Journal of Industrial Information Integration 51 (2026) 101068

Contents lists available at ScienceDirect

Journal of Industrial Information Integration

journal homepage: www.elsevier.com/locate/jii

Full length article

Prior knowledge-embedded first-layer interpretable paradigm for rail transit vehicle human–computer collaboration fault monitoring ☆

Chao He [a,b], Hongmei Shi [a,b,*], Jing-Xiao Liao [c], Bin Liu [d], Qiuhai Liu [a,b], Jianbo Li [a,b], Zujun Yu [a,b]

[a] State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China
[b] School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China
[c] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom 999077, Hong Kong Special Administrative Region
[d] School of Mechanical and Electronical Engineering, Lanzhou University of Technology, Lanzhou 730050, China

ARTICLE INFO

Keywords:
Embodied intelligence
Multi-view learning
First-layer interpretable paradigm
Attention fusion routing
Noise threshold amplitude ratio
Rail transit vehicles

ABSTRACT

Rail transit vehicles endure large loads, high speeds, and harsh environment, leading to component failure. The first-layer interpretable paradigm (FLIP) embeds human prior knowledge into smart equipment, which is one of intelligent paradigms guided by customized manufacturing and embodied intelligence. It consists of first-layer interpretable modules, backbones, loss metrics. However, existing efforts rely on single-source information, an absence of interpretable backbones, an inability to feature fusion, thereby struggling with multi-excitation, coupled signals. To bridge this gap, a FLIP-based one-stage multi-view capsule fusion network (PIFCapsule) is proposed. Firstly, a signal processing prior-empowered first-layer interpretable module is devised to realize automatic parameter optimization and highlight the complementarity between multi-view features from different signal processing algorithms. Secondly, an interpretable capsule network serves as the backbone. To overcome the inefficiency and shortage of information fusion, an efficient attention fusion routing (AFR) is proposed to reduce the parameters (about 5.72 times) and the complexity (about 2.93 times) in contrast to the vanilla capsule-based networks. In response to the lack of physics-based constraints during training, a noise threshold amplitude ratio (NTAR) is posed as a regularization, which enhances weak periodic transient pulses by suppressing learned noises. The effectiveness and reliability are verified through three real-world rail transit vehicle datasets: PIFCapsule outperforms the state-of-the-art by 6.77% in accuracy with only ten samples. Given the lightweight nature, it holds substantial promise to be deployed in intelligent edge devices. Code is available at https://github.com/liguge/PIFCapsule.

1. Introduction

The safety of railway transportation is closely intertwined with daily routine and production activities. This study focuses on monitoring and maintenance technologies for high-speed, heavy-haul freight, and subway trains. The running gear of vehicles is a vital subsystem. Bearings, key components of the running gear, directly impact safety and reliability. However, they face challenges such as high loads, high speeds, strong noises, and complex operating conditions, resulting in failure modes like pitting, spalling, and fracture.

Embodied Intelligence (EI) serves as the guiding principle for realizing intelligent operation and maintenance of rail transit vehicles. EI integrates perception, cognition, and action into manufacturing systems, enabling a on-device ‘‘perception-cognition-execution-feedback’’ closed loop [1]. Prognostics and Health Management (PHM) platforms have also become standard equipment in current rail transit vehicle maintenance. Within the PHM system, fault prediction and diagnosis of critical components is a crucial link. Through early warning information generated by fault prediction and diagnosis models, health status assessment models automatically evaluate the condition and formulate the scientific maintenance decisions, thereby effectively implementing condition-based maintenance plans. Subsequently, the maintenance level and effectiveness are assessed based on post-maintenance monitoring data, forming a closed-loop process of “data collection - fault prediction - health assessment - maintenance decision-making and execution - effective feedback”, which ensures continuous health management throughout the entire lifecycle. Fault prediction and diagnosis constitutes a vital component of the intelligent operation and maintenance system.

☆ This article is part of a Special issue entitled: ‘EI - Smart Manufacturing’ published in Journal of Industrial Information Integration.

* Corresponding author.
E-mail addresses: chaohe@bjtu.edu.cn (C. He), hmshi@bjtu.edu.cn (H. Shi), jingxiaoliao@hit.edu.cn (J.-X. Liao), b_liu163@163.com (B. Liu), 25110780@bjtu.edu.cn (Q. Liu), jbli@bjtu.edu.cn (J. Li), zjyu@bjtu.edu.cn (Z. Yu).

https://doi.org/10.1016/j.jii.2026.101068
Received 4 August 2025; Received in revised form 8 January 2026; Accepted 12 January 2026
Available online 6 February 2026
2452-414X/© 2026 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

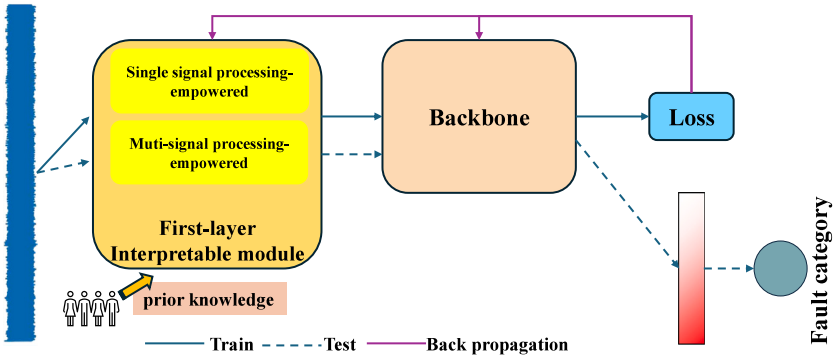

Fig. 1. The first-layer interpretable paradigm.

In fault prediction and diagnosis algorithms, some algorithms suitable for common rotating machinery fail to achieve satisfactory performance. Despite these challenges, intelligent fault diagnosis (IFD) utilizes advanced signal processing and data-driven algorithms, and can achieve precise fault identification and localization, thereby enhancing the safety and reliability of railway transportation [2–5]. To date, various algorithms have been developed for these specialized running gears [6–12]. These approaches cover fault mechanisms, signal processing, traditional machine learning, and deep learning. Deep learning, with excellent generalization capabilities, has effectively addressed the challenges of complex fault localization. However, existing algorithms lack explicit interpretability and fail to incorporate human prior knowledge. This poses significant challenges to realize Customized Manufacturing and Embodied Intelligence (CMEI) [1], which plays an irreplaceable role in the intelligent monitoring paradigm of rail transit vehicles.

Interpretable IFD with human prior knowledge has garnered widespread attention [13]. While computers can learn the statistical patterns in data, it is uncontrollable and may fail to align with human expectations. Intrinsic interpretability aims to design transparent and understandable mappings grounded in domain-specific priors. Therefore, human prior knowledge is embedded into a system, and the optimization process is constrained by the interpretable theory and physics laws [14]. In human–computer collaborative monitoring, humans are responsible for offering prior knowledge, while computers are responsible for learning the data distribution pattern in accordance with human needs. FLIP represents a typical intelligent paradigm of smart equipment. It embeds signal processing prior knowledge into deep models to address complex industrial problems.

As shown in Fig. 1, FLIP is a simple yet effective approach. FLIP is defined as a framework that either modifies or replaces the first-layer weights using prior knowledge, or incorporates a differentiable signal processing layer at the front of backbones to simulate signal processing operations. FLIP can also be called a signal processing-empowered or physics-based paradigm. It primarily encompasses two approaches. One approach constructs interpretable prior layers using domain-specific knowledge, such as signal processing, to either replace or connect the initial backbone, such as wavelet convolutional layers [15], differentiable short-time Fourier transform (STFT) layers [6], and multiplicative convolutional layers [16]. Another approach entails assigning prior knowledge-based initialization weights to the unaltered first-layer structure, such as wavelet weight initialization [7] and cyclo-stationarity weight assignments [17].

FLIP abandons complex architectures and training strategies, and simply relies on human prior knowledge to reduce the dependence on data quality and quantity. However, according to the time–frequency uncertainty principle (Gabor limit), the time–frequency resolution is difficult to fully express the weak fault features. This is because these algorithms rely solely on a single view, which reflects the local state of signals and is susceptible to signal distortion and ambiguity. Inspired by multi-view learning [18], the features of vibration signals can be described from a variety of different signal processing techniques, which constitute multiple views of raw signals. Apparently, the physics-based multi-view fusion can capture the different features, and provide a more reliable basis for subsequent fault decision-making [19,20]. It addresses the interpretability while mitigating the unreliability and ambiguity inherent in any single view. To the best of our knowledge, multi-view physics-based feature fusion remains little attention.

In addition, another problem with FLIP is that CNN and Transformer-based backbones are often non-interpretable. Capsule network (CapsNet) [21] provides interpretability for complex IFD. Capsules contain multiple feature vectors, which can capture finer changes in physics-based time-frequency representations, especially the precise position, tilt, and size parameters of the energy bands, while preserving spatial information. CapsNet is challenging to be directly applied as the backbone because of the dynamic routing mechanism. It has a large number of parameters, and it can obscure the intrinsic generalization capability, leading to lower efficiency. Beyond the fault diagnosis field, although attention-based routing [22,23] can achieve non-iterative routing and reduce parameters, most solutions adopt the Softmax attention mechanism, which faces huge computational complexity. Moreover, when faced with physics-based multi-view information, non-iterative Softmax attention routing [24] requires additional bottlenecks to achieve feature fusion, increasing parameters and weakening the efficiency. Therefore, it is crucial to develop a high-performance, efficient and low-complexity dynamic routing algorithm with physics-based multi-view feature fusion capabilities.

Finally, some first-layer interpretable algorithms overlook the critical role of physics-based constraints. As a regularization term, it can alleviate overfitting and enhance robustness. Therefore, it is essential to incorporate the physics-based loss into the paradigm. To our knowledge, integrating physics-based loss is not a novel concept; its effectiveness has been demonstrated in numerous studies [25,26]. However, loss functions should be tailored to specific tasks and target systems rather than simply copying existing metrics. Conversely, inappropriate physics-based losses can degrade performance and reliability. This necessitates considering not only the first-layer structure and backbone but also the target application when selecting an appropriate loss function. In this work, the key components under consideration include wavelet layer, STFT layer, blind convolution (BD) layer, CapsNet, and bearings of rail transit vehicle running gears.

In summary, the current research deficiencies on the first-layer interpretable paradigm can be summarized as the following fourfold problems:

- As highlighted in Ref. [6], many so-called "physics-based" methods adopt a two-stage strategy–requiring preprocessing prior to model input. However, it resorts to precisely handcrafted hyperparameters to capture discriminative features.

- The single physics-based information is not only difficult to balance the time and frequency resolution because of Gabor limit, but also can only reflect partial signal characteristics, and is susceptible to various factors, leading to ambiguous and distorted information.
- In interpretable capsule network, vanilla or Softmax attention routing is computationally intensive and parameter-heavy, which impair internal knowledge transformation and lead to poor generalization. Moreover, routing mechanism is inherently incapable of performing multi-view feature fusion.
- For one thing, some first-layer interpretable methods overlook the critical role of physics-based constraints. For another thing, although some physics-informed metrics exist, their blind adoption – without considering the model architecture, equipment and application scenarios – can degrade the performance and reliability.

