

Article

An Intelligent Bearing Fault Diagnosis Model with Physics-Information Fusion: Design and Interpretability Mechanism Research

Zhenyang Yu ^{1,2} , Fei Shao ^{2,*}, Qian Xu ², Xingkun Xie ² , Lixiang He ² and Yinchuan Hou ^{2,*}

¹ State Key Laboratory of Disaster Prevention and Mitigation of Explosion and Impact, Army Engineering University of PLA, Nanjing 210001, China; yuzy020704@163.com

² Field Engineering Institute, Army Engineering University of PLA, Nanjing 210001, China; xuqian@aeu.edu.cn (Q.X.); jfxie2020@sina.com (X.X.); helx@aeu.edu.cn (L.H.)

* Correspondence: shaofei2022@sina.com (F.S.); houyc0521@163.com (Y.H.)

Abstract

To address the issue of limited interpretability in current deep learning-based bearing fault diagnosis methods, this study proposes a convolutional neural network model incorporating spectral physical constraints. The model integrates data-driven learning with physical constraints by embedding the spectral features of bearing faults as prior knowledge into the network architecture during training. This guides the model to adaptively learn spectral features closely associated with fault mechanisms. Experimental results on both public datasets and self-collected data show that the proposed model not only maintains high diagnostic accuracy but also provides intuitive and credible justification for fault classification through the visualization of spectral responses at the network output layer. This enhances the spectral interpretability of the model's decisions, achieving a transition from a “black box” to a “white box” and effectively improving the reliability of deep diagnostic models.

Keywords: intelligent fault diagnosis; adaptive learning; interpretability; physical constraints

1. Introduction

As a core component of rotating machinery, the operational state of rolling bearings directly impacts the reliability, safety, and production efficiency of the entire equipment [1]. Bearing failures leading to unplanned downtime can cause significant economic losses and even trigger safety incidents [2,3]. Therefore, conducting accurate and efficient research on bearing fault diagnosis is of paramount importance for achieving intelligent equipment operation and maintenance, as well as ensuring the safe and stable operation of industrial systems [4,5].

In recent years, with the rapid advancement of sensor technology and big data acquisition capabilities, data-driven intelligent diagnostic methods, particularly deep learning, have become a research hotspot in this field [6–9]. Mansouri [10] et al. proposed a simplified RNN based on hierarchical K-means clustering and two interval-valued data simplification techniques, enabling deeper extraction of signal features while reducing computation time and improving accuracy. Niu [11] et al. introduced a CNN variant model that enhances the extraction of fault-relevant information by incorporating a channel attention mechanism to explore intra-channel feature relationships, achieving an end-to-end diagnostic mode and effectively improving bearing diagnostic performance. Cao [12] et al. proposed an unsupervised domain-shared convolutional neural network for fault transfer diagnosis



Academic Editor: Zhuming Bi

Received: 3 July 2026

Revised: 21 August 2026

Accepted: 23 August 2026

Published: 7 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

from constant to variable speed conditions, demonstrating high diagnostic accuracy and faster convergence in bearing fault diagnosis cases. Deep learning models represented by convolutional neural networks and recurrent neural networks, with their powerful automatic feature extraction and nonlinear mapping capabilities, have shown outstanding performance in processing high-dimensional and complex bearing vibration signals, significantly improving fault identification accuracy and generalization ability, thereby advancing fault diagnosis technology toward intelligence and automation [13–15].

However, mainstream deep learning models, known as “black boxes,” though effective in diagnosis, have decision-making processes that are difficult for humans to interpret [16,17]. This lack of interpretability poses a critical challenge in the safety-critical field of industrial diagnostics: engineers cannot trust diagnostic results that lack reasoning, and it is difficult to uncover new fault mechanisms or perform root cause analysis of misjudgments [18]. In response, explainable artificial intelligence research aimed at revealing model decision logic and establishing human–machine trust has emerged [19,20]. In the field of fault diagnosis, to enhance the credibility of deep learning models, current interpretability research primarily unfolds in two directions: first, developing post hoc interpretability tools to visualize the decision basis of existing “black-box” models. For example, Pang [21] et al. proposed an interpretable convolutional neural network model that directly visualizes the spectral features of samples output from the final layer, verifying the model’s ability to effectively learn and extract frequency-domain components closely related to the physical mechanisms of faults, thereby providing clear post hoc interpretability for the decision-making process; second, exploring inherently interpretable model structures to directly provide readable reasoning clues during diagnosis. For instance, Lao [22] et al. proposed an interpretable network architecture that reconstructs the signal processing workflow of fault diagnosis as the first four layers of the network, making the entire processing flow from raw vibration signals to fault features transparent within the model. Wang [23] et al. incorporated classical physical sparsity metrics as the direct basis for model decisions, ensuring that the output health index has a clear physical measurement source, thereby guaranteeing the interpretability of the decision path.

Although existing research has made progress in enhancing model interpretability, the improvement approaches have mainly focused on algorithmic and model-level optimization [24,25]. Whether developing visualization tools to display the model’s areas of attention or designing inherently interpretable network structures, these methods, while helpful in understanding how the model makes judgments, overlook the physical laws and prior knowledge inherent to the diagnostic objects in the field of fault diagnosis. This results in interpretability outcomes that may appear statistically reasonable but lack a solid physical foundation, limiting the generalization and trustworthy application of models in complex and variable industrial scenarios.

To address the above issues, this paper proposes a physics-informed intelligent diagnosis model for bearing faults (PIIDM). The model transforms the spectral physical knowledge of bearing faults into differentiable constraints and structured network modules, integrating them into the model’s architectural design and training process. By fusing physical laws with data-driven mechanisms, the model not only adaptively learns bearing fault features at the data level but also adheres to the spectral physical mechanisms of bearing faults in its decision-making path. This ensures that the generated feature representations and diagnostic results are consistent with data statistical characteristics while strictly satisfying the physical constraints of faults. This study aimed to construct an intelligent diagnostic framework with transparent internal logic, reliable diagnosis, and physical verifiability, thereby advancing the practical application of interpretable deep learning in industrial fault diagnosis.

Table 1 shows the essential differences in architectural design between the model in this paper and the ELCNN in reference [21], the Causal Invariant GNN in reference [22], and the Sparsity-constrained ELM in reference [23]. Both Sparsity-constrained ELM and Causal Invariant GNN hard-code prior knowledge into specific network architectures, resulting in the methods being limited by specific backbone models. ELCNN achieves structural simplification and feature enhancement by removing the bias, activation, and pooling layers and combining the square characteristics of S-GAP. However, its physical properties originate from the signal processing logic, and mechanical dynamic parameters are not explicitly introduced. In contrast, this paper constructs an architecture-independent physical information loss function, injecting the bearing failure frequency as an explicit constraint into the optimization process. This kind of physical embedding loss avoids the complexity of network reconstruction and is an effective supplement to the existing research on the embedding architecture of physical prior knowledge.

Table 1. Comparison of different interpretability model architectures.

	ELCNN [21]	Causal Invariant GNN [22]	Sparsity-Constrained ELM [23]	PIIDM
Core Paradigm	Interpretability by structure	Causality in architecture	Physics in architecture	Physics in loss
Physical priori source	—	Structural causal model/Graphically coupled module	Wavelet transform layer/Kurtosis/Gini sparse neurons	Spectral energy aggregation constraint
Backbone network limitations	Strongly dependent on the GNN architecture	Strongly dependent on the GNN architecture	Strongly dependent on the ELM architecture	Architectural independence
Optimization method	Reduces the number of parameters to enhance noise immunity	Modifies the definition of the graph structure	Modifies the definition of the network layer	Only the loss function term is added, while the network remains unchanged

The main structure of the paper is organized as follows. The introduction section elaborates on the development of intelligent fault diagnosis for bearings and the current status and challenges in interpretability research. The second chapter systematically introduces the spectral characteristics of bearing faults and their physical implications. The third chapter proposes the overall architecture of PIIDM and provides a detailed explanation of its module design and operational mechanisms. The fourth chapter validates the classification performance and interpretability of the proposed model through experiments. The fifth chapter summarizes the entire work and outlines directions for future research.

2. Background

In the fault diagnosis of rolling bearings, local defects (such as spalling of the inner ring, outer ring, or rolling elements) induce periodic impact excitations. These impact excitations can be monitored through vibration signals and exhibit distinct characteristic patterns in the frequency domain [26]. Figure 1 shows the time-domain waveforms and frequency spectra of a bearing under normal and different fault conditions. The analysis reveals the following:

(1) In the frequency spectrum, bearing fault characteristics manifest as a series of uniformly spaced discrete spectral lines, with the characteristic frequency serving as the fundamental frequency. The characteristic frequency components (including the fundamental and its harmonics) of the inner ring, outer ring, and rolling elements are all clearly observable in the spectrum. The fundamental frequency can be calculated using Formula (1).

(2) Due to modulation by the system's frequency response function, the amplitudes of these discrete spectral lines are no longer constant. Instead, they exhibit significant amplitude modulation and frequency modulation effects, forming sidebands centered around the system's natural frequency in the spectrum, characterized by amplitude fluctuations.

Therefore, spectrum analysis serves as an effective criterion for extracting fault features and diagnosing rolling bearing faults. By directly observing the characteristic frequency components strongly correlated with bearing geometry and rotational speed and analyzing their amplitude levels, harmonic structures, and interactions with the system's natural frequency, the type, location, and severity of faults can be accurately determined.

$$\begin{cases} f_{IR} = \frac{Z}{2}(1 + \frac{d}{D}\cos\alpha)f_r \\ f_{OR} = \frac{Z}{2}(1 - \frac{d}{D}\cos\alpha)f_r \\ f_{Ball} = \frac{D}{d}[1 - (\frac{d}{D}\cos\alpha)^2]f_r \end{cases} \quad (1)$$

where Z represents the number of rolling elements, d is the diameter of the rolling elements, D is the pitch circle diameter of the bearing, α is the contact angle, and f_r is the rotational frequency of the shaft.

