

Research paper

A hybrid CNN-Transformer model with optimized VMD for fault diagnosis of wave energy converter gearboxes

Zhimin Cheng^a, Jichuan Kang^a, Yu Sun^{a,**}, Dandan Peng^d, Mingxin Li^{b,c,*}

^a College of Shipbuilding Engineering, Harbin Engineering University, Harbin, Heilongjiang, China

^b Department of Civil Engineering, The University of Tokyo, Tokyo, Japan

^c Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom

^d School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, China

ARTICLE INFO

Keywords:

Wave energy converters
Gearboxes
Fault diagnosis
Deep learning
Convolutional neural network
Transformer

ABSTRACT

This paper presents a robust multi-stream hybrid convolutional neural network (CNN) and Transformer architecture for WEC gearbox fault diagnosis. To suppress severe background interference, an adaptive preprocessing method is developed by coupling Bayesian-optimized variational mode decomposition (VMD) with a windowed short-time Fourier transform to enable the reconstruction of task-relevant components and informative time-frequency representations. Within the diagnostic backbone, the CNN extracts localized morphological transients associated with gearbox defects, whereas a Transformer encoder captures global and long-range dependencies in the sequence representations. A channel-attention module is further employed to recalibrate discriminative feature maps, and a multi-stream fusion strategy integrates deep time-frequency features with statistical and envelope descriptors to enhance complementary fault evidence. A WEC gearbox testbed is built to acquire real-world vibration signals under multiple fault states. Experimental results demonstrate the superiority of the proposed framework, achieving a 99.14% diagnostic accuracy under nominal conditions. Comparative studies verify that the model maintains exceptional robustness, yielding an 89.66% accuracy even under severe interference, thereby providing a highly accurate and robust diagnostic solution for WECs deployed in complex marine environments.

1. Introduction

To mitigate the environmental crises induced by fossil fuel reliance (Li et al., 2023), the global energy landscape is rapidly shifting toward renewables. Within this evolving paradigm, wave energy has emerged as a highly promising renewable energy resource. It is characterized by high energy density and sustained operational readiness (Li et al., 2024a), superior forecasting predictability (Shadmani et al., 2023), and the capacity to generate power while mitigating coastal erosion (Berrio et al., 2023; Ozkan et al., 2020).

Wave energy converters (WECs) have been developed to harness this marine potential. WECs encompass various designs based on operating principles and deployment locations (Roach et al., 2025). Among these designs, point-absorber wave energy converters (PA-WECs) have emerged as a particularly promising WEC type. The compact dimensions

relative to dominant incident wavelengths enable near-omnidirectional wave energy absorption, which significantly simplifies structural design and reduces overall costs. Advanced rigid-flexible composite structures have further enhanced the structural performance and reliability of such offshore marine assets (Ye et al., 2025). As illustrated in Fig. 1, the core energy conversion mechanism of these assets relies on a vertically integrated Power Take-Off (PTO) system, wherein structural loads are transmitted directly from the hydrodynamic float through the ballscrew to critical rotating components, notably the gearbox. Furthermore, these characteristics make PA-WECs well suited to large-scale array deployments, facilitating more continuous power generation and improved conversion efficiency (He et al., 2023; Pastor and Liu, 2012). For clarity purposes, all subsequent references to WECs in this paper specifically refer to PA-WECs.

Despite the vast potential of wave energy, large-scale deployment of

This article is part of a special issue entitled: OE Maintenance published in Ocean Engineering.

* Corresponding author. Department of Civil Engineering, The University of Tokyo, Tokyo, Japan.

** Corresponding author.

E-mail addresses: sunyu9@hrbeu.edu.cn (Y. Sun), m.li@bridge.t.u-tokyo.ac.jp (M. Li).

<https://doi.org/10.1016/j.oceaneng.2026.125527>

Received 28 February 2026; Received in revised form 4 April 2026; Accepted 8 April 2026

Available online 11 April 2026

0029-8018/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

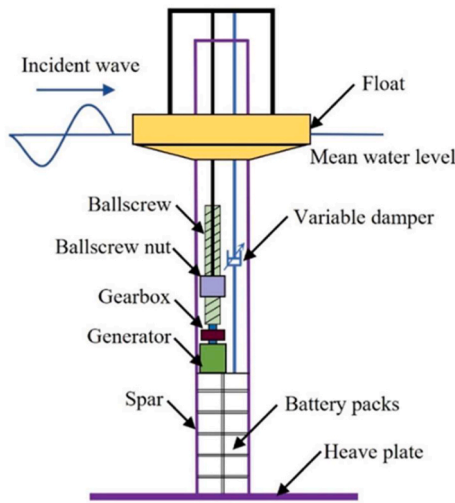


Fig. 1. Structural schematic of a PA-WEC, detailing the internal mechanical PTO transmission chain (Zhang et al., 2023).

WECs still lags behind solar and wind technologies, largely due to limited technological convergence, high leveled costs of energy (LCOE), and the huge challenges of the harsh marine environment (Hasan Maheen and Yang, 2023). Ensuring system survivability under harsh ocean conditions remains a primary barrier to commercialization, and, accordingly, design and operation are often constrained more by extreme loading and environmental degradation than by energy conversion efficiency (Guo et al., 2022; Guo and Ringwood, 2021; Said et al., 2025). This challenge is reflected in high operation and maintenance (O&M) burden, which may account for approximately 27% of overall LCOE for WECs (Tang et al., 2020). The O&M burden is further amplified by severe, location-dependent degradation mechanisms. Recent failure analysis of floating offshore renewable energy devices has revealed that degradation modes are strongly correlated with specific working conditions (Sun et al., 2023a). Specifically, shoreline devices commonly experience gear wear due to stochastic mechanical motions. Nearshore systems are prone to pitting faults induced by continuous wave stress and seawater immersion. Offshore converters endure extreme loads from typhoons and salt spray erosion, frequently resulting in catastrophic broken tooth faults (Li et al., 2021).

To address these localized vulnerabilities, the development of advanced and robust fault diagnosis models capable of identifying multiple failure modes is essential for reducing downtime and accelerating the commercialization of WEC technologies. Achieving system-level reliability for marine renewable energy systems must be built upon reliable monitoring and diagnosis of critical components, such as gearboxes. In this context, early studies on WEC reliability, which have mainly relied on qualitative or semi-quantitative assessments, need to evolve toward intelligent fault diagnosis and Remaining Useful Life (RUL) estimation, which are recognized as core components of Prognostics and Health Management (PHM) (Lai et al., 2024, 2025). Fault diagnosis and prognostics provide timely triggers for maintenance intervention, support forward-looking maintenance planning and structural health prediction, and enable the efficient use of maintenance-related information for optimized decision-making (Xu, 2025; Li et al., 2025). Diagnostic methods typically include model-based and data-driven methods (Li et al., 2026a,b). For example, as a model-based method, Kolios et al. (2018) integrated finite element analysis with response surface modeling to quantitatively assess the fatigue reliability of WEC floaters. Smoothed excitation force estimations for fault detection (González-Esculpi et al., 2022) and fault-tolerant control for sensors and actuators (Zhang et al., 2022) were developed to monitor specific PTO components. However, while model-based methods offer high physical interpretability, they struggle

to capture the highly non-linear and stochastic dynamics of critical components, e.g., gearbox degradation in harsh marine environments (Li et al., 2019b; Wang et al., 2025).

To overcome these modeling bottlenecks, data-driven approaches are increasingly dominant in fault diagnosis (Ahmad et al., 2024). Among them, deep learning has emerged as a major class because it can autonomously extract degradation-related signatures directly from operational data. Within deep learning, existing studies can be broadly categorized into two methodological trajectories according to how information is represented and learned. The first focused on temporal sequence modeling, utilizing recurrent neural networks (RNNs) and long short-term memory (LSTM) networks to capture time-series dependencies in signals (Cao et al., 2019; Peng et al., 2017). These methods have been widely validated in fault detection tasks for offshore energy systems, effectively modeling sequential degradation trends (Zhao et al., 2024; Zhu et al., 2025). The second trajectory focused on spatial feature extraction, employing convolutional neural networks (CNNs) to identify localized fault patterns within mapped time-frequency domains, such as the Wigner-Ville distribution (Li et al., 2019a). Complementary signal processing techniques, such as wavelet-based high-order spectrum analysis and encoder-based health indicators, have also been developed to extract subtle fault features for early gear damage detection (Verwimp et al., 2025; Zhu et al., 2026).

Despite their respective strengths, monolithic deep learning methods often fail to simultaneously capture both localized impact transients and long-range degradation trends. This limitation has motivated increasing interest in hybrid architectures that combine complementary feature-learning mechanisms. However, its application to WECs remains very scarce, with only a small number of studies reported to date. Despite this, the available results suggest substantial potential. For example, Kang et al. (2024) integrated a CNN-LSTM method for WEC fault classification and achieved diagnostic accuracies above 92% by jointly leveraging spatial and temporal feature extraction. Similar optimization-enhanced deep learning schemes have also been validated in related offshore tasks, demonstrating the effectiveness of integrating advanced optimization strategies to boost diagnostic performance under complex conditions (Li et al., 2024b).

Despite this progress, current diagnostic frameworks exhibit limitations when applied to WEC gearboxes operating under harsh marine conditions. Specifically, two critical methodological gaps remain unresolved. First, existing models fail to isolate weak, localized fault transients from severe, wave-induced background hydrodynamic interference. Standard methods process highly non-stationary signals inadequately, necessitating a deterministic frequency decoupling mechanism to penetrate the wave noise. Second, standard monolithic or simple hybrid deep learning models typically dilute high-frequency spatial transients during global sequence modeling. Parallel architectures or unconstrained attention mechanisms inadvertently smooth over these critical impact signatures, demanding a structurally isolated spatial-temporal learning framework capable of preserving transient fault features.

In response to these challenges, this study develops a hybrid CNN-Transformer model for WEC fault diagnosis, in which signal denoising, global dependency modeling, and local feature extraction are addressed. First, to mitigate the severe hydrodynamic noise and non-stationarity inherent to raw vibration measurements, a Bayesian-optimized Variational Mode Decomposition (VMD) strategy is introduced. Because standard VMD is highly sensitive to the empirical selection of its hyperparameters, Bayesian optimization is employed to deterministically search for the global optimum based on a composite energy-correlation index. This autonomously calibrated VMD, coupled with a windowed short-time Fourier transform (STFT), serves as a robust preprocessing stage to mathematically decouple localized fault transients from stochastic background interference. Second, to overcome the limitations of purely sequential or monolithic networks in capturing multiscale signal dynamics, a CNN branch is utilized to efficiently extract

localized morphological fault transients from time-frequency representations. Finally, to model global contextual relations and long-range sequential dependencies, a Transformer encoder is integrated, and the overall framework is further strengthened through multi-stream feature fusion and channel attention mechanisms.

In summary, the contributions of this study are summarized as follows.

- (1) A novel signal preprocessing framework is proposed by synergistically coupling adaptive VMD with windowed STFT. This method effectively isolates discriminative time-frequency features while mitigating random background interference, generating high-fidelity inputs for downstream analysis.
- (2) A hybrid CNN-Transformer diagnostic architecture is developed to address the complex dynamics of WECs. This model innovatively unites the localized spatial pattern extraction of CNNs with the global sequence modeling of Transformers to yield superior feature representation.
- (3) A robust multi-stream feature fusion and channel attention mechanism is integrated into the model. This structural design selectively isolates physically meaningful fault signatures to provide exceptional robustness against intense noise interference in low signal-to-noise ratio scenarios.
- (4) A physical WEC gearbox testbed is constructed to acquire real-world vibration signals under varying fault states, which is used for demonstrating the model's deployment potential in complex marine environments.

The remainder of this paper is organized as follows. In Section 2, the theoretical foundations of the proposed method are introduced. Section 3 presents the hybrid architecture, including a pipeline from Bayesian-optimized VMD to a multi-stream fusion strategy integrating STFT deep features with handcrafted statistics. In Section 4, a PTO testbed serves as a representative case study to highlight the model's superior performance and robustness across varying noise intensities. Moreover, the interpretability of the model is examined. Finally, Section 5 concludes the paper and discusses implications for improving WEC reliability.

2. Preliminaries

2.1. Variational mode decomposition

Proposed by Dragomiretskiy and Zosso (2014), VMD is a data-driven signal decomposition method that serves as an alternative to the well-established empirical mode decomposition (EMD). Unlike EMD's recursive sifting, VMD operates on a non-recursive, variational constraint framework to decompose a multi-component signal into a series of quasi-orthogonal, band-limited intrinsic mode functions (IMFs). The core of the algorithm is a constrained optimization problem designed to identify modes that are maximally compact around their respective central frequencies. This characteristic is particularly valuable in machinery fault diagnosis, where fault signatures often manifest at unique frequencies (Jiang et al., 2013), enabling VMD to effectively isolate these specific fault characteristic modes for subsequent analysis. The specific steps of the VMD algorithm to decompose the original signal into K modal components are as follows:

- (1) Select the number K of the decomposed modes and the balance parameter α , and initialize the modes $u_k(t)$ ($k = 1, \dots, K$).
- (2) Compute the analytic signal $H\{u_k(t)\}$ corresponding to $u_k(t)$ via the Hilbert transform as follows:

$$H\{u_k(t)\} = u_k(t) + \frac{j}{\pi} \int_{-\infty}^{+\infty} \frac{u_k(\tau)}{t - \tau} d\tau = \left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \quad (1)$$

where $j = \sqrt{-1}$.

- (3) Mix a complex exponential term with the analytic signal and shift the frequency spectrum of the mixed signal to the corresponding estimated central frequency ω_k . Then, a constrained variational problem, in other words, the squared norm of the gradient, is established as follows:

$$\min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] \exp(-j\omega_k t) \right\|_2^2 \right\} \quad (2)$$

subject to

$$\sum_k u_k(t) = f(t) \quad (3)$$

where $f(t)$ represents the raw data, δ is the Dirac distribution and denotes the convolution operator.

- (4) The quadratic penalty term and Lagrangian multiplier λ are introduced to render the reconstruction problem in Equation (2). Hence, the new constrained problem with the quadratic penalty term and Lagrangian multiplier is constructed as:

$$L(\{u_k\}, \{\omega_k\}, \lambda) = \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] \exp(-j\omega_k t) \right\|_2^2 + \left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \langle \lambda(t), f(t) - \sum_k u_k(t) \rangle \quad (4)$$

where $\lambda(t)$ is the coefficient of the Lagrangian multiplier and can be set as zero without the requirement of enforcing constraints strictly.

- (5) Solve Equation (4) by utilizing the alternate direction method of multipliers optimization algorithm introduced to obtain the decomposed modes u_k and their central frequency ω_k , which are expressed by

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{u}_i^n(\omega) + \frac{\hat{\lambda}^n(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k^n)^2} \quad (5)$$

and

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{u}_k^{n+1}(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k^{n+1}(\omega)|^2 d\omega} \quad (6)$$

respectively.

2.2. CNN and Transformer

CNNs are capable of capturing local spatial features and effectively handling time-invariant patterns, which makes them particularly suitable for analyzing complex sensory data. By automatically learning hierarchical and discriminative feature representations, they greatly reduce the dependence on manual feature engineering. CNNs have been widely applied in various fault diagnosis scenarios, such as fault detection in photovoltaic systems (Ali and Ding, 2025), fault diagnosis in rotating machinery (Janssens et al., 2016) and motor fault detection (Sun et al., 2023b). In this paper, a one-dimensional CNN (1D-CNN) is employed to extract localized temporal features from pre-processed vibration signals before subsequent feature fusion and classification.

