

Physics-Informed Modulated Fusion ResNet for Noise-Robust Acoustic Fault Diagnosis of Railway Wheelset Bearings

Peihai Li

China University of Mining and Technology

Zheng Li

China University of Mining and Technology

Dengyang Cao

China University of Mining and Technology

Bo Zhang

Gu1angzhou GRG Metrology & Test Group (Shanghai) Technology Co., Ltd

Chaoying Li

Gu1angzhou GRG Metrology & Test Group (Shanghai) Technology Co., Ltd

Houguang Liu

liuhg@cumt.edu.cn

China University of Mining and Technology

Research Article

Keywords: Railway wheelset bearings, Acoustic fault diagnosis, Auditory-inspired acoustic representation, Physics-informed learning, Noise robustness

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11348519/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

**Physics-Informed Modulated Fusion ResNet for Noise-Robust Acoustic Fault
Diagnosis of Railway Wheelset Bearings**

Peihai Li^{a,b,†}, Zheng Li^{a,†}, Dengyang Cao^{a,†}, Bo Zhang^{c,d}, Chaoying Li^c, Houguang Liu^{1,*}

^a*School of Mechanical and Electrical Engineering, China University of Mining and Technology, Xuzhou
221116, China;*

^b*Chongqing Changzheng Heavy Industry Co., Ltd. Chongqing 400083, China;*

^c*Gulangzhou GRG Metrology & Test Group (Shanghai) Technology Co., Ltd. Shanghai 201899, China;*

^d*Guangzhou GRG Metrology & Test Co., Ltd, Guangzhou 510600, China;*

[†]*Peihai Li, Zheng Li, Dengyang Cao should be considered co-first authors.*

*Corresponding author: Houguang Liu

Email: liuhg@cumt.edu.cn

Address: School of Mechanical and Electrical Engineering, China University of Mining and
Technology, NO.1 Daxue Road, Xuzhou 221116, Jiangsu Province, China.

Phone: 086-18761435299

Abstract: Reliable condition monitoring of railway wheelset bearings is essential for train safety, yet acoustic fault diagnosis in practical wayside environments remains challenging due to strong background noise. Although acoustic monitoring provides a flexible and non-contact solution, fault-related signatures can be easily masked under low signal-to-noise ratio (SNR) conditions. Moreover, purely data-driven models lack physical guidance, which limits their noise robustness in complex environments. To address these issues, this paper proposes a Physics-Informed Modulated Fusion ResNet (PIMF-ResNet) that combines auditory-inspired acoustic representation with physics-informed learning. Acoustic feature maps are first constructed to capture fault-related information from noisy signals. Meanwhile, 15-dimensional envelope-spectrum peak-amplitude features are extracted around characteristic frequencies predicted from bearing geometry and rotational speed and their first three harmonic orders. A physics attention mechanism adaptively weights these priors and maps them to channel-wise coefficients for ResNet34 feature modulation. A physics-consistency loss further encourages class-conditioned attention to the corresponding fault-frequency families. Experiments on laboratory railway wheelset bearing acoustic data collected at 250 r/min with added ESC-50 environmental noise show that PIMF-ResNet achieves accuracies of 98.83% on the original recordings and 79.24% at -15 dB under the reported validation protocol. At -15 dB, accuracy exceeds that of the auditory-only baseline by 6.64 percentage points. Comparisons with representative CNN and Transformer models demonstrate improved diagnostic performance and noise robustness under the evaluated added-noise conditions.

Keywords: Railway wheelset bearings; Acoustic fault diagnosis; Auditory-inspired acoustic representation; Physics-informed learning; Noise robustness

1. Introduction

Railway wheelset bearings are critical components of the running gear, and their condition directly affects train safety and maintenance costs. Localized bearing defects generate periodic impacts and modulation features in vibration and acoustic signals, providing an important basis for fault diagnosis. Compared with vibration monitoring, acoustic monitoring is non-contact, flexible in sensor deployment, and well suited to wayside inspection [1,2]. Existing studies have explored several techniques for railway bearing acoustic diagnosis. He et al. enhanced weak fault information through resampling and stochastic resonance [3]. Zhang et al. addressed Doppler distortion and multi-source coupling using source separation and correction methods [4], and later extracted fault components from wayside acoustic signals through adaptive sparse decomposition [5]. Other studies employed enhanced spline-kernel Chirplet transform and parametric Doppler correction to improve feature extraction under moving-source conditions [6,7]. These methods demonstrate the potential of acoustic signals for railway bearing monitoring. However, strong background noise, propagation attenuation, harmonic interference, and moving-source distortion remain major challenges in practical environments.

With the development of intelligent diagnosis, research has gradually shifted from handcrafted features to deep representation learning. Zhang et al. constructed graph representations of bearing acoustic signals and used a deep graph convolutional network for fault identification [8]. Hasan et al. combined acoustic spectral imaging with transfer learning to improve diagnosis under variable-speed conditions [9]. Auditory perception mechanisms have also been introduced to improve feature robustness. For example, Yang et al. fused the Mel spectrum with auditory neurotransmitter release rate to construct an auditory perception map, achieving improved recognition under low signal-to-noise ratio (SNR) conditions [10,11]. Zhao et al. systematically reviewed deep learning applications in machine health monitoring and pointed out that deep models can automatically learn fault features from large-scale monitoring data [12]. Guo et al. developed a hierarchical adaptive convolutional neural network for bearing fault diagnosis [13]. Zhang et al. proposed a deep convolutional network training strategy for noisy environments and varying load conditions, improving the robustness of end-to-end models [14]. These studies indicate that perceptual representations can help capture weak acoustic patterns in noisy environments.

Nevertheless, perceptual features alone cannot fully address the problem. Bearing faults follow explicit kinematic mechanisms determined by bearing geometry, contact angle, and rotational speed. Defects in the inner race, outer race, rolling element, and cage produce characteristic fault frequencies and their harmonics. Without such physical constraints, deep models may focus on residual noise, harmonic interference, or unrelated operating components. This problem becomes more serious when the noise distribution changes. Purely data-driven models may then learn spurious patterns instead of fault-related information. Their decision basis is also difficult to relate to known bearing mechanisms, limiting interpretability in safety-critical applications.

Physical knowledge provides a possible solution to this limitation. Xie et al. combined fault-frequency priors with a self-learning branch to improve the interpretability of wind turbine bearing diagnosis [15]. Gao et al. developed MPINet with physics-informed blocks for few-shot fault diagnosis [16]. Zeng et al. introduced a frequency-aware physics-informed convolutional network that directs convolution kernels toward characteristic fault frequencies [17]. From an acoustic perspective, Ma et al. used probabilistic sparse modeling to recover bearing impact features under low-SNR conditions [18]. These studies show that physical priors can reduce dependence on purely statistical correlations and provide more interpretable fault representations.

Recent studies have further explored mechanism–data hybrid diagnosis. Herwig et al. embedded signal-processing operations, including envelope analysis and order tracking, into neural networks [19]. You et al. introduced acoustic–vibration multi-physics constraints into a digital twin framework [20], while Qin et al. incorporated bearing dynamic models into an inverse physics-informed neural network [21]. Zhong et al. developed a multi-source physics-informed network based on improved SincNet [22]. Simulation-driven learning and domain adaptation have also been investigated for transferring physical knowledge to measured bearing data [23]. Despite this progress, physics-guided learning for railway wheelset bearing acoustic signals under strong noise remains insufficient. In particular, the connection among auditory features, fault-frequency priors, and deep acoustic representations has not been fully explored.

This issue is especially important for acoustic diagnosis. Unlike vibration signals, acoustic signals are strongly affected by spatial propagation, reflection, sensor position, and environmental noise. Therefore, the energy distribution of fault-related frequencies can vary considerably. Simply adding theoretical frequencies to the classifier does not guarantee that the network will focus on the corresponding acoustic features. Post-hoc visualization methods such as Grad-CAM can reveal model attention but cannot control noise-related responses during training. Physical priors should therefore participate directly in feature learning and guide the network toward mechanism-related information.

To this end, this paper proposes a Physics-Informed Modulated Fusion ResNet (PIMF-ResNet) for noise-robust acoustic fault diagnosis. Acoustic feature maps generated by an auditory model are used as data-driven inputs. In parallel, a 15-dimensional physical feature vector is constructed from envelope-spectrum peak amplitudes around kinematically predicted characteristic frequencies and their 1X–3X harmonic orders. A physical attention layer adaptively weights FR, BPFO, BPFI, BSF, FTF, and their harmonic components. The weighted priors are then projected into channel attention weights to modulate the deep features of ResNet34. In addition, a physical consistency loss is introduced to reduce attention drift under strong noise. It encourages stronger responses to fault-related frequency families while suppressing irrelevant components. In this way, bearing kinematic knowledge directly participates in feature learning and network optimization.

The main contributions of this paper are summarized as follows.

(1) Bearing kinematic priors are incorporated into deep acoustic feature learning to construct a physics-guided framework for wheelset-bearing diagnosis under strong noise. Rather than treating physical

information as additional classifier inputs, the proposed framework uses it to guide the learning of fault-related representations.

(2) To establish a connection between physical priors and deep acoustic features, fault-frequency information is adaptively weighted and mapped to ResNet channels for feature modulation. A physical consistency constraint is further imposed to strengthen fault-related responses and reduce noise-induced attention drift.

(3) The proposed PIMF-ResNet is evaluated on measured wheelset-bearing acoustic data with ESC-50 environmental noise augmentation. Representative CNN and Transformer models are used for comparison across multiple SNR conditions to assess diagnostic accuracy and noise robustness.

2. Methodology

2.1. Overview of the Fault Diagnosis Framework

Traditional CNN-based bearing diagnosis relies mainly on data-driven feature learning, which limits physical interpretability and can lead to performance degradation under strong noise. To address these issues, PIMF-ResNet integrates bearing kinematic priors with deep acoustic features through a dual-stream architecture, improving the physical relevance and noise robustness of fault representations.

The overall framework consists of three stages, as shown in Figure 1. First, raw acoustic signals are preprocessed and transformed into two-dimensional feature maps. Meanwhile, a 15-dimensional physical prior vector is constructed from envelope-spectrum peak amplitudes around bearing characteristic frequencies and their 1X–3X harmonic orders.