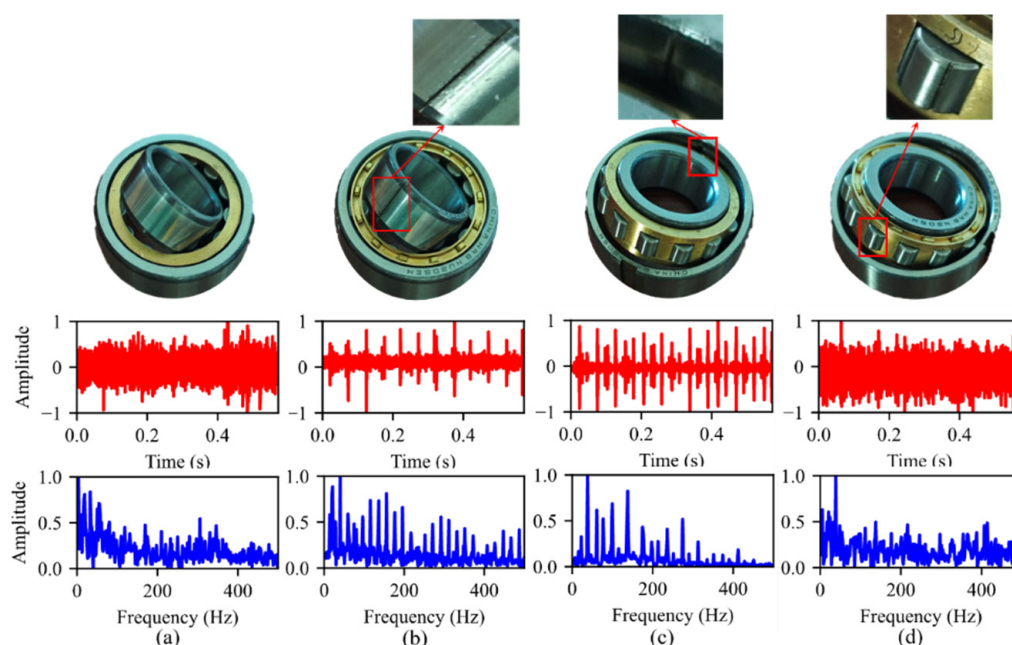


Figure 1. Time-domain and frequency-domain diagrams of bearings in different states. (a) Normal bearing. (b) Inner ring spalling. (c) Outer ring spalling. (d) Rolling element spalling.

3. Design of Interpretable Models Based on Spectral Physical Constraints

The PIIDM model proposed in this paper empowers the frequency-domain features of bearing faults into the network in the form of physical constraints, guiding the model to learn features with clear physical meanings and achieving an interpretability leap from “black box” to “white box”. The model architecture is shown in Figure 2, which includes: (1) a multi-scale CNN feature extractor to capture discriminative features from the original signal; (2) the loss function was improved by introducing a physical-penalty loss on the basis of the cross-entropy loss function. The significance of the bearing fault spectrum features was used as the regularization term to enhance the transparency and physical credibility of the model's diagnostic decisions.

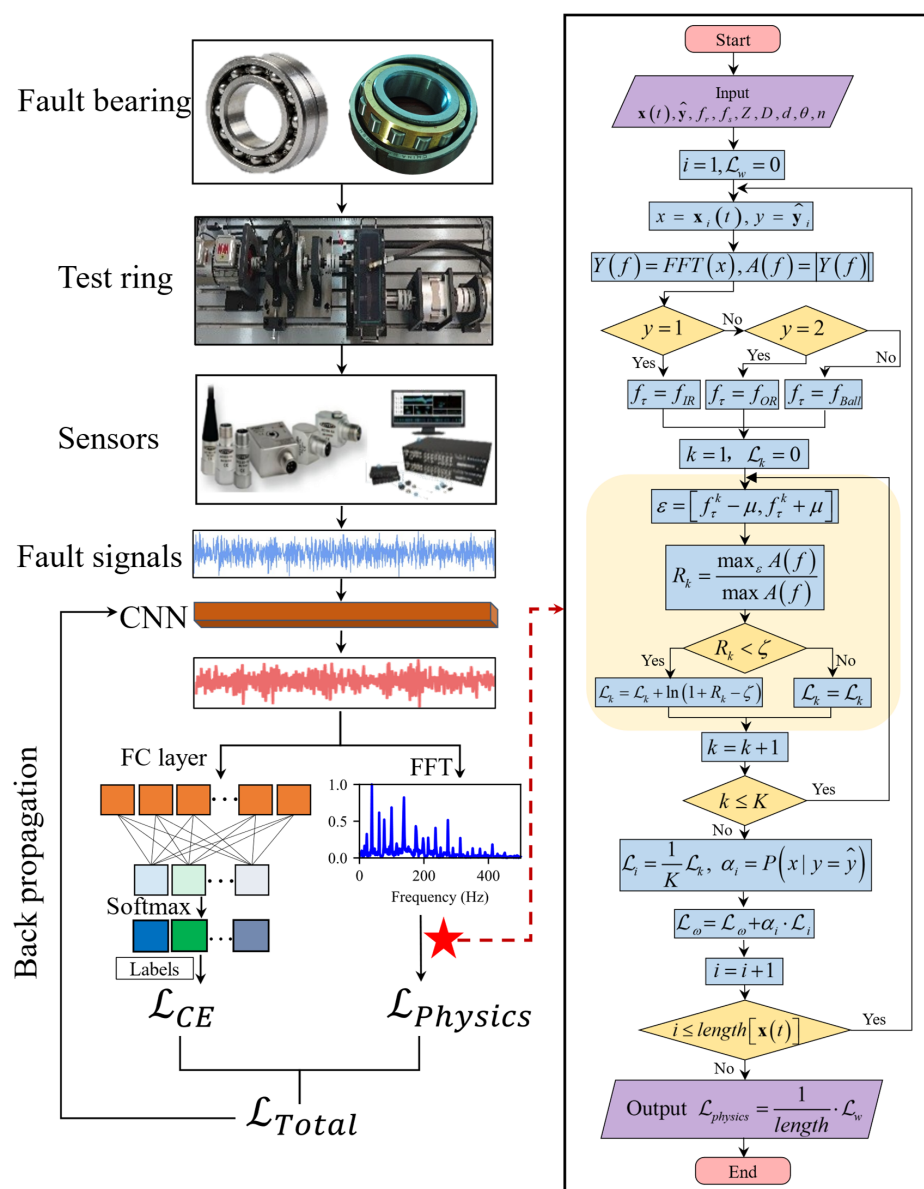


Figure 2. PIIDM model framework.

3.1. CNN Feature Extractor

In bearing fault diagnosis, the frequency-domain characteristics of vibration signals are crucial for fault identification. Traditional methods often rely on Fourier transform to extract spectral features. Its essence is the convolution operation of signal $x(t)$ with a set of orthogonal basis functions $e^{-j\omega t}$, as shown in Formula (2). It extracts the static features of the signal in the frequency domain through global convolution.

$$X(\omega) = \int_{-\infty}^{\infty} x(t) \cdot e^{-j\omega t} dt \quad (2)$$

Inspired by this, the convolution operation in convolutional neural networks has a high degree of structural homology with it, and both achieve feature extraction through convolution operations: In CNN, the calculation of the k -th feature map in the l layer can be expressed as

$$y_k^{(l)} = \sigma \left(\sum_i \omega_{k,i}^{(l)} * y_i^{(l-1)} + b_k^{(l)} \right) \quad (3)$$

where $*$ represents the convolution operation, $\omega_{k,i}^{(l)}$ is the learnable convolution kernel weight, and σ is the nonlinear activation function.

Unlike the fixed basis functions ($e^{-j\omega t}$) in the Fourier transform, the convolution kernel ω in CNN adaptively learns and updates through backpropagation and gradient descent algorithms during the training process:

$$\omega \leftarrow \omega - \eta \frac{\partial \mathcal{L}}{\partial \omega} \quad (4)$$

where $\mathcal{L}$ is the loss function, and η is the learning rate.

This mechanism enables the convolution kernel to evolve autonomously, focusing on key features such as resonance bands and modulation sidebands caused by faults, rather than relying on artificially designed fixed band divisions. Therefore, as shown in Figure 3, this study selects CNN as the feature extraction model, which not only inherits the mathematical unity of convolution operations in frequency-domain analysis but also possesses the ability to independently evolve the optimal feature expression from the data, providing a theoretical and structural basis for the construction of an end-to-end intelligent fault diagnosis system.

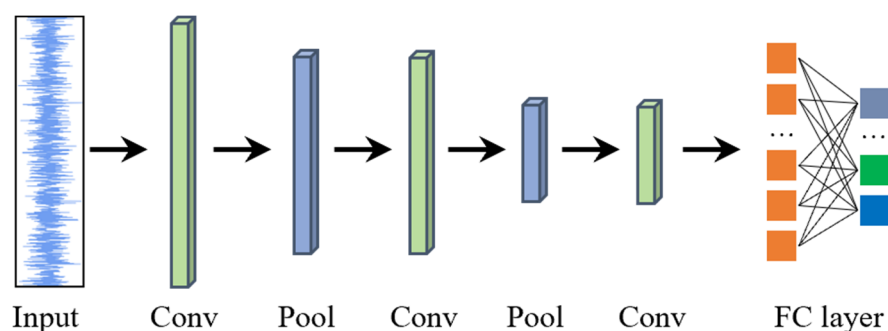


Figure 3. CNN feature extractor.

From the morphological changes of the signal before and after convolution in Figure 4, it can be intuitively seen that after convolution processing, the fault frequency-domain components in the signal are highlighted, indicating that the convolution operation in the convolutional neural network also has the ability to extract the spectral features of the signal. This provides an intuitive basis for interpreting the fault feature extraction mechanism of CNN in the frequency domain.

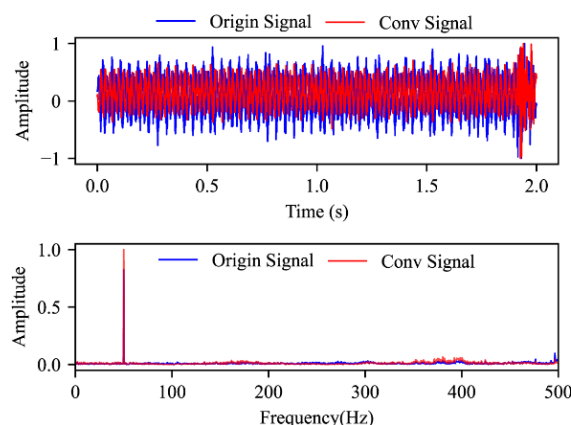


Figure 4. Comparison of the time-domain and frequency-domain graphs of the signal before and after convolution.