Driven by the aforementioned motivations, a one-stage physics-based multi-view feature fusion CapsNet is proposed. In bearing diagnosis, vibration signals are highly sensitive to faults and adequately reflect fault features [27]. First, to address the limitations of single-view information, we employ three complementary physics-informed views to devise signal processing-empowered fusion model: wavelet, STFT, and blind convolution. BD [28] is adept at extracting periodic transient pulse components, whereas STFT and WT exhibit relatively limited capability in this regard [29]. STFT, limited by the fixed window function, lacks adaptive adjustment and has relatively low resolution. In contrast, wavelet transform (WT), offers multi-scale analysis and higher resolution but may introduce aliasing and potential signal distortion [30]. By processing the same signal input from three different views, we obtain complementary multi-view features to design the first-layer interpretable module.

Notably, the prior STFT layer uses complex structural simulations [6,9] and could not integrate into standard convolution. Here, we treat the channel as time and the last dimension as frequency resolution, embedding into 1D-CNN to boost performance and reduce complexity.

Second, CapsNet faces issues like large dynamic routing parameters, weaker generalization, and inability of multi-view fusion. To resolve these, we propose a highly parallel AFR that fuses the three physics-based views. This algorithm enhances efficiency and reduces parameters. As far as we know, attention mechanism for feature fusion in fault diagnosis exists [31], but the self-attention mechanism has not been fully explored. Unlike traditional attention, AFR is a high-level variant of the self-attention mechanism and enables cross-view interaction, and establishes global-local dependencies within modalities. It adaptively adjusts attention weights according to tasks. In addition, norm-activation operation and depthwise separable convolution is introduced to refine the self-attention mechanism, emphasizing information-rich features and improving multi-view fusion.

The physics-based loss guides optimization and improves interpretability. For collected signals of the rail transit vehicles, we propose a novel loss function – the noise threshold amplitude ratio (NTAR) – defined as the ratio of noise intensity to the root mean square. By reducing learned noise, it amplifies discriminative fault-related features. In NTAR, noise level is an important parameter, and it is usually necessary to know the fault period, which is difficult to obtain in advance. To avoid it, we employ deep learning to automatically learn soft thresholding, treating features below thresholdings as noise.

This paper presents several significant contributions to the field:

1. Guided by human–computer collaboration, a one-stage physics-based multi-view fusion capsule network (PIFCapsule) is proposed, which covers three signal processing-empowered first-layer designs, an attention fusion routing mechanism, physics-based regularization terms. It has improved the interpretability of fused features.
2. The attention routing fusion mechanism is introduced into CapsNet. This mechanism significantly reduces the parameter count of vanilla dynamic routing while improving generalization. Through it, PIFCapsule fully considers the global dependencies among the physics-based multi-view information, adaptively adjusts attention weights, captures long-range dependencies, increases feature diversity, and ultimately enables efficient cross-view information interaction.
3. We introduce the noise threshold amplitude ratio as a physics-informed regularization term. This approach can suppress noise using a differentiable soft thresholding module without relying on fault prior periods. By weakening noise, it enhances signal impacts. As a regularization term, it constrains optimization of model, enhances robustness, and extracts discriminative features.
4. Comprehensive results from the running gear datasets of high-speed trains, heavy-haul freight trains, and subway trains demonstrate the superior performance. Extensive ablations validate the effectiveness of each component and structure. This method shows broad application prospects in the rail transit vehicle bearing health monitoring.

Subsequent sections are organized into the following structure. We defines the main problem, presents the motivation and review the prior efforts. The various components of PIFCapsule are elaborately outlined in Section 3. Section 4 incorporates multiple empirical trials and assessments to emphasize the inherent advancements. Finally, Section 5 conducts a thorough interpretability analysis. It concludes in Section 6.

2. Related work

2.1. Definition of physics-based multi-view information

Multi-view feature fusion is a method to improve performance by utilizing complementary information from multiple views or features [18,32]. For the definition of physics-based multi-view information, different researchers may have different opinions. These views can be different features of the same dataset. In this paper, different signal processing techniques are considered to be different views.

Although these three signal processing methods all target vibration signals, they emphasize distinct physical characteristics: the wavelet layer uses wavelet weights and time-view convolution to obtain the time-frequency spectrum from a time-view perspective; the STFT layer uses window weights and frequency-view multiplication to obtain the frequency-view representation; and the BD layer does not require preset filter parameters, but only uses optimized weights by NTRA to obtain the time-view representation.

Although the adaptive filtering weights guided by prior knowledge can achieve a relative balance between time and frequency resolution, it still cannot solve the Gabor limit. To further alleviate this limitation, different signal processing methods focus on different physical aspects, and then achieve global and local feature fusion through capsule fusion network. The fused features are three different physical views, beyond what the raw vibration signal alone can provide.

2.2. The human–computer collaborative intelligent paradigm in fault diagnosis

In the field of industrial diagnosis, human–computer collaboration is emerging and warrants re-examination within the overarching framework of Circular Embodied Intelligence Manufacturing. Within it, the intelligence paradigms of smart equipment – encompassing key capabilities such as Prognostics and Health Management and human–computer collaboration – serves as a core component of Customized Manufacturing and Embodied Intelligence, providing systematic support for fault diagnosis. Refs. [33,34] mainly add human intervention

in the decision-making process. If the fault confidence level given by the algorithm is low, the human needs to make the final decision based on other indicators and experience. Wen et al. [35] introduced the fault causal knowledge graph and used a large language model to realize the human–computer collaborative decision. Li et al. [36] also borrowed this thinking and sent the suspicious results to human for secondary judgment.

The above scheme only introduces human knowledge in the decision-making stage to make the final decision [19], while ignoring the guiding role of human knowledge for computer decision-making. As Kim said [37], humans excel at developing prior knowledge from experience while machines are good at calculating data. Therefore, this paper interprets human–computer collaborative diagnosis from different angles, classifies the interpretable first-layer paradigm embedded with signal processing prior knowledge as one of the typical human–computer collaborative paradigms, and solves the limitations of FLIP.

2.3. The first-layer interpretable bottlenecks guided by customized manufacturing

The proposed first-layer interpretable paradigm belongs to the broader class of signal processing-empowered technology. It modifies and optimizes deep models using signal processing, thereby enhancing interpretability. Inspired by it, WaveletKernelNet [15] integrates continuous wavelet transform into convolution. Since then, research on the wavelet kernel has garnered significant attention [7,38,39]. SincNet [40] highlights the significance of the first-layer module for extracting features. The physics-based convolutional neural network (PCNN) [41] generates convolution kernels based on bearing speed and fault feature frequency. However, PCNN can only identify single fault. M-PINet [42] combines multiple PCNN branches for multi-fault recognition but suffers from high computational complexity. M-IPISincNet [43] utilizes band-pass filters and inverse Fourier transform to design convolution to extract multi-scale information but relies on only a single source of physics-based information. PICNN [44] designs a physical feature weighting layer based on different feature frequencies corresponding to different faults. It assigns higher weights to the feature frequencies of a certain fault and reduces weights for other frequencies to resist interference. Liu et al. [45] treats the finite impulse response filter as a kernel, considering the center frequency and bandwidth as polynomials of the frequency, developing a first-layer interpretable model [46]. The Multiplication Convolutional Network (MCN) [47] consists of a series of multiplicative filtering cores that extract differential patterns from spectrum samples. The time-frequency network [48] utilizes the STFT, Chirplet, and Morlet wavelets to simulate time-frequency transform, acquiring abundant features [49]. Although the random convolution layer [50] does not emphasize interpretability, it randomly changes the kernel sizes to generate new data and keeps global information, while improving sample diversity and model generalization. Envelope spectrum neural network [51] guides the design of the first-layer module according to integrated empirical mode decomposition, extracting physical features. Although the aforementioned methods have achieved promising results, they are inherently limited by the unreliable single information, which is susceptible to uncertainties. This limitation results in an incomplete analysis, which in turn diminishes the robustness of model.

2.4. Capsule network in fault diagnosis

CapsNet has extensive applications in IFD [52]. Some reports [53] have employed spectrograms as the input, and depict the heatmaps from post-hoc explanations. However, this method is high-computational costs and high-parameter quantities. In contrast, one-dimensional signals offer certain advantages. CapsNet also has the capability to address diagnosis under small samples. For instance, CapsFormer [54] integrates CapsNet with self-attention mechanism. CapsNet is used for feature extraction, while the self-attention mechanism focuses on the spatial and sequential aspects of spectrograms. Wang et al. [55] initialize capsules via knowledge-informed convolution and propose a spectrum template method to establish a mapping between capsules and fault types. Taken as a whole, most studies still focus on the key role of CapsNet in feature extraction, such as using the feature outputs by CapsNet for view adaptation to compute similarity metrics between source and target domains [56–58], or combining them with other neural networks [59–62]. However, the high-parameter dynamic routing mechanism lacks multi-source feature fusion capabilities. Moreover, ignoring physics-based information not only reduces the interpretability but also increases dependence on quality and quantity of data. Relying on a single data modality or view often yields ambiguous or unreliable diagnostic decisions, ultimately undermining generalization performance.

2.5. Physics-based loss function

The physics-based loss is a pivotal component in physics-based machine learning [63–65]. pyDSN [9] introduces a balanced spectrum quality metric to assess STFT spectrogram quality. Qin et al. [66] employs boundary loss to constrain dynamic model parameters, thereby minimizing frequency discrepancies between real-world and simulated data. ClassBD [67] optimizes blind deconvolution filters by utilizing kurtosis and l_2/l_4 norm. PGNN [68] considers the imbalance amplitude and phase angle of rotors as loss function. Jia et al. [69] extracts state characteristic frequencies and signal spectral energy to derive physical pseudo-labels and formulate physics-based loss. In conclusion, physics-based metrics are typically tailored to specific signal processing pipelines or equipment dynamics. So, the physics-based loss should be customized for different systems and algorithms.