Second, the acoustic feature maps and physical priors are processed by the acoustic and physics-guided branches, respectively. ResNet-34 extracts deep acoustic features, while the physical priors are adaptively weighted and mapped to channel attention weights. These weights modulate the acoustic features to emphasize fault-related information. The model is jointly optimized using cross-entropy loss and physics consistency loss.

Finally, the trained model is evaluated under eight bearing health conditions and different noise levels at a fixed rotational speed. Its diagnostic accuracy and added-noise robustness are compared with representative fault diagnosis methods.

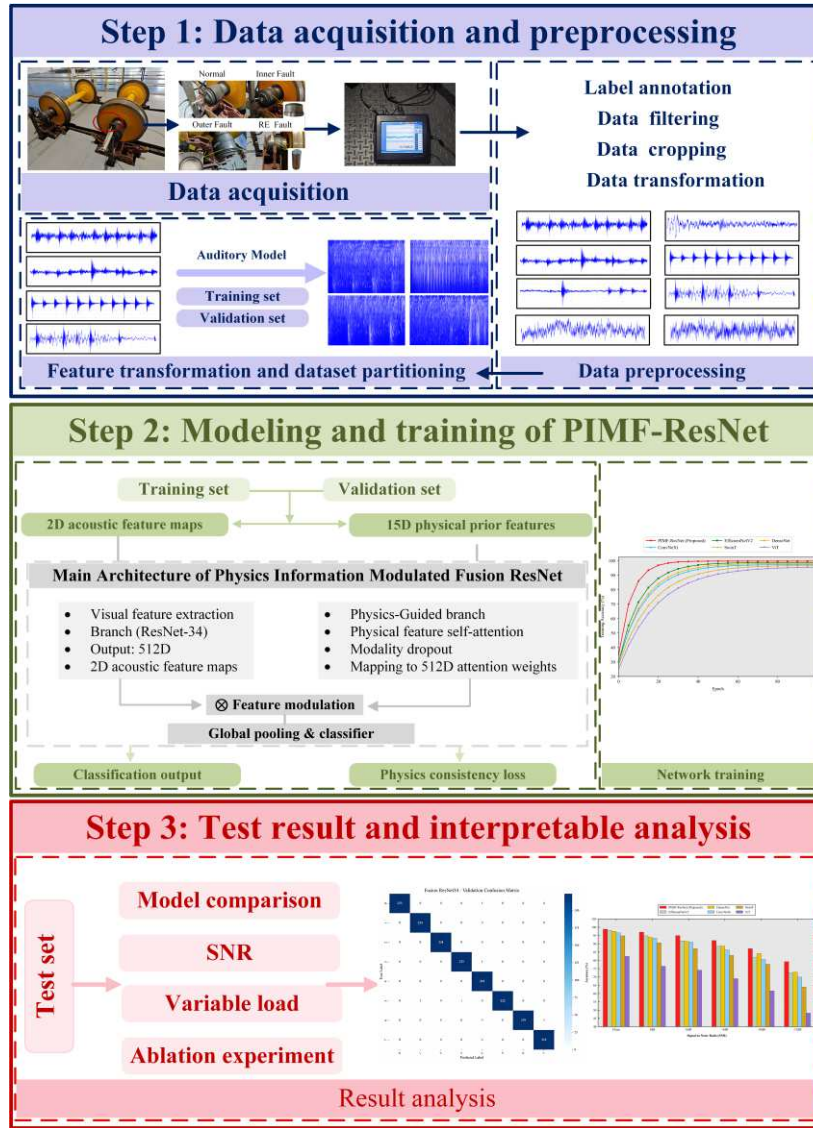


Figure 1. Flowchart of the PIMF-ResNet-based bearing fault diagnosis framework.

In the following section, we describe the proposed fault diagnosis model and feature extractors, as well as the methods used to train each feature extractor and the overall model.

2.2 Auditory-Inspired acoustic feature representation

The acoustic feature maps used in this study are generated using the auditory-perception-based acoustic feature extraction method developed in our previous work [24–26]. Inspired by the cascaded processing mechanisms of the human auditory periphery and brainstem, this method converts one-dimensional acoustic segments into multi-feature auditory spectrograms across frequency channels by simulating middle-ear transmission, cochlear frequency selectivity, inner hair cell transduction, auditory nerve synaptic responses, and modulation-sensitive processing in the early auditory brainstem. Compared with conventional energy-based time–frequency representations, this auditory representation not only characterizes variations in acoustic energy over time and frequency, but also incorporates key auditory processing properties, including cochlear nonlinear compression, frequency selectivity, temporal integration, synaptic adaptation, and modulation-sensitive responses. In our previous study, this auditory front-end was used to construct multi-feature auditory spectrograms and was integrated with a neural network for acoustic fault identification of

1 railway wheelset bearings, where its effectiveness was experimentally validated. The present study builds on
 2 this auditory front-end by introducing physics-guided channel modulation and class-conditioned consistency
 3 regularization.

4 To match the frequency organization of the human auditory system and the spectral characteristics of
 5 railway wheelset bearing acoustic signals, the characteristic frequency (CF) channels are distributed using a
 6 logarithmic frequency mapping strategy [27,28]. Specifically, the center frequency of the i -th auditory channel
 7 is defined as:

$$CF_i = f_L \left(\frac{f_H}{f_L} \right)^{\frac{i}{N_{cf}-1}}, \quad i = 0, 1, \dots, N_{cf} - 1 \quad (1)$$

8 where CF_i represents the characteristic frequency of the i -th auditory channel, f_L and f_H denote the lower
 9 and upper frequency boundaries, respectively, and N_{cf} represents the total number of characteristic-frequency
 10 channels. This logarithmic allocation strategy provides denser frequency resolution in the low-frequency
 11 region while maintaining effective coverage of the high-frequency components associated with bearing fault-
 12 related acoustic responses.

13 After the auditory peripheral and brainstem processing stages, the input acoustic signal $x(t)$ is
 14 transformed into a multi-channel auditory representation. The output auditory feature map can be expressed
 15 as [29–31]:

$$\mathbf{A} = [A_1(t), A_2(t), \dots, A_{N_{cf}}(t)] \in \mathbb{R}^{N_{cf} \times T} \quad (2)$$

16 where $A_i(t)$ denotes the auditory response corresponding to the i -th characteristic frequency channel and
 17 T represents the temporal dimension. Each channel contains the time-varying response extracted through the
 18 auditory processing cascade, enabling the representation to preserve both frequency-selective information and
 19 temporal modulation characteristics of bearing acoustic signals. Furthermore, nonlinear compression is
 20 introduced in the auditory processing pipeline to reduce the influence of large-amplitude interference and
 21 enhance weak fault-related responses. The compressed auditory representation can be described as [32–34]:

$$Y_i(t) = \Psi(x(t), CF_i) \quad (3)$$

22 where $\Psi(\cdot)$ denotes the auditory periphery–brainstem cascade including cochlear filtering, inner hair
 23 cell transduction, auditory nerve synaptic processing, and brainstem modulation.

24 The obtained auditory feature maps are subsequently used as the input of the acoustic feature extraction
 25 branch in the proposed PIMF-ResNet framework. By combining the auditory-inspired representation with
 26 physics-guided feature modulation, the model is able to exploit both data-driven acoustic patterns and bearing
 27 mechanism-related prior information for robust fault diagnosis.

28 A schematic illustration of the auditory-inspired acoustic feature construction process is shown in Figure
 29 2, and the resulting auditory spectrograms are presented in Figure 3.

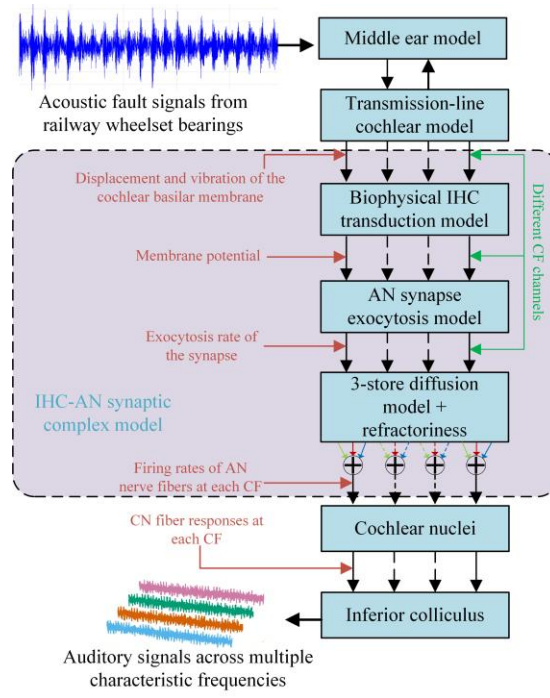


Figure 2. Schematic Illustration of the Auditory-Inspired Acoustic Feature Construction Process.

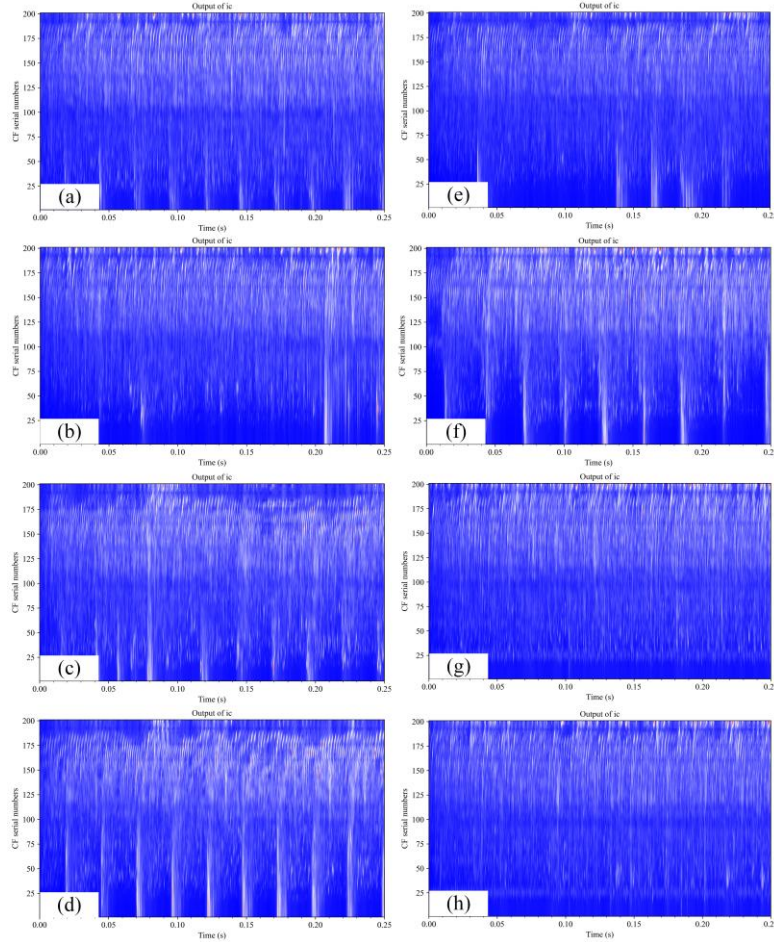


Figure 3. Auditory spectrograms generated by the auditory model under different bearing health conditions: (a) healthy bearing; (b) moderate inner-race fault; (c) severe inner-race fault; (d) moderate outer-race fault; (e) severe outer-race fault; (f) rolling-element fault; (g) slight inner-race fault; and (h) slight outer-race fault.