The Transformer model, introduced by Vaswani et al. (2017), is a sequential modeling architecture founded on attention mechanisms. Its central principle is the use of self-attention to capture dependencies between all positions within an input sequence. In this paper, we focus

solely on the Transformer encoder, the component responsible for learning a rich representation from the input sequence. Fig. 2 depicts its architecture, and the major components are detailed below.

(1) Self-attention mechanism

The self-attention mechanism weighs the representation of each position by learning the correlation between the positions in the sequence. Given input $\mathbf{X}$, that is the length is n and the feature dimension is d . The information of the input elements is represented by three vectors Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$), respectively. The formulas are as follows

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \mathbf{K} = \mathbf{X}\mathbf{W}_K, \mathbf{V} = \mathbf{X}\mathbf{W}_V \quad (7)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ represents the learnable parameter matrices. The self-attention formula is as follows

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (8)$$

where d_k is the dimension of the key and $\sqrt{d_k}$ is the scaling factor, $\text{softmax}(\cdot)$ denotes the normalization of the result to obtain the attention weights, and $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ is the output of the self-attention layer.

(2) Multi-head attention

In order to allow the model to attend to distinct positional relationships in the input sequence, Transformer uses a multi-head self-attentive regime to map Query, Key, and Value to multiple sub-spaces, respectively, as follows

$$\text{head}_i = \text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) \quad (9)$$

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)\mathbf{W}^O \quad (10)$$

where h denotes the number of attention heads, $\mathbf{W}^O$ is the output linear transformation matrix.

(3) Transformer encoder

The Transformer encoder is constructed from a stack of N identical layers. Each layer is composed of two primary sub-layers, i.e., a multi-

head self-attention mechanism and a position-wise, fully connected feed-forward network. Furthermore, a residual connection followed by layer normalization is applied to the output of each sub-layer. The computational procedure of the encoder layer can be described as

$$\mathbf{Z} = \text{LN}(\mathbf{X} + \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V})) \quad (11)$$

$$\mathbf{Y} = \text{LN}(\mathbf{Z} + \text{FFN}(\mathbf{Z})) \quad (12)$$

where $\mathbf{Z}$ is the intermediate output matrix of the self-attention sub-layer, $\text{LayerNorm}(\text{LN}(\cdot))$ denotes the layer normalization that helps to keep the gradient stable during training and accelerates convergence. Position-wise Feed-Forward Networks ($\text{FFN}(\cdot)$) denotes the feed forward network used for nonlinear transformation. $\mathbf{Y}$ denotes the output matrix of the coding layer.

3. Methodology

In this paper, a comprehensive and robust framework for gearbox fault diagnosis is established, as visually summarized in Fig. 3. The framework is structured into two stages: (1) signal processing and feature extraction, and (2) a hybrid architecture for classification. The first stage initiates with an optimized VMD for adaptive signal denoising, where key parameters are determined via Bayesian optimization to minimize residual energy. The signal is then reconstructed using only the top-ranked modes, selected based on a hybrid score of correlation and energy. Subsequently, this pipeline generates a multi-modal feature set: (1) STFT spectrograms that illustrate the evolution of frequency content; (2) time and frequency-domain statistics that summarize the signal characteristics; (3) an envelope spectrum that highlights periodic fault impacts. The second stage introduces a multi-branch hybrid deep learning model. The main branch, designed for STFT data, serially integrates a CNN with a Transformer encoder. The CNN component is designed to extract characteristic local patterns from the embedded STFT spectrograms. This output is then fed into the Transformer branch, which processes the feature sequences to capture long-term temporal dependencies and global structural patterns. In parallel, auxiliary branches process the time-, frequency-, and envelope features using smaller Multi-Layer Perceptron (MLP). The fusion of these complementary feature sets, i.e., the deep features from the STFT branch and the engineered features from the auxiliary branches, enables the model to achieve accurate and robust fault diagnosis.

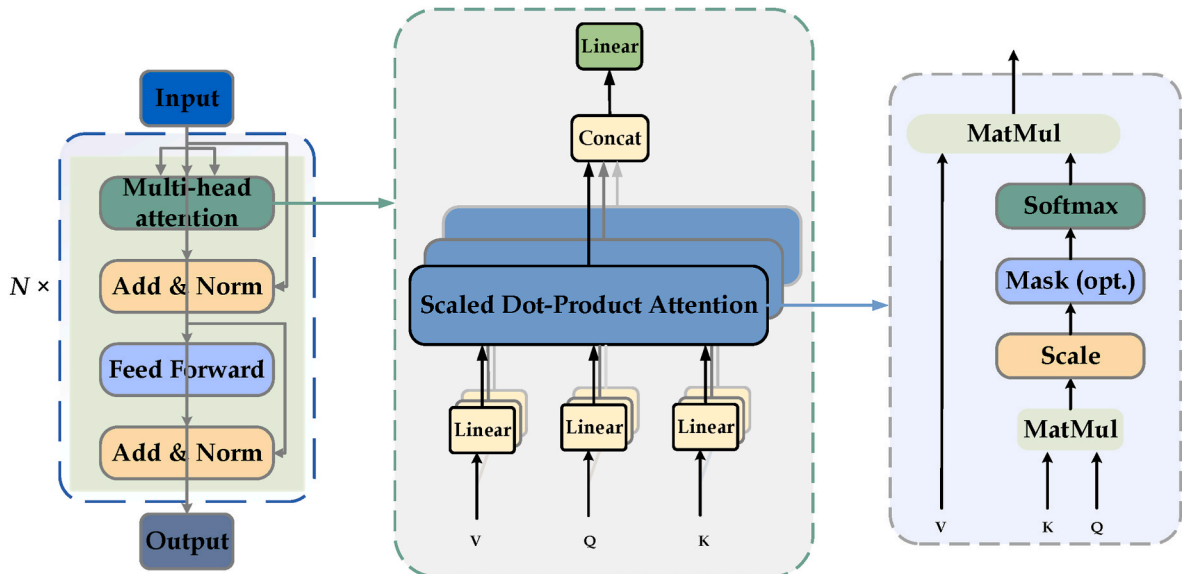


Fig. 2. Representative structure of Transformer encoder.

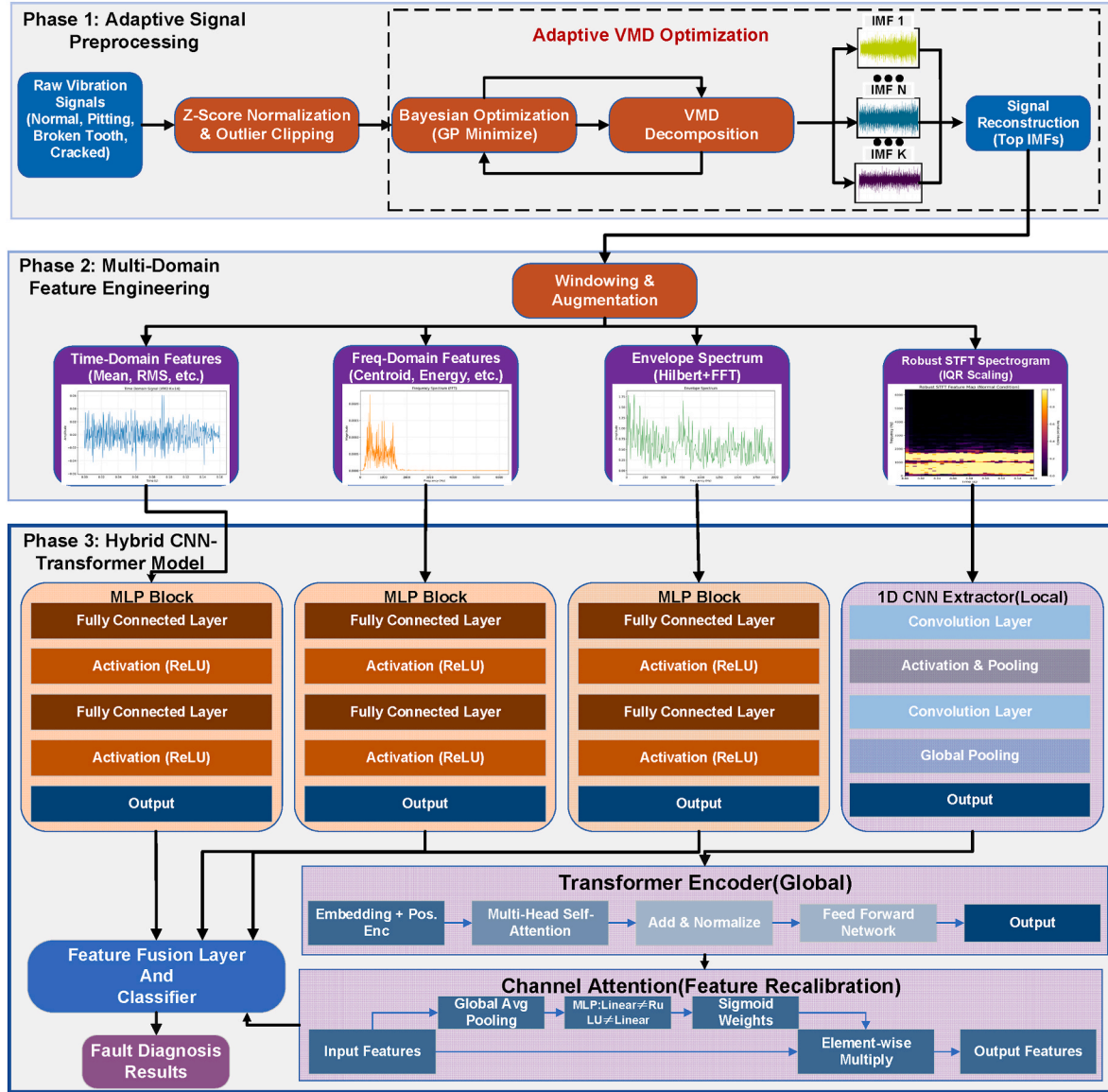


Fig. 3. Overall Structure of the proposed hybrid model.

3.1. Adaptive signal denoising with optimized VMD

To enhance the signal-to-noise ratio and extract salient fault features, this study implements an adaptive VMD method for signal denoising. Compared with other typical methods, e.g., EMD and ensemble empirical mode decomposition (EEMD), VMD has several key advantages for fault diagnosis. It has a solid theoretical foundation, is more robust to sampling and noise, and can effectively avoid mode mixing (Gu et al., 2022; Lian et al., 2018). Despite these advantages, the efficacy of VMD is highly dependent on two key hyperparameters: the number of modes K and the penalty factor α . To reduce the subjectivity and time burden associated with manual parameter tuning, various data-driven optimization strategies have been employed, such as using a correlation coefficient criterion to determine the number of modes (Liu et al., 2016). Building on this trend, a class-specific Bayesian optimization strategy was employed in this work to determine both optimal parameters simultaneously.

Specifically, we utilize Gaussian Process minimization to search the hyperparameter space defined by $K \in [5, 15]$ and $\alpha \in [500, 3000]$. Unlike methods that maximize correlation, our objective function is designed to minimize the residual energy between the raw signal $f(t)$ and the sum of

the extracted modes. This ensures that the decomposition preserves the maximum information content of the original vibration signature while avoiding over-decomposition. The optimization seeks to solve:

$$\{K_{\text{opt}}, \alpha_{\text{opt}}\} = \underset{K, \alpha}{\operatorname{argmin}} \left\| f(t) - \sum_{k=1}^K u_k(t) \right\|_2^2 \quad (13)$$

where $u_k(t)$ represents the k -th IMF obtained under the current parameter set.

Once the optimal K and α are determined, the entire signal is decomposed using these parameters. A subsequent intelligent filtering process is applied to discern substantive modes from noise and artifacts. To quantify the relationship between the decomposed modes and the original signal, the Pearson Correlation Coefficient (PCC) is initially calculated. For the k -th IMF, the PCC (ρ_k) is defined as:

$$\rho_k = \frac{\sum_{t=1}^{N_s} (u_k(t) - \bar{u}_k)(f(t) - \bar{f})}{\sqrt{\sum_{t=1}^{N_s} (u_k(t) - \bar{u}_k)^2} \sqrt{\sum_{t=1}^{N_s} (f(t) - \bar{f})^2}} \quad (14)$$

where $\bar{u}_k$ and $\bar{f}$ denote the mean values of the IMF and the raw signal, respectively, and N_s represents the total number of sample points.

However, using PCC as the sole selection criterion can be prone to errors, as wide-band noise may occasionally exhibit high correlation with the raw signal. To enhance robustness, this study employs a weighted composite index that incorporates both the correlation and the energy contribution of each mode. This ensures that the selected IMFs are both structurally relevant and energetically significant. The composite score S_k is formulated as:

$$S_k = \gamma_1 \cdot \left(\frac{\|u_k\|_2^2}{\sum_{i=1}^K \|u_i\|_2^2 + \epsilon} \right) + \gamma_2 \cdot |\rho_k| \quad (15)$$

In this work, the weighting factors are set to $\gamma_1 = 0.6$ and $\gamma_2 = 0.4$ to prioritize high-energy fault transients. The term ϵ (set to 10^{-10}) is a small regularization constant introduced to prevent division by zero in the event of vanishing signal energy, thereby ensuring numerical stability during computation. The signal is subsequently reconstructed by aggregating the top-ranked IMFs (specifically, the top 4 modes based on S_k to yield the denoised signal $f_{\text{rec}}(t)$.

To validate this parameter configuration, a sensitivity analysis was conducted on the composite score function (Fig. 4). The grid search results confirm that the aforementioned weighting coefficients yield the optimal diagnostic accuracy of 89.66%, with performance degrading by 18.97% and 8.63% when γ_1 deviates by ± 0.2 . Furthermore, the regularization constant ϵ was systematically evaluated at adjacent orders of magnitude (10^{-9} and 10^{-11}), yielding strictly identical outcomes. This insensitivity empirically confirms that ϵ serves exclusively as a numerical safeguard rather than a tunable hyperparameter affecting model behavior.

3.2. Multi-domain feature construction

To facilitate a comprehensive fault diagnosis under varying operating conditions, this study constructs a hybrid feature space that integrates high-dimensional time-frequency representations with expert-defined statistical descriptors. This multi-view approach ensures that the model captures both complex non-stationary patterns and specific statistical properties indicative of early-stage failure. While deep neural networks, such as the proposed CNN-Transformer architecture, exhibit strong spatial and temporal modeling capacities, relying exclusively on STFT-derived latent representations presents fundamental algorithmic

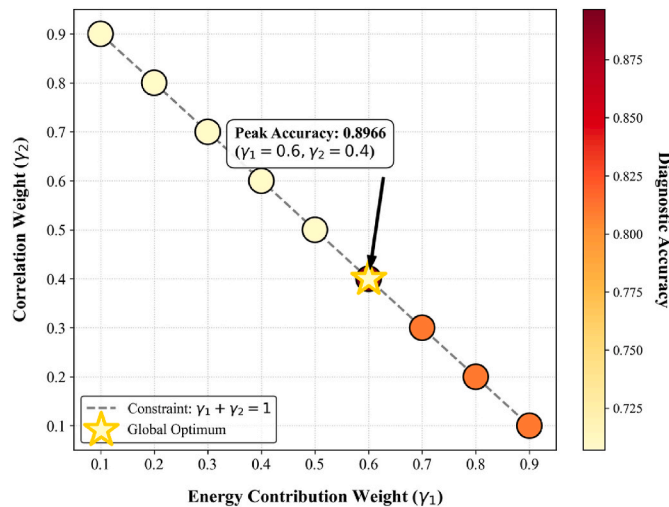


Fig. 4. Sensitivity analysis of the weighting coefficients γ_1 and γ_2

limitations. Specifically, the STFT is mathematically constrained by the Heisenberg-Gabor uncertainty principle, which dictates an unavoidable trade-off between time and frequency resolution (Feng et al., 2013). This strict limitation inherently blurs low-energy transient impacts and high-frequency modulation sidebands within the 2D spectrogram. Consequently, forcing a deep learning backbone to autonomously map precise statistical distributions and signal demodulation kinematics solely from raw spectrograms is highly inefficient and prone to overfitting under severe background noise. The feature extraction process is therefore executed in three distinct phases: first, a robust STFT generates 2D spectrograms to resolve non-stationary signal characteristics; second, a 19-dimensional vector of explicit statistical metrics quantifies impulsive faults and spectral shifts; and third, an enhanced envelope spectrum analysis isolates the periodic impulses characteristic of localized component defects.