2.6. Physics-based feature fusion

Multi-source feature fusion has been extensively studied in IFD [31,70]. However, multi-source physics-based feature fusion remains an emerging field, which is investigated by noteworthy works. MPINet [42] combines multiple physics-based blocks (PIBs) and achieves fusion through multi-branch feature concatenation. However, this approach requires training multiple PIBs. Similarly, M-IPISincNet [43] integrates current and vibration signals, designs physics-based convolutional layers using inverse Fourier transforms and bandpass filters, and extracts multi-scale features through multi-scale convolution. Nevertheless, the fusion methods in MPINet and M-IPISincNet are relatively simplistic, lacking full interaction between complementary information views and only considering single physics-based information. Sun et al. [71] introduce a physics-based multi-modal feature fusion network (PMFN) that combines acoustic or vibration signals with infrared images. PMFN integrates two attention mechanisms: frequency-view attention, which emphasizes modulation relationships under specific faults, and region attention, which highlights more significant features in spectrograms, which are fused through self-attention. However, the two-dimensional encoding relies on signal processing, the effectiveness of which heavily depends on manual parameter selection. Besides, the fusion process through two self-attention branches is high-computational, and PMFN also overlooks physics-based loss. Ying et al. [72] proposes a trustworthy fusion module based on the Dirichlet distribution to ensure the reliability of dynamic fusion. However, TMFEFN lacks physical information and requires a large amount of training data. PFCG-Transformer [73] proposes a physical framework that integrates acoustic and vibration signals, constructing consistency and uncertainty quantification losses by comparing the data distribution of the physical model with actual operating signals. Nevertheless, the acoustic-vibration fusion demands extensive manual experience

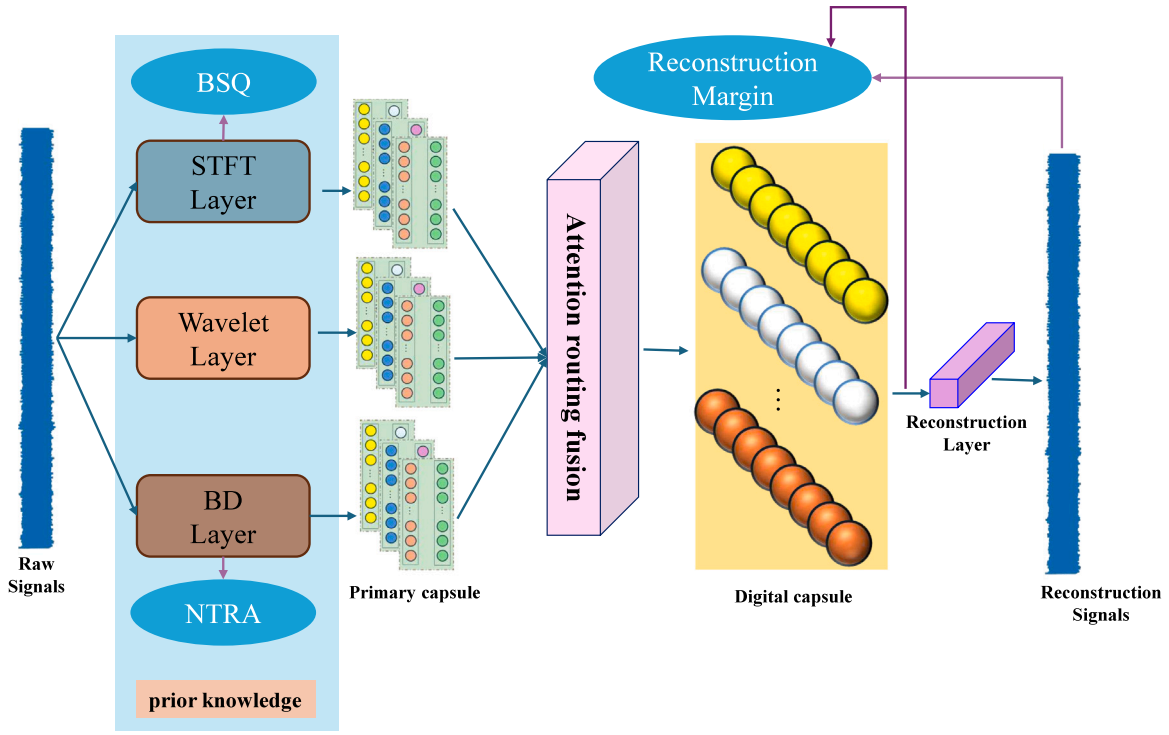

Fig. 2. The overall of PIFCapsule.

and involves complex computations, and collecting acoustic signals is challenging, and image signal processing is highly complex.

In practical bearing diagnosis, vibration signals are most sensitive to faults and can adequately reflect fault features [27,74]. PIFCapsule treats the signal processing-empowered weights as trainable, and finds them by data-driven algorithms, rather than relying on the manual parameter. Additionally, PIFCapsule achieves physics-based multi-view feature fusion through attention fusion routing mechanism, focuses on the most informative features using a norm-activation operation, and constrains the optimization direction using physics-based regularization.

3. The proposed method: PIFCapsule

3.1. The overview of PIFCapsule

The prior first-layer interpretable paradigm has been limited to single-view information, resulting in inefficient information interaction and neglecting the benefits of physics-based regularization. Therefore, this paper focuses on studying the physics-based multi-view feature fusion based on CapsNet within the framework of first-layer interpretable paradigm. As shown in Fig. 2, we propose an attention fusion routing mechanism that enhances cross-view information interaction efficiency and retains discriminative features. At the same time, we realize capsule routing, and AFR promotes the efficient information flow from the primary capsule to the digital capsule. Furthermore, a physics-based regularization is incorporated into the optimization process to enhance the physical fidelity of fused representations, yielding more generalizable fault feature embeddings.

The specific framework is shown in Fig. 3. The diagnostic workflow comprises the following steps:

1. Three types of rail transit vehicles damage datasets are collected: high-speed train traction motor bearing dataset, heavy-haul freight train wheelset bearing dataset, and railway train bogie dataset. The collected datasets are divided into training, validation, and test sets, with the training set size not exceeding 50 per class, consistent with the definition of a small-sample problem.
2. The collected training set data is input into PIFCapsule, and all physics-based losses are calculated to guide model training. The optimal model weights are saved after training.
3. In test stage, the optimal weights are loaded, and the performance is evaluated.

3.2. Multiple physics-based interpretable layers with signal processing prior knowledge

Single-source information suffers from poor reliability and ambiguous decision-making, whereas multi-source feature fusion can capture features with anti-interference capabilities and robustness. Therefore, multi-source feature fusion is an excellent solution. However, multi-source feature fusion remains constrained by poor interpretability. Thus, we propose a physics-based multi-view feature fusion framework. This scheme integrates STFT, WT, and BD information views: BD effectively extracts transient pulses from bearing faults, while STFT and WT are less capable. STFT has limited adaptability and resolution because of fixed windows. WT offers higher resolution and enables multi-scale analysis but may introduce aliasing and distortion.

3.2.1. Dual-damped Laplace wavelet weight initialization

Within the context of wavelet source information, the dual-damped Laplace wavelet is employed, which has not been previously explored in the realm of wavelet weight initialization. The fault pulse response observed in real-world vibration signals frequently manifests as a bilaterally asymmetric attenuation waveform, and this characteristic is precisely aligned with the inherent properties of the dual-damped Laplace wavelet [75].

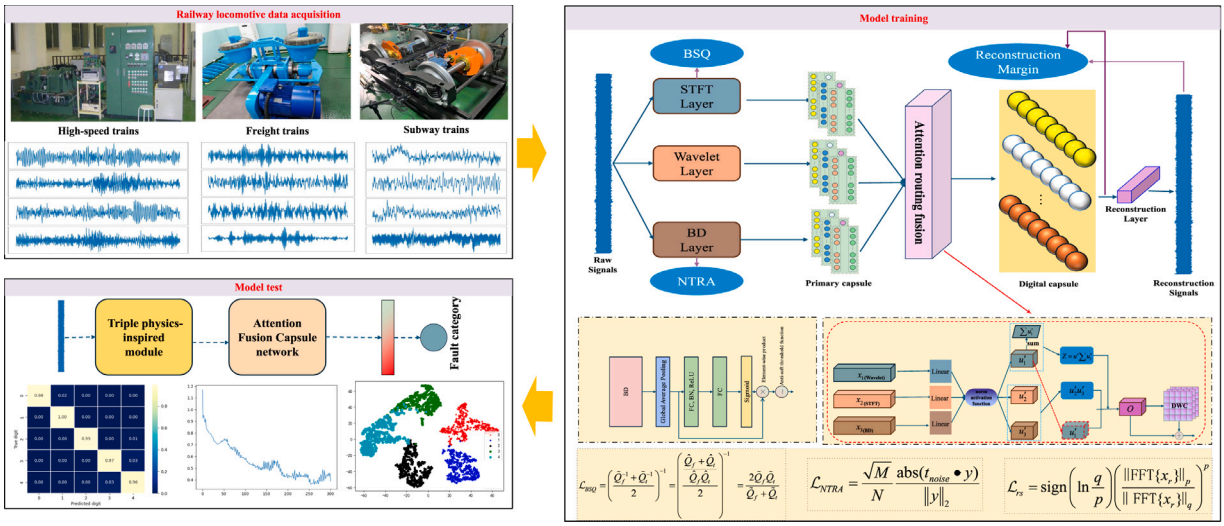

Fig. 3. The diagnostic process of the system.

The formula for the dual-damped Laplace wavelet is as follows:

$$\psi(\omega,\epsilon_1,\epsilon_2,t) = A_1 \cdot \exp\left(-\frac{\epsilon_1}{\sqrt{1-\epsilon_1^2}} \cdot \omega(t-\tau)\right) \cdot \cos(\omega(t-\tau)) - A_2 \cdot \exp\left(\frac{\epsilon_2}{\sqrt{1-\epsilon_2^2}} \cdot \omega(t-\tau)\right) \cdot \cos(\omega(t-\tau))$$
$$where \quad A_1 = \begin{cases} 1, & if \quad M > 0 \\ 0, & \text{otherwise} \end{cases}, \; A_2 = \begin{cases} 1, & if \quad M \leqslant 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\omega = 2\pi f$, f is the sampling frequency, t is time, A is the coefficient of the dual-damped Laplace wavelet, τ is the time constant, and ϵ_1 and ϵ_2 is the damping ratio.

Therefore, the final wavelet-view layer can be expressed as:

$$\begin{aligned} y_{wavelet} &= \psi_{u,s}(T) \odot x \\ &= \psi(\tfrac{t-u}{s}) \odot x \\ &= \psi(\omega,\epsilon_1,\epsilon_2,\tfrac{t-u}{s}) \odot x \end{aligned} \tag{2}$$

where s is scale factor, and u is translation factor. $\odot$ is represented as convolution operation.

3.2.2. Window weight initialization

Prior research has demonstrated that weight initialization methodologies exhibit superior performance compared with the design of signal-processing layers [7]. The signal-processing layer focuses on transforming key parameters into differentiable form. However, this approach suffers from substantial computational overhead, as the sliding window operation is not replaced by convolution. Inspired by weight initialization, we treat the channels of weights as time-view axes and the last dimension as the frequency-view axes. STFT kernels are used as the initialization weights for CNN to effectively simulate STFT. The window weight initialization is introduced:

Assuming the input signal is $\{x_t\}$, with a sampling frequency of f_s, the number of output channels of the filter is N, and the length of the filter is K. The defined bandwidth Δf is $\Delta f = \frac{0.1 f_s}{K}$. For each channel i, a filter h_i is set within the range $[\Delta f \cdot i + 0.01, \Delta f \cdot (i+1) - 0.01]$, and each filter h_i is designed using a window function, such as the Blackman window. By normalizing the cutoff frequency and considering the filter length, the ideal impulse response $h_{\text{ideal}}[n]$ is calculated, and the FIR filter response can be expressed as the product of $h_{\text{ideal}}[n]$ and the window function $w[n]$. To ensure that the initial weights are more closely aligned with random initialization, standardization is employed to restrict the range of the filters.

$$h_i = \text{standardization}(h_{\text{ideal}}[n] \cdot w[n]) \tag{3}$$

The principle of this operation is that each filter captures the frequency components of the raw signal, and treating the channels as sliding windows along the time axis captures the temporal variations of these frequency components to simulate the STFT:

$$y_{stft} = h_i \odot x \tag{4}$$

In STFT, a Fourier transform is performed on the windowed time period. To simulate this process, we use the Fourier convolution rather than the ordinary convolution:

$$y_{stft} = \mathcal{F}^{-1}\left\{\mathcal{F}\{x_t\} \cdot \mathcal{F}\{h_i\}\right\} \tag{5}$$

3.2.3. Blind convolution module

In the realm of BD, the time-view component is executed through the utilization of quadratic convolutional neural networks (QCNN), whereas the frequency-view component is handled via FFT. This process is further enhanced by integrating the kurtosis loss and the l_2/l_4 norm. In contrast, we design Noise threshold amplitude ratio, which serves to quantify the essential information within BD.