2.3 Physics-Based Priors for Bearing Fault Diagnosis

Bearing fault frequencies refer to specific vibration frequency components generated by bearing defects as a result of the bearing geometry and operating characteristics. These components manifest as characteristic patterns in the time-domain waveform or as distinct peaks in the frequency spectrum [35].

When a fault occurs, periodic impacts generated during bearing rotation induce vibrations, thereby producing characteristic fault-frequency components in the measured signal [36].

For a given bearing, the parameters associated with its fault frequencies, such as the contact angle, rolling-element diameter, and number of rolling elements, generally remain constant. Consequently, variations in the fault frequencies are governed primarily by the rotational speed. Under constant-speed operating conditions, these frequencies can therefore guide the extraction of diagnostic spectral-amplitude features. The characteristic fault frequencies of the bearing components are calculated as follows:

$$f_{BPFO} = \frac{zf_r}{2} \left(1 - \frac{d}{D} \cos \alpha \right) \quad (4)$$

$$f_{BPFI} = \frac{zf_r}{2} \left(1 + \frac{d}{D} \cos \alpha \right) \quad (5)$$

$$f_{BSF} = \frac{Df_r}{2d} \left[1 - \left(\frac{d}{D} \cos \alpha \right)^2 \right] \quad (6)$$

where f_{BPFI} , f_{BPFO} , and f_{BSF} denote the characteristic fault frequency of the inner raceway, the characteristic fault frequency of the outer raceway, and the rolling-element spin frequency, respectively. z is the number of rolling elements, f_r is the shaft rotational frequency, α is the contact angle, and D and d denote the pitch diameter and rolling-element diameter, respectively.

The probability density distributions of the characteristic bearing fault frequencies are shown in Figure 4. Different fault types exhibit distinct density peaks at specific frequencies, reflecting their respective physical characteristics. Owing to their strong discriminative capability, these characteristic frequencies can serve as reliable diagnostic indicators and facilitate the identification of different bearing fault modes.

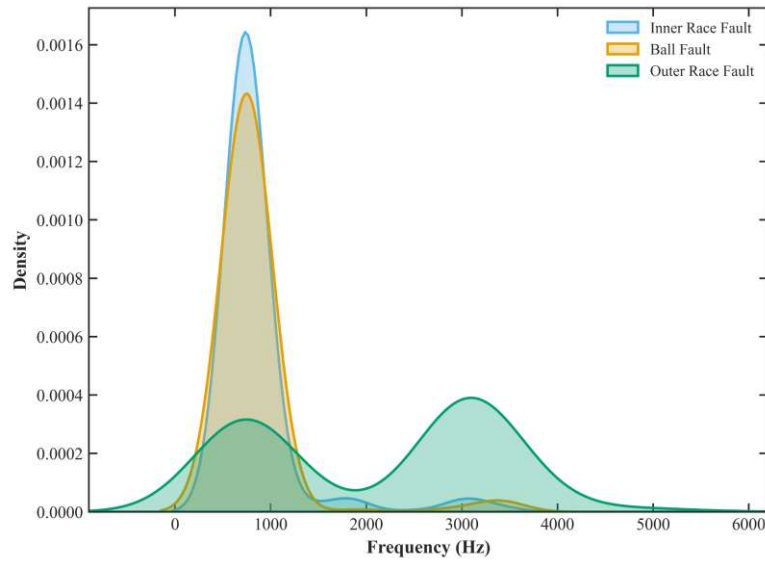


Figure 4. Probability density curves of characteristic bearing fault frequencies.

During the operation of mechanical equipment, fault-induced vibration responses are often modulated, while localized bearing defects generate periodic impacts whose repetition rates lie predominantly in the low-frequency range. Envelope-spectrum analysis demodulates these low-frequency fault-related components from the modulated response, thereby enhancing the detectability of characteristic fault signatures [37].

To explicitly incorporate the kinematic mechanisms of railway wheelset bearings into the network, a physics-informed feature purification operator based on resonance-sensitive locations in the envelope spectrum is developed, as illustrated in Figure 5. First, the amplitude envelope of the raw time-domain signal is obtained using the Hilbert transform, after which the envelope spectrum is calculated via the fast Fourier transform (FFT). This demodulation process reveals low-frequency components associated with periodic fault impacts, although the resulting spectrum remains susceptible to noise. Subsequently, to prevent the complete high-dimensional spectrum from introducing redundant information and noise interference, frequency tolerance bands are defined to selectively capture the local peak amplitudes around five theoretically derived characteristic frequencies—namely, the shaft rotational frequency (FR), outer-race fault frequency (BPFO), inner-race fault frequency (BPFI), rolling-element spin frequency (BSF), and cage rotational frequency (FTF)—together with their first three orders from 1X to 3X.

Through this procedure, the high-dimensional continuous spectrum is condensed into a 15-dimensional feature vector with explicit and well-defined physical meaning, which is then used as the input to the physics-guided branch of the network. This feature-purification mechanism not only preserves the fault-induced kinematic characteristics of the bearing but also substantially reduces the learning burden and noise sensitivity of the physics-guided branch.

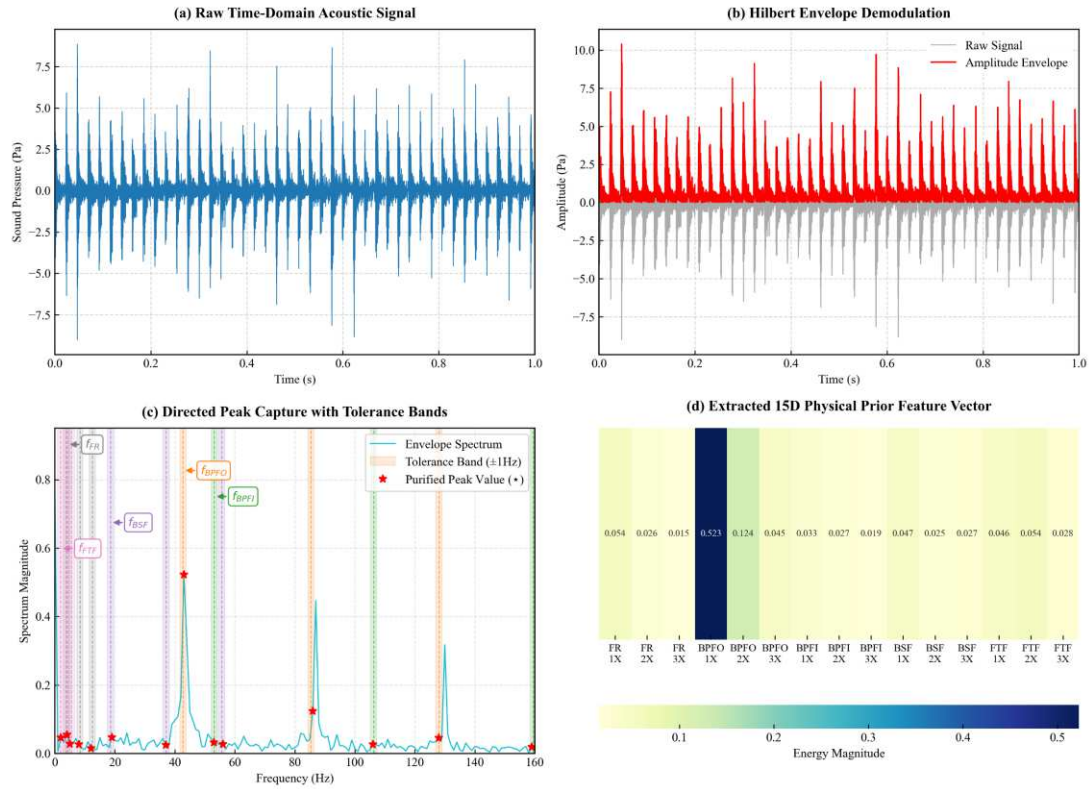


Figure 5. Schematic illustration of the physics-based feature extraction mechanism for bearing faults.

2.4 Physics-Informed Modulated Fusion ResNet

2.4.1 Overall Network Architecture

To overcome the limited interpretability of conventional purely data-driven methods and their vulnerability to complex background noise in acoustic/vibration-based fault diagnosis, a physics-informed residual neural network, referred to as Physics-Informed Modulated Fusion ResNet (PIMF-ResNet), is proposed. Its overall architecture is illustrated in Figure 6.