To resolve the non-stationary characteristics of the vibration signals, we generate 2D spectrograms using a Robust STFT. A Blackman window ($N = 256$, overlap = 75%) is employed to minimize spectral leakage. To address the impulsive nature of industrial noise, the magnitude response is logarithmically transformed and subjected to an adaptive outlier suppression mechanism. This involves clipping pixel intensities based on the interquartile range (IQR) to a standardized scale of $[0, 3]$, so that subsequent convolutional layers emphasize coherent frequency-evolution patterns rather than isolated noise spikes.

While deep learning is effective at learning implicit representations, explicit statistical metrics remain highly effective for quantifying impulsive faults and spectral shifts. Standard convolutional pooling and normalization operations inherently dilute discrete, non-Gaussian impulses, creating a representation blind spot for localized mechanical shocks. Explicit statistical priors bypass this bottleneck by providing deterministic scalar metrics that strictly quantify probability distribution shifts independent of spatial textures. We extract a 19-dimensional vector of scalar features from each signal segment to capture distinct physical degradation phenomena in the gearbox. First, time-domain statistics quantify the physical impacts of gear tooth deterioration via 11 metrics. Metrics such as Kurtosis and Skewness are highly sensitive to the non-Gaussian shock characteristics generated by early-stage pitting and spalling, while the root mean square reflects the overall vibration energy increase associated with distributed wear. Furthermore, the Crest Factor and Zero-Crossing Rate are computed to quantify signal impulsiveness and oscillation frequency. Second, frequency-domain statistics characterize the fault-induced energy redistribution across the spectrum; we extract 8 metrics including the Spectral Centroid, Spectral Spread, Spectral Rolloff (85% energy threshold), and Spectral Flux. Additionally, the spectrum is divided into five specific sub-bands (0–100 Hz, 100–200 Hz, 200–300 Hz, 300–400 Hz, and 400–500 Hz), and the energy ratio of each band is calculated to detect structural resonance shifts triggered by stiffness degradation in the faulty gear teeth.

To isolate the periodic impulses characteristic of rolling element bearing and gear defects, we employ envelope spectrum analysis. Localized defects typically generate repetitive transient impulses with pronounced non-stationary and cyclostationary characteristics (Chen et al., 2023). Such responses are commonly governed by amplitude- and frequency-modulation effects, where high-frequency structural resonances are modulated by low-frequency fault characteristic components (Cheng et al., 2010). CNN-Transformer models operating solely on STFT magnitudes cannot inherently execute the non-linear Hilbert demodulation required to physically decouple these carrier and modulating frequencies. Consequently, applying a Fourier transform directly to the raw vibration signal often obscures defect-related characteristic frequencies; demodulation is therefore required prior to spectral analysis (Chen et al., 2023).

In this study, we utilize the Hilbert transform to construct the analytic signal $z(t)$ and extract its instantaneous amplitude envelope. This process is defined as:

$$\mathbf{z}(t) = \mathbf{f}_{\text{rec}}(t) + j\mathcal{H}[\mathbf{f}_{\text{rec}}(t)] \quad (16)$$

where $\mathcal{H}[\cdot]$ denotes the Hilbert transform operator, and $j = \sqrt{-1}$ is the imaginary unit. To enhance the feature extraction, the raw envelope is first smoothed using a moving average filter to suppress high-frequency noise artifacts. Subsequently, the smoothed envelope is transformed into the frequency domain via Fast Fourier Transform (FFT). Finally, a logarithmic transformation ($\log(1+x)$) is applied to the magnitude spectrum, followed by normalization. This logarithmic scaling is critical for managing the wide dynamic range of spectral amplitudes and suppressing random impulse noise to enhance the visibility of the modulation sidebands relevant to fault diagnosis (Chen et al., 2023).

3.3. Hybrid CNN-Transformer architecture with multi-stream fusion

Vibration signals exhibit complex characteristics, containing both high-frequency local transients, indicative of incipient faults, and long-range temporal dependencies, indicative of evolving degradation. To simultaneously capture these disparate features, this study proposes a serial CNN-Transformer hybrid network augmented with channel attention and a multi-stream feature fusion strategy. The adoption of a strict serial CNN-Transformer hierarchy is theoretically motivated by the geometric and temporal properties of STFT spectrograms derived from mechanical vibrations. Gearbox fault signatures, such as pitting or broken tooth, manifest in the STFT domain as highly localized, high-frequency transient impacts. Transformers, which rely on global self-attention mechanisms, lack the inductive bias required to process localized grid topologies efficiently. Feeding raw or linearly projected STFT matrices directly into a Transformer forces the attention matrix to compute unconstrained global interactions across fine-grained local noise, inevitably leading to feature dilution and high computational overhead. Conversely, convolutional neural networks possess stringent inductive biases—specifically, translational invariance and localized receptive fields. By positioning the 1D-CNN frontend prior to the global attention mechanism, the architecture establishes a requisite inductive bottleneck. The CNN operates as a localized feature aggregator, compressing sparse, high-dimensional spectral transients into a sequence of low-dimensional latent feature vectors. This architectural precedence is therefore mandatory: the CNN must first extract and condense the localized morphological features of transient impacts, rigorously enabling the subsequent Transformer encoder to map the global, long-range cyclostationary dependencies of the degradation process without interference from localized spectral noise. Coupling these structures in series allows the network to leverage the CNN for localized feature extraction while utilizing the Transformer to capture global semantic dependencies across the sequence (Chen et al., 2024). Unlike conventional parallel architectures, this model processes the time-frequency representation sequentially to learn hierarchical abstractions, while concurrently integrating domain-specific engineered features for enhanced diagnostic robustness.

The robust STFT spectrogram is used as the primary input to the deep learning backbone. The frequency dimension is first projected into a high-dimensional latent space ($d_{\text{model}} = 192$) via a linear embedding layer. To capture local spectral correlations and transient impulse patterns, the embedded sequence is processed by a 1D-CNN block. This block comprises two stacked 1D convolutional layers with a kernel size of 3 and padding of 1, preserving the temporal resolution. Each convolution is followed by Batch Normalization and a GELU activation function. This design allows the model to extract local morphological features from the spectrogram before global context modeling. To capture the long-range temporal dependencies and global structural evolution of the signal, the output of the CNN is fed into a Transformer Encoder.

Positional encodings are added to the feature sequence to retain temporal order information. The encoder consists of $N = 4$ stacked layers, each employing a multi-head self-attention mechanism with 8

attention heads and a feed-forward network with a hidden dimension of 576 ($3 \times d_{\text{model}}$). A dropout rate of 0.4 is applied to mitigate overfitting. By allowing the model to assign attention weights adaptively across time steps, the self-attention mechanism facilitates the extraction of global cyclostationary fault-related patterns.

Following the Transformer encoder, a channel attention model is introduced to adaptively recalibrate the feature maps. This model functions as a squeeze-and-excitation block, designed to model the inter-channel dependencies explicitly. Let $\mathbf{H} \in \mathbb{R}^{n \times B \times C}$ denote the output feature map from the Transformer, where n is the sequence length and C is the channel dimension.

First, the model aggregates global spatial information via global average pooling known as the “squeeze” operation, to generate a channel descriptor $\mathbf{v} \in \mathbb{R}^C$. The c -th element of $\mathbf{v}$ is calculated as:

$$v_c = F_{\text{sq}}(\mathbf{H}_c) = \frac{1}{n} \sum_{i=1}^n H_c(i) \quad (17)$$

Next, the model captures channel-wise dependencies through a two-layer MLP, referred to as the “excitation” operation, with a reduction ratio of $r = 16$. This structure employs a gating mechanism with a Sigmoid activation to generate the attention weights $\mathbf{s} \in \mathbb{R}^C$:

$$\mathbf{s} = F_{\text{ex}}(\mathbf{v}, \mathbf{W}) = \sigma(\mathbf{W}_2 \text{ReLU}(\mathbf{W}_1 \mathbf{v})) \quad (18)$$

where $\text{ReLU}(\cdot)$ refers to the ReLU activation function, σ refers to the Sigmoid function, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ is the dimensionality-reduction layer, and $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ is the dimensionality-increasing layer.

Finally, the generated attention weights are applied to the input features via element-wise multiplication, referred to as the ‘scale’ operation:

$$\tilde{\mathbf{H}}_c = s_c \cdot \mathbf{H}_c \quad (19)$$

This process results in the recalibrated feature map $\tilde{\mathbf{H}}$, selectively emphasizing informative feature channels that correlate strongly with fault modes while suppressing irrelevant noise channels.

To leverage the complementarity between learned representations and domain-informed descriptors, the network employs a multi-stream fusion strategy. While the main branch extracts deep features from the STFT, three auxiliary MLP branches independently process the envelope spectrum, time-domain statistics, and frequency-domain statistics, mapping them to dimensions of 64, 32, and 32, respectively. These auxiliary vectors are concatenated with the global average-pooled output of the main CNN-Transformer branch, forming a comprehensive hybrid feature vector. This fused representation is finally passed through a classification head consisting of batch normalization, dropout, and fully connected layers to predict the probability distribution over the four states: normal, pitting, broken tooth and cracked.

4. Experimental validation and result analysis

4.1. Experiment and data acquisition

To empirically validate the proposed diagnostic framework, this study utilizes vibration datasets acquired from a custom-engineered experimental testbed. Rather than employing the steady-state rotation typical of standard machinery testbeds, this apparatus replicates the transient kinematics of a point-absorber wave energy converter gearbox. The kinematic drive system couples a variable-speed motor with a crank-linkage mechanism, converting continuous rotary input into the reciprocal vertical heave motion characteristic of a wave buoy. A programmable control circuit modulates this rack motion within a 5 to 8 s cycle, strictly mapping to the fundamental excitation periods of ocean waves. Torque is subsequently transferred via a ratchet overrunning clutch to the transmission chain. To emulate the damping torque exerted

by a generator, a fixed load is applied to the output shaft, ensuring the gearbox is subjected to dynamic mechanical stresses equivalent to real-world hydrodynamic loading.

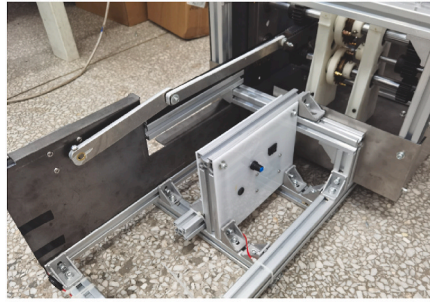
The experimental protocol evaluated four distinct gear health states: a healthy baseline and three critical failure modes (broken tooth, pitting, and cracked) prevalent in harsh offshore environments. These fault states were physically injected into the transmission chain using pre-manufactured defective gear kits. To ensure dynamic stability and eliminate transient start-up artifacts, continuous data acquisition commenced only after the rotor achieved steady-state oscillation under each predefined speed and torque condition. The vibration signatures were measured utilizing two tri-axial ICP/IEPE piezoelectric accelerometers (Model CT1010SLFP) rigidly mounted to the gearbox housing, capturing six independent acceleration channels. Simultaneously, the instantaneous rotational speed was tracked via a CZ400 Hall-effect tachometer. The analog signal chain was conditioned by a CT5210 ten-channel constant current adapter and digitized through an MCC USB-1608G acquisition card. Data was sampled at 1 kHz, with each operational state recorded continuously for 20 min. This extended duration guarantees the sufficient capture of low-frequency amplitude-modulation effects and cyclostationary fault features essential for robust deep learning model training. The specific experimental configuration is illustrated in Fig. 5.

4.2. Implementation description

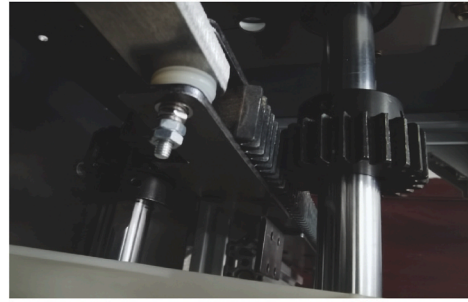
4.2.1. Data configuration and partitioning

In this study, vibration datasets collected at a wave period of $T = 6.7s$ were selected for analysis. For each of the four states, i.e., normal, pitting, broken tooth, and cracked, a continuous vibration sequence consisting of 100,500 data points was retained.

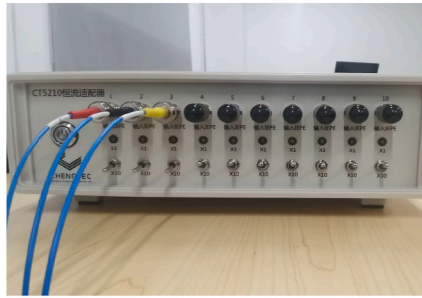
To generate sufficient samples for model training, these continuous time-series signals were segmented using a sliding window technique. A window length of 2048 points was utilized to capture adequate spectral resolution, with an overlap ratio of 75% applied to maximize feature extraction and data volume. Strict data partitioning was enforced to prevent data leakage. We employed stratified shuffle splitting to randomly allocate 15% of the samples to an independent test set, ensuring the class distribution remained balanced. The remaining data were subdivided into a training set and a validation set, comprising 18% of the training partition for hyperparameter tuning and model checkpointing. Furthermore, to enhance the model robustness, an online data augmentation strategy was integrated into the training pipeline. With a probability of 0.7, input samples underwent random perturbations, including Gaussian noise injection, i.e., signal-to-noise ratio (SNR) ranging from -4 to 20 dB, amplitude scaling ($0.9 \times$ to $1.1 \times$), and circular time-shifting.



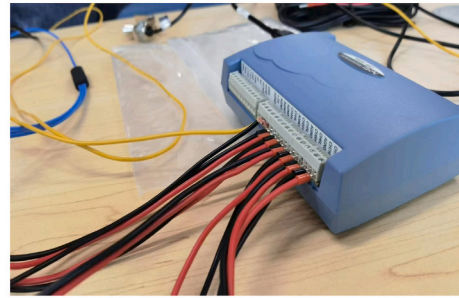
(a) Experimental platform



(b) Internal structure



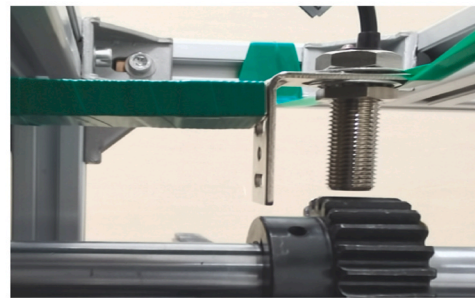
(c) CT 5200 series constant current adapter



(d) USB-1608G data acquisition card



(e) ICP/IEPE piezoelectric accelerometer



(f) CZ400 Hall speed sensor

Fig. 5. Experimental setup and sensor locations(Kang et al., 2024).