In contrast, QCNN engenders a considerable quantity of parameters is difficult to optimize [67]. Consequently, we opt for standard CNN, Analogously, we identify the resonance bands by diminishing the periodic noise within the time view and eliminating superfluous frequency components.

$$y_{BD} = Conv(Conv(x)) \tag{6}$$

In particular, the filters of $Conv(x)$ are also learned under the constraints of kurtosis loss and the l_2/l_4 norm metric.

3.3. Attention fusion routing mechanism

As illustrated in Fig. 4, the capsules in layer $l+1$ are predicted from all capsules U_i in layer l, which is achieved through a weight matrix W and the routing coefficient Z.

Given that the three information sources possess distinct physical properties, employing a single weight matrix W is insufficient for extracting highly complementary features. Therefore, three linear layers are utilized to accomplish this transformation. In these three information sources, the channels represent the time axis and the last

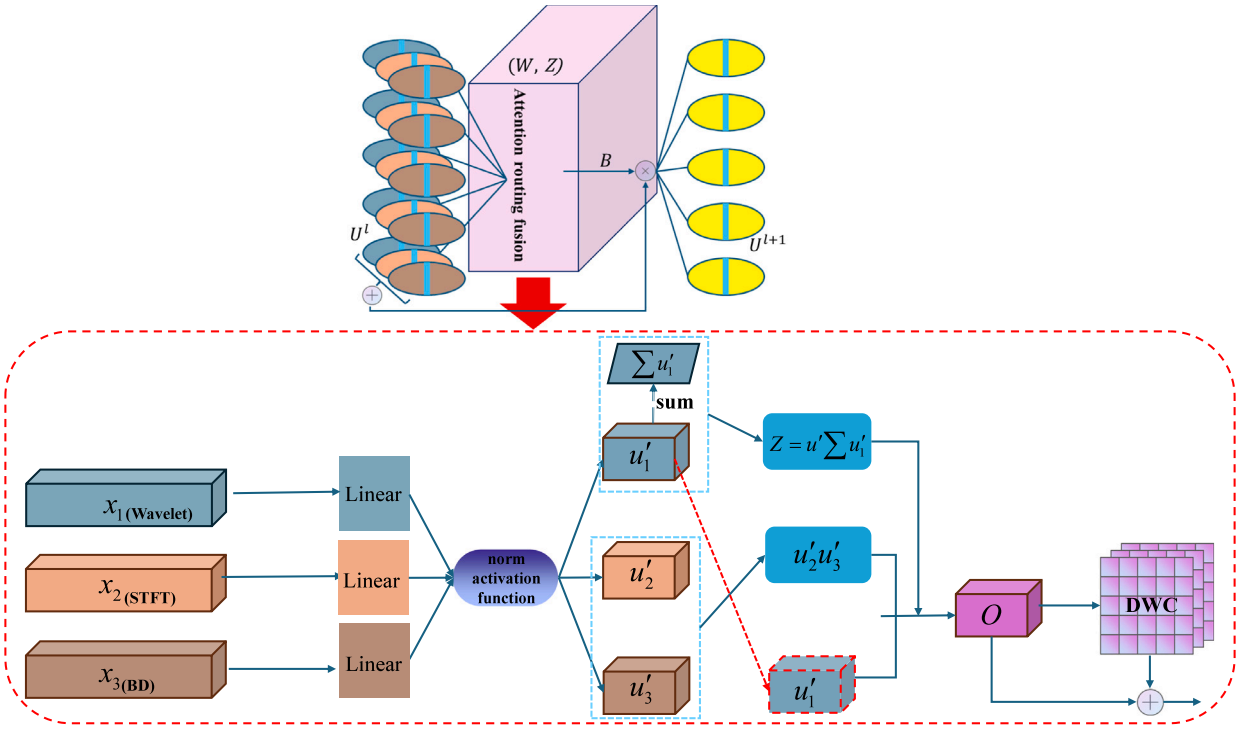

Fig. 4. The attention fusion routing mechanism.

dimension represents the frequency axis, which are collectively mapped to the new space:

$$\begin{cases} u_1 = FC(x_1) = FC(y_{wavelet}) \\ u_2 = FC(x_2) = FC(y_{stft}) \\ u_3 = FC(x_3) = FC(y_{BD}) \end{cases} \tag{7}$$

where FC refers to the full connection layer.

Subsequently, feature fusion is realized through norm linear attention mechanism.

The self-attention mechanism, which often employs the Softmax activation function and has a complexity of $O(N^2)$, is contrasted with the linear self-attention mechanism $O(N)$. The latter is considered an effective approach to reduce computational complexity, approximating the similarity function with a designed kernel function. In the process of feature fusion, calculating similarity helps adjust the fusion weights. Given such a kernel with the feature representation $\phi(\bullet)$:

$$sim(u_1, u_2) = \phi(u_1)\phi(u_2)^T \tag{8}$$

where $sim(\bullet)$ indicates similarity function.

According to the definition of the linear attention mechanism, norm linear attention can be expressed as follows:

$$O_i = \sum_{j=1}^{N} \frac{\phi(u_{1j})\phi(u_{2j})^T}{\sum_{j=1}^{N} \phi(u_{1j})\phi(u_{2j})^T} u_{3j} \tag{9}$$

Based on the properties of matrix associative property, the order of computation can be changed from $(u_1 u_2^T)u_3$ to $u_1(u_2^T u_3)$, that is:

$$O_i = \frac{\phi(u_{1j})(\sum_{j=1}^{N} \phi(u_{2j})^T u_{3j})}{\phi(u_{1j})(\sum_{j=1}^{N} \phi(u_{2j})^T)} \tag{10}$$

The Softmax attention mechanism provides a nonlinear weighting mechanism, thus easily focusing on the most important features. However, the output of the linear attention mechanism is closer to the average value and struggles to focus. Therefore, the norm activation function is employed as the kernel function to mitigate this drawback, namely:

$$u_i' = \phi(u_i) = \frac{\| |\mathrm{GELU}(u_i)| \|}{\| |\mathrm{GELU}(u_i)|^p \|} |\mathrm{GELU}(u_i)|^p \tag{11}$$

where $\phi(u_i) = f(|\mathrm{GELU}(u_i)|)$ and $f(u_i) = \frac{\|u_i\|}{\|u_i{}^p\|} u_i{}^p$. $\Phi(u')$ is the cumulative distribution function of the Gaussian normal distribution. $\|\cdot\|$ denotes the norm, p is used to control the degree of focus, generally $p = 2$.

Compared to the ReLU, GELU offers smoothness, which aids in gradient optimization, and maintains non-negativity for negative inputs, thereby preventing neuron deactivation. However, to meet the requirement of the denominator, the output must be kept non-negative, necessitating the use of absolute values. Finally, the direction of adjustment is regulated through feature norm mapping.

In reality, O_i acquires the global fused features of the three information sources through the self-attention mechanism, but local features are also required. To this end, depthwise separable convolution is adopted, which can be understood as a convolutional attention mechanism. It focuses on adjacent local information in the spatial view, thereby obtaining different outputs from each local region and enhancing feature diversity. As shown in Eq. (12):

$$O = \phi(u_1)\phi(u_2)^T u_3 + DWC(\phi(u_1)\phi(u_2)^T u_3) \tag{12}$$

Among them, $DWC(\cdot)$ represents the depthwise separable convolution operation.

The representation of the next capsules can be obtained through the attention weights:

$$U^{l+1} = \sum_{j=1}^{N} [U^l O] = \sum_{j=1}^{N} [(x_1 + x_2 + x_3)O] \tag{13}$$

AFR has two advantages. Firstly, it implements a routing mechanism in CapsNet, achieving non-iterative, high-speed parallel routing operations, which reduces model parameters and improves the efficiency of the algorithm. Secondly, unlike traditional attention, AFR supports cross-view interaction and establishes local and global dependencies (long-distance dependence) within the modality, and it has efficient information adaptability. It dynamically assigns features according to the task and focuses on key information, improving fusion efficiency.

Algorithm 1 Attention fusion routing mechanism

Require: x_1, x_2, x_3
Ensure: U^{l+1}
1: **while** epoch < max_epoch **do**
2: Calculate the feature mappings of different physics-based views u_1, u_2, u_3 by Eq. (7).
3: Calculate the kernel function feature expression u_1', u_2', and u_3' by Eq. (11).
4: The routing coefficient $Z = u_1' \sum u_1'$.
5: Achieve the norm linear self-attention fusion by Eq. (10).
6: Achieve global-local feature fusion by Eq. (12).
7: Implement the fused information transmission between primary capsules and digital capsules by Eq. (13).
8: **end while**

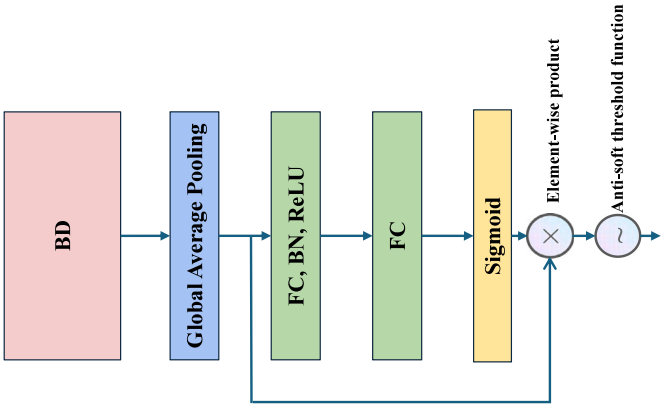

Fig. 5. Anti-noise soft-thresholding module.

3.4. Noise threshold amplitude ratio

Blind convolution serves to amplify repetitive impacts by searching for filters through maximizing the statistical metrics of signals. Fang et al. proposed the Periodic Noise Amplitude Ratio (PNAR), defined as the ratio of the average amplitude of noise points to the root square. As shown in Eq. (14):

$$\mathrm{PNAR}(\boldsymbol{y}, t_{\mathrm{noise}}) = \frac{\sqrt{M}\, t_{\mathrm{noise}} \cdot \mathrm{abs}(\boldsymbol{y})}{N \|\boldsymbol{y}\|_2} \tag{14}$$

where $\|\cdot\|$ denotes the Euclidean norm, M is the signal length, and N is the noise length. t_{noise} can be interpreted as the distribution of noise locations.

Clearly, t_{noise} is a pivotal parameter, and its value determines the effectiveness of PNAR. In PNAR, the noise distribution t_{noise} is continuous, and the noise length is determined by the noise ratio ρ and the period T, such that $N = \rho T$. It is evident that the assumption of continuous noise distribution is overly idealized. In reality, due to environmental interference, noise may exist throughout the entire period, with only intensity difference. Similarly, the predefined noise length is also impractical.