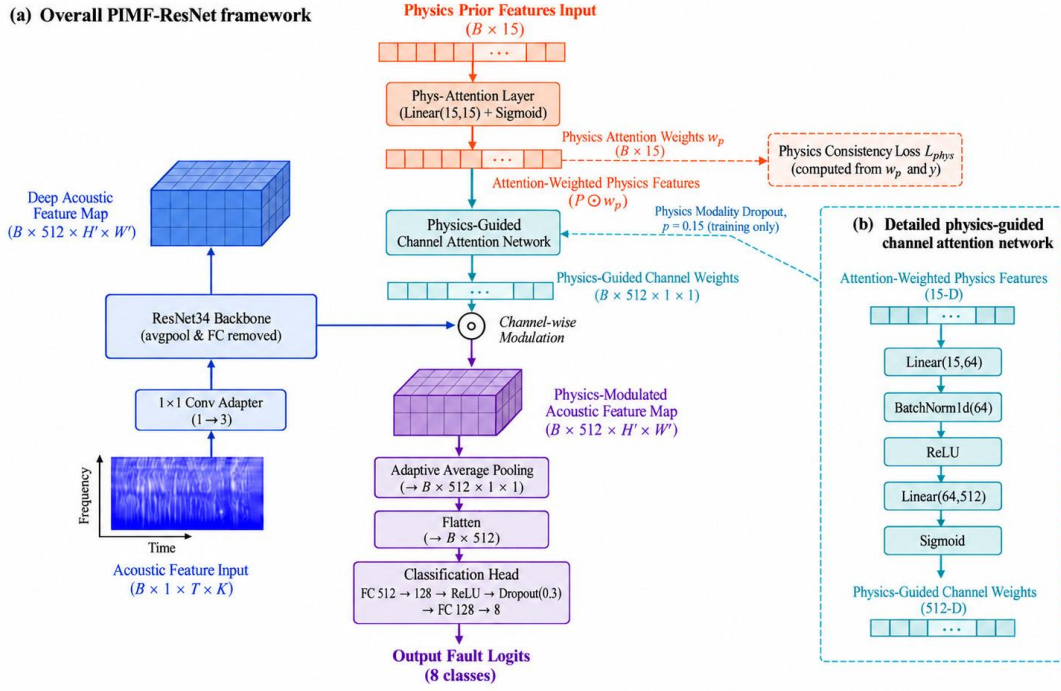


Figure 6. Architecture of the PIMF-ResNet Framework.

The proposed architecture consists of an acoustic feature extraction branch and a physics-guided modulation branch. The acoustic feature map is first converted to three channels through a 1×1 convolutional adapter and then fed into ResNet34 to extract high-level acoustic representations. In parallel, the 15-dimensional physics prior features are adaptively weighted by the Phys-Attention Layer and subsequently mapped to 512-dimensional channel weights through two linear layers with batch normalization and nonlinear activations. These physics-guided weights modulate the ResNet34 feature map in a channel-wise manner, enabling explicit incorporation of bearing fault priors into deep acoustic representations. The modulated features are finally processed by adaptive average pooling and a fully connected classification head with ReLU activation and dropout to produce logits for the eight bearing conditions.

2.4.2 Physics Attention Generation Mechanism

To inject bearing kinematic priors into deep feature learning, a physics attention mechanism is employed to adaptively reweight the characteristic fault-frequency components. The physics prior vector $P \in R^{15}$ comprises FR, BPFO, BPFI, BSF, and FTF, together with their first three harmonics. The corresponding attention weights are computed as:

$$W_p = \sigma(W_1 P + b_1) \quad (7)$$

where $W_1 \in R^{15 \times 15}$ and $b_1 \in R^{15}$ are learnable parameters, and $\sigma(\cdot)$ denotes the Sigmoid function. The resulting $W_p \in (0, 1)^{15}$ characterizes the relative importance of the individual physics-based frequency components and provides the basis for subsequent physics-guided feature modulation.

The physics attention weights W_p are applied to the original physics prior vector P through element-wise multiplication, yielding the attention-weighted physics representation:

$$\tilde{P} = P \square W_p \quad (8)$$

To prevent the network from becoming overly dependent on the physics priors, a physics modality dropout strategy is introduced during training. Specifically, the attention-weighted physics vector is randomly masked as a zero vector with a probability of 0.15:

$$\tilde{P}_{attended} = \begin{cases} \mathbf{0}, & \text{with probability 0.15} \\ \tilde{P}, & \text{with probability 0.85} \end{cases} \quad (9)$$

The resulting 15-dimensional representation is then mapped to the 512-dimensional channel space of the ResNet34 feature map through a two-layer Physics-Guided Channel Attention Network, providing physics-guided weights for subsequent channel-wise modulation.

The attention-weighted physics representation is further projected into the channel space of the ResNet34 backbone through a two-layer mapping network. The first layer expands the 15-dimensional physics vector to a 64-dimensional latent representation, followed by batch normalization (BN) and ReLU activation:

$$H = \text{ReLU}(\text{BN}(W_2 \tilde{P}_{attended} + b_2)) \quad (10)$$

where $W_2 \in R^{64 \times 15}$ and $b_2 \in R^{64}$ are learnable parameters, and $H \in R^{64}$ denotes the intermediate representation.

The second layer maps H to a 512-dimensional channel-weight vector, followed by a Sigmoid function:

$$W_v = \sigma(W_3 H + b_3) \quad (11)$$

where $W_3 \in R^{512 \times 64}$ and $b_3 \in R^{512}$. The resulting $W_v \in (0, 1)^{512}$ provides physics-guided channel weights that are dimensionally aligned with the deep acoustic feature map. These weights are subsequently applied through channel-wise modulation to relatively emphasize fault-relevant responses while suppressing physically irrelevant or noise-dominated channels. Since the Sigmoid coefficients lie between 0 and 1, this multiplication reweights or attenuates channels rather than amplifying their absolute magnitudes.

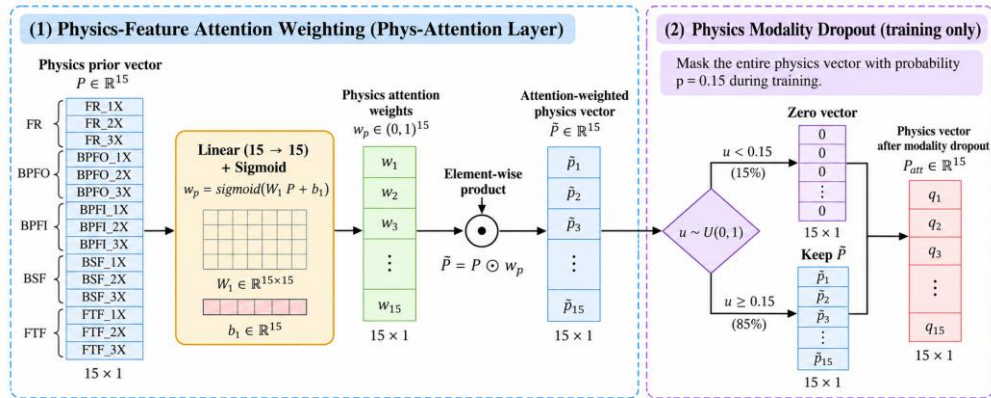


Figure 7. Physics-Feature Gating and Modality Dropout.

2.4.3 Physics-Guided Channel Modulation Fusion

After mapping the 15-dimensional physical prior features into a 512-dimensional channel attention space,

the PIMF-ResNet further employs a physics-guided channel modulation fusion mechanism to couple the bearing kinematic priors with the deep acoustic representations extracted by ResNet34. Unlike conventional feature-level fusion strategies that directly concatenate physical descriptors with deep features before classification, the proposed approach transforms the physical priors into channel-wise modulation coefficients. In this manner, the physical information does not merely serve as an additional input to the classifier but directly participates in the selection and recalibration of deep acoustic features. Specifically, let the deep acoustic feature map extracted by the ResNet34 backbone, with its classification layers removed, be denoted as $F_a \in R^{C \times H \times W}$. Where $C=512$ denotes the number of feature channels, and H and W represent the spatial dimensions of the feature map. As described in Section 2.4.2, the physical attention branch converts the 15-dimensional bearing physical prior into a channel attention vector $W_c \in R^C$, whose elements characterize the relative importance of individual deep feature channels under the guidance of bearing fault-related physical information.

To enable channel-wise modulation of the acoustic feature map, the attention vector is first reshaped by introducing two singleton spatial dimensions:

$$\hat{W}_c = \text{Unsqueeze}(W_c) \in R^{C \times 1 \times 1} \quad (12)$$

The reshaped attention vector is then broadcast along the spatial dimensions and applied to the deep acoustic feature map through element-wise multiplication. The resulting physics-guided feature map is expressed as:

$$F_f(c, h, w) = F_a(c, h, w) \cdot \hat{W}_c(c, 1, 1) \quad (13)$$

where $c \in \{1, 2, \dots, C\}$ denotes the channel index, while $h \in \{1, 2, \dots, H\}$ and $w \in \{1, 2, \dots, W\}$ denote the spatial coordinates of the feature map.

As illustrated in Figure 8, the 15-dimensional physical prior is first adaptively reweighted by the physics attention layer and subsequently projected into a 512-dimensional channel attention vector aligned with the output channels of the ResNet34 backbone. The generated attention coefficients then modulate the deep acoustic representation on a channel-by-channel basis. Consequently, channels that are more consistent with bearing fault-related physical characteristics can be selectively emphasized, whereas responses dominated by noise or physically irrelevant information can be relatively suppressed.

This modulation strategy differs fundamentally from direct feature concatenation. Instead of treating the physical prior and the acoustic representation as two independent feature sources, the proposed mechanism uses the physical prior to regulate the internal representation learned by the acoustic backbone. Therefore, the bearing kinematic knowledge is incorporated directly into the feature-selection process without altering the spatial structure of the ResNet34 feature map. Such a soft modulation mechanism preserves the data-driven representation capability of the backbone while introducing physically meaningful inductive guidance, thereby reducing the tendency of the network to rely on noise-induced spurious patterns under low-SNR

1 conditions.

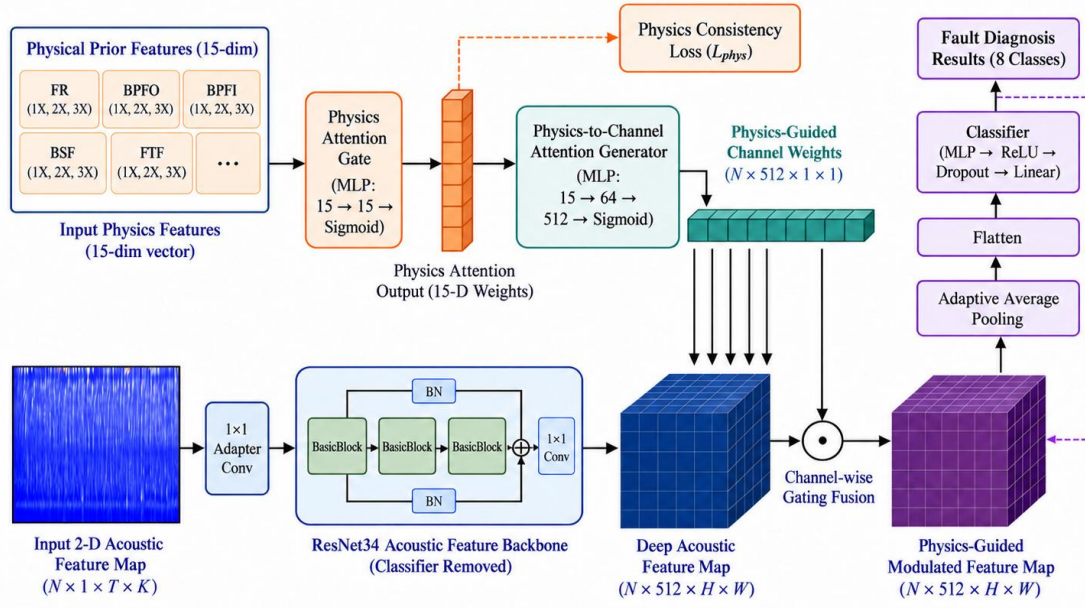


Figure 8. Schematic illustration of the proposed physics-guided channel modulation fusion mechanism.