3.2. Design of Composite Loss Function Integrating Spectral Physical Constraints

Based on the cross-entropy classification loss function, PIIDM integrates a physical penalty loss function and constrains the model's learning direction in the form of physical regularization, making the model's decision-making process transparent and "white-boxed". Concretely, it can be introduced as follows.

(1) Through supervision and the learning mechanism driven by classification loss, the CNN model is guided to extract discriminative fault features from raw bearing vibration signals for accurate identification of multiple fault modes. Specifically, high-dimensional features extracted by the CNN network are mapped to a category space via the fully connected layer. The cross-entropy loss function (Formula (5)) quantitatively minimizes the discrepancy between the predicted distribution and ground-truth labels. This process compels the network to learn intrinsic distinctions among different fault types and improves the robustness of classification decisions:

$$\mathcal{L}_{ce} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \cdot \log(\hat{y}_{i,c}) \quad (5)$$

(2) Physical-penalty loss in bearing fault diagnosis field—the depth model because of its black-box nature is often unplanned and learning results cannot be explained and are constrained. The proposal of physical information neural networks provides a new idea for enhancing the credibility of model classification. In practical engineering applications, spectral analysis is a reliable method for diagnosing bearing faults. Therefore, in this paper, the spectral characteristics of faulty bearings are introduced as physical priors in model training to constrain the model learning process. The specific steps are as follows.

① Bearing spectral feature extraction.

As shown in Formulas (6) and (7), the Fourier transform is first performed on the signal sample features $x_i(t)$ output by the CNN to obtain the predicted spectrum features of the model. Subsequently, the amplitude spectrum of this predicted spectrum is extracted to characterize the frequency energy distribution:

$$X_i(f) = \int_{-\infty}^{\infty} x_i(t) \cdot e^{-j2\pi f t} dt \quad (6)$$

$$A_i(f) = |X_i(f)| = \sqrt{\text{Re}(X_i(f))^2 + \text{Im}(X_i(f))^2} \quad (7)$$

$\text{Re}(\cdot)$ is the real part of $\cdot$, $\text{Im}(\cdot)$ imaginary part of $\cdot$.

② Calculation of the peak ratio of fault characteristic frequencies.

The peak ratio of the fault characteristic frequency refers to the proportion of the amplitude of the fault frequency component (including the fundamental frequency and its harmonics) in the spectrum of the faulty bearing to the global maximum amplitude. This ratio can reflect the significance of the fault frequency component in the spectrum and is an important indicator for measuring whether the fault characteristics are clear. The calculation process is as follows.

First, as shown in Formula (8), a constraint neighborhood is set for the fault characteristic frequency to compensate for the characteristic frequency offset of the frequency resolution and the rotational speed fluctuation derivative:

$$\begin{cases} \varepsilon_{start} = f_{\tau}^k - \mu \\ \varepsilon_{end} = f_{\tau}^k + \mu \end{cases} \quad (8)$$

Among them, f_{τ}^k represents the k -th characteristic frequency of the τ -th bearing fault category (when $k = 1$, it represents the fundamental frequency of the fault feature; when $k > 1$, it represents the $(k - 1)$ -th harmonic component), and the window range ε restricts the size of the left and right neighborhoods of the fault characteristic frequency.

Second, the local maximum amplitude and the full-frequency maximum amplitude of the bearing fault characteristic frequency within this constraint neighborhood are calculated:

The maximum local amplitude is the following:

$$P_{i,k}^{local}(f_{\tau}^k) = \max_{\varepsilon \in [\varepsilon_{start}, \varepsilon_{end}]} A_i(f) \quad (9)$$

The maximum amplitude value across the full frequency band is the following:

$$P_{i,k}^{global}(f_{\tau}^k) = \max A_i(f) \quad (10)$$

Finally, the peak ratio of the k -th fault characteristic frequency of the i -th sample is obtained:

$$R_{i,k} = \frac{P_{i,k}^{local}}{P_{i,k}^{global}}, R_{i,k} \in [0, 1] \quad (11)$$

③ Calculation of physical penalty losses.

A physical threshold ζ is set. By comparing the ratio of this threshold to the peak frequency of the fault feature in the model's predicted spectrum, it can be determined whether the model effectively captures the physical patterns related to the fault. Based on this, the physical loss is calculated.

The single frequency penalty loss $\mathcal{L}_{i,k}$ is calculated as shown in Formula (12), representing the penalty loss of the k -th feature frequency of the i -th sample:

$$\mathcal{L}_{i,k} = \begin{cases} \ln(1 + \zeta - R_{i,k}), & R_{i,k} < \zeta \\ 0, & R_{i,k} \geq \zeta \end{cases} \quad (12)$$

The calculation of the penalty loss for a single sample is shown in Formula (13), which represents the average loss of all valid fault characteristic frequencies of a single sample and is used to characterize the degree of physical constraint violation of the sample. In the formula, K represents the number of effective fault characteristic frequencies:

$$\mathcal{L}_i = \frac{1}{|K|} \sum_{k \in K} \mathcal{L}_{i,k} \quad (13)$$

This paper introduces the confidence probability to weight the single-sample penalty loss (Formula (14)), making the loss positively correlated with the prediction credibility. The prediction confidence weighting mechanism enables the physical penalty loss to impose stronger penalties on high-confidence incorrect predictions. In the formula, $P(y = \hat{y}|x)$ represents the probability that the model predicts category $\hat{y}$:

$$\mathcal{L}_{\omega} = P(x|y = \hat{y}) \cdot \mathcal{L}_i \quad (14)$$

The final physical loss is shown in Formula (15). This physical penalty loss is based on the frequency-domain feature prior of the bearing fault, ensuring that the fault category

predicted by the model is consistent with the physical law of the frequency-domain energy distribution. In the formula, N represents the number of samples:

$$\mathcal{L}_{physics} = \frac{1}{N} \sum_{\omega=1}^N \mathcal{L}_{\omega} \quad (15)$$

Ultimately, the complete loss function of the proposed PIIDM is defined in Formula (16). By integrating cross-entropy classification loss and physical constraint loss, this work constructs a network framework that synergistically combines data-driven learning and physical fault mechanisms. The cross-entropy loss supervises the model to extract highly discriminative features from vibration signals, enabling accurate identification of various bearing faults, including inner-race faults, outer-race faults, and rolling element faults. Meanwhile, the physical penalty term, derived from prior bearing fault frequency knowledge, guides the network to capture frequency-domain feature patterns consistent with real fault physical mechanisms. This design transforms the traditional black-box model into an interpretable white-box paradigm, which substantially improves the physical consistency, interpretability, and reliability of the model diagnostic decisions:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{physics} \quad (16)$$

In Formula (16), λ_1 and λ_2 are the weights of the cross-entropy classification loss and the physical constraint loss, respectively; The larger the weight value, the higher the proportion of the corresponding loss term in the total loss, and the stronger the model training's tendency to fit this loss.

4. Test

After systematically expounding the theoretical framework and method design of PICNN in this paper, multi-dimensional empirical research is now conducted. All experiments were independently repeated six times. The average value of the indicators was adopted to reduce the interference caused by the randomness of training and ensure the reliability of the experimental results, thereby verifying the superiority and interpretability of PICNN in the intelligent fault diagnosis task.

4.1. Experiment Preparation

4.1.1. Dataset

(1) The Huazhong University of Science and Technology (HUST) bearing vibration dataset [27].

HUST datasets were collected under the conditions of 11 kinds of speed (ten kinds of constant speed). The vibration signals of bearings in different states (with a time-varying speed) were sampled. The load during the sampling process was set to a constant value of 20 Nm, and the sampling frequency was 25.6 KHz. This dataset contains eight bearing failure states and one health state. The fault conditions of bearings respectively included severe damage and minor damage to the inner ring, outer ring, rolling elements, and compound faults. The model of the bearing used was ER-16K. Its specific parameters are shown in Table 2. The left side of Figure 5 is the HUST dataset acquisition test ring, and the right side of Figure 5 is the schematic diagram of the faulty bearing.

Table 2. Parameters of ER-16K bearings.

Ball Diameter	Number of Balls	Pitch Diameter
7.94 mm	9	35.73 mm

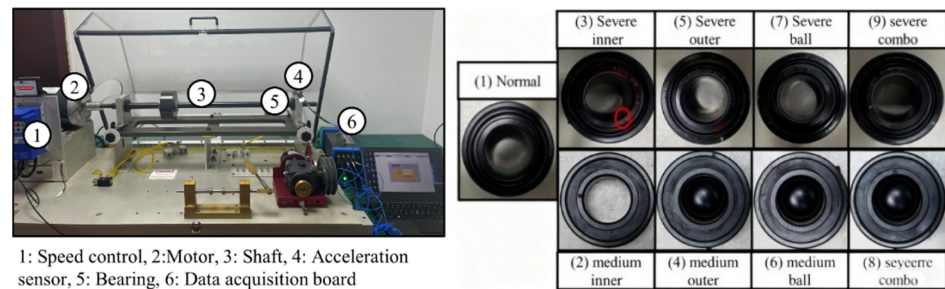


Figure 5. HUST dataset test ring (left), faulty bearing (right).

As shown in Table 3, the vibration signals of three states (inner ring, outer ring, and rolling elements) of the bearings collected under two rotational speed conditions (20 Hz and 30 Hz) in the HUST bearing dataset were selected as experimental data. The respective fault characteristic frequencies can all be obtained by Formula (1).

Table 3. HUST dataset.

Condition of Loading	Fault Type	Fault Character Frequency	Label
20 Hz	Inner-race fault	110.09 Hz	0
	Outer-race fault	70.72 Hz	1
	Ball fault	42.78 Hz	2
30 Hz	Inner-race fault	165.16 Hz	3
	Outer-race fault	106.08 Hz	4
	Ball fault	64.17 Hz	5

(2) The bearing fault dataset (BFD) was collected independently by our laboratory [28].