To rectify this limitation, the Noise Threshold Amplitude Ratio (NTAR) is proposed. We introduce a novel perspective, considering the noise position as the weight occupied by the noise—-noise coefficient. Since the noise is distributed throughout the entire period, the noise length is the same as that of signals($M = N$). Assuming that noise exists throughout the entire period, our goal is to determine the proportion of noise intensity to the total signal intensity. Thus, Eq. (14) can be reformulated into:

$$\mathrm{NTAR}(\mathbf{y}, t_{\mathrm{noise}}) = \frac{t_{\mathrm{noise}} \cdot \mathrm{abs}(\mathbf{y})}{\sqrt{M}\|y\|_2} \tag{15}$$

NTAR mainly obtains noises through the noise coefficient, facilitating the gradual reduction of noise to approach zero in the process of training. Inspired by wavelet coefficients, soft-thresholding [76] can also be utilized to reduce noise coefficients, where the thresholding λ is crucial. As shown in Fig. 5, we employ the attention-based architecture to automatically search for λ and retain the noise coefficient $S(y, \lambda)$, as shown in Eq. (16):

$$\mathrm{NTAR}(y, \lambda) = \frac{S(y, \lambda) \cdot \mathrm{abs}(y)}{\sqrt{M}\|y\|_2} \tag{16}$$

$$S(y, \lambda) = \mathrm{abs}[\mathrm{sign}(y) \cdot \min(|y| - \lambda, 0)] \tag{17}$$

where $\mathrm{sign}(y)$ is the sign function, and λ is the thresholding. Retaining the noise coefficient means preserving all values less than the thresholding.

The proportion of noises is reduced through NTRA to improve SNR, so as to extract more robust fault features.

3.5. The loss function

The margin loss L_c for class C is shown in Eq. (18). T_c is an indicator variable, where $T_c = 1$ if class C exists, and $T_c = 0$ otherwise. The lower bound $m^- = 0.1$, the upper bound $m^+ = 0.9$, $\|v_c\|$ denotes the norm of the output, and λ is a down-weighting factor, typically set to 0.5. The margin loss aims to output a true label distribution for each class while suppressing the activation of nonexistent categories.

$$L_c = T_c \max(0, m^- - \|v_c\|)^2 + \lambda(1 - T_c)\max(0, \|v_c\| - m^+)^2 \tag{18}$$

Secondly, the vanilla CapsNet incorporates a reconstruction loss to measure the similarity between the generated signals and the raw signals. However, our objective is to enhance the signal-to-noise ratio (SNR) rather than reconstructing the raw signals with low SNR. Consequently, $G - (l_p/l_q)$ norm is employed to assess the sparsity, which facilitates the construction of high-SNR signals, effectively suppressing noise and thereby bolstering robustness. This is represented in Eq. (19) as follows:

$$L_r = \mathrm{sign}\left(\log\left(\frac{q}{p}\right)\right)\left(\frac{\|x_{es}\|_p}{\|x_{es}\|_q}\right)^p \tag{19}$$

where $p = 2$, $q = 4$, and $\|x_{es}\| = \|\mathrm{FFT}(x_r)\|$.

In addition, both wavelet and STFT generate spectrograms. In our previous work, the balanced spectrum quality (BSQ) [9] is proposed to balance information from time-frequency views and enhance energy concentration. As mentioned in Section 1, wavelet and STFT information are complementary, with different sensitivities to different components of signals, which yields more sensitive energy-concentrated regions. Therefore, BSQ is used to measure the quality of spectrograms, as shown in Eq. (20):

$$L_b = \frac{1}{2} \times \left(\frac{2Q_f^w Q_t^w}{Q_f^w + Q_t^w} + \frac{2Q_f^s Q_t^s}{Q_f^s + Q_t^s}\right) \tag{20}$$

where Q_f and Q_t are the quality coefficients in the frequency and time views, respectively.

Subsequently, the NTAR (L_n) is employed to determine the optimal parameters for the BD filter. This process aims to amplify repetitive features while mitigating noise. Ultimately, the comprehensive loss function is expressed in Eq. (21) as follows:

$$L = L_c + \lambda_1 L_r + \lambda_2 L_b + \lambda_3 L_n \tag{21}$$

Determining $\lambda_1, \lambda_2, \lambda_3$ are of paramount importance. To tackle this challenge, we utilize uncertainty-aware weighted loss [67] to automatically balance the importance of each loss component for the learning problem and automatically search for the specific parameter configures.

4. Application to bearings of rail transit vehicles

4.1. The description of datasets and tasks

4.1.1. Data description of the real-world scenario

This study encompasses three datasets from rail transit vehicles, namely the high-speed train traction motor dataset, the heavy-haul freight train wheelset bearing dataset, and the subway train bogie gearbox dataset. A brief overview of these datasets (Table 1) is provided as follows:

$BJTU_1$ [9]: A dataset is acquired from the NTN traction motor bearing test bench, a specialized equipment with international advanced standards for conducting experiments on high-speed train traction motor bearings, as illustrated in Fig. 6(a). The equipment covers data and information collection and analysis of the dynamic train drive system, experimental verification on the ground platform, and the full life cycle. It serves as a specific experimental platform for realizing Prognostics and Health Management of traction motor bearings in high-speed trains. The test platform is composed of an electrical control cabinet, accelerometers, multiple test bearings, and signal transmission devices. Vibration signals of the bearings are collected through sensors and the signal transmission system. To align with actual application scenarios, the bearing model utilized for the experiment is the HRB NU214 EM 32214H, which shares identical dimensions with the bearings employed in real-world applications. The four mechanical states are included: Inner Fault (IF), Outer Fault (OF), Ball Fault (BF), and Normal (NC), with a sampling frequency of 100 kHz under 200, 250, 300, and 350 km/h.

$BJTU_2$ [77]: The dataset is sourced from the heavy-haul freight train wheelset bearing platform developed by CRRC Qingdao Sifang, as depicted in Fig. 6(b). This test bench is designed for fault simulation, performance testing, and remaining useful life assessment of various bearing models. It can simulate the load, speed, and corresponding environmental conditions of vehicle bearings during actual operation, thereby maximizing alignment with real-world railway scenarios. Specifically, the vertical load is used to simulate the axle load; the lateral load is employed to mimic different track conditions, such as lateral forces on the track during turning, passing through switches, and traveling up or down slopes; and fans are utilized to simulate crosswinds during operation, with the wind speed set between 8 m/s and 10 m/s. The experimental samples are faulty bearings confirmed through disassembly inspection after being removed from service—these bearings are initially identified as abnormal either by alarms from the online monitoring system of in-service railway freight cars or through manual inspection. Specifically, standard bearings are installed on one side of the wheelset, with experimental bearings mounted on the opposite side. A vibration acceleration sensor is affixed to the exterior of the axle box, operating at a sampling frequency of 16 kHz. The tested bearing is 352226X2-2RZ. The dataset comprises seven fault types: inner ring peeling (IRP), outer ring peeling (ORP), rolling element peeling (REP), rolling element crack (REC), cage fracture (CF), cage half crack (CHC), and compound faults (outer ring peeling + rolling element peeling) (CFs). Additionally, data from healthy bearings (H) is included. The bearing rotation speed is set within the range of 60 km/h to 180 km/h.

4.1.2. Public data description of the subway train scenario

$BJTU_3$ [78]: As depicted in Fig. 6(c), the experimental platform is designed based on the bogie of the subway train, with a scale ratio of 1:2 compared to the real-world bogie. Also, the sensor deployment on the experimental bench highly reproduces the real measuring points. The collected signal comprises 21 channels, sampled at a frequency of 64 kHz, considered six working conditions, with rotational speeds set at 20 Hz, 40 Hz, and 60 Hz, and lateral loads of 0 kN and 10 kN. Different speeds are employed to simulate various train speeds, while distinct Lateral loads were used to replicate straight-line driving and turning conditions. The gearbox of train transmission system is evaluated under nine different health states: Normal, gear cracked tooth(GCT), gear worn tooth(GWT), gear missing tooth(GMT), gear chipped tooth(GCPT), bearing inner race fault(BIR), bearing outer race fault(BOR), bearing cage fault(BC), bearing rolling element fault(BFE).

Table 1
Data description of the rail transit vehicle.

Dataset name	Data source	Speed	Sampling frequency	Health status
H	$BJTU_1$	150 km/h 200 km/h 250 km/h 300 km/h 350 km/h	100 kHz	N,IR,OR,BF
F	$BJTU_2$	60 km/h 100 km/h 120 km/h 140 km/h 180 km/h	16 kHz	IRP, ORP, REP, REC, CF, CHC, CFs
G	$BJTU_3$	20 Hz/0 kN 40 Hz/0 kN 60 Hz/0 kN 20 Hz/10 kN 40 Hz/10 kN 60 Hz/10 kN	64 kHz	GCT, GWT, GMT, GCPT, BIR, BOR, BC, BFE

Table 2
The parameter configures.

Parameter	batch_size	Optimizer	max_epoch	lr	weight_decay
Value	256	Adam	300	0.001	0.00001

4.1.3. Task description

Rotating mechanical equipment in rail transit vehicles encounters several challenges, including heavy load, high speed, strong noise, and complex operating conditions, compounded by small samples. This scarcity often hinders the training of a sufficiently generalized model. To address it, a multi-source physics-based feature fusion CapsNet is utilized, particularly under limited samples. This study, on the one hand, covers three distinct types of rail transit vehicles; on the other hand, it addresses various fault types and equipment, including single-point faults, compound faults, as well as gearboxes and axle boxes.

The signal is segmented into samples of 2048 using a sliding window. There is no overlap between sliding windows to ensure that no data leakage occurs. Based on previous definitions, if there are fewer than 50 samples per category, the problem is classified as a small-sample problem [6,79]. With small samples, the proposed method is sufficient to train a generalized model, enabling effective fault identification on the test set. In comparison, other mainstream methods exhibit relatively weaker performance in this regard. The training sample sizes are 5, 10, 20, 30, 40, and 50. The settings of training samples vary across different datasets and experiments, but all of them conform to the definition of small sample. The test set is uniformly set to 1000 samples per fault.

The experimental procedures are executed on PyTorch 2.5, leveraging two GTX 4090 24 GB GPUs. Each experiment is repeated five times to obtain the mean and standard deviation. The parameter configure is detailed in Table 2.