The physics-guided feature map F_f is subsequently compressed using adaptive global average pooling and fed into the classification head to obtain the final fault-category prediction. Through this procedure, the proposed PIMF-ResNet combines the adaptive representation capability of deep neural networks with the prior guidance provided by bearing fault-related physical information, thereby improving the stability and robustness of fault discrimination in complex acoustic environments.

2.5 Physics-Consistency Regularization

Although the physics-guided channel modulation mechanism introduces bearing kinematic priors into the deep feature learning process, the physics attention weights are still learned in a data-adaptive manner. Under severe background noise, these learned weights may be affected by noise-induced spurious correlations, causing the attention distribution to deviate from physically meaningful fault-related components. Such attention drift may weaken both the robustness and physical interpretability of the diagnostic model.

To alleviate this problem, a physics-consistency regularization mechanism is introduced to explicitly constrain the 15-dimensional physics attention distribution during training. The basic principle is that, for a given bearing fault class, the attention weights associated with its corresponding characteristic-frequency family should be higher than those assigned to unrelated fault-frequency components. In this way, the physical prior is not only used to modulate the deep acoustic representation but also serves as an explicit regularizer for the internal decision process of the network.

2.5.1 Physics-Consistency Loss

According to rolling-bearing kinematics, a localized defect on a specific bearing component produces quasi-periodic impacts whose repetition rate is related to the bearing geometry and rotational speed. These impacts are expected to induce characteristic components around the corresponding defect frequencies and their harmonics in the measured acoustic response. Therefore, when the model identifies a particular fault

type, the physics attention gate should preferentially emphasize the physical-prior components associated with the corresponding fault-frequency family.

As described in Section 2.3, the physical-prior vector contains 15 envelope-spectrum peak-amplitude components associated with harmonic orders 1X–3X of the rotational frequency (FR), ball-pass frequency of the outer race (BPFO), ball-pass frequency of the inner race (BPFI), ball spin frequency (BSF), and cage rotational frequency (FTF). Let the physics attention vector of the i -th sample be denoted by

$$\mathbf{w}_p^{(i)} = [w_{p,1}^{(i)}, \dots, w_{p,15}^{(i)}] \in (0,1)^{15}$$

which is the output of the physics attention gate before the subsequent 15-to-512-dimensional channel projection. A class-to-physics mapping $M : y_i \rightarrow I_{y_i}$ is defined for the eight fault classes, where I_{y_i} denotes the index subset corresponding to the characteristic-frequency family associated with the ground-truth label y_i . Since FR primarily characterizes the rotational operating condition and is shared across different health states, its three components are excluded from the class-specific physics-consistency constraint. Accordingly, the candidate characteristic-frequency index set is defined as $I_{char} = \{3, 4, \dots, 14\}$. For a sample with label

y_i , the non-target index set is then given by $I_{other}^{(i)} = I_{char} \setminus I_{y_i}$. The mean attention weight assigned to

the target fault-frequency family is calculated as $\bar{w}_{target}^{(i)} = \frac{1}{|I_{y_i}|} \sum_{j \in I_{y_i}} w_{p,j}^{(i)}$, whereas the mean attention weight

of the remaining non-target characteristic-frequency components is $\bar{w}_{other}^{(i)} = \frac{1}{|I_{other}^{(i)}|} \sum_{k \in I_{other}^{(i)}} w_{p,k}^{(i)}$. To

encourage a clear separation between the target and non-target physical components, a margin-based hinge loss is introduced. Let $V = \{i | y_i \in \text{Dom}(M)\}$ denote the set of valid fault samples in the current mini-batch and let $N_{valid} = |V|$. The proposed physics-consistency loss is formulated as:

$$L_{phys} = \frac{1}{N_{valid}} \sum_{i=1}^B \max(0, \bar{w}_{other}^{(i)} - \bar{w}_{target}^{(i)} + m) \cdot I(y_i \in \text{Domain}(M)) \quad (14)$$

where m denotes the predefined margin and is set to 0.1 in this study.

The loss becomes zero when the mean attention assigned to the target fault-frequency family exceeds that of the non-target components by at least the prescribed margin. Otherwise, a penalty is introduced to increase the separation between them. Therefore, the loss acts as a soft physical regularizer rather than a hard deterministic constraint, allowing the network to retain sufficient data-driven flexibility while encouraging its internal attention distribution to remain consistent with bearing kinematics.

The healthy class is not included in the physics-consistency loss because it does not correspond to a unique defect-frequency family. Likewise, the FR components are retained as physical inputs to the network but are excluded from the class-specific consistency loss because they mainly represent the common rotational

condition. Although FTF components are retained in the physical prior vector to preserve complete bearing kinematic information, they are not assigned as target components in the physics-consistency loss because no cage fault category is considered in this study. The class-to-physics mapping used in the proposed method is summarized in Table 1.

Table 1. Mapping between fault classes and target physical-prior components.

Label	Bearing condition	Target physical-prior components
0	Healthy	Not included in the physics-consistency loss
1	Moderate inner-race fault	BPFI, 2BPFI, 3BPFI
2	Severe inner-race fault	BPFI, 2BPFI, 3BPFI
3	Moderate outer-race fault	BPFO, 2BPFO, 3BPFO
4	Severe outer-race fault	BPFO, 2BPFO, 3BPFO
5	Rolling-element fault	BSF, 2BSF, 3BSF
6	Slight inner-race fault	BPFI, 2BPFI, 3BPFI
7	Slight outer-race fault	BPFO, 2BPFO, 3BPFO

2.5.2 Joint Optimization Objective

To jointly optimize fault classification performance and physical consistency, the standard cross-entropy loss L_{CE} is combined with the proposed physics-consistency loss L_{phys} . The overall training objective of PIMF-ResNet is defined as:

$$L_{total} = L_{CE} + \alpha_{physics} \cdot L_{phys} \quad (15)$$

where α_{phys} controls the contribution of the physics-consistency regularization and is set to 0.10 in this study. The cross-entropy term drives the network to learn discriminative acoustic representations, whereas L_{phys} encourages the physics attention distribution to remain consistent with the characteristic-frequency semantics of the corresponding bearing fault. Through joint optimization, the model retains the flexibility of data-driven feature learning while incorporating physically meaningful constraints into the internal representation process.

3. Experimental verification

3.1 Experimental setup

To validate the effectiveness of the proposed PIMF-ResNet for acoustic-based railway wheelset bearing fault diagnosis, experiments were conducted using acoustic signals collected from a practical bearing test platform. The experimental specimens were railway wheelset bearings provided by Chongqing Changzheng Heavy Industry Co., Ltd., with the bearing model designated as 353130B.

The geometric configuration of the tested bearing is represented as $(d \times D \times C_1 \times B_1)$, where d , D , C_1 , and B_1 denote the inner diameter, outer diameter, outer-race width, and assembly height, respectively. The corresponding dimensions are $150 \times 250 \times 160 \times 180.6$ mm. The rolling element diameter and effective axial

length are 22.2 mm and 50.1 mm, respectively. The bearing rotational speed was maintained at 250 r/min during the experiments. These geometric parameters not only characterize the tested bearing structure but also provide essential information for calculating the theoretical characteristic frequencies and constructing the proposed physics-prior representations.

The acoustic signals were acquired using a free-field measurement microphone (Type 4189-A-021, Brüel & Kjær). During data acquisition, the microphone was rigidly mounted on a mechanical support and positioned near the tested bearing with its sensing direction aligned toward the bearing surface, ensuring stable signal collection and reliable acoustic measurements. The acquired signals were digitized using a MI-7008D dynamic signal analyzer with a sampling frequency of 51.2 kHz, and the microphone sensitivity was set to 50 mV/Pa. The experimental platform configuration and microphone placement are illustrated in Figure 9.

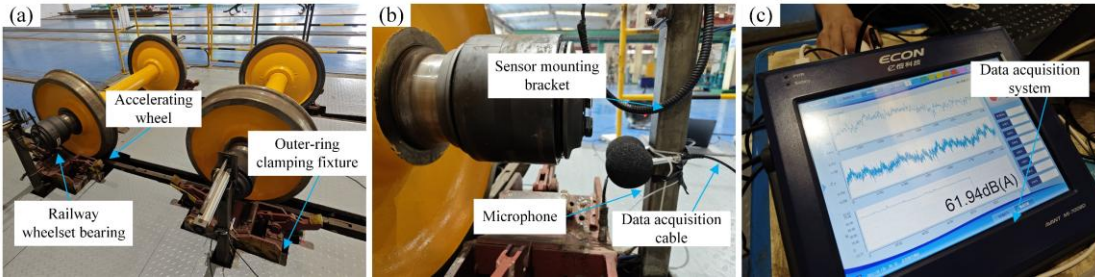


Figure 9. Experimental setup for acoustic signal acquisition.

The raw acoustic signals collected under different health conditions were first labeled according to their corresponding fault categories. In this study, eight bearing operating conditions were considered, including the healthy state and various fault states with different fault types and severity levels. The detailed category names, label indices, original recording durations, and fault sizes are summarized in Table 2, while the photographs of the inner- and outer-race faults are provided in Figure 10. Here, “slight” denotes a smaller artificial defect relative to the other defects in this experiment, rather than a naturally occurring early-stage fault.

To ensure consistency during sample construction, all raw acoustic recordings were segmented using a fixed-length window. The duration of each segment was set to 0.25 s, and no overlap was applied during the segmentation process. Given the sampling frequency of 51.2 kHz, each acoustic segment contains 12,800 sampling points.

