390/app11146564

[40] Yang, X., Bai, M., Liu, J., Liu, J., & Yu, D. (2021). Gas path fault diagnosis for gas turbine group based on deep transfer learning. *Measurement*, *181*, 109631. https://doi.org/10.1016/j.measurement.2021.109631

[41] Song, Q., Yang, J., Fu, Q., Zhang, S., Shi, K., & Zhang, H. (2025). A transformer-based multi-modal fault diagnosis model for power metering terminals. In *Proceedings of the 2025 5th International Conference on New Energy and Power Engineering*, 878–882. https://doi.org/10.1109/ICNEPE67923.2025.11384664

[42] Zhang, L., Liu, Y., Zhou, J., Luo, M., Pu, S., & Yang, X. (2022). An imbalanced fault diagnosis method based on TFFO and CNN for rotating machinery. *Sensors*, *22*(22), 8749. https://doi.org/10.3390/s22228749

How to Cite: Debnath, S., Mia, M. U., Abubakkar, M., Biswas, A. K., Hosain, M. S., & , O. (2026). FFT-Based Spectral–Temporal Deep Learning Framework for Wind Turbine Fault Diagnosis and Predictive Maintenance. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE620210516

Pre-proof Version