4.2. Case analysis

4.2.1. PIFCapsule performance with different data sets under different working conditions

As illustrated in Fig. 7, the robustness of the PIFCapsule model is evaluated across various datasets and operating conditions. Specifically, the sample size per class is set to 5 for Data H, and 10 for

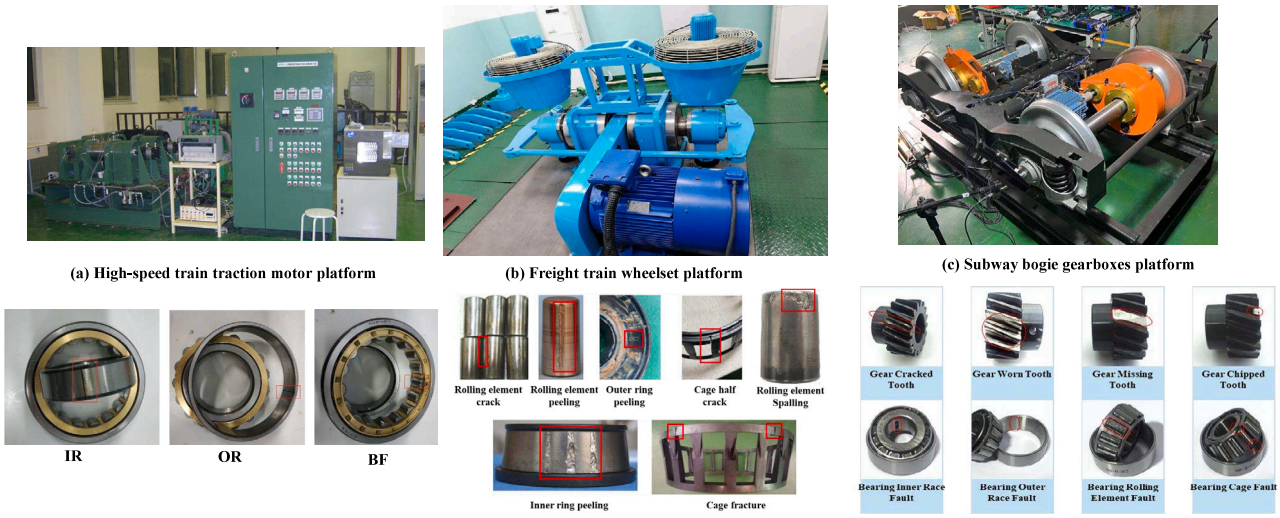

Fig. 6. Equipment: (a) high-speed train traction motor; (b) heavy-haul freight train wheelset; (c) subway train bogie gearbox.

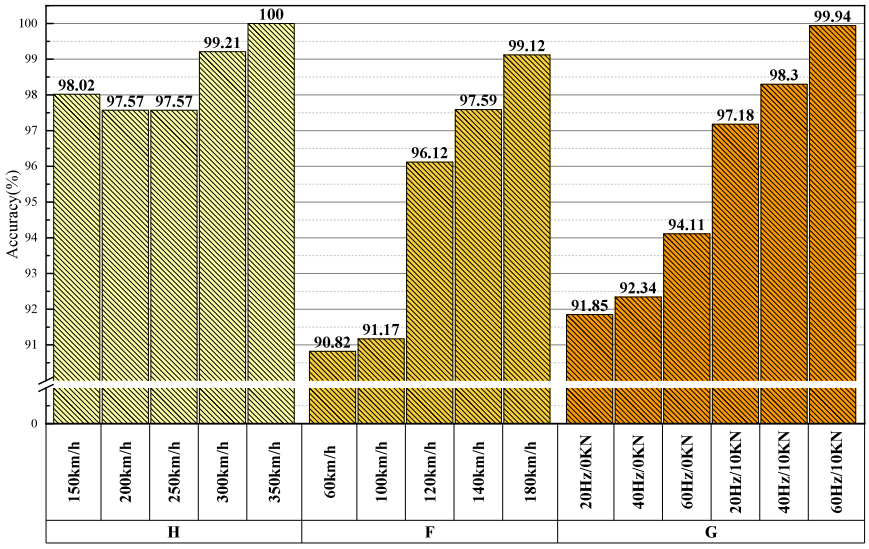

Fig. 7. Performance ability under different working conditions with small samples.

Data F and G. For the vehicles, the slower the rotational speed, the lower the accuracy. Compared to other conditions, under low-speed and high-load scenarios, the signal-to-noise ratio is lower, and noise interference has more significant impacts under small samples. (Subsequent evaluations are benchmarked against this most challenging condition.) Even under such a condition, with the assistance of modules such as physics-based information and attention fusion routing mechanism, PIFCapsule achieves an accuracy exceeding 90%, demonstrating robust performance.

4.2.2. Performance evaluation of PIFCapsule under different sample sizes

Consequently, we evaluated the capability of PIFCapsule in addressing small sample challenges. As illustrated in Fig. 8, the performance improves steadily with an increase in the sample size. Specifically, for the dataset pertaining to traction motors of high-speed trains, when the sample size reaches 10, PIFCapsule achieves 100%. This exceptional performance may be attributed to the high SNR. In contrast, for freight train wheelset or subway bogie, the SNRs are relatively low, resulting in poor-quality signals. Since the performance is contingent upon the quality and quantity of the data, it is imperative to devise methods for enhancing data/feature quality under small samples. To achieve it, PIFCapsule extracts fault feature frequencies from multiple views, thereby further reducing noise in the raw signals and identifying the primary bands. The single information view may not fully capture the characteristic frequencies. Therefore, the wavelet view, blind deconvolution view, and STFT view complement and reinforce each other, separately extracting multi-scale high-resolution features, periodic pulse components, and features that are less prone to aliasing and of high fidelity. Subsequently, these features are fully integrated through attention routing feature fusion and physics-based loss functions, enabling the model to learn more discriminative and high-quality features. As a result, even with 10 samples, PIFCapsule attains 91.54% and 95.62%, respectively.

4.2.3. Performance with different routing mechanisms

PIFCapsule is a capsule-based architecture, whose performance is compared with vanilla CapsNet and EfficientCapsule. EfficientCapsule employs the self-attention mechanism to replace the vanilla dynamic routing for vision. As shown in Table 3 and Fig. 9, PIFCapsule improves 27.35% and 24.21%. Notably, for Dataset F, the two models are below 50%. However, by leveraging physics-based multi-source information and attention fusion routing mechanism, PIFCapsule effectively mitigates noises and extracts valuable information, performing at over 90%. Additionally, PIFCapsule significantly reduces parameter number

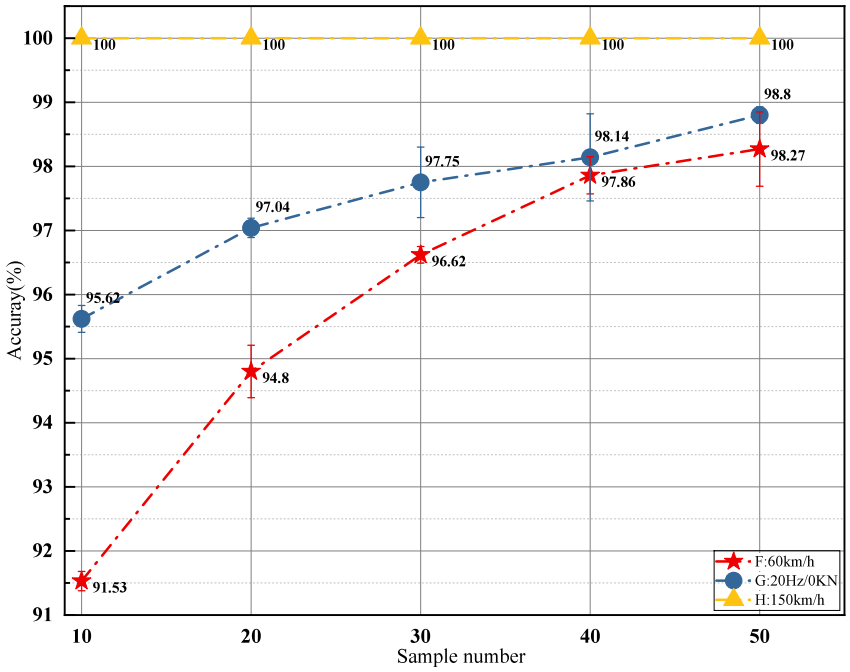

Fig. 8. Performance capability of PIFCapsule under small samples.

Table 3
Comparison of performance, parameters and computational complexity of three models.

Models	Parameters (MB)	FLOPs (M)	Accuracy
PIFCapsule	**0.4824**	**22.73**	**95.72%**
EffientCapsule [23]	2.7626	68.80	68.37%
Vanilla CapsNet	2.7624	68.80	71.51%

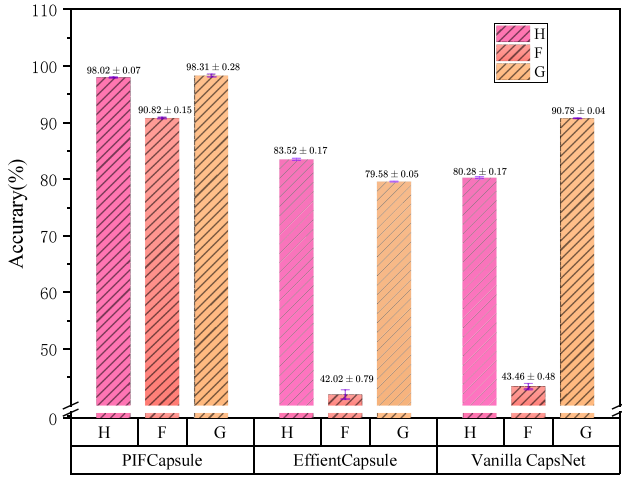

Fig. 9. The performance of three models.

and computational complexity. It reduces the parameters by 82% and the FLOPs by 66%, achieving a parameter size of 0.4824 MB and a computational cost of 22.73 MFLOPs, respectively.

The attention fusion routing is employed to compute routing coefficients, thus reducing repeated parameter updates during iterative processes. Additionally, a norm activation function is utilized to optimize attention weights, mitigating the high-parameter count and complexity associated with Softmax. Multi-view global feature fusion is directly implemented within the attention fusion routing module, eliminating the necessity for complex multi-branch convolutional fusion structures. Depthwise separable convolution is implemented within the attention fusion routing framework to achieve local feature fusion, reducing the computational complexity inherent in traditional convolutional operations.

Furthermore, by the confusion matrix and T-SNE as shown in Fig. 10, we analyze the performance differences among the three models. EfficientCapsule achieves only 68% for broken teeth, with confusion with bearing cage faults. The vanilla CapsNet fails to fully identify both normal and abnormal, with the accuracy below 62% for tooth surface wear and bearing cage faults. Instead, PIFCapsule integrates physics-based multi-view information and successfully fuses these features through AFR. To precisely extract robust physical features, a physics-based loss is employed to constrain the optimization direction.

4.2.4. *Comparison with other state-of-the-art (SOTA) models*

As presented in Table 4, we conducted a comprehensive evaluation of the performance comparison with SOTAs. The models encompass three categories. ① The fundamental models: ResNet, WDCNN, RCL, and MSResNet. ② The first-layer interpretable paradigm models: EWSNet, WaveletKernelNet, SincNet, FCC, MCNWFK, TFN, ClassBD, DFAWNet [81], GTFENet[82], wave_convNext[83], and RaVEL[84].③ The well-performing models under small samples: MCSwinT [85], Convformer_NSE[86], CLFormer[87], Liconvformer[88].

ResNet, WDCNN, RCL, and MSResNet are data-driven. They are significantly influenced by the sample size. Under small samples, it is impossible to obtain a sufficiently generalized model. Consequently, their average is all below 82%. Although ResNet exhibits relatively better performance, it has approximately 3.8 million parameters and high computational complexity. On the other hand, RCL performs random convolutional data augmentation on the input raw data. Compared with WDCNN, although it increases the computational complexity, it improves 7.62%. Overall, these models are constrained by the limited samples and fail to extract discriminative features, resulting in poor accuracies.