As shown in Figure 6, it is the test ring for signal collection. This test bench is composed of a platform base plate, a three-phase AC asynchronous motor, a gearbox fault simulation module, a rolling bearing fault simulation module, a control system, test pieces, etc. The acceleration sensor used for data acquisition is shown in Figure 7, with the model being HF7051-1. Figure 8 shows the HFSS signal acquisition system, which is responsible for collecting and analyzing the signals collected by this sensor.

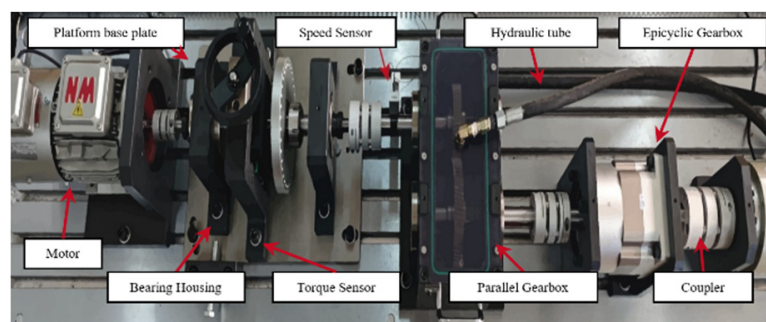


Figure 6. BFD dataset test ring (left), faulty bearing (right).

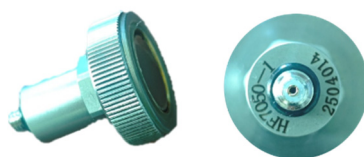


Figure 7. HF7051-1 sensor.

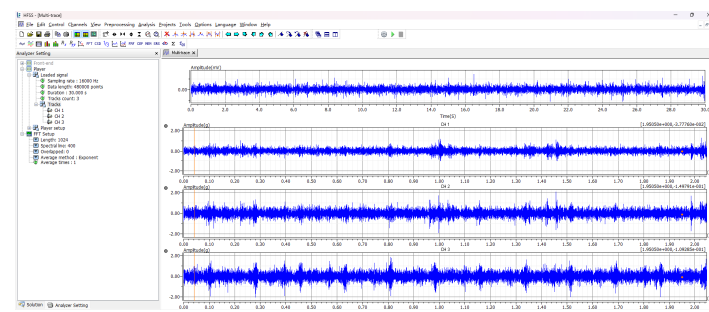


Figure 8. HFSS schematic diagram of signal acquisition software.

As shown in Table 4, the dataset used in this paper is the vibration signal of the bearing in the three states (inner ring, outer ring and rolling elements) collected by the rolling bearing fault simulation module under two rotational speed conditions (20 Hz and 30 Hz). This fault simulation module adopts a standardized rolling bearing, model NU205, and its specific parameters are shown in Table 5. The fault characteristic frequencies of each can be calculated by combining Formula (1). The sampling process load is set to a constant value of 20 Nm, and the sampling frequency is 16 KHz.

Table 4. BFD dataset.

Condition of Loading	Fault Type	Fault Character Frequency	Label
20 Hz	Inner-race fault	155.3 Hz	0
	Outer-race fault	104.65 Hz	1
	Ball fault	49.4 Hz	2
30 Hz	Inner-race fault	232.1 Hz	3
	Outer-race fault	156.9 Hz	4
	Ball fault	74.1 Hz	5

Table 5. Parameters of NU205 bearings.

Inner Diameter	Outer Diameter	Ball Diameter	Width	Number of Balls	Pitch Diameter
25 mm	52 mm	7.5 mm	15 mm	13	38.5 mm

4.1.2. Sample Selection and Training Parameters

Vibration signals contain rich fault characteristics. A reasonable signal sampling length is a basic condition to ensure that the sample truly reflects the original signal and that the sampling length refers to the time recording length required to analyze the lowest frequency in the signal. To ensure the integrity of the data cycle and avoid marginal effects, we selected samples with a length of 1024. The number of various types of fault samples is 100, among which the training set accounts for 60%, the validation set for 20%, and the test set for 20%.

Table 6 shows the PIIDM core training parameters. Details are as follows.

Table 6. PIIDM training parameters.

Optimizer	Number of Epochs	Learning Rate	Patience
Adam	200	0.01	20

(1) The optimizer and initial vector adopt the Adam optimizer, with initial vector set to 0.01. Adam combines the advantages of momentum gradient descent and adaptive learning rate to achieve a good balance between training stability and convergence efficiency. A

learning rate of 0.01 ensures rapid convergence while avoiding problems such as gradient oscillations caused by excessively high learning rates or slow convergence and becoming stuck in local optima due to excessively low learning rates.

(2) The model of the training set has a maximum of 200 rounds, and uses the patient value of the 20 early stop mechanism. The upper limit of 200 rounds ensures that the model has sufficient iterations to fully converge and avoid underfitting. The early stop mechanism monitors the performance of the validation set. If there is no significant improvement in performance indicators (such as accuracy and loss) for 20 consecutive rounds, the training will be automatically terminated. This strategy not only avoids ineffective iterations and saves training costs, but also effectively prevents overfitting and maintains the generalization ability of the model.

4.1.3. Model Evaluation Index

As shown in Table 7, based on the evaluation criteria of classification tasks, a four-dimensional evaluation system including Accuracy (Acc), Precision (Pre), Recall (Rec), and F1-score (F1) is constructed. The system quantifies the classification efficiency of the bearing fault diagnosis model.

Table 7. Model evaluation index.

Metric	Expression
Acc	$\frac{tp+tn}{tp+fp+tn+fn}$
Pre	$\frac{tp}{tp+fp}$
Rec	$\frac{tp}{tp+fn}$
F1	$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

In the model evaluation index system shown in Table 7, tp refers to the positive class samples correctly identified by the model, tn refers to the negative class samples correctly identified by the model, fp refers to samples where the model wrongly predicts negative class samples as positive class samples, fn refers to samples where the model wrongly predicts positive class samples as negative class samples.

4.2. Model Classification Performance Verification and Interpretability Analysis

PIIDM minimizes classification loss and physical constraint loss during backpropagation through the gradient descent algorithm, promoting the model to enhance feature discriminability while improving physical interpretability. In Formula (16), the loss weight ratio is used to regulate the relative interaction intensity between classification loss and physical constraint loss. This study is based on the HUST and BFD datasets and conducts multi-dimensional experiments by setting different loss weight ratios ($R = \lambda_2 / \lambda_1$) to systematically verify the comprehensive performance of the model in terms of classification performance and spectral interpretability.

In the spectral interpretability experiments, this paper adopts the peak to background ratio (PBR) to quantitatively characterize the prominence of spectral fault features. The calculation is defined in Formula (17):

$$\begin{cases} \text{PBR}_k = \frac{A_{f_k}}{A_{\text{bg},k}}, k = 1, 2, \dots, n \\ \text{PBR} = \frac{1}{n} \sum_{k=1}^n \text{PBR}_k \end{cases} \quad (17)$$

where k denotes the order of fault frequency, $k = 1$ represents the fundamental fault frequency, and $k \geq 2$ corresponds to harmonic frequencies. A_{f_k} is the spectral amplitude of

the k -th theoretical fault frequency, and $\bar{A}_{bg,k}$ is the average background spectral amplitude in the adjacent interval of the target frequency. Notably, this neighboring interval adopts a relative bandwidth of $[0.8f, 1.2f]$ centered at f , with the fault frequency peak excluded when computing the background amplitude. PBR_k quantifies the prominence of individual single-order fault features, while PBR integrates the results of fundamental components and all harmonic components to comprehensively measure the overall saliency of superimposed multi-order fault features. A higher PBR indicates stronger model focusing on inherent fault frequencies and more prominent regularization guidance of physical constraints. A lower PBR reflects weak constraint regulation, where fundamental and harmonic fault features cannot be effectively highlighted, resulting in insufficient physical interpretability and diagnosis reliability.

4.2.1. Experiments Based on the HUST Dataset

This paper selected the publicly available and representative HUST bearing fault dataset for experiments to verify the classification performance and interpretability of the proposed model. The experimental results are displayed as follows.

(1) The classification performance of Figure 9 shows the model under different weight ratios of each classification performance indicator. Figure 10 shows the characteristics of the model under different weight classification visual results. It can be seen that as the weight ratio increases (the proportion of physical loss increases), the classification indicators of the model show a trend of first rising and then falling. When the weight ratio is $R = 0$, the model has no physical constraints, and its classification performance is the worst. When R increases to 1, all indicators reach the optimum. If the R value is further increased, the classification performance will instead decline.

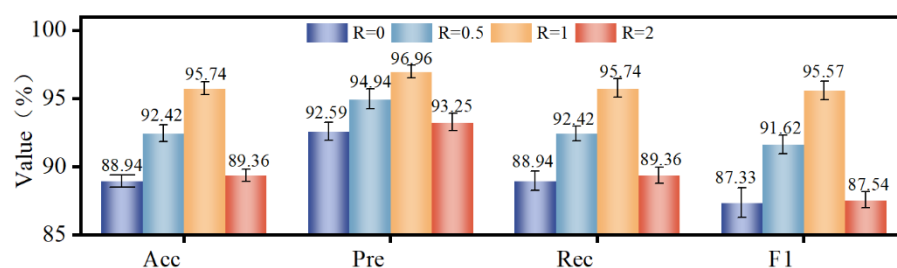


Figure 9. Classification performance of the PIIDM under different weight ratios (HUST dataset).

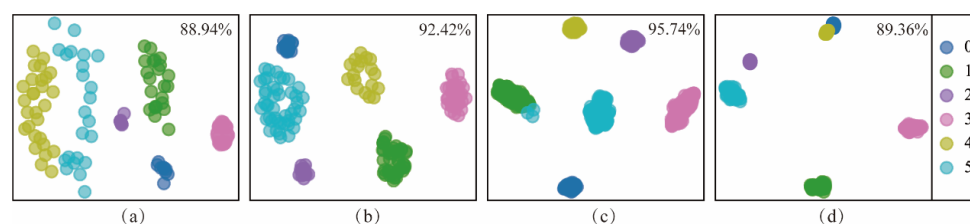


Figure 10. Visualization of model classification features under different weight ratio settings (HUST dataset). (a) $R = 0$. (b) $R = 0.5$. (c) $R = 1$. (d) $R = 2$.