4.2.2. Network architecture and training protocol

The proposed method was implemented using PyTorch. The hybrid architecture was configured with a latent feature dimension (d_{model}) of 192. The Transformer encoder consisted of 4 stacked layers, employing 8 parallel attention heads to capture global temporal dependencies, with a feed-forward dimension of 576. To mitigate overfitting during the learning of complex spectral patterns, a dropout rate of 0.4 was applied throughout the network.

The network parameters were optimized using the AdamW algorithm with a decoupled weight decay of 0.01 to improve generalization. The training process was governed by a cosine annealing with warm restarts learning rate scheduler ($T_0 = 20$ epochs, $T_{\text{mult}} = 2$, $\eta_{\text{min}} = 10^{-6}$), initiated at a learning rate of 5×10^{-4} . This cyclic schedule allows the model to periodically escape local minima in the loss landscape.

To address potential class imbalances inherent in fault diagnosis tasks, the focal loss function ($\gamma = 2.0$) was utilized as the objective function, assigning higher weights to hard-to-classify samples compared to standard cross-entropy. The model was trained for 100 epochs with a batch size of 32 on a GPU-accelerated workstation. The optimal model

parameters were determined based on the highest weighted F1-score achieved on the validation set.

4.3. Experimental results analysis and discussion

4.3.1. Evaluation of signal decomposition and time-frequency representation

To assess the effectiveness of the proposed preprocessing strategy, we conducted a comparative analysis between a standard VMD setup using generic parameters $K = 10$, $\alpha = 1000$ and the proposed Bayesian optimized VMD. The analysis focuses on hyperparameter convergence, spectral orthogonality, and signal-noise separation performance. The optimization results for four different health states are presented in Fig. 6. Each subfigure illustrates the time-domain signal reconstruction (top), IMF spectral separation (middle), and residual error analysis (bottom). Based on these results, the convergence of hyperparameters and their physical interpretation are analyzed as follows.

(1) Hyperparameter convergence and physical interpretation

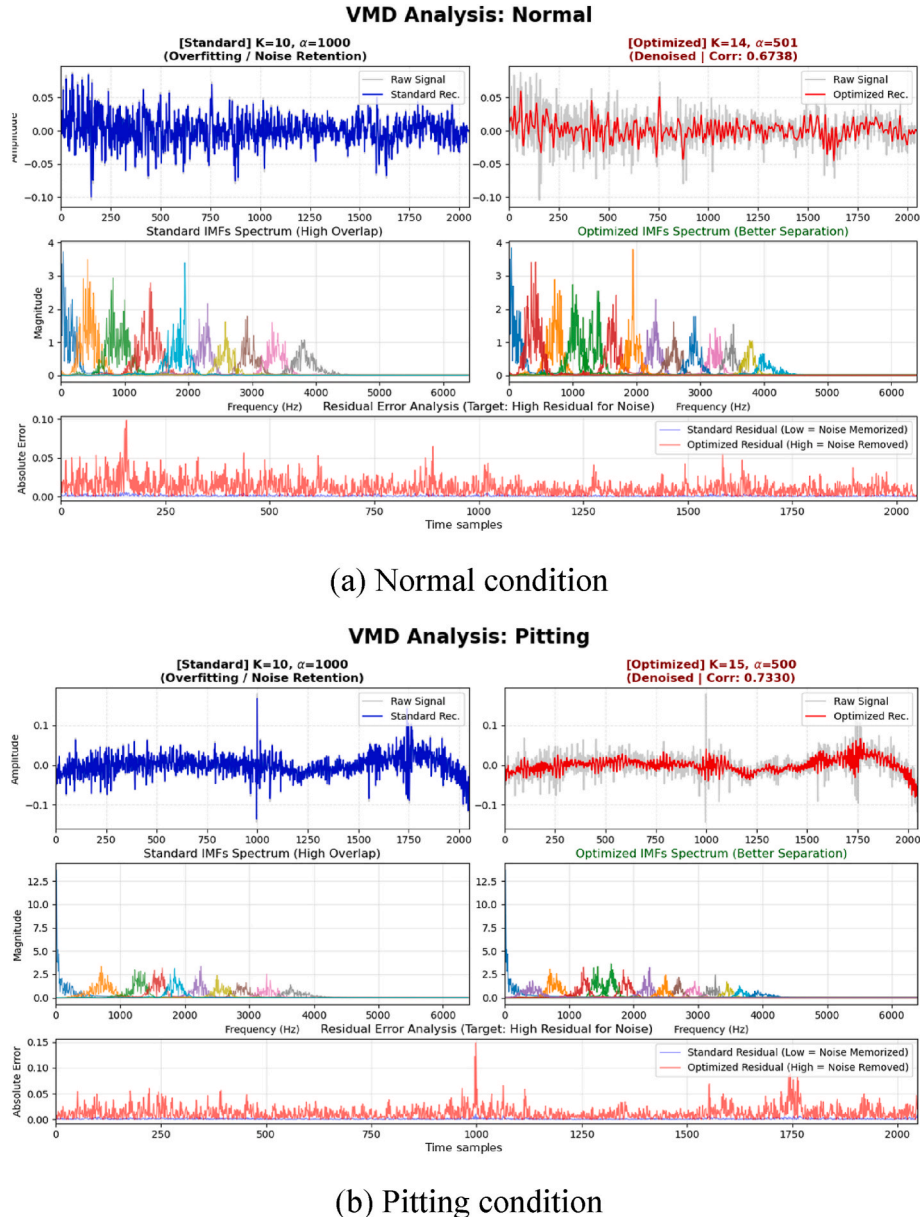


Fig. 6. Optimization results for four different health states

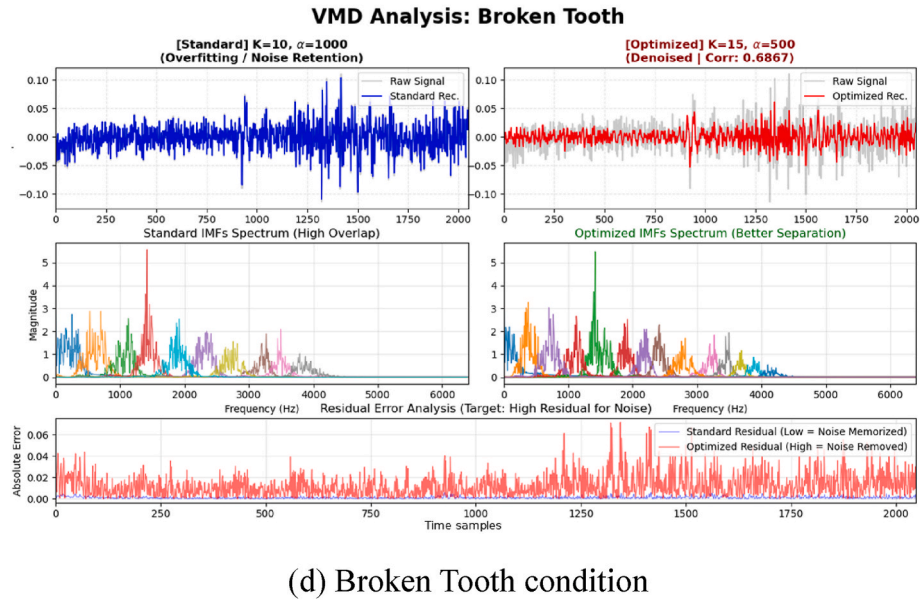
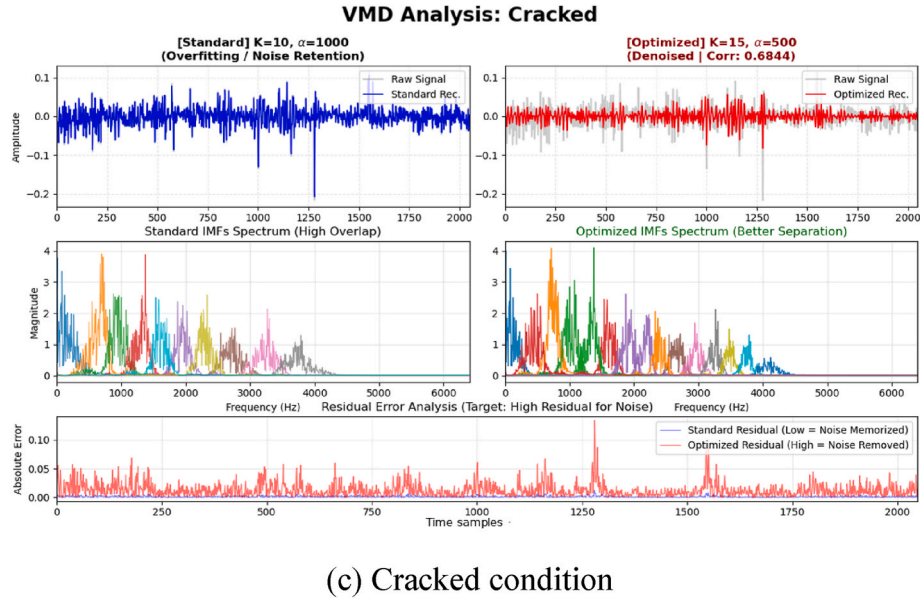


Fig. 6. (continued).

The Bayesian optimization algorithm, employing Gaussian Process priors to model the objective function, successfully converged to a global optimum across all health conditions. The optimized parameter sets obtained are: normal ($K = 14, \alpha = 501$), broken tooth ($K = 15, \alpha = 500$), cracked ($K = 15, \alpha = 500$), and pitting ($K = 15, \alpha = 500$). Analyzing this convergence behavior yields critical physical insights into the dynamic response of the machinery, specifically regarding spectral complexity and transient recovery.

A distinct deviation is observed in the optimization of the mode count, K . While standard heuristics typically suggest that a low number of modes (e.g., $K \in [3, 5]$) suffices to capture fundamental harmonics, the optimizer consistently identified a higher resolution ($K \approx 15$) as requisite. This convergence indicates a dense spectral architecture within the gearbox. Beyond fundamental meshing frequencies and shaft orders, the spectrum exhibits abundant modulation sidebands and structural resonances. Physically, the selection of a high K for the normal condition challenges the assumption that healthy signals represent low-complexity, stationary waveforms. Instead, the healthy state possesses a distinctive harmonic signature. Preserving this spectral richness pro-

vides a higher-fidelity baseline for subsequent diagnosis and reduces the likelihood of false positives, where benign but unmodeled harmonics could otherwise be mistaken for fault-related anomalies.

Furthermore, the convergence of the penalty factor α , which governs the bandwidth of the extracted modes, reveals significant insights into the physics of fault impulses. The optimization favored a moderate value ($\alpha \approx 500$) over the strictly narrowband constraints ($\alpha = 1000$) often applied in standard VMD. This relaxation is physically justified by the nature of localized faults, such as broken tooth or pitting, which manifest as periodic, short-duration impulses. In the frequency domain, these sharp time-domain transients translate into broadband spectral features. A high α value would impose excessive smoothing, effectively suppressing this diagnostic high-frequency energy. Consequently, the optimized moderate α strikes a critical balance. It retains sufficient bandwidth to capture the sharp rise-times of fault-induced transients while maintaining adequate constraints to decouple these signatures from the carrier signal.

(2) Spectral separation and mitigation of mode mixing

The clearest benefit of the proposed optimization is observed in the frequency domain, specifically regarding the mitigation of mode mixing i.e., the leakage of physical components across multiple modes or the collapse of multiple components into a single mode. In the unoptimized state, the decomposition suffers from severe spectral overlaps, as evidenced by the standard IMF spectra (middle-left panels in Fig. 6). This visual aliasing implies that the center frequencies (ω_k) failed to align with distinct physical harmonics, causing distinct fault features to become conflated with wideband noise or unrelated components. This feature entanglement results in a noisy, redundant input space, which obscures the decision boundaries required by the downstream classifier and significantly complicates the feature extraction process.

By contrast, the optimized decomposition (middle-right panels) demonstrates enhanced spectral sparsity and reduced feature entanglement, effectively resolving these ambiguities. This improvement is not solely due to the relaxation of α , but rather the symbiotic optimization of a higher K combined with a moderate α . The increased number of modes provides sufficient discrete center frequencies (ω_k) to align with distinct physical harmonics, preventing a single mode from erroneously spanning multiple components. Consequently, the spectral peaks in this state are discrete and isolated with minimal cross talk, indicating that the VMD has functioned as an adaptive, non-overlapping filter bank. This induced spectral sparsity is instrumental for the hybrid CNN-Transformer architecture. It yields a cleaner and more structured frequency representation where the attention mechanism can precisely assign weights to fault-specific harmonics without the confounding interference of spectral leakage. By ensuring that the input tokens represent physically meaningful features rather than decomposition artifacts, the optimization significantly enhances the interpretability and accuracy of the diagnostic model.

(3) Residual analysis and denoising efficacy

A comparative analysis of the residual signals (bottom panels in Fig. 6) serves as a critical metric for evaluating the algorithm capacity for noise rejection and information preservation. In the case of the standard VMD, the residual signal exhibits negligible amplitude, approaching zero across all datasets (blue line). While such minimization is theoretically advantageous for data compression, within the context of fault diagnosis, it is symptomatic of overfitting. This behavior indicates that the decomposition has erroneously assimilated stochastic background noise, such as sensor thermal noise or random friction components, into the extracted IMFs. Consequently, the downstream model is compelled to encode these spurious noise patterns, thereby severely compromising its ability to generalize to unseen data.

In contrast, the residual associated with the optimized approach (red line) displays pronounced amplitude variability and substantial energy content. This is achieved because the optimized reconstruction does not simply sum all extracted modes; instead, it selectively discards specific IMFs that exhibit low computational metrics indicative of noise or irrelevant artifacts. In this framework, a high-energy residual is a positive indicator of effective feature sifting. It signifies that the optimization algorithm and subsequent mode selection has successfully segregated non-deterministic, Gaussian noise components and spurious, high-energy macro-shocks (such as the isolated artifact visible at time sample 1000 in the pitting signal and time sample 1250 in the cracked signal) into the residual term, thereby preserving the underlying, continuous deterministic and non-Gaussian fault modulations within the reconstructed signal. By aggressively filtering out these distracting, isolated transients, the optimized reconstruction provides a cleaner cyclostationary representation for the classifier. This denoising capability is quantitatively corroborated by the correlation coefficients between the reconstructed and raw signals, which reached 0.7330 for the pitting condition and 0.6867 for the broken tooth condition, as well as 0.6738 for the normal condition and 0.6844 for the cracked condition. These moderate correlation metrics confirm that the optimized

reconstruction successfully preserves the fundamental deterministic signal structure while rejecting a significant portion of the non-deterministic variance and noise floor. This demonstrates its efficacy as a robust and non-linear denoising filter that significantly augments the effective SNR of the input features, laying a clean foundation for the downstream model to extract diagnostic information.