Furthermore, the first-layer interpretable models are based on signal-processing empowered algorithms such as wavelet, STFT, BD, and time-frequency transform to reduce the dependence on data quality and quantity. Compared with data-driven models, these algorithms have achieved improvements. Among them, EWSNet and RaVEL perform the best, with 86.51% and 86.79%, respectively. However, compared with PIFCapsule, their performances lag by 7.05% and 6.77%,

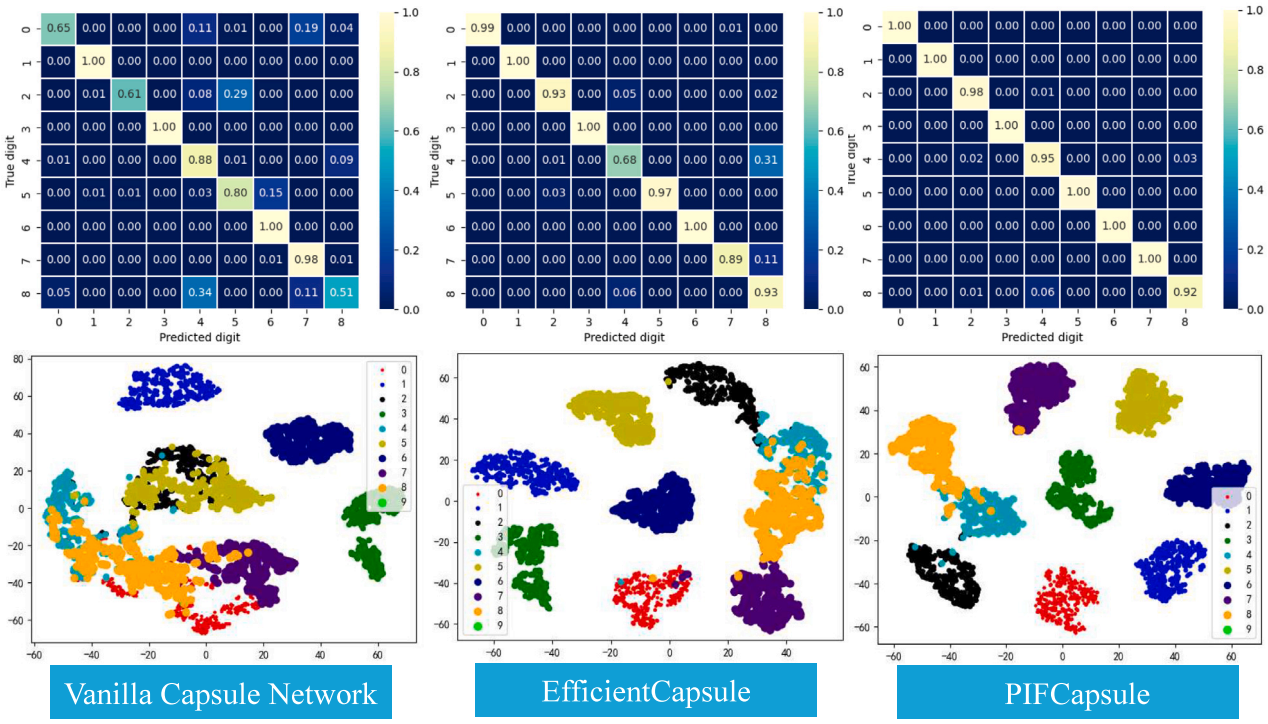

Fig. 10. Confusion matrix and T-SNE.

Table 4
Performance comparison with other methods (%).

	$BJTU_1$(H)	$BJTU_2$(F)	$BJTU_3$(G)	Average	Parameters	FLOPs (M)
ResNet	77.24 ± 0.06	78.31 ± 0.66	89.20 ± 0.41	81.58	3,853,321	351.38
WDCNN	47.08 ± 0.39	45.47 ± 0.59	91.43 ± 0.26	61.33	53,868	1.49
RCL [50]	82.87 ± 0.47	41.86 ± 0.51	82.11 ± 0.04	68.95	53,969	5.75
MSResNet [80]	72.57 ± 0.98	84.30 ± 0.78	90.35 ± 0.82	82.41	2,112,964	177.73
EWSNet [7]	88.63 ± 0.66	80.35 ± 0.25	90.56 ± 0.36	86.51	66,044	31.11
WaveletKernelNet [15]	80.95 ± 0.74	59.78 ± 0.11	84.35 ± 0.45	75.03	39,401	13.35
SincNet [40]	81.20 ± 0.82	82.66 ± 0.32	86.08 ± 0.20	83.31	39,592	13.45
FCC [46]	96.10 ± 0.12	56.88 ± 0.86	88.55 ± 0.24	80.51	237,259	97.44
MCNWFK [47]	91.89 ± 0.17	72.20 ± 0.16	71.38 ± 0.32	78.49	51,748	3.45
TFN [48]	–	–	–	–	185,128	86.70
ClassBD [67]	78.07 ± 0.74	80.17 ± 0.03	89.52 ± 0.38	82.59	8,094,014	711.42
DFAWNet [81]	84.91 ± 0.41	83.50 ± 0.10	87.29 ± 0.70	85.23	39,529	13.45
GTFENet [82]	82.32 ± 0.89	82.71 ± 0.02	90.96 ± 0.87	85.33	670,408	250.13
wave_convNext [83]	83.84 ± 0.11	47.33 ± 0.82	80.68 ± 0.86	70.62	35,481	3.38
RaVEL [84]	89.88 ± 0.63	84.69 ± 0.59	85.79 ± 0.33	86.79	47,064	0.70
MCSwinT [85]	93.73 ± 0.28	67.50 ± 0.01	89.16 ± 0.91	83.46	1,937,034	452.99
Convformer_NSE [86]	62.63 ± 6.65	64.60 ± 0.93	81.80 ± 0.64	69.68	245,005	12.39
CLFormer [87]	46.55 ± 0.36	38.56 ± 0.17	59.04 ± 0.12	48.05	**4991**	**0.13**
Liconvformer [88]	61.02 ± 0.67	72.31 ± 0.17	89.06 ± 0.47	74.13	322,263	28.79
PIFCapsule	**98.02 ± 0.07**	**90.82 ± 0.15**	**91.85 ± 0.88**	**93.56**	482,407	22.73

respectively. EWSNet mainly adopts the wavelets but overlooks the other physics-based information, and it employs a multi-scale structure, which extracts multi-scale wavelet features. RaVEL can be regarded as injecting wavelets into RCL. It improves the SNR during the data preprocessing stage, enhancing the quality of the dataset. As a result, RaVEL has a lower requirement for data, and is the extremely low computational complexity 0.70 MB. Compared with the aforementioned two models, PIFCapsule incorporates three physics-based information and successfully fuses them using AFR, achieving 93.56%.

Several lightweight Transformer-based models for small sample diagnosis are tested. Although MCSwinT achieves the highest accuracy of 83.46%, its parameter quantity and computational complexity are 4 times and 17 times those of PIFCapsule, respectively. CLFormer is sufficiently lightweight, but the average accuracy is less than 50%. PIFCapsule shows an improvement of 10.1% compared with MCSwinT. It indicates that even when facing the most popular Transformer-based architecture, the proposed CapsNet-based model is competitive.

4.2.5. Comparison with other state-of-the-art multi-source fusion methods

As illustrated in Table 5, we conducted a comparative analysis of several SOTA multi-source feature fusion models to evaluate the advantages of AFR. CSST_Net and MSIFT share similarities with PIFCapsule, as they all employ self-attention structures. CSST_Net primarily achieves data fusion by concatenation, while MSIFT utilizes a cross-attention mechanism to realize the feature fusion.

In this paper, CSST_Net performs data fusion by concatenating three features from the wavelet, STFT, and BD views. MSIFT is input the wavelet and BD views. As demonstrated by the ablation experiments in Table 6, these two views have the most significant impacts.

Table 5
A comparison with advanced multi-source feature fusion algorithms under ten samples.

Model	Fusion algorithm	G	F	Parameters	FLOPs (M)
CSST_Net [89]	Concatenation	86.79 ± 0.27	53.13 ± 0.33	657,192	169.61
MSIFT [90]	Cross-attention	94.64 ± 0.12	83.95 ± 0.52	1,185,144	1025.38
HSE_ResNet [91]	TAP+CDPF	53.45 ± 0.81	77.82 ± 0.11	3,891,944	716.67
PIFCapsule	AFR	**95.62 ± 0.21**	**91.53 ± 0.15**	482,407	22.73

Table 6
Function comparison of each module in PIFCapsule under dataset G.

STFT-based	BD-based	Wavelet-based	NTRA	BSQ	Accuracy (%)
✓	✓	✓	✓	✓	91.53 ± 0.15
	✓	✓	✓	✓	90.25 ± 0.91
✓	✓		✓	✓	48.50 ± 0.27
			✓	✓	61.35 ± 0.13
✓		✓	✓	✓	85.28 ± 0.26
✓	✓	✓	✓		89.59 ± 0.16
✓	✓	✓		✓	87.85 ± 0.15

The results indicate that PIFCapsule achieves improvements of 0.98% and 7.58% on different datasets. Although MSIFT exhibits competitiveness, its parameter count is 2.46 times that of PIFCapsule, and computational complexity is 45.11 times that of PIFCapsule. Consequently, the efficiency is far inferior to that of PIFCapsule. Besides, both CSST_Net and MSIFT utilize the relatively complex Transformer architecture without lightweight.

4.2.6. Ablation experiment

To assess the impact of different modules, as illustrated in Table 6, the improvements attributed to the STFT view, wavelet view, and BD view are 1.28%, 43.03%, and 6.25%, respectively. It is evident that the wavelet view exerts the most significant influence, indicating that multi-scale and high-resolution time-frequency features can capture the vast majority of fault features, thereby substantially enhancing the performance. The STFT view and BD view play supplementary roles, with the BD view particularly extracting periodic characteristics.

Furthermore, we investigated the influence of different metrics. The primary objective of BSQ is to enhance the energy concentration of the time-frequency representation, while NTRA is primarily employed to reduce noise and extract periodic impulse features. When BSQ is adopted, the performance improves by 1.94%. BSQ enhances the energy concentration of the extracted time-frequency representation, and accentuates the feature bands. The performance is boosted by 3.68% utilizing NTRA. NTRA primarily learns the noise in the periodic signals through the differentiable soft-thresholding module, thereby amplifying the periodic impulses and highlighting the fault feature frequencies. In summary, the two primary loss functions in this study complement each other, improving the quality of raw signals.

5. The interpretability analysis of PIFCapsule

5.1. The interpretability of the multi-view first-layer bottleneck

To verify the interpretability of the first-layer bottleneck, we consider the freight wheelset bearing as a case, where the fault frequency is predominantly below 1000 Hz. Based on the ablation experiment presented in Table 6, the wavelet view information is the most critical factor. Consequently, it will be on analyzing the wavelet weight view.

Initially, the amplitude frequency response $h(f)$ of the wavelet weights is computed [48]:

$$h(f) = \frac{1}{m}\sum_{i}^{m} \left| freqz(w_i) \right| \tag{22}$$

where w_i denotes the weight of each channel, and m represents the channel number.