Analysis of the cause, in the absence of the physical loss constraint model through data driven to extract fault features, only failed to introduce spectrum physical characteristics as an a priori restraint, giving characteristics of limited understanding. As the weight ratio R increases, the spectral physical features are gradually integrated into the optimization constraints of the model, guiding the model to learn the spectral features of bearings with clear physical meanings. This effectively enhances the discriminability of the features and thereby improves the classification accuracy. However, as R continues to increase, the weight of classification loss is too low, the role of the data-driven method is weakened,

and the model has difficulty in fully utilizing the statistical patterns in the data, ultimately leading to a decline in overall performance.

(2) The interpretability analysis.

Under the condition of rotating speed of 30 Hz, according to Formula (1) and the ER-16K bearing parameter listed in Table 2, the fundamental frequency of the inner ring fault is 165 Hz, that of the outer ring fault is 105 Hz, and that of the ball bearing fault is 64 Hz.

Table 8 presents the quantitative comparison results of PBR for various faults under different physical constraint intensities. As can be seen from the table, with the increase of the constraint weight coefficient R and the gradual strengthening of the physical constraint effect, the PBR values of the inner ring, outer ring, and rolling element faults all show a stable upward trend. This result indicates that the spectral physical constraints introduced in this paper can effectively highlight the fundamental frequency and harmonic characteristics of various faults, significantly enhancing the physical interpretability of the model's spectral representation.

Table 8. PBR comparison of different fault types under varying physical constraint strengths (HUST dataset).

	$R = 0$	$R = 0.5$	$R = 1$	$R = 2$
IF	2.84	5.37	8.62	8.92
OF	3.12	5.91	9.25	10.15
BF	2.56	4.82	7.74	8.31

Figures 11–13 respectively show the spectral comparison of the PIIDM model before and after aggregating fault features at this rotational speed with different weight settings.

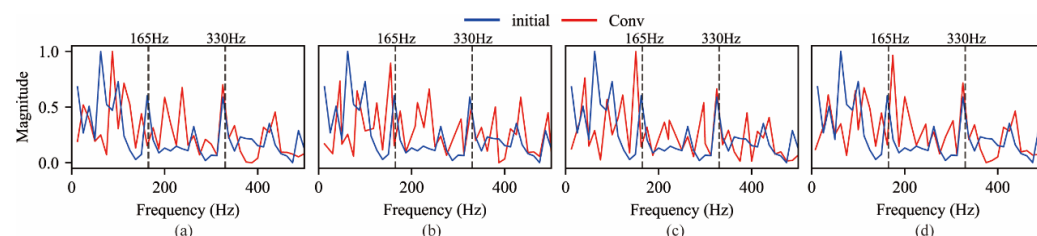


Figure 11. The spectrum features of the IF before and after aggregation (HUST). (a) $R = 0$. (b) $R = 0.5$. (c) $R = 1$. (d) $R = 2$.

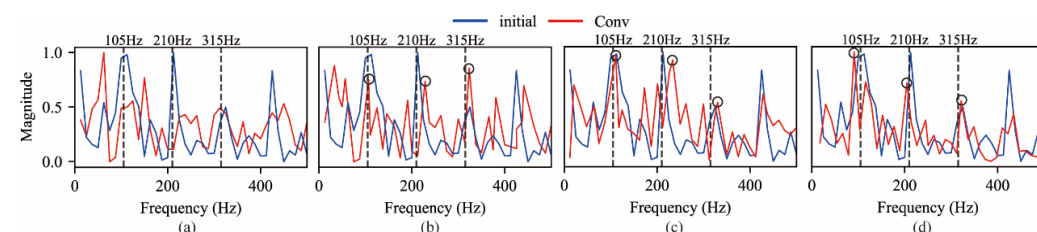


Figure 12. The spectrum features of the OF before and after aggregation (HUST). (a) $R = 0$. (b) $R = 0.5$. (c) $R = 1$. (d) $R = 2$.

On analysis of Figure 11, when $R = 0$ (without physical constraints), the spectrum obtained after model-based feature extraction fails to produce a prominent peak at the fundamental frequency (165 Hz), indicating that the model failed to effectively extract the fault features at this frequency. As R increases, the spectral peaks of the model at the fundamental frequency and its second-order harmonics significantly enhance, indicating that under the guidance of spectral physical constraints, the model successfully extracts the fundamental frequency and harmonic components of the inner loop fault.

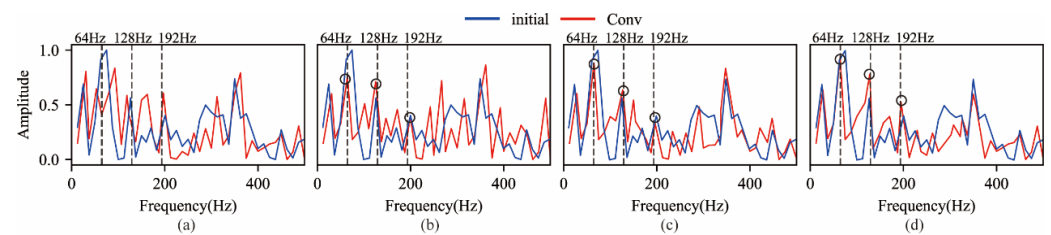


Figure 13. The spectrum features of the BF before and after aggregation (HUST). (a) $R = 0$. (b) $R = 0.5$. (c) $R = 1$. (d) $R = 2$.

As shown in Figure 12, when $R = 0$, the spectral peaks of the OF at the fundamental frequency (105 Hz), the second harmonic (210 Hz), and the third harmonic (315 Hz) are not obvious, and the characteristics are submerged in noise and other frequency components. When $R = 0.5$, the peak values of the fundamental frequency of the outer ring and its harmonic components have become significantly prominent. If R is further increased, the peak at the above frequency is more obvious than that of the signal before aggregation, indicating that the model effectively extracts the fundamental frequency and harmonic components of the outer loop fault.

On observing Figure 13, it can be found that when $R = 0$, the fault characteristics of the BF are almost completely submerged. When $R = 0.5$, the peak values of its fundamental frequency (64 Hz) and harmonic components (128 Hz, 192 Hz) start to show slightly. As R further increases, these peaks gradually intensify and reach their most significant level at $R = 2$. This result indicates that under the influence of spectral physical constraints, the model gradually extracts and intensifies the fundamental frequency and harmonic components of the ball bearing fault.

4.2.2. Experiments Based on the BFD Dataset

To verify the generalization performance of the model, in addition to using public datasets, this paper also selected the self-collected bearing fault dataset (BFD dataset) from the laboratory for the experiments, in order to further investigate the adaptability of the model under different data sources. The relevant experimental results are shown as follows.

(1) Classification performance.

Figure 14 shows the model under different weight ratio BFD datasets on the classification performance indicators. Figure 15 shows the corresponding weight on the model under the classification characteristics of the visual results. Consistent with the results on the HUST dataset, the classification performance of the model shows a trend of first improving and then decreasing with the increase of the weight ratio R . It reaches the optimal level at $R = 1$, while the performance indicators are at the lowest level, and the values are the same at $R = 0$ and $R = 2$.

Combined with Figure 15a,d, the characteristics of visualization, it can be seen that although the performance on classification performance indicators are the same, the characteristics of the distribution patterns show obvious differences. The reason lies in the fact that Figure 15a corresponds to the results without physical constraints and is completely dependent on data-driven methods, while Figure 15d shows the classification features obtained under the condition that the weights of physical constraints are dominant. Under the guidance of physical constraints, the model can learn more discriminative feature representations, making it easier to separate different categories in the feature space and thereby enhancing the model's separability.

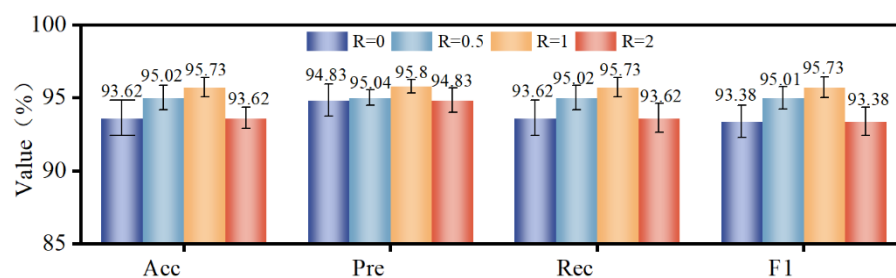


Figure 14. Classification performance of the PIIDM under different weight ratios (BFD dataset).

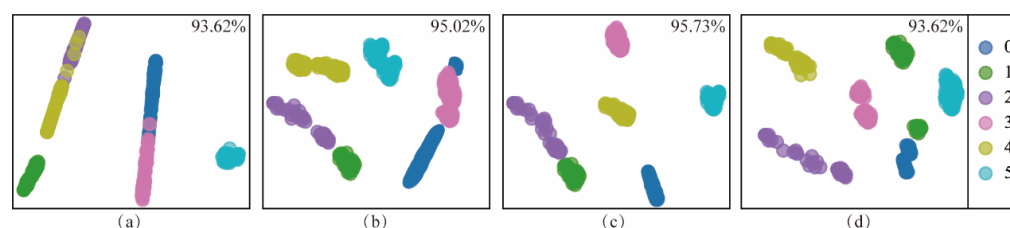


Figure 15. Visualization of model classification features under different weight ratio settings (BFD dataset). (a) $R = 0$. (b) $R = 0.5$. (c) $R = 1$. (d) $R = 2$.

(2) Interpretability analysis.

Table 9 shows the PBR comparison results of the BFD dataset under different physical constraint intensities, and the obtained patterns are consistent with those of the HUST dataset. With the increase of the weight coefficient R and the gradual strengthening of physical constraints, the PBR values of various faults generally show an upward trend. It is worth noting that when R is increased from 1 to 2, the growth rate of PBR significantly slows down, indicating that within this weight range, the enhancement effect of physical constraints on fault characteristics has tended to be saturated, and the constraint gain is limited.