(4) Generalizability of VMD optimization trends

The convergence of the Bayesian optimization toward a high K and a moderate α is a direct mathematical reflection of the physical kinematics inherent to WEC gearboxes, aligning precisely with established signal processing mechanics. WECs operate under complex, multi-harmonic excitation, simultaneously generating low-frequency hydrodynamic noise, structural resonances, and high-frequency gear mesh components. As demonstrated by Wang et al. (2015), in complex rotating machinery, a high K value is strictly required to sufficiently partition this broadband spectrum and prevent the aliasing of high-energy background noise into the fault-induced high-frequency bands. Simultaneously, the foundational mathematics of VMD dictate that the penalty factor α enforces the bandwidth constraint of the extracted modes. A moderate α is mandatory to ensure the extracted intrinsic mode functions maintain an optimal spectral bandwidth. Narrow bandwidths (induced by an excessively high α) mathematically truncate the critical frequency-modulation sidebands generated by localized gear faults, while overly wide bandwidths (low α) induce severe mode mixing (Dragomiretskiy and Zosso, 2014; Wang et al., 2015). Because the physical phenomena of low-speed, high-torque hydrodynamic excitation and fault-induced sideband modulation govern virtually all marine point absorbers, these hyperparameter optimization trends are highly generalizable across different WEC architectures and rotating component failure modes.

4.3.2. Comparative analysis on standard dataset

To validate the diagnostic superiority of the proposed hybrid CNN-Transformer framework, a comprehensive comparative study was conducted against a spectrum of established deep learning architectures. The baseline models were categorized into three distinct groups to evaluate specific architectural advantages.

- Convolutional-recurrent hybrids, i.e., CNN-LSTM, CNN-Gated Recurrent Unit (CNN-GRU), CNN-Bidirectional LSTM (CNN-BiLSTM), CNN-Bidirectional GRU (CNN-BiGRU), representing the traditional state-of-the-art baselines for time-series forecasting.
- Transformer-recurrent hybrids, including Transformer-GRU, Transformer-BiLSTM, Transformer-BiGRU, evaluating the efficacy of combining attention mechanisms with recurrence.
- Advanced multi-element hybrids, including CNN-LSTM-AM (Attention Mechanism), CNN-GRU-AM, CNN-BiLSTM-AM, and CNN-BiGRU-AM, evaluating the diagnostic performance improvements gained by integrating attention mechanisms into sequential convolutional-recurrent architectures under severe noise interference.
- Vision Transformer (ViT), assessing the performance of pure attention-based architecture without inductive bias.

(1) Quantitative performance assessment

The experimental results on the noise-free test set are summarized in Table 1. The proposed hybrid CNN-Transformer achieved an exceptional accuracy of 99.14% and a weighted F1-score of 99.14%, indicating near-ceiling diagnostic performance under nominal operating conditions.

(2) Analysis of recurrent and Transformer hybrids

Within the group of recurrent hybrids, the CNN-GRU architecture

Table 1

Performance comparison of different deep learning models on the clean dataset.

Model Category	Model Architecture	Accuracy (%)	Precision (%)	Recall (%)	weighted F1-Score (%)
Recurrent Hybrids	CNN-LSTM	94.83	94.66	95.31	94.83
	CNN-GRU	100.00	100.00	100.00	100.00
	CNN-BiLSTM	97.41	97.41	97.75	97.50
	CNN-BiGRU	98.28	98.50	98.25	98.27
Transformer Hybrids	Transformer-GRU	87.07	87.25	87.25	86.87
	Transformer-LSTM	98.28	98.27	98.50	98.25
	Transformer-BiGRU	85.34	86.00	85.00	84.99
	Transformer-BiLSTM	98.28	98.50	98.25	98.27
Advanced multi-element hybrids	CNN-LSTM-AM	96.55	96.50	96.55	96.58
	CNN-BiLSTM-AM	97.44	97.61	97.44	97.44
	CNN-GRU-AM	100.00	100.00	100.00	100.00
	CNN-BiGRU-AM	99.14	99.25	99.25	99.14
Pure Attention	ViT	86.32	86.25	86.25	86.26
Proposed	Hybrid CNN-Transformer	99.14	99.25	99.25	99.14

achieved perfect classification accuracy of 100% on the clean dataset. This result underscores that for high-SNR signals, the gating mechanisms inherent to the GRU are sufficient to model temporal dependencies effectively. In contrast, the CNN-LSTM and CNN-BiLSTM variants exhibited lower performance degradation, yielding accuracies of 94.83% and 97.41%, respectively. Although the CNN-GRU excels in this idealized scenario, its dependence on sequential processing inherently constrains computational parallelization, and as discussed in Section 4.3.3, limits its capacity to filter stochastic noise relative to attention-based mechanisms.

Conversely, the Transformer hybrids group displays significant performance instability. While the Transformer-BiLSTM model attained a competitive accuracy of 98.28%, the Transformer-BiGRU and Transformer-GRU architectures suffered substantial performance collapses, dropping to 85.34% and 87.07%, respectively. This marked variance suggests that the arbitrary superposition of recurrent layers upon Transformer blocks introduces optimization challenges and architectural redundancy, rather than guaranteeing performance enhancements.

(3) Analysis of advanced multi-element models

To evaluate against highly complex contemporary architectures, a set of three-element models incorporating attention mechanisms, specifically CNN-LSTM-AM, CNN-BiLSTM-AM, CNN-GRU-AM, and CNN-BiGRU-AM, were tested. The CNN-LSTM-AM and CNN-BiLSTM-AM achieved accuracies of 96.55% and 97.44%, respectively. Notably, the heavily stacked GRU-based variants attained exceptional performance on the clean dataset, with the CNN-BiGRU-AM matching the proposed Hybrid CNN-Transformer at 99.14%, and the CNN-GRU-AM reaching 100.00%. However, achieving absolute accuracy under ideal, nominal conditions is not the definitive metric for WEC diagnostics. In clean environments, highly parameterized multi-element architectures easily saturate the classification boundaries, rendering the clean dataset insufficient for distinguishing true diagnostic superiority. The critical engineering challenge lies in maintaining this feature separability under severe wave-induced hydrodynamic noise. Therefore, the true efficacy of the proposed structural bottleneck versus these highly stacked recurrent variants must be evaluated through a rigorous noise

robustness analysis, as detailed in Section 4.3.3.

(4) Comprehensive advantages of proposed architecture

The proposed architecture outperforms most baseline models while remaining competitive with the strongest recurrent hybrids. A critical ablation insight is provided by the comparison with the pure ViT, which achieved an accuracy of only 86.32%. This gap suggests that, without convolutional inductive bias, ViT models may be less effective at capturing localized fault transients. By integrating a CNN frontend, the proposed framework effectively bridges this gap, yielding an accuracy improvement exceeding 12%. This substantial gain validates the efficacy of the proposed local-global representation strategy, in which local convolutions capture high-frequency local patterns while the Transformer models long-range dependencies.

Beyond raw accuracy, the proposed model exhibits a highly balanced decision boundary, evidenced by precision and recall metrics both reaching 99.25%. Although CNN-GRU achieves perfect classification on this clean dataset, the proposed architecture offers a complementary advantage through attention-based feature selection. Specifically, self-attention can adaptively emphasize informative time-frequency regions while down-weighting irrelevant components, which is beneficial for improving robustness and generalization under non-ideal operating conditions.

4.3.3. Robustness and noise immunity analysis

In practical industrial scenarios, gearboxes operate under complex conditions where vibration signals are inevitably contaminated by strong background noise, including electromagnetic interference, structural resonance, and environmental stochastic perturbations. Consequently, the performance of the diagnostic model depends not only on its peak accuracy under nominal conditions but, more critically, on its stability and interference resistance in low SNR environments. To systematically evaluate the robustness of the proposed framework alongside the comparative models, a comprehensive noise immunity experiment was designed.

To accurately simulate hardware-level sensor corruption in real-world operating environments, additive white Gaussian noise was injected directly into the raw 1D time-domain vibration signals strictly prior to the VMD-STFT preprocessing stage. This specific injection sequence is mandatory to rigorously evaluate the empirical mode decomposition and noise-filtering capabilities of the optimized VMD algorithm before downstream deep learning classification. The experiment spans a broad dynamic range of noise intensities, with SNR levels set at -4 dB, -2 dB, 0 dB, 2 dB, 4 dB, 6 dB, 8 dB, 10 dB, 12 dB, 15 dB, and 20 dB. This gradient allows for a granular assessment of model performance, ranging from signal-dominated regimes to noise-dominated regimes, characterized by negative SNR, where the fault characteristic energy is submerged below the noise floor.

The relationship between the raw signal and the added noise is quantified by the SNR, defined as follows:

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) \quad (20)$$

where P_{signal} represents the power of the original effective vibration signal, and P_{noise} represents the power of the injected Gaussian noise.

By exposing all models, including the proposed hybrid CNN-Transformer and the high-performing benchmarks like CNN-GRU, to this identical noise spectrum, we can construct performance degradation curves. This comparison is designed to assess whether attention-based feature selection provides stronger implicit soft-thresholding against noise than the hard gating mechanisms inherent to recurrent networks.

All candidate models were subjected to this identical noise spectrum without retraining. This protocol specifically tests the generalization capability of the feature extraction layers. By comparing the degradation

rates of the proposed model against the benchmarks, we aim to verify whether the local-global feature extraction strategy offers superior interference suppression compared to the gating mechanisms of recurrent neural networks.

To evaluate structural robustness against non-stationary hydrodynamic interference, this analysis prioritizes benchmarking the proposed Hybrid CNN-Transformer against a triad of the highest-performing baselines: the CNN-GRU, the CNN-GRU-AM, and the CNN-BiGRU-AM. As established in Section 4.3.2, these specific architectures define the absolute ceiling of nominal performance, achieving 100.00%, 100.00%, and 99.14% accuracy on the noise-free dataset, respectively. Consequently, they serve as the most demanding empirical targets for stability testing. By subjecting this specific structural progression, from a standard recurrent network to an attention-augmented and bidirectional variant, to progressively severe SNRs, the evaluation transitions from assessing ideal feature saturation to authentic diagnostic generalization. This comparative isolation explicitly reveals whether structural overparameterization (via BiGRU) or auxiliary modules (via AM) genuinely enhance noise immunity or inadvertently compound susceptibility to hydrodynamic interference. The comparative performance degradation trajectories, alongside the absolute accuracy preservation of the proposed structural bottleneck over these heavily stacked baselines under deteriorating environmental conditions, are quantitatively illustrated in Fig. 7.

The experimental results reveal a divergence in model stability as the intensity of stochastic Gaussian noise increases. In the high-SNR environment (e.g., 20 dB), the baselines exhibit a marginal advantage, with both the CNN-GRU and CNN-GRU-AM achieving 99.14% accuracy compared to the proposed model's 98.28%. This confirms that for pristine or lightly corrupted signals, the rigorous sequential gating of

recurrent architectures efficiently models time-series dependencies. However, as the SNR drops to an intermediate level (10 dB), a significant stability gap emerges. While the proposed Hybrid CNN-Transformer sustains a robust accuracy of 97.41%, the CNN-GRU degrades to 92.24%, and the attention-augmented CNN-GRU-AM drops to 93.97%. This divergence indicates that recurrent baselines begin to overfit to the broadband stochastic noise, causing their decision boundaries to falter once the clean fault impulses are moderately masked.

This performance gap intensifies in the severe low-SNR regime (−4 dB), where the injected noise energy substantially exceeds the signal energy. Under these extreme conditions, the proposed model maintains an accuracy of 89.66%, exhibiting a flat degradation curve with a total absolute drop of only 9.48% from its clean baseline. In stark contrast, the highly parameterized recurrent models suffer catastrophic performance collapses. The standard CNN-GRU drops to 85.34% (a 14.66% decrease), while the complex CNN-BiGRU-AM and CNN-GRU-AM plummet to 82.76% and 81.90% (decreases of 16.38% and 18.10%), respectively. This definitively proves that arbitrary structural stacking and parameter bloat inadvertently compound a model's susceptibility to random Gaussian interference.

The superior noise immunity of the proposed framework is governed by a synergistic hard-soft denoising mechanism that directly mitigates the fundamental mathematical limitations of recurrent networks. Recurrent architectures process data sequentially; consequently, when random Gaussian perturbations corrupt the input sequence, the error propagates forward step-by-step, irreversibly distorting the hidden state updates throughout the entire subsequent temporal window. To overcome this sequential vulnerability, the proposed framework employs a dual-layer isolation strategy. First, the optimized VMD acts as a deterministic "hard" filter, stripping broadband stochastic components from the residual before feature extraction. Second, the self-attention mechanism functions as a dynamic "soft" filter. By calculating global dependencies simultaneously rather than sequentially, it assigns high attention weights to deterministic gear fault impulses while forcing the weights of noise-dominated segments to near-zero. This mathematical architecture bypasses sequential error accumulation entirely, directly resulting in the flattened degradation curve observed in the −4 dB environment.

4.3.4. Ablation study

To rigorously verify the contribution of each component within the proposed hybrid CNN-Transformer framework, a systematic multi-level ablation study was conducted. The experiments were designed to isolate the impact of four critical factors: (i) the signal decomposition and reconstruction strategy, (ii) the feature fusion strategy, (iii) the hybrid architectural backbone, and (iv) the channel attention.

(1) Impact of decomposition and screening strategy

To verify the necessity of the proposed Bayesian-optimized VMD with IMF selection, we evaluate four input configurations. This design isolates the contribution of the IMF screening step, i.e., ranking and selecting modes based on correlation and energy, relative to directly using the decomposed components without selection.

- Raw Signal: Normalized raw vibration data directly fed into the network.
- Bandpass Filter: Standard bandpass filtering (0 Hz - 500 Hz) applied to isolate resonance bands.
- VMD (All IMFs): The signal is decomposed using VMD, but all IMFs are summed together for reconstruction, effectively removing the selection step.
- Optimized VMD (selected IMFs): The signal is decomposed, and only the optimal IMFs (selected via the hybrid energy-correlation index) are used for reconstruction.

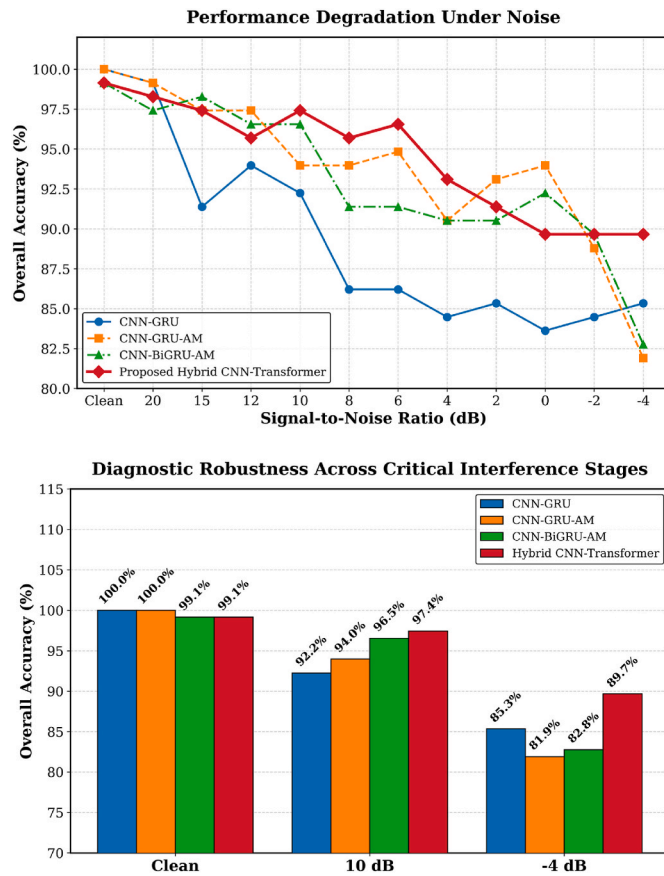


Fig. 7. Robustness comparison between the proposed Hybrid CNN-Transformer and three baseline models (CNN-GRU, CNN-GRU-AM, and CNN-BiGRU-AM) under varying SNR conditions.