The non-overlapping segmentation strategy reduces the information redundancy among adjacent samples and mitigates the risk of data leakage caused by highly correlated samples being simultaneously included in the training and validation sets. Consequently, the constructed dataset provides a more reliable evaluation of the generalization capability of the proposed model under different acoustic noise conditions.

Table 2. Bearing fault categories, labels, and artificially introduced defect dimensions.

Label	Fault Type	Defect Dimensions (Length ×	Original Recording
		Width × Depth)	Duration
0	Healthy	/	301s

1	Moderate Inner-Race Fault	$50 \times 5 \times 1 \text{mm}$	296s
2	Severe Inner-Race Fault	$50 \times 10 \times 1 \text{mm}$	283s
3	Moderate Outer-Race Fault	$50 \times 6 \times 1 \text{mm}$	296s
4	Severe Outer-Race Fault	$50 \times 12 \times 1 \text{mm}$	311s
5	Rolling-Element Fault	$50 \times 1 \times 0.8 \text{mm}$	294s
6	Slight Inner-Race Fault	$6 \times 2 \times 1 \text{mm}$	598s
7	Slight Outer-Race Fault	$8 \times 2 \times 1 \text{mm}$	616s



Figure 10. Schematic illustrations of bearing fault types.

The proposed PIMF-ResNet adopts a dual-input architecture consisting of a two-dimensional acoustic feature map and its corresponding 15-dimensional physical-prior feature vector. For each sample, the acoustic representation and physical-prior vector are paired according to the file name and label index, ensuring a consistent correspondence between data-driven features and physics-based information. To satisfy the input size requirements of the network, all acoustic feature maps are uniformly cropped or zero-padded along the temporal dimension.

The dataset was divided into training and validation sets with a ratio of 8:2. The training set was used for parameter optimization, while the validation set was employed for model performance evaluation. During the splitting process, the proportion of samples from each fault category was maintained to reduce the influence of class imbalance on performance comparison. The same preprocessing procedure was applied to all experiments to ensure fair evaluation among different models.

The proposed PIMF-ResNet was implemented using the PyTorch framework. The acoustic feature extraction branch adopts ResNet34 as the backbone network. Since the input acoustic feature maps contain a single channel, they were expanded to three channels before being fed into the ResNet34 backbone. In parallel, the physical-prior branch employs a physics attention gate to generate 15-dimensional physical attention weights, which are subsequently projected into 512-dimensional channel modulation weights for guiding the learning of deep acoustic representations. The final classifier consists of a fully connected layer, ReLU

activation function, and Dropout layer to perform eight-class bearing condition recognition.

The training configurations are summarized in Table 3. The batch size was set to 8, and the model was trained for 60 epochs with an initial learning rate of 0.001. The AdamW optimizer was adopted with a weight decay coefficient of 1×10^{-4} . The overall optimization objective combines the classification loss and the proposed physics-consistency loss, where the weighting coefficient of the physics-consistency term was set to $\alpha_{\text{phys}}=0.10$. Furthermore, a physics dropout strategy with a probability of 0.15 was introduced during training to randomly suppress weighted physical features. This strategy reduces the model's excessive dependence on individual physical priors and improves robustness against acoustic noise interference and physical-feature perturbations.

Table 3. Training parameter settings of the PIMF-ResNet model.

Parameter	Value
Batch size	8
Epoch	60
Learning rate	0.001
Optimizer	AdamW
Weight decay	(1×10^{-4})
Classification loss	Cross-Entropy Loss
Physics loss weight	0.10
Physics-consistency loss margin	0.10
Dropout	0.30
Physics dropout probability	0.15

3.2 Feature Ablation and Noise Robustness Evaluation

Acoustic-based railway wheelset bearing fault diagnosis is often affected by complex operating environments, where the collected acoustic signals are inevitably contaminated by non-stationary environmental noise. To evaluate the robustness of the proposed model under noisy conditions, the ESC-50 environmental sound dataset [38,39] was introduced as an external noise source. This dataset contains diverse real-world environmental sounds and provides a suitable basis for simulating complex acoustic interference encountered in railway wheelset bearing monitoring scenarios.

The ESC-50 dataset consists of 2000 environmental audio recordings, each with a duration of 5 s. For noise generation, a 2-s segment was first randomly extracted from each recording, followed by random cropping of a 0.25-s noise segment to match the duration of the bearing acoustic samples. All noise signals were resampled to 51.2 kHz and subsequently mixed with the original bearing acoustic signals according to predefined signal-to-noise ratios (SNRs) to construct noisy evaluation samples.

The SNR is defined as:

$$\text{SNR} = 10 \log_{10} \left(\frac{P_s}{P_n} \right) \quad (16)$$

where P_s and P_n denote the power of the original bearing acoustic signal and the added environmental noise, respectively. In this study, the SNR levels were selected from 5 dB to −15 dB to evaluate the diagnostic performance under different noise intensities.

To avoid reusing the same ESC-50 files across training and validation, disjoint noise-file subsets were employed. Specifically, the 2000 environmental recordings were divided into 1400 training noise samples and 600 validation noise samples. During training, random noise augmentation was applied, where noise segments and target SNR values were randomly selected for each sample. During validation, fixed noise samples were generated under clean, 5, 0, −5, −10, and −15 dB SNR conditions, ensuring a fair comparison among different models.

All experiments were conducted using the same data partition strategy and preprocessing procedure, with 80% of the samples used for training and the remaining 20% used for validation. The time-domain waveforms of the noise-augmented samples are presented in Figure 11. To investigate the contribution of different information sources to diagnostic performance, comparisons were performed among three feature configurations: physical-prior features alone, acoustic deep features alone, and the proposed fusion representation combining both. The model performance was evaluated using four metrics, including Accuracy, Precision, Recall, and F1-score, as summarized in Table 4.

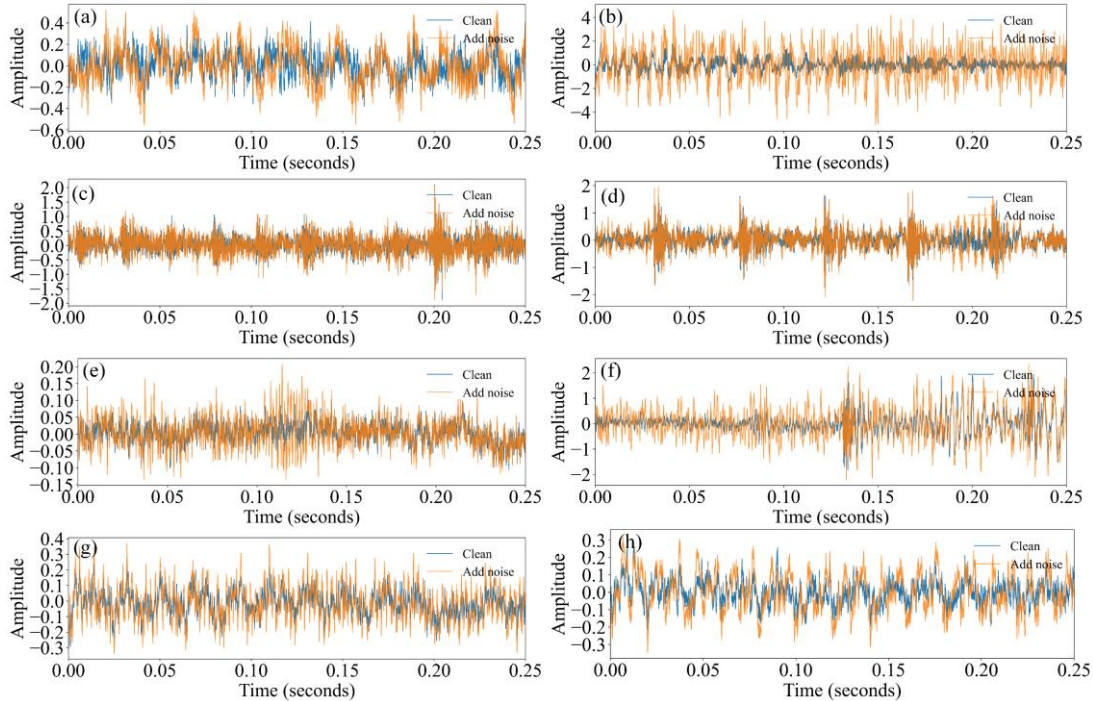


Figure 11. Time-domain waveforms of noise-augmented samples.

Table 4. Comparison of different feature configurations under various SNR conditions.

Feature configuration	SNR	Accuracy	Precision	Recall	F1-score
Physics-guided fusion	clean	0.9883	0.9884	0.9883	0.9883
	5dB	0.9704	0.9708	0.9704	0.9704

	0dB	0.9498	0.9500	0.9498	0.9498
	-5dB	0.9202	0.9217	0.9202	0.9205
	-10dB	0.8711	0.8718	0.8711	0.8709
	-15dB	0.7924	0.8010	0.7924	0.7919
	clean	0.9727	0.9732	0.9727	0.9728
	5dB	0.9414	0.9434	0.9414	0.9412
Auditory acoustic	0dB	0.9191	0.9200	0.9191	0.9185
features only	-5dB	0.8867	0.8910	0.8867	0.8864
	-10dB	0.8354	0.8420	0.8356	0.8360
	-15dB	0.7260	0.7414	0.7260	0.7275
	clean	0.8097	0.8248	0.8097	0.8139
Feature configuration	5dB	0.6390	0.7163	0.6390	0.6456
	0dB	0.3376	0.4205	0.3376	0.3255
Physics-prior features	-5dB	0.1987	0.2800	0.1987	0.1681
only	-10dB	0.1635	0.2216	0.1635	0.1121
	-15dB	0.1535	0.2485	0.1535	0.0912

As summarized in Table 4, different feature configurations exhibit distinct performance variations under different noise conditions. The proposed fusion strategy achieves the best performance among all evaluated configurations across the entire range of evaluation conditions. Specifically, the proposed method obtains accuracies of 0.9883, 0.9704, 0.9498, 0.9202, 0.8711, and 0.7924 under clean, 5 dB, 0 dB, -5 dB, -10 dB, and -15 dB conditions, respectively. Compared with the model using acoustic deep features alone, the fusion model consistently improves the diagnostic performance under all SNR levels. Particularly, under the extremely noisy condition of -15 dB, the Accuracy increases from 0.7260 to 0.7924, corresponding to an improvement of 6.64 percentage points. This result demonstrates that incorporating physical-prior information can effectively enhance fault discrimination capability under low-SNR environments.