Table 9. PBR comparison of different fault types under varying physical constraint strengths (BFD dataset).

	$R = 0$	$R = 0.5$	$R = 1$	$R = 2$
IF	3.11	4.52	7.15	7.92
OF	2.72	3.06	7.52	8.26
BF	1.33	2.16	4.87	5.09

Figures 16–18 respectively present the comparison of the spectral characteristics of the PIIDM model before and after aggregation in the BFD dataset under the condition of a speed of 20 Hz for the faulty inner and outer rings and under the condition of a speed of 30 Hz for the ball faults. Based on Formula (1) and the NU205 bearing parameters listed in Table 4, it can be found that under the above two working conditions, the fault fundamental frequency of the inner ring is 155.3 Hz, the fault fundamental frequency of the outer ring is 104.65 Hz, and the fault fundamental frequency of the ball is 74.1 Hz. The specific analysis of the experimental results is as follows.

As shown in Figure 16, for the IF of the BFD bearing, when $R = 0$, the model lacks spectral physical constraints and only relies on data-driven methods. At the fault fundamental frequency of 155.3 Hz, the spectral peak of the signal after feature extraction via PIIDM is less prominent and even attenuated compared with that before feature extraction. This indicates that it is difficult to obtain interpretable fault features if solely relying on data-driven methods. When $R \neq 0$, the peak value at the fundamental frequency of the inner loop fault is significantly higher than that before aggregation. Moreover, as the R

value increases, the peak values of the second harmonic (310.7 Hz) and the third harmonic (466.1 Hz) components also gradually increase. This indicates that under the guidance of spectral physical constraints, the model can effectively extract the fundamental frequency and harmonic components of the inner loop fault, enhancing the physical interpretability of the features.

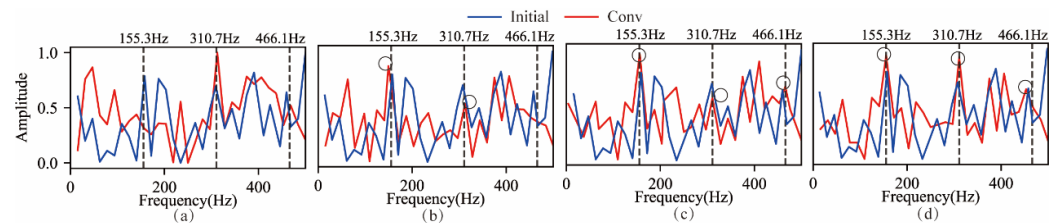


Figure 16. The spectrum features of the IF before and after aggregation (BFD). (a) $R = 0$. (b) $R = 0.5$. (c) $R = 1$. (d) $R = 2$.

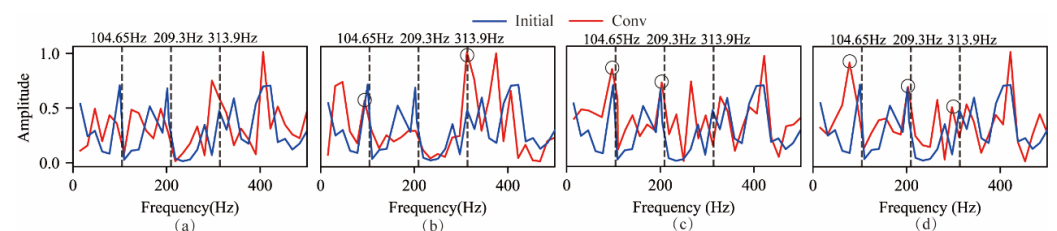


Figure 17. The spectrum features of the OF before and after aggregation (BFD). (a) $R = 0$. (b) $R = 0.5$. (c) $R = 1$. (d) $R = 2$.

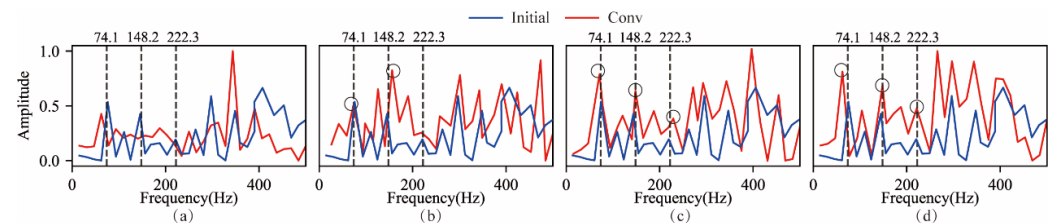


Figure 18. The spectrum features of the BF before and after aggregation (BFD). (a) $R = 0$. (b) $R = 0.5$. (c) $R = 1$. (d) $R = 2$.

Figure 17 shows the spectral aggregation effect of the OF under the BFD dataset and different weight ratio conditions. It can be seen that when $R = 0$ (without physical constraints and purely data-driven), at the outer-race fault fundamental frequency (104.65 Hz) as well as its second harmonic (209.3 Hz) and third harmonic (313.9 Hz), the spectral peaks exhibit no obvious increase after model-based aggregation; furthermore, the peak amplitude of the fundamental frequency is even attenuated. This indicates that relying solely on data-driven approaches fails to effectively extract fault features with physical significance. As R increases ($R = 0.5, 1, 2$), the peaks of the fundamental frequency and each harmonic component gradually intensify. Particularly at $R = 1$ and $R = 2$, the peaks of these frequency components become significantly prominent and are notably higher than the spectrum before aggregation. This indicates that the introduction of physical constraints effectively guides the model to extract the fundamental frequency and its harmonic components of the outer loop fault. The physical interpretability and discriminability of the features have been enhanced.

Figure 18 presents the spectral comparison results of BF before and after PIIDM aggregation on the BFD dataset and under different weight ratios. When the fault characteristics are relatively not obvious, the model can still effectively extract the fundamental frequency (74.1 Hz) as well as its second harmonic (148.2 Hz) and third harmonic (222.3 Hz) compo-

nents of the ball bearing fault under the guidance of spectral physical constraints. With the increase of R , the amplitude responses of the spectrum obtained after model-based feature extraction gradually increase at the fundamental frequency and each harmonic component. Especially under the conditions of $R = 1$ and $R = 2$, the peak at the characteristic frequency is significantly higher than that before aggregation. This result indicates that even in the face of weak fault features such as ball bearings, spectral physical constraints can still guide the model to extract frequency components with physical significance, thereby enhancing the discriminative power and interpretability of the features, demonstrating the importance of physical constraints in fault diagnosis.

Through the above experimental verification based on the public dataset (HUST) and the laboratory self-collected bearing fault data (BFD), the spectral physical constraints, as an important component of the loss function during the model training process, can guide the model to effectively learn spectral features with physical priors during the backpropagation process. This gives the model a natural basis for interpretability at the structural design level. Subsequent experiments showed that the spectral feature extraction results of the model were highly consistent with the spectral mechanism. This not only verifies the theoretical and experimental consistency of the model but also gradually realizes the transparent evolution of the model from black box to white box, providing physical interpretability support for fault diagnosis methods based on deep learning.

4.3. Physical Constraint Threshold Optimization Experiment

In the design of the loss function of PIIDM, physical constraint loss is introduced. This loss term is mainly constructed by the difference between the peak ratio at the characteristic frequency of the faulty bearing and the preset threshold ζ . The setting of the threshold needs to simultaneously satisfy the interpretability of the fault mechanism and the convergence stability of the learning process. A higher threshold tends to strengthen the physical constraints, which can enhance the interpretability of the features, but may increase the difficulty of optimization. Lower thresholds relax the constraints, which is conducive to model convergence, but may weaken the physical understanding coupling ability of fault features. To explore the influence of threshold settings on the convergence performance and classification performance of the model, this paper conducted comparative experimental analyses under different threshold settings based on the BFD dataset. By cross-validating the classification performance of the model and the convergence curve of the constraint loss, the balance between physical consistency and discriminative performance of the model can be obtained, thereby determining the optimal threshold configuration.

The results in Figure 19 indicate that the setting of the threshold (ζ) in the physical constraint loss affects the classification performance of the model. At the low threshold ($\zeta = 0.1$), the spectral physical constraints on the model are relatively weak, and it fails to fully learn the discriminative feature of the spectral characteristics of the bearing faults, resulting in low classification indicators for all items. As the threshold increases ($\zeta = 0.2$), stronger spectral physical constraints prompt the model to learn more interpretable discriminative features, and all its classification indicators improve. This proves that moderately strong constraints are the key to enhancing the discriminative performance of the model. However, if the threshold is set too high ($\zeta = 0.3/0.4$), overly strict physical constraints will dominate the optimization process, causing the model to fall into a local optimum, resulting in oscillation or difficulty in convergence during the training process, thereby damaging its stability and final generalization performance.

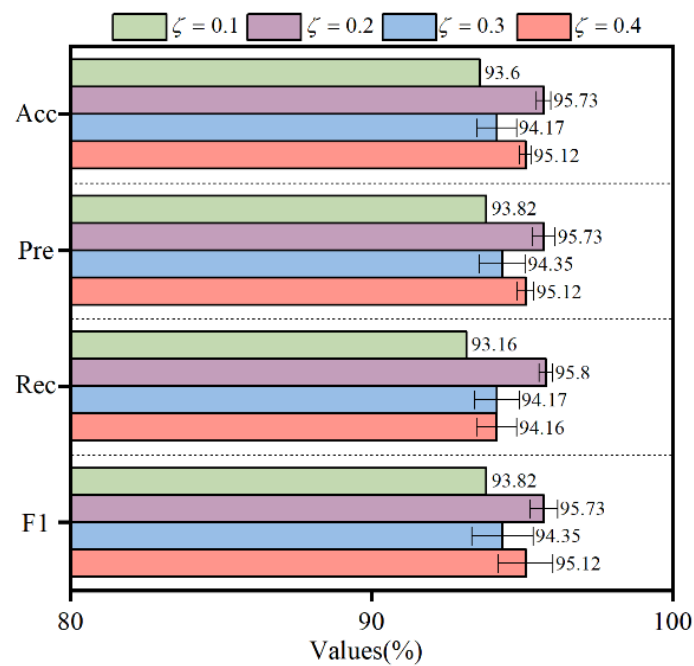


Figure 19. The influence of the thresholds ζ on the classification performance of PIIDM.