Fig. 8 indicates that while the pure end-to-end CNN-Transformer achieves a competitive accuracy of 97.41% on the clean dataset, its performance collapses to 56.03% at -4 dB, which demonstrates that explicit VMD preprocessing is indispensable for maintaining diagnostic robustness in high-noise engineering environments, despite the associated computational overhead. In contrast, the transition to VMD (all IMFs) provides a significant performance leap to 71.5%, demonstrating the advantage of adaptive multi-component decomposition over linear filtering. However, the most substantial gain is observed with the proposed optimized VMD, which achieves 89.7% accuracy. This leap validates that the screening process, utilizing Bayesian optimization to select IMFs based on correlation and energy, effectively reduces information contamination from redundant or noise-dominant modes. As evidenced by the high symmetry of the metrics in the radar plot, the proposed method ensures a robust balance between precision and recall, providing the high-reliability diagnostic capability necessitated by the complex, non-linear dynamics of offshore energy systems.

(2) Impact of multi-source feature fusion

The proposed architecture uniquely integrates deep representations (STFT branch) with handcrafted statistical features (envelope, time, frequency stats). To verify the necessity of this feature fusion, we compared two configurations.

- Deep features only: The model uses only the STFT branch, removing the engineered feature heads.
- Proposed Fusion: The complete dual-path framework.

Fig. 9 compares the proposed dual-path fusion framework (Fig. 9a) with the “Deep Features Only” baseline (STFT branch, Fig. 9b) under a severe -4 dB SNR. Despite the low-SNR condition, the dual-path framework retains robust discriminative capability, achieving 100.00% classification accuracy for both pitting and broken tooth faults. While there is minor confusion between normal and cracked states,

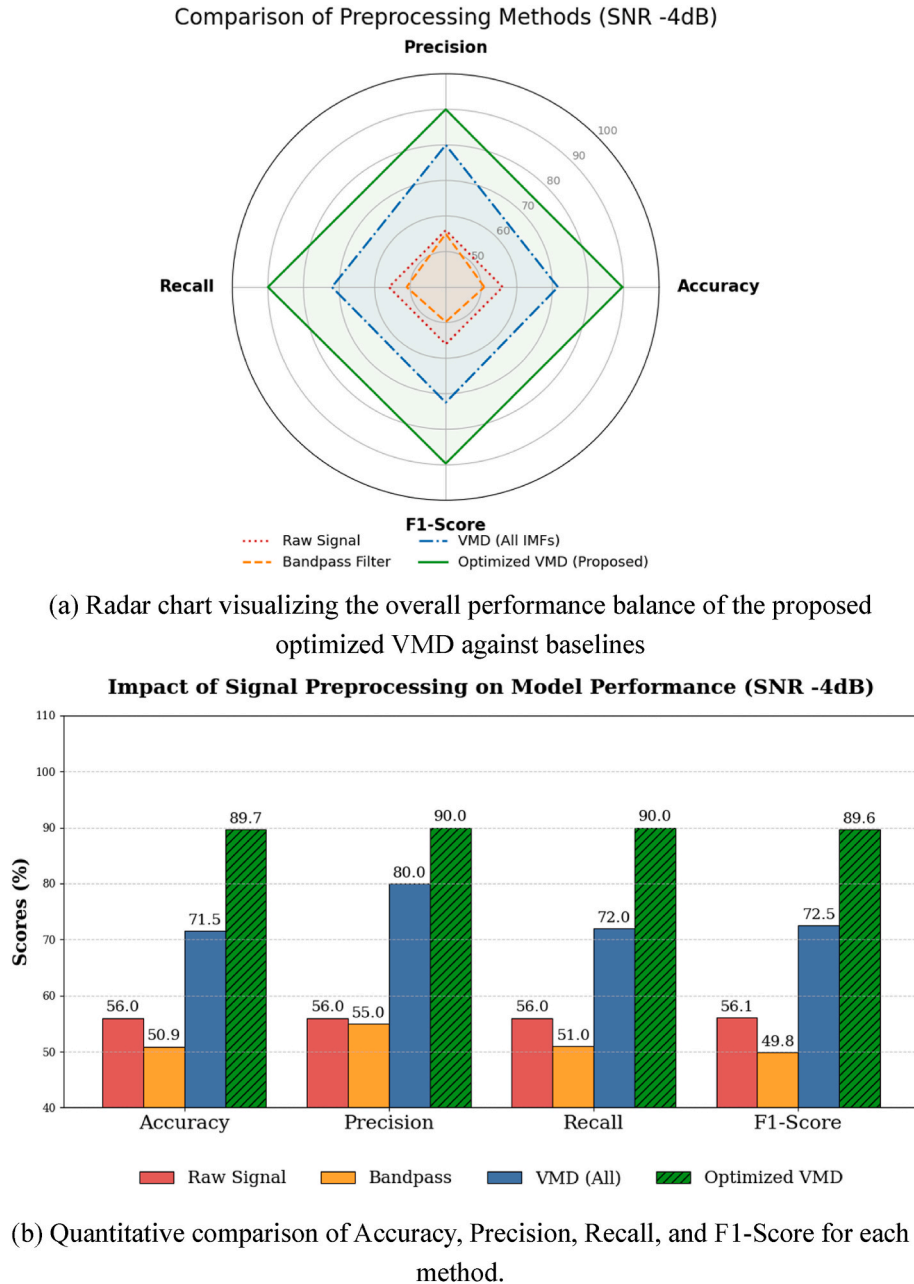


Fig. 8. Comparison of different signal preprocessing methods.

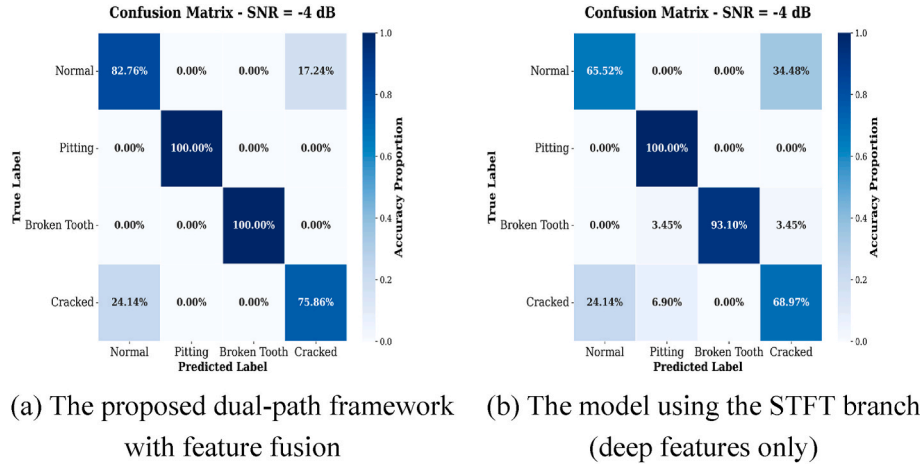


Fig. 9. Confusion matrices comparing diagnostic performance under -4 dB SNR.

where 17.24% of normal samples are misclassified as cracked and 24.14% of cracked samples as normal, the framework maintains an overall accuracy of 89.66%, driven by its precision in identifying distinct fault signatures.

In contrast, the “Deep Features Only” model exhibits a marked degradation in reliability, yielding an overall accuracy of 81.90%. Most notably, the diagnosis of broken tooth faults deteriorates from 100.00% to 93.10%, with misclassifications dispersed across pitting and cracked categories, indicating reduced feature separability. Furthermore, the confusion between normal and cracked conditions is exacerbated, with the error rate for normal samples nearly doubling to 34.48%. Additionally, 6.90% of cracked samples are misidentified as pitting, highlighting the baseline model's vulnerability to noise-induced feature overlap.

The divergence in accuracy (89.66% vs. 81.90%) substantiates the advantage of the proposed framework in adverse environments. These results indicate that integrating complementary information streams through feature fusion significantly enhances noise resilience, whereas relying exclusively on STFT-derived features fails to resolve noise-related ambiguity. Consequently, multi-modal feature fusion proves essential for ensuring robust fault diagnosis under conditions of high signal interference.

(3) Efficacy of hybrid architecture

To investigate the local-global modeling hypothesis, we decoupled the backbone architecture.

- VMD-CNN: Removing the Transformer encoder and relying solely on the CNN frontend.
- VMD-Transformer: Removing the CNN frontend and projecting STFT embeddings directly into the Transformer.
- Hybrid (the proposed model): The complete serial architecture.

Fig. 10 presents an ablation study comparing the diagnostic performance of the proposed dual-path fusion framework against standalone VMD-CNN and VMD-Transformer variants under a -4 dB SNR condition. The proposed framework (Fig. 10a) achieved a superior overall accuracy of 89.66%, effectively leveraging feature fusion to maintain discriminative precision. In contrast, the VMD-CNN model (Fig. 10b) exhibited a measurable decline in reliability (87.07% accuracy). Notably, the misclassification of normal samples as cracked increased to 34.48%, nearly double the rate of the fused model, suggesting that the CNN-only architecture lacks the long-range temporal modeling capabilities required to distinguish subtle fault features from noise. Performance deteriorated further in the VMD-Transformer model (Fig. 10c),

which yielded an accuracy of only 64.66%. Despite effectively identifying broken tooth and cracked states, this variant completely failed to identify the normal state, misclassifying 100% of these samples as cracked, and failed to diagnose pitting faults reliably with an accuracy of only 62.07%. This severe misclassification highlights the insufficiency of global attention mechanisms alone in capturing critical local spatial features. Ultimately, the significant accuracy gap between the fused architecture and the single-branch baselines substantiates the necessity of integrating CNN-driven local feature extraction with Transformer-based temporal modeling to ensure robust fault diagnosis in adverse signal environments.

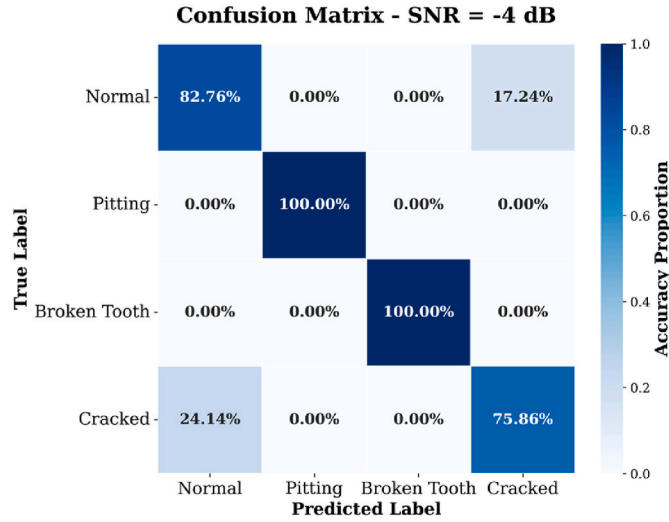
(4) Contribution of channel attention

To assess the specific contribution of the channel attention mechanism, we conducted an experiment removing the attention block from the network bottleneck.

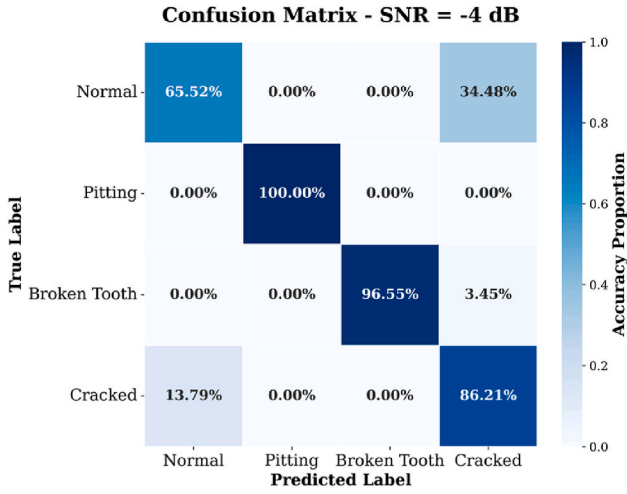
To quantify the impact of the channel attention mechanism on noise robustness, Fig. 11 contrasts the proposed hybrid CNN-Transformer model against a baseline lacking this module. The efficacy of the channel attention integration is most obvious under severe noise conditions characteristic of marine engineering environments. Specifically, at a challenging SNR of -4 dB, the proposed model maintains a resilient accuracy of 89.66%, surpassing the baseline's 78.45% by a substantial margin of over 11%. This significant performance gap indicates that the channel attention mechanism effectively recalibrates channel-wise feature responses, enabling the model to selectively emphasize informative fault signatures while suppressing noise-induced activations. Conversely, as signal quality improves ($\text{SNR} \geq 6$ dB), the performance metrics of both architectures begin to converge. Even under favorable conditions (e.g., the clean test set), the proposed method retains a slight superiority (99.14% vs. 98.28%). Overall, these trajectories validate the necessity of channel-wise feature refinement for ensuring reliable WEC monitoring in variable and harsh operational settings.

4.3.5. Feature representation and interpretability analysis

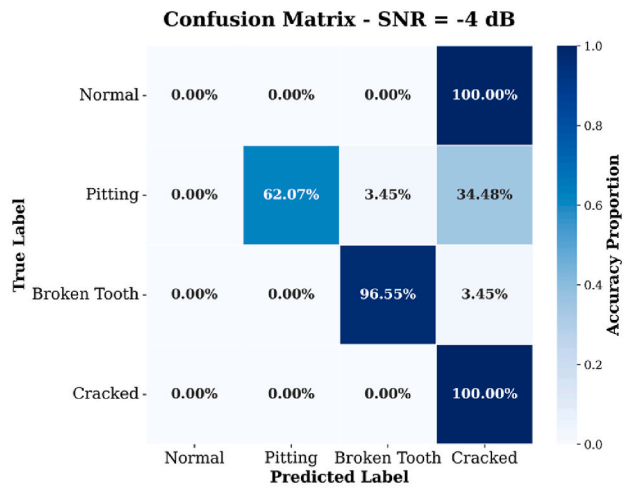
Following the component-wise ablation study, this section visualizes and evaluates the learned feature representations to verify their discriminative capability and physical interpretability. We utilize t-distributed stochastic neighbor embedding (t-SNE) to map the topological evolution of the feature space, illustrating the gradual separation of fault categories across network stages. Clustering metrics and statistical tests support this qualitative assessment, providing a quantitative benchmark for class separability. To clarify the diagnostic mechanism, we incorporate gradient-weighted class activation mapping (Grad-CAM) and channel attention visualizations. These methods isolate the



(a) Proposed Hybrid (VMD-CNN-Transformer)



(b) VMD-CNN



(c) VMD-Transformer

Fig. 10. Confusion matrices comparing the diagnostic performance of the proposed framework and ablation variants under -4 dB SNR.