The physics-prior features alone achieve limited diagnostic performance, especially under noisy conditions. These features contain measured envelope-spectrum peak amplitudes around theoretical fault-frequency locations, rather than only the frequency locations themselves. Their reduced accuracy under added noise suggests that these compact spectral descriptors alone are insufficient for reliable fault recognition in the evaluated conditions.

The model using acoustic deep features alone shows better robustness under different noise levels, suggesting that deep acoustic representations can effectively capture discriminative patterns from complex acoustic environments and preserve fault-related information. Nevertheless, without explicit physical constraints, its performance still deteriorates under extremely low-SNR conditions.

Overall, the physical-prior features and acoustic deep representations exhibit strong complementarity. The physical priors provide mechanism-related constraints associated with bearing fault characteristics, while

the acoustic deep features enhance the adaptability of the model to complex noise environments. By integrating both information sources, the proposed method maintains high diagnostic accuracy under clean conditions and achieves robust performance under severe noise interference, demonstrating its effectiveness for railway wheelset bearing fault diagnosis in complex acoustic scenarios.

3.3 Comparison of Diagnostic Performance and Noise Robustness among Multi-Architecture Baseline Models

Based on the aforementioned noise robustness experiments, further comparative experiments were conducted to evaluate the diagnostic performance of the proposed PIMF-ResNet and the effectiveness of its architectural design. Several representative deep neural network architectures were selected as comparison models for comprehensive evaluation.

The compared models include two Transformer-based vision models, Vision Transformer (ViT) and Swin Transformer, which represent current self-attention-based feature modeling approaches. In addition, three convolutional neural network architectures, namely ConvNeXt, EfficientNetV2, and DenseNet, were selected to represent modern convolutional networks, efficient lightweight CNNs, densely connected networks, and mobile-oriented lightweight architectures, respectively.

All baseline models were trained and evaluated using the same data partition strategy and experimental configurations. The diagnostic performance and noise robustness of different architectures were assessed under six testing conditions, including clean data, 5 dB, 0 dB, -5 dB, -10 dB, and -15 dB SNR environments. The classification results of different models under various noise conditions are summarized in Figure 12 and Table 5. In addition, the average F1-score across different SNR conditions is further calculated to provide a comprehensive evaluation of classification stability, as shown in Figure 13.

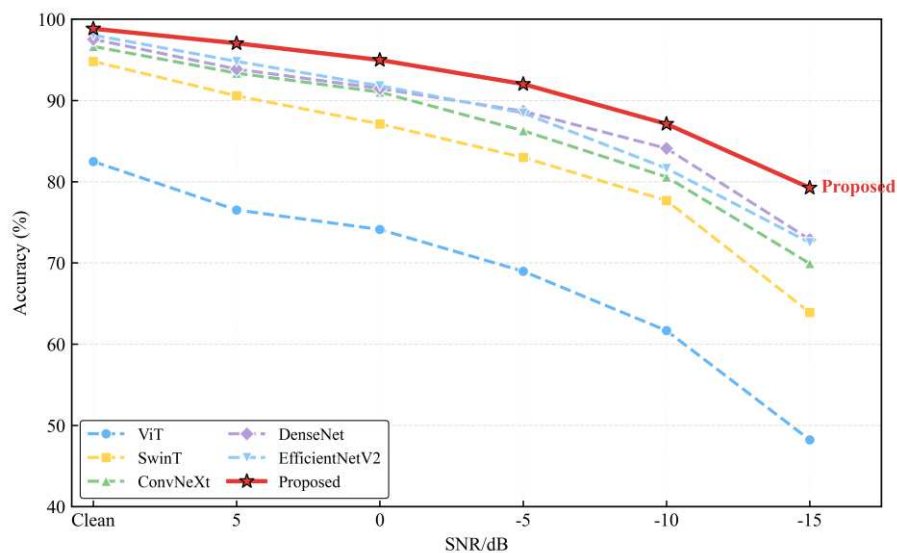


Figure 12. Comparison of diagnostic accuracy among different models under various signal-to-noise ratio conditions.

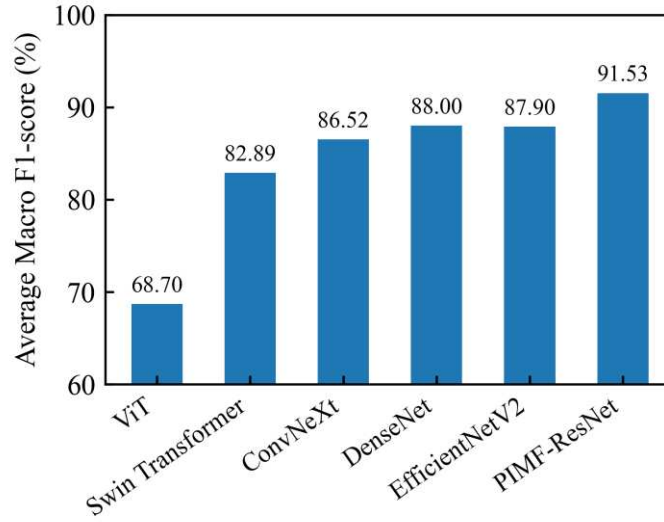


Figure. 13. Comparison of average Macro F1-score among different models under various SNR conditions.

Table 5. Recognition accuracy results of different models under various noise scenarios.

SNR/dB	Accuracy					
	ViT	SwinT	ConvNeXt	DenseNet	EfficientNetV2	Proposed
Clean	0.8248	0.9481	0.9665	0.9754	0.9805	0.9883
5	0.7651	0.9057	0.9336	0.9386	0.9481	0.9704
0	0.7411	0.8711	0.9102	0.9146	0.9180	0.9498
-5	0.6897	0.8298	0.8627	0.8867	0.8845	0.9202
-10	0.6166	0.7768	0.8058	0.8410	0.8164	0.8711
-15	0.4821	0.6390	0.6992	0.7288	0.7249	0.7924

As shown in Table 5, all models exhibit certain fault identification capability under clean conditions, while noticeable performance differences exist among different architectures. ViT achieves an accuracy of only 82.48%, which is lower than those of the other models, under the present training configuration; this comparison alone does not isolate the effect of global self-attention from data size, pretraining, or optimization settings. In contrast, Swin Transformer, ConvNeXt, DenseNet, and EfficientNetV2 achieve accuracies of 94.81%, 96.65%, 97.54%, and 98.05%, respectively, showing higher accuracy under the same evaluation protocol.

The proposed PIMF-ResNet achieves the highest accuracy of 98.83% under clean conditions, indicating that the introduced physics-informed modulation mechanism not only preserves the feature learning capability of deep networks but also enhances the representation of fault-related information.

As the signal-to-noise ratio decreases, the performance of all models deteriorates, but with different degradation rates. Under the extreme noise condition of -15 dB, ViT and Swin Transformer decrease to 48.21% and 63.90%, respectively, whereas ConvNeXt, DenseNet, and EfficientNetV2 maintain accuracies of 69.92%, 72.88%, and 72.49%, showing greater robustness in this comparison.

In comparison, PIMF-ResNet achieves the best performance across all evaluation conditions. The accuracies under Clean Data, 5 dB, 0 dB, -5 dB, -10 dB, and -15 dB conditions reach 98.83%, 97.04%, 94.98%, 92.02%, 87.11%, and 79.24%, respectively. Particularly, under the -15 dB condition, PIMF-ResNet improves the accuracy by 6.36, 6.75, 9.32, 15.34, and 31.03 percentage points compared with DenseNet, EfficientNetV2, ConvNeXt, Swin Transformer, and ViT, respectively.

Furthermore, PIMF-ResNet achieves an average accuracy of 91.54% across six noise conditions, outperforming DenseNet (88.09%), EfficientNetV2 (87.87%), ConvNeXt (86.30%), Swin Transformer (82.84%), and ViT (68.66%). Meanwhile, the accuracy degradation from Clean Data to -15 dB is only 19.59 percentage points, which is significantly lower than that of other models, demonstrating its superior robustness under severe noise environments.

The observed improvement is consistent with a benefit from physics-guided fusion, although the individual contributions of channel modulation, consistency loss, and modality dropout have not been isolated. By incorporating bearing fault mechanism priors into the channel attention generation process, PIMF-ResNet enhances fault-related feature responses while suppressing noise-induced interference, thereby improving the reliability of acoustic fault diagnosis under complex noise conditions.

Besides accuracy, the average Macro F1-score under six SNR conditions was further calculated to evaluate the classification stability of different models. As shown in Fig. 13, PIMF-ResNet achieves the highest average Macro F1-score of 91.53%, outperforming all comparison methods. This result indicates that the proposed physics-informed modulation mechanism improves not only overall recognition accuracy but also the consistency of multi-class fault identification under noisy environments.

4. Conclusion

This paper addresses the acoustic fault diagnosis problem of train wheelset bearings under strong background noise and proposes a Physics-Informed Modulated Fusion ResNet (PIMF-ResNet). The proposed method utilizes acoustic deep features as the data-driven representation and incorporates fault-frequency physical priors derived from bearing geometric parameters and rotational speed information. Through physics-informed attention modulation and channel feature reconstruction, the proposed framework integrates bearing fault mechanisms into deep feature learning, thereby improving diagnostic performance and robustness under complex noise environments:

The main conclusions are summarized as follows:

(1) The proposed PIMF-ResNet effectively integrates acoustic deep features with bearing fault-frequency priors. By guiding the network to focus on fault-related sensitive feature regions through physical information modulation, the proposed method enhances the representation capability of characteristic patterns associated with inner-race, outer-race, and rolling-element faults.

(2) Extensive experiments under multiple noise conditions demonstrate that PIMF-ResNet consistently outperforms comparative models under different signal-to-noise ratio levels. From clean conditions to the

extremely noisy condition of -15 dB, the recognition accuracies reach 98.83%, 97.04%, 94.98%, 92.02%, 87.11%, and 79.24%, respectively, indicating superior noise robustness.

(3) Across six noise scenarios, PIMF-ResNet achieves an average accuracy of 91.54% and exhibits a relatively smaller performance degradation from clean data to -15 dB conditions, validating the effectiveness of physics-informed guidance in improving the robustness of acoustic fault diagnosis.