Figure 20 shows the influence of different thresholds (ζ) in physical constraint losses on the convergence behavior of the model training. Under the low threshold setting ($\zeta = 0.1$), the initial value of the physical loss term is relatively low and decreases gently, reflecting that the constraint conditions are relatively loose, and the model optimization process shows good stability. As the threshold increases, the initial value of physical loss significantly rises. Although it provides a more explicit physical gradient signal for the optimization process, due to overly strict constraints, obvious fluctuations occur during the training process, and the stability of the convergence trajectory correspondingly decreases. The experimental results show that a reasonable setting of the threshold is the key to achieving a balance between training stability and convergence quality.

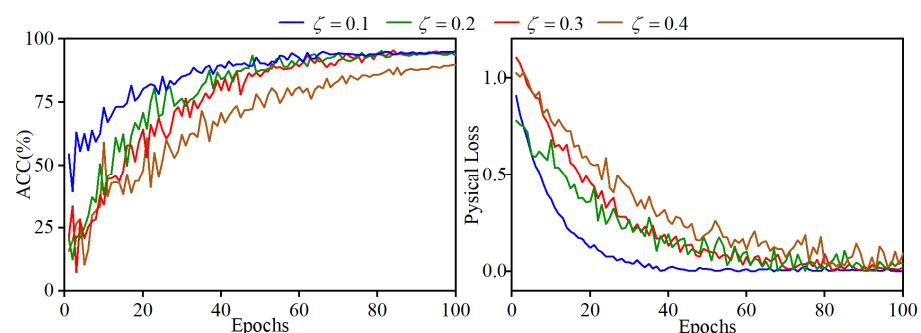


Figure 20. The influence of thresholds ζ on the convergence of model training.

5. Conclusions and Outlook

5.1. Conclusions

This paper proposes an interpretable deep learning framework integrated with spectral physical constraints to address the trust barriers caused by the “black-box” nature of deep learning models in bearing fault diagnosis. The method systematically constructs an interpretability mechanism encompassing both ante hoc and post hoc interpretability.

(1) Ante hoc interpretability: At the model design level, spectral characteristics of bearing faults are encoded as structural priors in the form of physical constraints, guiding

the network to explicitly learn frequency-domain representations closely related to fault mechanisms. This design achieves intrinsic integration of domain knowledge and data-driven learning, endowing the model with interpretable fault perception capabilities during the training phase.

(2) Post hoc interpretability: At the model inference level, experimental validation using public datasets and real industrial measurement data demonstrate that the model can extract spectral features strongly correlated with different fault states. Through feature visualization and decision traceability analysis, an interpretable mapping relationship between model outputs and physical fault characteristics is confirmed, thereby verifying the model's credibility and traceability from the result perspective.

The interpretable diagnostic framework developed in this study not only enhances diagnostic accuracy but also improves the model's reliability and acceptability. It provides a solution that combines high precision and high interpretability for intelligent fault diagnosis in industrial applications.

Comprehensive hyperparameter configurations, the detailed architecture of the proposed model, and the corresponding pseudo-code are provided in Appendix A for reproducibility.

5.2. Outlook

The proposed PICNN improves the physical rationality and interpretability of bearing fault diagnosis by embedding spectrum-based physical constraints. However, this method has clear application boundaries and may suffer performance degradation under complex industrial operating conditions. The detailed limitations and corresponding improvement directions are presented as follows.

(1) Limited working condition adaptability of spectrum constraints.

The spectrum constraint constructed based on Fourier transform relies on signal stationarity and a minimum signal-to-noise ratio, making it difficult to adapt to non-stationary conditions such as variable loads and speed fluctuations. In scenarios with incipient weak faults, impulsive noise interference, and overlapping fault spectra, fault features are easily aliased and submerged, leading to constraint failure and decreased diagnosis accuracy. In future work, time–frequency analysis methods, including wavelet transform and short-time Fourier transform, will be introduced to break the stationary signal limitation and enhance feature adaptability under complex working conditions.

(2) Negative suppression effect of fixed physical constraints.

Although fixed spectrum prior constraints can regularize model learning under ideal conditions, the rigid constraint mechanism limits the adaptive feature mining capability of CNNs. In complex interference scenarios, the model overly relies on theoretical fault frequencies and fails to capture latent fault features, resulting in reduced generalization performance. Future studies will develop a dynamic adaptive spectrum constraint mechanism to replace the fixed threshold strategy and avoid negative constraint effects in complex scenarios.

(3) Restricted engineering generalization capability.

The model is trained and validated using laboratory and public datasets with stable operating conditions and single interference sources, which differ greatly from real industrial signals characterized by strong noise and coupled disturbances. This limits the model's adaptability to complex real-world faults. In the future, real industrial field data will be adopted for iterative model optimization, and anti-interference and weak fault feature extraction modules will be integrated to enhance the engineering practicability of the proposed method.

Author Contributions: Writing—Original Draft, Y.H. and Z.Y.; Writing—Review & Editing, F.S.; Visualization, Q.X.; Data curation, X.X.; Software, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Jiangsu Province, China (Grant No. KYGYGKQTSZ25001). General Equipment Research Program (Grant No. KY-GYJKZBXM24024).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is unavailable due to privacy or ethical restrictions.

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

Nomenclature

f_{IR}	Frequency of faults in the inner ring of the bearing
f_{OR}	Frequency of faults in the outer ring of the bearing
f_{Ball}	Frequency of bearing ball failure
Z	Number of balls
d	Ball diameter
D	Bearing pitch diameter
α	Contact angle
f_r	Rotational frequency
$x(t)$	Raw vibration signal
ω	Learnable parameters of convolution kernel
η	Learning rate
f_{τ}^k	The k -th characteristic frequency of bearing fault category τ
$p_{i,k}^{local}$	Local maximum amplitude of fault characteristic frequency within the constrained neighborhood
$p_{i,k}^{global}$	Global maximum amplitude of fault characteristic frequency within the constrained neighborhood
ζ	Physical threshold
$\mathcal{L}_{ce}$	Classification loss
$\mathcal{L}_{i,k}$	Single-frequency penalty loss
$\mathcal{L}_i$	Single-sample penalty loss
$\mathcal{L}_{physics}$	Physical penalty loss
$\mathcal{L}_{total}$	Total loss
λ_1	Weight of classification loss
λ_2	Weight of physical penalty loss
R	Loss weight ratio
ACC	Accuracy
Pre	Precision ratio
Rec	Recall rate
F1	F1 Score
PBR	Peak to background ratio

Appendix A

Appendix A.1. Multi-Scale CNN Network Architecture

Table A1 illustrates the complete architecture of multi-scale CNN. Parallel multi-scale convolutional branches are adopted, where kernels of different sizes capture short-, medium- and long-period fault impulses from vibration signals. The output features of each branch are concatenated along the channel dimension. After feature dimensionality

reduction via global average pooling, the network outputs category-aware features for fault classification.

Table A1. Multi-scale CNN network architecture.

Network Layer	Kernel Size	Stride	Padding	Output Channels
Input layer	—	—	—	1
Conv1	16	2	7	16
MaxPool1	8	2	3	16
Multi-scale Conv-branch1	8	1	3	16
Multi-scale Conv-branch2	16	1	7	16
Multi-scale Conv-branch3	32	1	15	16
Concat	—	—	—	48
FC layer	—	—	—	64
Output layer	—	—	—	Number of categories

Appendix A.2. Experimental Setup and Computational Overhead

All experiments in this paper were implemented based on the PyTorch 2.2.1 framework and conducted on a desktop workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM). The consistent hardware and software environment ensured the stability and reproducibility of all comparative experiments and ablation experiments.

To further illustrate the computational efficiency of the proposed PIIDM, we quantitatively analyzed its training cost and the extra overhead introduced by the FFT operation in the physics-informed constraint loop. Based on the experimental statistics of the ordinary CNN backbone model, the total training time for the full training process was approximately 24.6 s, with an average single-epoch training time of 0.8 s (The above values are the average of multiple experiments). Compared with the baseline ordinary CNN model without physical constraints, the embedded FFT calculation for post hoc spectrum construction and physics-informed loss solution brings negligible extra computational overhead. The frequency-domain transformation operation runs efficiently on the GPU, and the overall training efficiency of the model is not significantly affected.

Appendix A.3. Model Pseudo-Code

The complete pseudocode of the proposed PIIDM is presented below to illustrate the Algorithm A1 logic and ensure experimental reproducibility. The algorithm mainly includes three core procedures: forward propagation of the multi-scale convolutional network, post hoc interpretable spectrum calculation, and physics-informed loss $\mathcal{L}_{Physics}$ optimization.

Specifically, the interpretable spectrum is constructed by pooling the output features of the multi-scale feature concatenation layer along the channel dimension and subsequently applying Fast Fourier Transform (FFT), which is utilized for post hoc interpretability analysis. The computation of $\mathcal{L}_{Physics}$ is strictly implemented following the algorithm flowchart in this paper. The hyperparameter λ_1, λ_2 controls the weight of the physics-informed loss, while ζ denotes the threshold of the spectrum peak ratio. These two parameters are crucial for regulating the physical constraint performance and interpretability of the model.