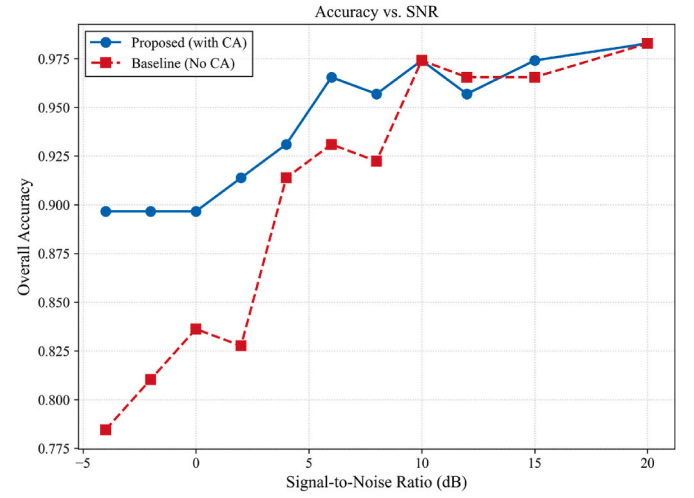


Fig. 11. Ablation study isolating the contribution of channel attention to noise robustness.

temporal and spectral features driving the model inference, verifying that predictions rely on physically meaningful signal patterns necessary for robust health state estimation.

Fig. 12 visualizes the progressive improvement in feature separability by projecting high-dimensional representations onto a two-dimensional manifold, supported quantitatively by the clustering metrics in Fig. 13. The initial raw signal domain (stage 1) lacks discriminative structure, reflected by a low Silhouette score of 0.09 and a Davies-Bouldin Index (DBI) of 3.79. While VMD processing (stage 2) successfully isolates the distinct signatures of severe macroscopic faults (pitting and broken teeth), severe topological entanglement persists between the normal and cracked samples, resulting in negligible metric improvement (Silhouette: 0.10, DBI: 3.72). This confirms that global spectral decomposition alone is insensitive to the subtle, low-energy transient dynamics characteristic of early-stage cracks. Introducing handcrafted statistical features (stage 3) successfully imposes physically meaningful structure onto the data, capturing essential dynamic behaviors. Visually, this stage successfully isolates the pitting and broken tooth faults into tight, distinct clusters. However, this physical relevance comes at the cost of topological complexity for the remaining states. A pronounced coalescence between normal and cracked samples illustrates that these physical descriptors are highly coupled and linearly inseparable in a low-dimensional space, an inherent non-linearity quantified by a negative Silhouette score of -0.03 and a sharp rise in the DBI to 22.81. Therefore, the parallel deep learning backbone (stages 4-6) is critical for extracting complementary, separable representations. CNN processing (stage 4) maintains the structural isolation of the broken teeth and pitting faults while initiating the disentanglement of the highly coupled normal and cracked classes, drastically improving the Silhouette score to 0.59 and reducing the DBI to 0.62. Further refinement via Transformer-based modeling (stage 5) successfully achieves full inter-class separation of these challenging states, elevating the Silhouette score to 0.73. Finally, the channel attention mechanism (stage 6) acts as a critical feature refinement step. Rather than performing the initial separation, it enhances intra-class compactness and tightens the cohesive clusters. This reduces feature variance within each fault category, maximizing the discriminative margins and optimizing the overall metrics to a peak Silhouette score of 0.74 and a DBI of 0.38. Finally, to confirm that this observed enhancement is not an artifact of stochastic variance, one-way analysis of variance (ANOVA) and pairwise significance tests were conducted across fault categories, as shown in Fig. 14. The resultant ANOVA p-value of 1.43×10^{-90} strongly rejects the null hypothesis, confirming that the learned representations exhibit statistically significant variance among classes. Furthermore, all pairwise comparisons

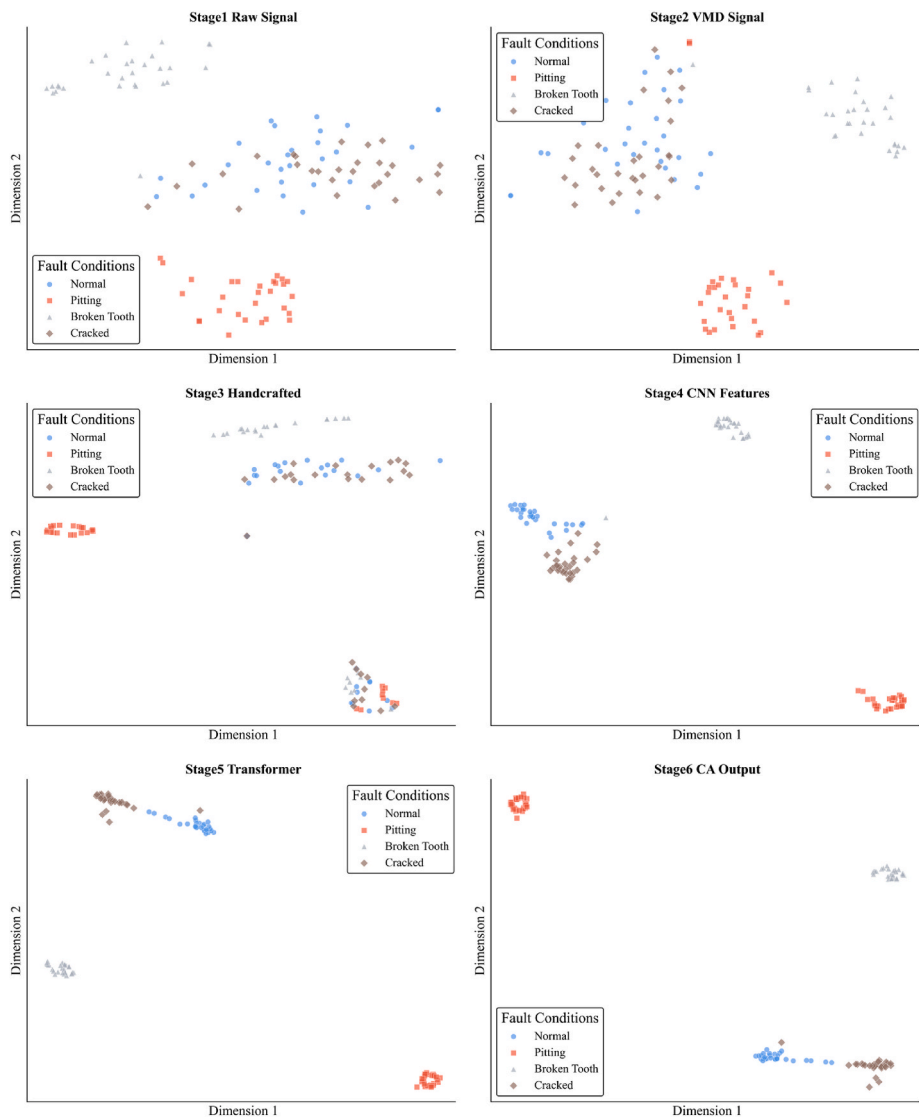


Fig. 12. t-SNE visualization of feature distributions at different stages of the proposed framework.

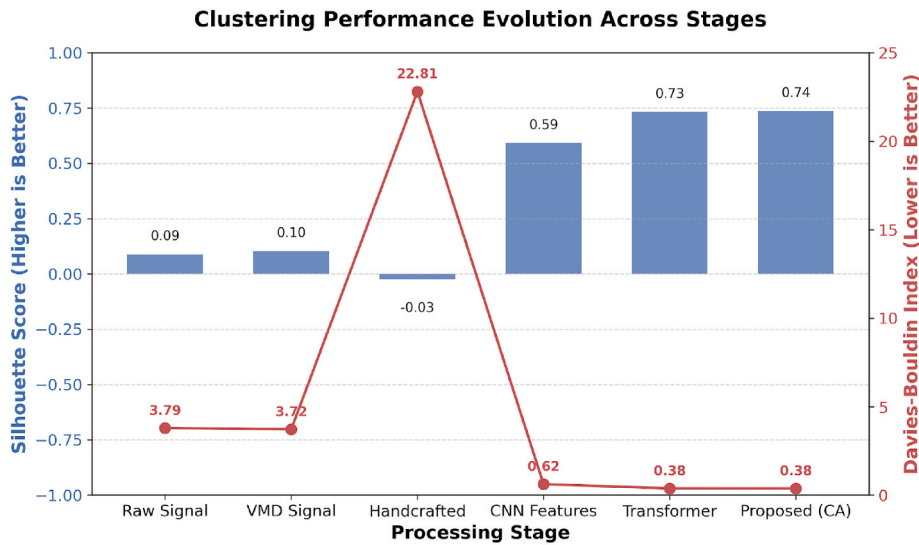


Fig. 13. Silhouette and Davies-Bouldin index evolution across stages.

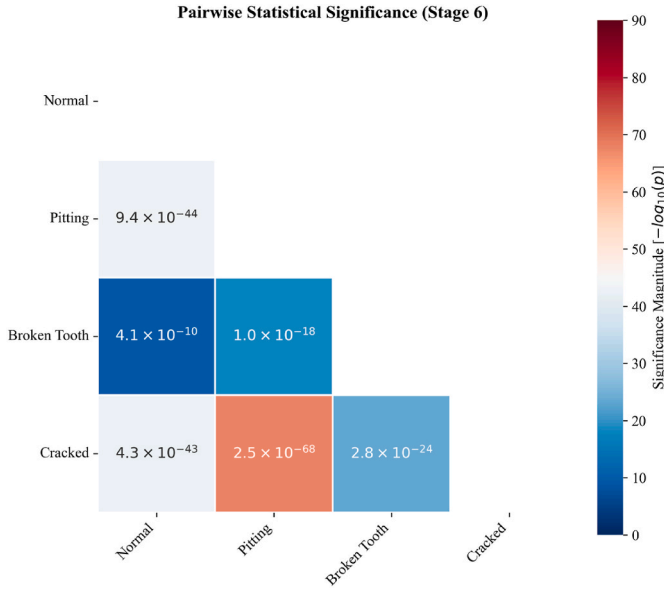


Fig. 14. Pairwise statistical significance of class separation (stage 6).

yield p -values markedly below the $\alpha = 0.01$ threshold, substantiating the model capacity to robustly distinguish between any two fault conditions. These results confirm that the observed enhancement in feature separability is not an artifact of stochastic variance, but a consistent, reproducible improvement conferred by the proposed framework.

To elucidate the physical interpretability governing the classification boundaries, Grad-CAM was utilized to map salient transient features back to the mechanical kinematics of the gearbox (Fig. 15). Because the convolutional backbone extracts features from 2D STFT embeddings, the raw activations are initially computed in the time-frequency domain. To achieve direct temporal interpretability, these 2D weights are aggregated and linearly interpolated to match the temporal resolution of the averaged STFT magnitude trace. In these visualizations, the solid black curve represents the averaged STFT magnitude, while the background heatmap displays the interpolated Grad-CAM weights, where warmer colors signify regions of high diagnostic importance.

The heatmaps demonstrate that the model successfully isolates distinct physical degradation mechanisms. For the broken tooth

condition, the model's attention is highly localized around sharp, secondary high-amplitude transients (temporal indices 19–22). Notably, the network actively bypasses the absolute maximum amplitude peak (near index 12). Physically, this proves the model does not naively correlate maximum amplitude with fault severity. The initial spike likely represents a low-frequency, wave-induced hydrodynamic shock, which is non-discriminative. Instead, the network targets the subsequent irregular, high-frequency amplitude modulations that physically correspond to the severe macro-mechanical shocks generated when the severely damaged gear tooth engages and disengages during the meshing cycle.

Conversely, the cracked state localizes its primary attention to an earlier temporal phase (indices 7–10). A root crack physically induces a periodic reduction in gear mesh stiffness rather than an explosive macro-shock. The Grad-CAM correctly reflects this by targeting moderate, broader amplitude variations, successfully differentiating the compliance-induced modulations of a crack from the impulsive volatility of a broken tooth. Furthermore, the pitting condition exhibits an activation pattern in the 1D temporal projection that mirrors the normal state. Micro-spalling (pitting) generates high-frequency structural resonances but negligible macroscopic amplitude deviations. The absence of temporal attention shifts indicates that the model correctly relies on the spectral distribution within the 2D STFT embeddings (specifically, the high-frequency sidebands distributed around the gear mesh frequency, f_m) rather than gross temporal shocks to identify early-stage wear.

Furthermore, visualizing the learned channel attention weights physically validates the model's feature selection mechanism (Fig. 16). The heatmap reveals that while the network utilizes a shared set of globally relevant spatial channels, the activation magnitudes are highly class-dependent. For the normal condition, attention weights remain uniformly moderate, reflecting the absence of transient fault-induced perturbations in the latent space. However, the broken tooth condition induces a highly polarized attention profile, aggressively amplifying specific channels (distinct yellow bands, weights >0.7) while suppressing others (dark purple bands). Mechanically, the amplified channels act as tuned filters isolating the broadband impulse energy generated by catastrophic tooth failure, while the suppressed channels actively mask out regular gear mesh harmonics and low-frequency wave noise that carry no fault-specific information. The cracked and pitting states exhibit more uniform, muted attention distributions, physically corresponding to their subtler degradation signatures, where fault energy is distributed across modulation sidebands rather than

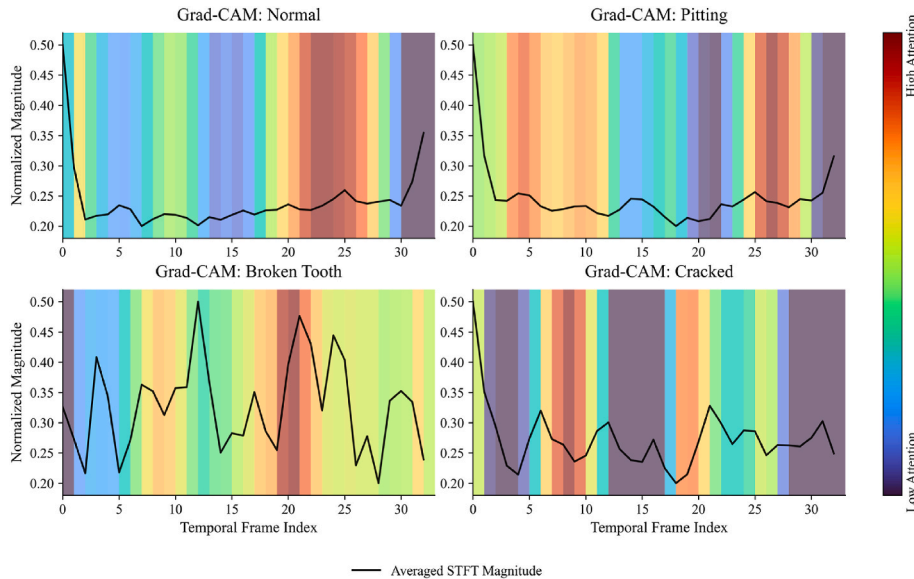


Fig. 15. Grad-CAM attention visualization for four fault conditions.