Although the proposed method achieves promising diagnostic performance, the current experiments are mainly conducted on laboratory datasets with relatively limited fault categories and operating conditions. Future work will extend the investigation to wheelset bearing data under variable rotational speeds, varying loads, and real railway operating environments to evaluate the generalization capability of the proposed physics-informed framework. Moreover, considering that the characteristic fault frequencies may vary under complex operating conditions, adaptive physics-prior modeling and dynamic mechanism-guided feature modulation will be further investigated. Furthermore, composite fault identification, fault severity estimation, and multi-source vibration-acoustic information fusion will be explored to enhance the applicability of the proposed method for practical online condition monitoring.

Statements and Declarations

Funding

This work was supported by the National Natural Science Foundation of China (Grant No.52275296), Graduate Innovation Program of China University of Mining and Technology (2026WLJCRCZL158), Postgraduate Research & Practice Innovation Program of Jiangsu Province (26CXJH4770) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Data availability

The data cannot be made publicly available upon publication because they contain commercially sensitive information. The data that support the findings of this study are available upon reasonable request from the authors.

CRedit authorship contribution statement

Peihai Li: Writing - review & editing, Writing - original draft, Methodology, Validation, Software, Formal analysis. **Zheng Li:** Writing - review & editing, Writing - original draft, Methodology, Validation, Software, Formal analysis. **Dengyang Cao:** Writing - review & editing, Writing - original draft, Methodology, Validation, Software, Formal analysis. **Bo Zhang:** Writing - review & editing, Formal analysis. **Chaoying Li:** Writing - review & editing, Formal analysis. **Houguang Liu:** Funding acquisition, Writing - review & editing, Supervision, Data curation, Investigation, Project administration, 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.

References

- [1] Tandon N and Choudhury A 1999 A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings *Tribology International* **32** 469–80
- [2] Al-Ghamd A M and Mba D 2006 A comparative experimental study on the use of acoustic emission and vibration analysis for bearing defect identification and estimation of defect size *Mechanical Systems and Signal Processing* **20** 1537–71
- [3] He Q, Wang J, Hu F and Kong F 2013 Wayside acoustic diagnosis of defective train bearings based on signal resampling and information enhancement *Journal of Sound and Vibration* **332** 5635–49
- [4] Anon Signal separation and correction with multiple Doppler acoustic sources for wayside fault diagnosis of train bearings - Shangbin Zhang, Qingbo He, Haibin Zhang, Kesai Ouyang, Fanrang Kong, 2017
- [5] Zhang D, Entezami M, Stewart E, Roberts C and Yu D 2018 Adaptive fault feature extraction from wayside acoustic signals from train bearings *Journal of Sound and Vibration* **425** 221–38
- [6] Zhang D, Entezami M, Stewart E, Roberts C, Yu D and Lei Y 2020 Wayside acoustic detection of train bearings based on an enhanced spline-kernelled chirplet transform *Journal of Sound and Vibration* **480** 115401
- [7] Ding X, Li Y, Xiao J, He Q, Yang X and Shao Y 2022 Parametric Doppler correction analysis for wayside acoustic bearing fault diagnosis *Mechanical Systems and Signal Processing* **166** 108375
- [8] Zhang D, Stewart E, Entezami M, Roberts C and Yu D 2020 Intelligent acoustic-based fault diagnosis of roller bearings using a deep graph convolutional network *Measurement* **156** 107585
- [9] Hasan M J, Islam M M M and Kim J-M 2019 Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions *Measurement* **138** 620–31
- [10] Zhang Z, Liu H, Yuan G, Yang J, Liu S, Shao Y and Zhang Y 2024 Gas-insulated switch-gear mechanical fault detection based on acoustic using feature fused neural network *Electric Power Systems Research* **230** 110226
- [11] Yang Z, Wang S, Yang S, Liu S, Zhang Z and Liu H 2025 Coal gangue recognition in the strong background noise using two-level auditory feature fusion with attention mechanism *Measurement* 117628
- [12] Zhao R, Yan R, Chen Z, Mao K, Wang P and Gao R X 2019 Deep learning and its applications to machine health monitoring *Mechanical Systems and Signal Processing* **115** 213–37
- [13] Guo X, Chen L and Shen C 2016 Hierarchical adaptive deep convolution neural network and its

application to bearing fault diagnosis *Measurement* **93** 490–502

- [14] Zhang W, Li C, Peng G, Chen Y and Zhang Z 2018 A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load *Mechanical Systems and Signal Processing* **100** 439–53
- [15] Xie T, Xu Q, Jiang C, Lu S and Wang X 2023 The fault frequency priors fusion deep learning framework with application to fault diagnosis of offshore wind turbines *Renewable Energy* **202** 143–53
- [16] Gao C, Wang Z, Guo Y, Wang H and Yi H 2024 MPINet: Multiscale Physics-Informed Network for Bearing Fault Diagnosis With Small Samples *IEEE Transactions on Industrial Informatics* **20** 14371–80
- [17] Zeng L, Zhang F, Lang G, Wang Y and Chen Q 2025 Application of Frequency Aware Mechanism-Based Physical Information Convolutional Neural Network in Rolling Bearing Fault Diagnosis *IEEE Sensors Journal* **25** 16797–811
- [18] Ma Z, Zhao M, Ou S, Ma B and Zhang Y 2024 Physics-Inspired Sparse Voiceprint Sensing for Bearing Fault Diagnosis *IEEE Transactions on Industrial Informatics* **20** 11273–84
- [19] Herwig N, Borghesani P, Smith W and Peng Z 2025 Signal processing- and physics-informed neural network for explainable bearing condition monitoring *Mechanical Systems and Signal Processing* **235** 112925
- [20] Keshun Y, Puzhou W, Peng H and Yingkui G 2025 A sound-vibration physical-information fusion constraint-guided deep learning method for rolling bearing fault diagnosis *Reliability Engineering & System Safety* **253** 110556
- [21] Qin Y, Liu H, Wang Y and Mao Y 2024 Inverse physics-informed neural networks for digital twin-based bearing fault diagnosis under imbalanced samples *Knowledge-Based Systems* **292** 111641
- [22] Zhong J, Zheng Y, Ruan C, Chen L, Bao X and Lyu L 2025 M-IPISincNet: An explainable multi-source physics-informed neural network based on improved SincNet for rolling bearings fault diagnosis *Information Fusion* **115** 102761
- [23] Anon 2023 Simulation Data-driven Enhanced Unsupervised Domain Adaptation for Bearing Fault Diagnosis *Journal of Mechanical Engineering* **59** 76
- [24] Ji H, Liu H, Wang J, Yuan G, Yang J and Yang S 2023 Mechanical fault diagnosis of gas-insulated switchgear based on saliency feature of auditory brainstem response under noise background *Meas. Sci. Technol.* **35** 015008
- [25] Li Z, Li P, Yang J, Cheng Z, Zhang B, Dong H, Jiang B and Liu H 2026 Auditory perception-based acoustic fault diagnosis for railway wheelset bearings under strong noise: a comparison with short-time Fourier transform, Mel spectrogram, and continuous wavelet transform *Engineering Research Express* **8**
- [26] Zhao Y, Liu H, Guo W, He Z, Yang J and Zhang Z 2024 Research on sound quality prediction of vehicle interior noise using the human-ear physiological model *J. Acoust. Soc. Am.* **156** 989–1003

- [27] Büchner A, Weller T, Penninger R T, Helpard L, Ladak H M, Agrawal S, Lenarz T and Schurzig D 2025 The Greenwood function shows close alignment with pitch perceived by cochlear implant patients with long, flexible electrode arrays and fine-structure stimulation *Front. Neurosci.* **19**
- [28] Schurzig D, Pietsch M, Erfurt P, Timm M E, Lenarz T and Kral A 2021 A cochlear scaling model for accurate anatomy evaluation and frequency allocation in cochlear implantation *Hearing Research* **403** 108166
- [29] Verhulst S, Altoè A and Vasilkov V 2018 Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss *Hearing research* **360** 55–75
- [30] Vecchi A O, Varnet L, Carney L H, Dau T, Bruce I C, Verhulst S and Majdak P 2022 A comparative study of eight human auditory models of monaural processing *Acta Acustica* **6** 17
- [31] Bruce I C, Erfani Y and Zilany M S 2018 A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites *Hearing research* **360** 40–54
- [32] Verhulst S, Dau T and Shera C A 2012 Nonlinear time-domain cochlear model for transient stimulation and human otoacoustic emission *The Journal of the Acoustical Society of America* **132** 3842–8
- [33] Verhulst S, Bharadwaj H M, Mehraei G, Shera C A and Shinn-Cunningham B G 2015 Functional modeling of the human auditory brainstem response to broadband stimulation *The Journal of the Acoustical Society of America* **138** 1637–59
- [34] Altoè A, Pulkki V and Verhulst S 2014 Transmission line cochlear models: improved accuracy and efficiency *The Journal of the Acoustical Society of America* **136** EL302–8
- [35] Zhou P, Chen S, He Q, Wang D and Peng Z 2023 Rotating machinery fault-induced vibration signal modulation effects: A review with mechanisms, extraction methods and applications for diagnosis *Mechanical Systems and Signal Processing* **200** 110489
- [36] Xu F, Ding N, Li N, Liu L, Hou N, Xu N, Guo W, Tian L, Xu H, Lawrence Wu C-M, Wu X and Chen X 2023 A review of bearing failure Modes, mechanisms and causes *Engineering Failure Analysis* **152** 107518
- [37] Chen B, Zhang W, Xi Gu J, Song D, Cheng Y, Zhou Z, Gu F and Ball A D 2023 Product envelope spectrum optimization-gram: An enhanced envelope analysis for rolling bearing fault diagnosis *Mechanical Systems and Signal Processing* **193** 110270
- [38] Piczak K J 2015 ESC: Dataset for Environmental Sound Classification *Proceedings of the 23rd ACM international conference on Multimedia MM '15: ACM Multimedia Conference (Brisbane Australia: ACM)* pp 1015–8
- [39] Dong H, Zhang Y, Zhang B, Zhang Z, Shao Y, Yuan G and Liu H 2025 A Noise-Robust xLSTM-Based Method for Fault Diagnosis in Gas-Insulated Switchgear Using Acoustic Signals Under Strong Noise Interference *Expert Systems with Applications* 130827