Algorithm A1 Overall procedure of the PIIDM

Require: Vibration signal: $x(t)$, label: $\hat{y}$, fault characteristic frequencies: f_{IR}, f_{OR}, f_{Ball} , hyper-parameters: $\lambda_1, \lambda_2, \zeta$

Ensure: Prediction, aggregated spectrum: $A_{agg}(f)$, total loss: $\mathcal{L}_{total}$

- 1: $F_{mid} \leftarrow \text{CNN}[x(t)]$
- 2: $\text{logits} \leftarrow \text{FC}(F_{mid})$
- 3: $\text{prediction} \leftarrow \text{ArgMax}(\text{logits})$
- 4: **▷Post hoc interpretability: aggregated spectrum**
- 5: $f_{ext}[t] \leftarrow \frac{1}{C} \sum_c F_{mid}[c, t]$
- 6: $A_{ext}(f) \leftarrow |\text{FFT}(f_{agg}[t])|$
- 7: **▷Physics-informed loss calculation**
- 8: $\mathcal{L}_w \leftarrow 0$
- 9: **for** each sample i **do**
- 10: $Y(f) \leftarrow \text{FFT}[x_i(t)], A(f) \leftarrow |Y(f)|$
- 11: Select f_τ according to label $\hat{y}_i$
- 12: $\mathcal{L}_k \leftarrow 0$
- 13: **for** $k = 1$ to K **do**
- 14: $\varepsilon \leftarrow [f_\tau^k - \mu, f_\tau^k + \mu]$
- 15: $R_k \leftarrow \frac{\max_{f \in \varepsilon} A(f)}{\max A(f)}$
- 16: **if** $R_k < \zeta$ **then**
- 17: $\mathcal{L}_k \leftarrow \mathcal{L}_k + \ln(1 + R_k - \zeta)$
- 18: **end if**
- 19: **end for**
- 20: $\mathcal{L}_i \leftarrow \frac{1}{K} \mathcal{L}_k$
- 21: $\mathcal{L}_w \leftarrow \mathcal{L}_w + \alpha_i \mathcal{L}_i$
- 22: **end for**
- 23: $\mathcal{L}_{Physics} \leftarrow \mathcal{L}_w / \text{length}(x)$
- 24: **▷Total loss for back-propagation**
- 25: $\mathcal{L}_{CE} \leftarrow \text{CrossEntropy}(\text{logits}, \hat{y})$
- 26: $\mathcal{L}_{total} \leftarrow \lambda_1 \mathcal{L}_{Physics} + \lambda_2 \mathcal{L}_{CE}$
- 27: **return** prediction, $A_{agg}(f)$, $\mathcal{L}_{total}$

References

1. Lee, J.; Wu, F.; Zhao, W.; Ghaffari, M.; Liao, L.; Siegel, D. Prognostics and Health Management Design for Rotary Machinery Systems—Reviews, Methodology and Applications. *Mech. Syst. Signal Process.* **2014**, *42*, 314–334. [\[CrossRef\]](#)
2. Xu, Z.; Zhao, K.; Wang, J.; Bashir, M. Physics-Informed Probabilistic Deep Network with Interpretable Mechanism for Trustworthy Mechanical Fault Diagnosis. *Adv. Eng. Inform.* **2024**, *62*, 102806. [\[CrossRef\]](#)
3. Sadoughi, M.; Hu, C. Physics-Based Convolutional Neural Network for Fault Diagnosis of Rolling Element Bearings. *IEEE Sens. J.* **2019**, *19*, 4181–4192. [\[CrossRef\]](#)
4. Adlen, K.; Ridha, K. Recurrent Neural Network Optimization for Wind Turbine Condition Prognosis. *Diagnostyka* **2022**, *23*, 2022301. [\[CrossRef\]](#)
5. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep Learning and Its Applications to Machine Health Monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [\[CrossRef\]](#)
6. Jin, Y.; Hou, L.; Chen, Y. A Time Series Transformer Based Method for the Rotating Machinery Fault Diagnosis. *Neurocomputing* **2022**, *494*, 379–395. [\[CrossRef\]](#)
7. Lee, H.; Jeong, H.; Koo, G.; Ban, J.; Kim, S.W. Attention Recurrent Neural Network-Based Severity Estimation Method for Interturn Short-Circuit Fault in Permanent Magnet Synchronous Machines. *IEEE Trans. Ind. Electron.* **2021**, *68*, 3445–3453. [\[CrossRef\]](#)
8. Li, T.; Zhao, Z.; Sun, C.; Yan, R.; Chen, X. Domain Adversarial Graph Convolutional Network for Fault Diagnosis under Variable Working Conditions. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–10. [\[CrossRef\]](#)

9. Xiang, S.; Qin, Y.; Zhu, C.; Wang, Y.; Chen, H. Long Short-Term Memory Neural Network with Weight Amplification and Its Application into Gear Remaining Useful Life Prediction. *Eng. Appl. Artif. Intell.* **2020**, *91*, 103587. [\[CrossRef\]](#)
10. Mansouri, M.; Dhibi, K.; Hajji, M.; Bouzara, K.; Nounou, H.; Nounou, M. Interval-Valued Reduced RNN for Fault Detection and Diagnosis for Wind Energy Conversion Systems. *IEEE Sens. J.* **2022**, *22*, 13581–13588. [\[CrossRef\]](#)
11. Niu, G.; Liu, E.; Wang, X.; Ziehl, P.; Zhang, B. Enhanced Discriminate Feature Learning Deep Residual CNN for Multitask Bearing Fault Diagnosis with Information Fusion. *IEEE Trans. Ind. Inform.* **2023**, *19*, 762–770. [\[CrossRef\]](#)
12. Cao, H.; Shao, H.; Zhong, X.; Deng, Q.; Yang, X.; Xuan, J. Unsupervised Domain-Share CNN for Machine Fault Transfer Diagnosis from Steady Speeds to Time-Varying Speeds. *J. Manuf. Syst.* **2022**, *62*, 186–198. [\[CrossRef\]](#)
13. Lu, F.; Tong, Q.; Jiang, X.; Du, S.; Xu, J.; Huo, J.; Zhang, Z. Envelope Spectrum Neural Network with Adaptive Domain Weight Harmonization for Intelligent Bearing Fault Diagnosis under Cross-Machine Scenarios. *Adv. Eng. Inform.* **2024**, *62*, 102787. [\[CrossRef\]](#)
14. Li, R.; Xia, T.; Luo, F.; Jiang, Y.; Chen, Z.; Xi, L. Hybrid Physics-Embedded Recurrent Neural Networks for Fault Diagnosis under Time-Varying Conditions Based on Multivariate Proprioceptive Signals. *Adv. Eng. Inform.* **2024**, *62*, 102851. [\[CrossRef\]](#)
15. Xu, J.; Ke, H.; Jiang, Z.; Mo, S.; Chen, Z.; Gui, W. OHCA-GCN: A Novel Graph Convolutional Network-Based Fault Diagnosis Method for Complex Systems via Supervised Graph Construction and Optimization. *Adv. Eng. Inform.* **2024**, *61*, 102548. [\[CrossRef\]](#)
16. Li, Q.; Liu, Y.; Sun, S.; Qin, Z.; Chu, F. Deep Expert Network: A Unified Method toward Knowledge-Informed Fault Diagnosis via Fully Interpretable Neuro-Symbolic AI. *J. Manuf. Syst.* **2024**, *77*, 652–661. [\[CrossRef\]](#)
17. Li, Q.; Chen, B.; Chen, Q.; Li, X.; Qin, Z.; Chu, F. HSE: A Plug-and-Play Module for Unified Fault Diagnosis Foundation Models. *Inf. Fusion* **2025**, *123*, 103277. [\[CrossRef\]](#)
18. Li, X.; Li, S.; Zhang, G.; Xie, Y.; Wang, T.; Chu, F. A Novel Interpretable Dynamic Weighted Domain Adaptation Network for Cross-Domain Fault Diagnosis of Bearings under Time-Varying Speeds. *Eng. Appl. Artif. Intell.* **2026**, *170*, 114173. [\[CrossRef\]](#)
19. Li, Q.; Qin, L.; Xu, H.; Lin, Q.; Qin, Z.; Chu, F. Transparent Information Fusion Network: An Explainable Network for Multi-Source Bearing Fault Diagnosis via Self-Organized Neural-Symbolic Nodes. *Adv. Eng. Inform.* **2025**, *65*, 103156. [\[CrossRef\]](#)
20. Li, Q.; Li, H.; Hu, W.; Sun, S.; Qin, Z.; Chu, F. Transparent Operator Network: A Fully Interpretable Network Incorporating Learnable Wavelet Operator for Intelligent Fault Diagnosis. *IEEE Trans. Ind. Inform.* **2024**, *20*, 8628–8638. [\[CrossRef\]](#)
21. Pang, P.; Tang, J.; Luo, J.; Chen, M.; Yuan, H.; Jiang, L. An Explainable and Lightweight Improved 1-D CNN Model for Vibration Signals of Rotating Machinery. *IEEE Sens. J.* **2024**, *24*, 6976–6997. [\[CrossRef\]](#)
22. Lao, Z.; Chen, G.; Zhang, Y.; Lu, P.; Jin, Z. An Interpretable Causal Invariant Graph Neural Network for Unseen Domain Gear Fault Diagnosis. *Eng. Appl. Artif. Intell.* **2026**, *170*, 114227. [\[CrossRef\]](#)
23. Wang, D.; Chen, Y.; Shen, C.; Zhong, J.; Peng, Z.; Li, C. Fully Interpretable Neural Network for Locating Resonance Frequency Bands for Machine Condition Monitoring. *Mech. Syst. Signal Process.* **2022**, *168*, 108673. [\[CrossRef\]](#)
24. Zhen, Z.; Shixi, Y.; Jun, H.; Wanchun, Z.; Yanxu, L. A Priori Fusion Fully Interpretable Transfer Learning Method for Rotating Machinery Fault Diagnosis. *Results Eng.* **2026**, *30*, 110685. [\[CrossRef\]](#)
25. Zhang, Y.; Shen, C.; Ye, X.; Shi, J.; Huang, W.; Zhu, Z. A New Physics-Guided Interpretable Continual Learning Method for Incremental Fault Diagnosis of Rotating Machinery. *J. Ind. Inf. Integr.* **2026**, *51*, 101118. [\[CrossRef\]](#)
26. Wu, H.; Ma, X.; Wen, C. Multilevel Fine Fault Diagnosis Method for Motors Based on Feature Extraction of Fractional Fourier Transform. *Sensors* **2022**, *22*, 1310. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Zhao, C.; Zio, E.; Shen, W. Domain Generalization for Cross-Domain Fault Diagnosis: An Application-Oriented Perspective and a Benchmark Study. *Reliab. Eng. Syst. Saf.* **2024**, *245*, 109964. [\[CrossRef\]](#)
28. Hou, Y.; Tang, J.; Zhang, J.; Pang, P.; Zhang, H.; Li, X. Mixed Fault Dataset of Bearings and Gearboxes under Variable Operating Conditions. *Data Brief* **2025**, *63*, 112187. [\[CrossRef\]](#) [\[PubMed\]](#)

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