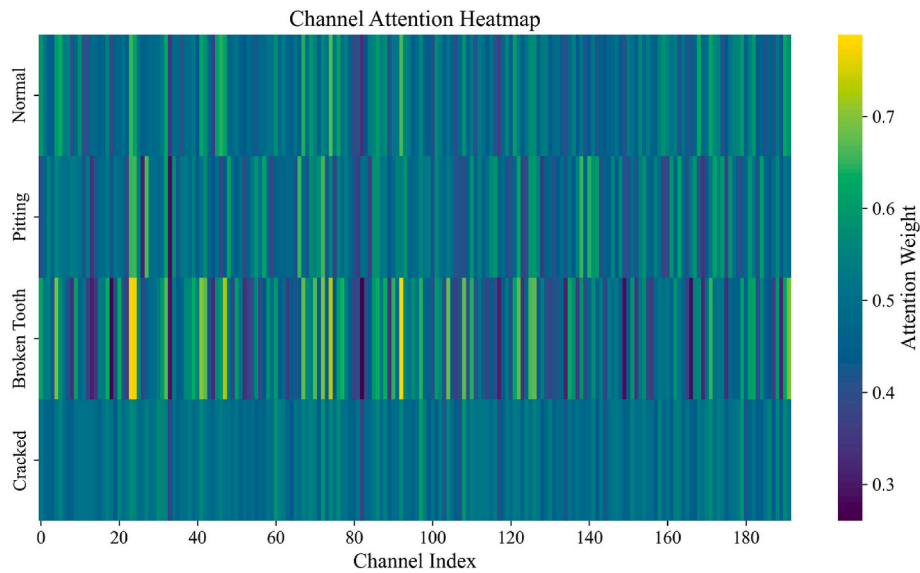


Fig. 16. Channel attention weight heatmap across fault conditions.

concentrated in violent, wide-band impulses. This dynamic calibration confirms that the attention module provides a physically sound, high-SNR representation for the final classification boundaries.

5. Conclusions

This study proposed a hybrid CNN-Transformer diagnostic framework, incorporating optimized VMD and multi-domain feature fusion, for fault identification in PA-WEC gearboxes. The main findings and engineering implications are summarized as follows.

- (1) The synergistic application of adaptive VMD and windowed STFT significantly mitigates random interference. By selectively reconstructing task-relevant components and applying overlapping segmentation, the framework drastically improves the discriminative resolution of time-frequency features. Ablation studies corroborate that this preprocessing stage is indispensable for generating the high-fidelity input required by downstream deep learning models.
- (2) The proposed diagnostic model unites the localized pattern extraction of CNNs with the long-range dependency tracking of Transformers. This integration delivers outstanding classification accuracy and class separability, particularly under fluctuating loads and variable speed regimes. Achieving a 99.14% accuracy in noise-free environments, the architecture establishes a highly competitive diagnostic baseline. While this performance effectively matches the optimal CNN-GRU benchmark under ideal conditions, the true superiority of the proposed hybrid structure emerges in complex operating environments.
- (3) The proposed model exhibits remarkable robustness against intense noise interference ($\text{SNR} = -4 \text{ dB}$), maintaining an accuracy of nearly 90% and exceeding the CNN-GRU benchmark by 4.32% with remarkably gradual performance degradation. Each functional module (i.e., optimized VMD, feature fusion, sequential CNN-Transformer, and channel attention) positively reinforces this robustness, especially in low-SNR scenarios. Meanwhile, visualization analyses verify that the model isolates physically meaningful fault signatures, successfully avoiding the pitfall of overfitting to background noise.
- (4) Validated on real-world WEC gearbox vibration data, the framework maintains stable and precise diagnostic performance across actual operating conditions. These empirical results

underscore the vast potential of the proposed model for field deployment in the highly complex and noise-intensive environments typical of marine engineering.

Despite these results, the computational complexity of VMD currently constrains real-time edge deployment. Future research will focus on developing lightweight frequency-decoupling algorithms for embedded monitoring, exploring cross-domain transfer learning to handle continuous variable sea states, and integrating the extracted health indicators for RUL prediction of WEC transmission systems.

CRedit authorship contribution statement

Zhimin Cheng: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jichuan Kang:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **Yu Sun:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Dandan Peng:** Methodology, Conceptualization. **Mingxin Li:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by the Oil & Gas Major Project (2025ZD1403405).

References

- Ahmad, H., Cheng, W., Xing, J., Wang, W., Du, S., Li, L., Zhang, R., Chen, X., Lu, J., 2024. Deep learning-based fault diagnosis of planetary gearbox: a systematic review. *J. Manuf. Syst.* 77, 730–745.
- Ali, Y.M., Ding, L., 2025. Fault diagnosis in photovoltaic arrays: a robust and efficient approach using feature engineering and 1D-CNN. *Int. J. Electr. Power Energy Syst.* 170, 110863.
- Berrio, Y., Rivillas-Ospina, G., Ruiz-Martínez, G., Arango-Manrique, A., Ricaurte, C., Mendoza, E., Silva, R., Casas, D., Bolívar, M., Díaz, K., 2023. Energy conversion and beach protection: numerical assessment of a dual-purpose WEC farm. *Renew. Energy* 219, 22.

- Cao, L.X., Qian, Z., Zareipour, H., Huang, Z.K., Zhang, F.H., 2019. Fault diagnosis of wind turbine gearbox based on deep Bi-Directional long short-term memory under time-varying non-stationary operating conditions. *IEEE Access* 7, 155219–155228.
- Chen, B., Zhang, W., Xi Gu, J., Song, D., Cheng, Y., Zhou, Z., Gu, F., Ball, A.D., 2023. Product envelope spectrum optimization-gram: an enhanced envelope analysis for rolling bearing fault diagnosis. *Mech. Syst. Signal Process.* 193, 110270.
- Chen, H., Wei, J., Huang, H., Wen, L., Yuan, Y., Wu, J., 2024. Novel imbalanced fault diagnosis method based on generative adversarial networks with balancing serial CNN and transformer (BCTGAN). *Expert Syst. Appl.* 258, 125171.
- Cheng, J., Yang, Y., Yu, D., 2010. The envelope order spectrum based on generalized demodulation time-frequency analysis and its application to gear fault diagnosis. *Mech. Syst. Signal Process.* 24 (2), 508–521.
- Dragomiretskiy, K., Zosso, D., 2014. Variational mode decomposition. *IEEE Trans. Signal Process.* 62 (3), 531–544.
- Feng, Z., Liang, M., Chu, F., 2013. Recent advances in time-frequency analysis methods for machinery fault diagnosis: a review with application examples. *Mech. Syst. Signal Process.* 38 (1), 165–205.
- González-Esculpi, A., Verde, C., Maya-Ortiz, P., 2022. Comparison of estimates of the excitation force for fault diagnosis in a wave energy converter. *IFAC-PapersOnLine* 55 (6), 396–401.
- Gu, J., Peng, Y., Lu, H., Chang, X., Chen, G., 2022. A novel fault diagnosis method of rotating machinery via VMD, CWT and improved CNN. *Measurement* 200, 111635.
- Guo, B., Wang, T., Jin, S., Duan, S., Yang, K., Zhao, Y., 2022. A review of point Absorber wave energy converters. *J. Mar. Sci. Eng.* 10 (10), 1534.
- Guo, B.Y., Ringwood, J.V., 2021. A review of wave energy technology from a research and commercial perspective. *IET Renew. Power Gener.* 15 (14), 3065–3090.
- Hasan Maheen, M., Yang, Y., 2023. Wave energy converters with rigid hull encapsulation: a review. *Sustain. Energy Technol. Assessments* 57, 103273.
- He, G., Luan, Z., Zhang, W., He, R., Liu, C., Yang, K., Yang, C., Jing, P., Zhang, Z., 2023. Review on research approaches for multi-point absorber wave energy converters. *Renew. Energy* 218, 119237.
- Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccufier, M., Verstockt, S., Van de Walle, R., Van Hoecke, S., 2016. Convolutional neural network based fault detection for rotating machinery. *J. Sound Vib.* 377, 331–345.
- Jiang, H., Li, C., Li, H., 2013. An improved EEMD with multiwavelet packet for rotating machinery multi-fault diagnosis. *Mech. Syst. Signal Process.* 36 (2), 225–239.
- Kang, J.C., Zhu, X., Shen, L., Li, M.X., 2024. Fault diagnosis of a wave energy converter gearbox based on an Adam optimized CNN-LSTM algorithm. *Renew. Energy* 231.
- Kolios, A., Di Maio, L.F., Wang, L., Cui, L., Sheng, Q., 2018. Reliability assessment of point-absorber wave energy converters. *Ocean Eng.* 163, 40–50.
- Li, H., Ding, Y., Sun, Y., Xie, M., Guedes Soares, C., 2025. An intelligent failure feature learning method for failure and maintenance data management of wind turbines. *Reliab. Eng. Syst. Saf.* 261, 111113.
- Li, H., Wu, J.Y., Shi, X.D., Kong, L.J., Kong, W.H., Zhang, Z.T., Pan, Y.J., Luo, D.B., Yan, J.Y., 2024a. A self-powered smart wave energy converter for sustainable sea. *Mech. Syst. Signal Process.* 220, 22.
- Lai, C., Baraldi, P., Zio, E., 2024. Physics-informed deep autoencoder for fault detection in new-design systems. *Mechanical Systems and Signal Processing* 215, 111420.
- Lai, C., Baraldi, P., Zio, E., 2025. Gradient-enhanced physics-informed long short-term memory networks for stable and accurate prediction of the RUL of electronic components. *Reliability Engineering & System Safety* 111485.
- Li, J., Deng, A.D., Yang, Y., Cheng, Q., 2019a. Fault diagnosis of wind turbine drive train using time-frequency estimation and CNN. In: 10th IEEE Prognostics and System Health Management Conference (PHM-Qingdao), Qingdao, PEOPLES R CHINA.
- Li, M., Xu, Z., Li, S., Kikuchi, Y., Dong, Y., Gryllias, K.C., Baraldi, P., Zio, E., Carroll, J., 2026. Health prognostics and maintenance decision-making for wind energy: A comprehensive overview. *Renewable and Sustainable Energy Reviews* 226, 116269.
- Li, M.-W., Xu, R.-Z., Yang, Z.-Y., Yeh, Y.-H., Hong, W.-C., 2024b. Optimizing berth-crane allocation considering tidal effects using chaotic quantum whale optimization algorithm. *Appl. Soft Comput.* 162, 111811.
- Li, M., Jiang, X., Carroll, J., Negenborn, R.R., 2023. A closed-loop maintenance strategy for offshore wind farms: incorporating dynamic wind farm states and uncertainty-awareness in decision-making. *Renew. Sustain. Energy Rev.* 184, 113535.
- Li, M., Kang, J., Sun, L., Wang, M., 2019b. Development of optimal maintenance policies for offshore wind turbine gearboxes based on the non-homogeneous continuous-time markov process. *J. Mar. Sci. Appl.* 18 (1), 93–98.
- Li, M., Kikuchi, Y., Ringsberg, J.W., Gryllias, K.C., Baraldi, P., Zio, E., Carroll, J., 2026. Empowering digital twins for wind energy operation and maintenance: A prospective framework and future directions. *Renewable and Sustainable Energy Reviews* 235, 116970.
- Li, S., Kim, D.K., Benson, S., 2021. A probabilistic approach to assess the computational uncertainty of ultimate strength of hull girders. *Reliab. Eng. Syst. Saf.* 213, 107688.
- Lian, J., Liu, Z., Wang, H., Dong, X., 2018. Adaptive variational mode decomposition method for signal processing based on mode characteristic. *Mech. Syst. Signal Process.* 107, 53–77.
- Liu, S., Tang, G., Wang, X., He, Y., 2016. Time-Frequency analysis based on improved variational mode decomposition and teager energy operator for rotor System fault diagnosis. *Math. Probl Eng.* 2016 (1), 1713046.
- Ozkan, C., Perez, K., Mayo, T., 2020. The impacts of wave energy conversion on coastal morphodynamics. *Sci. Total Environ.* 712.
- Pastor, J., Liu, Y., 2012. Hydrokinetic energy: overview and it's renewable energy potential for the Gulf of Mexico. In: 2012 IEEE Green Technologies Conference, pp. 1–3.
- Peng, Y., Cheng, F., Qiao, W., Qu, L., 2017. Fault prognosis of drivetrain gearbox based on a recurrent neural network. In: 2017 IEEE International Conference on Electro Information Technology (EIT), pp. 593–599.
- Roach, A., Meek, M., Ali, R., DuPont, B., Roberston, B., 2025. A state-of-the-art review of submerged wave energy converters. *Renew. Sustain. Energy Rev.* 222, 115901.
- Said, H.A., Sardá, A.C., Ringwood, J., 2025. Fault management in wave energy systems: diagnosis, prognosis, and fault-tolerant control. *Ocean Eng.* 316.
- Shadmani, A., Nikoo, M.R., Gandomi, A.H., Wang, R.Q., Golparvar, B., 2023. A review of machine learning and deep learning applications in wave energy forecasting and WEC optimization. *Energy Strategy Rev.* 49.
- Sun, Y., Li, H., Sun, L., Guedes Soares, C., 2023a. Failure analysis of floating offshore wind turbines with correlated failures. *Reliab. Eng. Syst. Saf.* 238, 109485.
- Sun, Z., Machlev, R., Wang, Q., Belikov, J., Levron, Y., Baimel, D., 2023b. A public dataset for synchronous motor electrical faults diagnosis with CNN and LSTM reference classifiers. *Energy AI* 14, 100274.
- Tang, Y., Huang, Y., Lindbeck, E., Lizza, S., VanZwieten, J., Tom, N., Yao, W., 2020. WEC fault modelling and condition monitoring: a graph-theoretic approach. *IET Electr. Power Appl.* 14 (5), 781–788.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.
- Verwimp, T., Zhu, R., Wen, H., Achilleos, A., Gryllias, K., 2025. Rotary encoder-based health indicators for early detection of gear pitting in commercial gearboxes. In: Proceedings of the Asia Pacific Conference of the PHM Society 2025. The Prognostics and Health Management Society.
- Wang, Y., Markert, R., Xiang, J., Zheng, W., 2015. Research on variational mode decomposition and its application in detecting rub-impact fault of the rotor system. *Mech. Syst. Signal Process.* 60–61, 243–251.
- Wang, Y., Zhu, C., Zhu, Y., Luo, X., 2025. Life extension of wind turbine gearboxes based on pitch control. *Results Eng.* 27, 106264.
- Xu, Z., 2025. An intelligent predictive maintenance framework for floating offshore wind turbine based on structural damage prediction. *Appl. Ocean Res.* 164, 104796.
- Ye, Y., Gan, J., Liu, H., Wu, W., Wang, L., Guo, J., Li, H., 2025. Bending behaviour of rigid-flexible combined inflatable membrane structures for floating bridge modules. *Ocean Eng.* 338, 122005.
- Zhang, Y., Zeng, T., Gao, Z., 2022. Fault diagnosis and fault-tolerant control of energy maximization for wave energy converters. *IEEE Trans. Sustain. Energy* 13 (3), 1771–1778.
- Zhang, Z., Qin, J., Wang, D., Huang, S., Liu, Y., Xue, G., 2023. Hybrid model predictive control of a two-body wave energy converter with mechanically driven power take-off. *J. Mar. Sci. Eng.* 11 (8), 1618.
- Zhao, T., Feng, S., Zhou, Y., Kang, Z., Kang, J., 2024. Data-driven fault detection framework for offshore wind-hydrogen systems. *Int. J. Hydrogen Energy* 70, 325–340.
- Zhu, R., Mousmoulis, G., Gryllias, K.C., 2025. Wavelet-based high order spectrum for local damage diagnosis of gears. *Mechanics & Industry* 6, 13.
- Zhu, R., Verwimp, T., Mauricio, A., Gryllias, K., 2026. An improved Instantaneous Wavelet Bispectrum Feature for gear fault diagnostics using a Wireless MEMS sensor. *Mech. Syst. Signal Process.* 244, 113729.