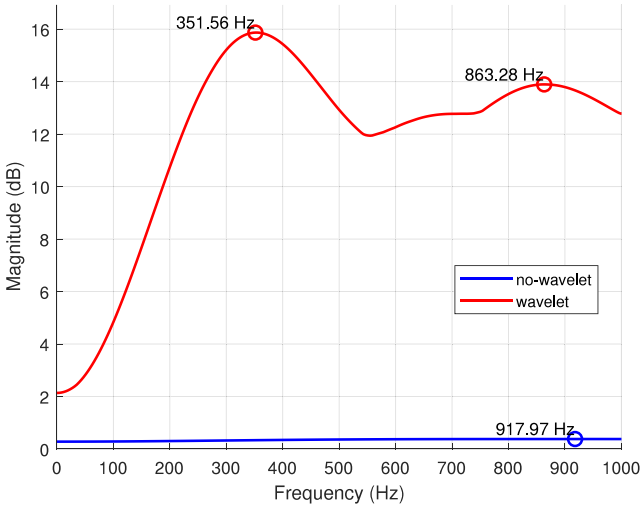

Fig. 11. The frequency response of the wavelet weights.

As illustrated in Fig. 11, the amplitude frequency response of the wavelet weights exhibits the highest peak at 351.56 Hz, that of the vanilla convolution approximates a straight line, which cannot reflect the physical characteristics of signals. Additionally, we conducted an analysis of raw signals using Fast Fourier Transform (FFT) across eight distinct states, as depicted in Fig. 12. Corresponding to the wavelet weights, the distribution of various states around 300 Hz is relatively dense. This observation indicates that the wavelet weights can discern differences between different states near 300 Hz and effectively extract the fault frequency features.

Furthermore, as demonstrated in Fig. 13, an analysis of the envelope spectrum amplitudes of the filtered signals by the BD layer corroborates this finding. It is evident that the amplitude attains its maximum around 300 Hz, thereby further substantiating that the mean frequency response of the wavelet kernel is primarily concentrated around 300 Hz. Concurrently, an interesting phenomenon was observed: as shown in panels (a), (e), (d), (h), the maximum or second amplitude points are located at the same frequency 304.96 Hz (289.06 Hz) and 664.06 Hz, which can lead to confusion and misclassification. Hence, additional view information – such as the wavelet or STFT views – is imperative to further distinguish fault types.

Finally, as depicted in Fig. 14, we analyze the frequency distribution from window weights across various channels. The frequencies in the initial-index channels are predominantly distributed around 300 Hz post-training, whereas these channels exhibit around 0 Hz before training, with no significant changes observed in other channels. This reveals three key phenomena: ① The window weights corroborate that the frequency distribution is concentrated near 300 Hz, and the discriminative features can be extracted. ② The window weight initialization approximates the optimal window weights, obviating the need for extensive adjustments. ③ In the initial-index channels, the filter is centered around 300 Hz, and different channels extract the primary features of raw signals at distinct frequencies, predominantly within the low-frequency range.

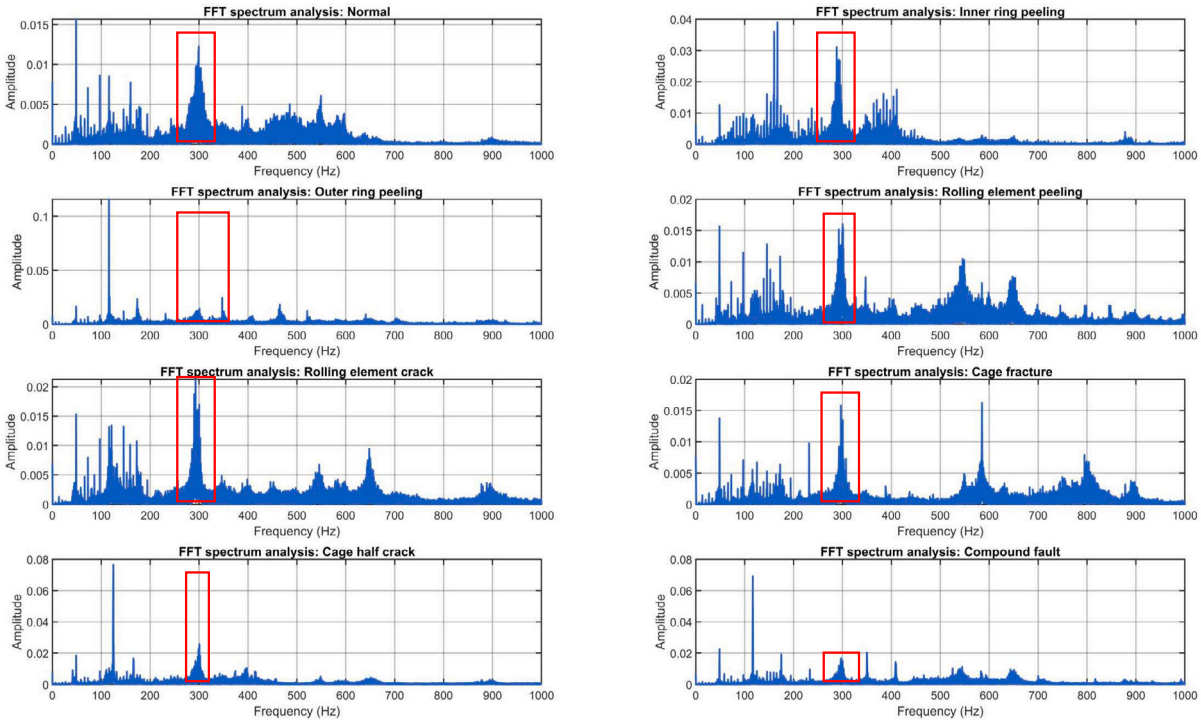

Fig. 12. FFT spectrum analysis of eight states.

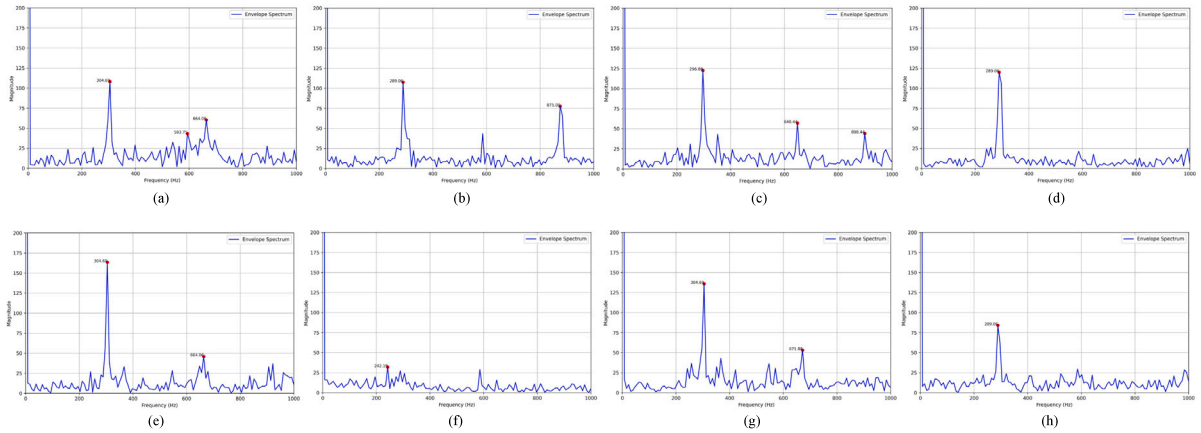

Fig. 13. Envelope spectrum analysis of the BD view output results.

5.2. The interpretability of the routing coefficient Z

With case study using Dataset F, in PIFCapsule, the low-level capsules are constructed from features extracted across three information views: wavelet, STFT, and BD. 16 groups of low-level capsules are utilized, which are subsequently aggregated into high-level capsules through an AFR mechanism, yielding 8 groups. To quantify the contributions of low-level capsules to high-level capsules, the following analysis is conducted.

As depicted in Fig. 4, the representation of low-level capsules can be expressed as m:

$$m = \sum_{k=1}^{16} x_1 \cdot KV \tag{23}$$

For any given low-level capsule i, the raw contribution can be represented as:

$$r_i = \|\mathbf{m}_i\|_2 \tag{24}$$

while the actual contribution can be expressed as:

$$a_i = \|\mathbf{m}_i \cdot Z_i\|_2 \tag{25}$$

where Z_i denotes the routing coefficient or attention weight associated with capsule i. PIFCapsule eschews pooling operations. Z is pivotal in facilitating the routing communication between low-level and high-level capsules, thereby enabling the fusion of three physics-based low-level capsules. Z offers insight into the advantages of PIFCapsule.

Consequently, the final heatmap is calculated as:

$$\text{heatmap}_{i,j} = \frac{a_{i\cdot 8+j}}{r_{i\cdot 8+j} + \epsilon} \tag{26}$$

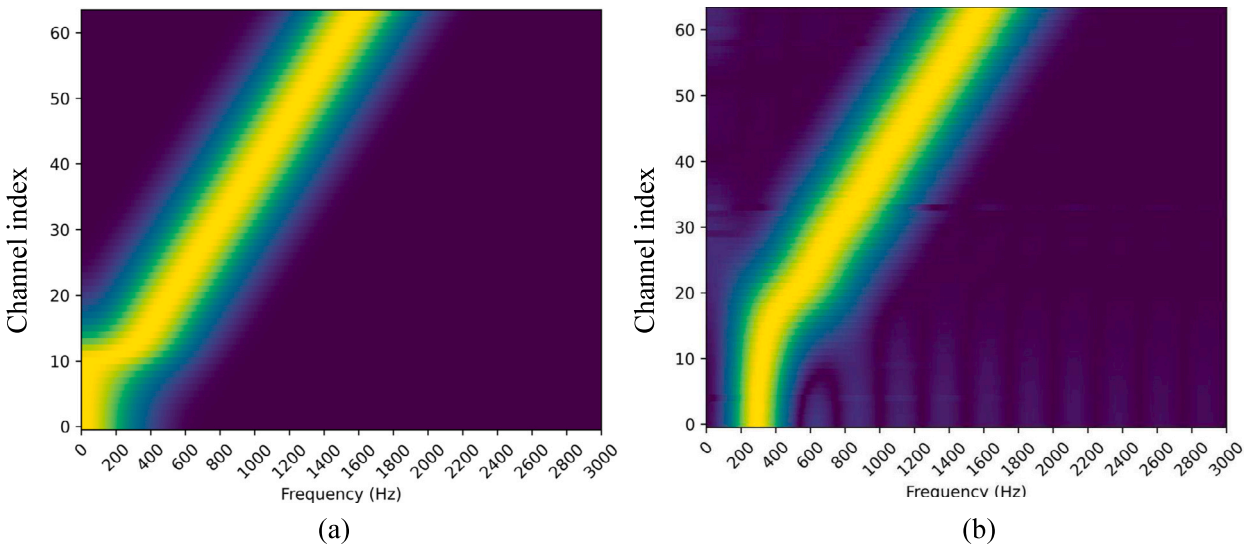

Fig. 14. The frequency distribution of window weights across various channels.

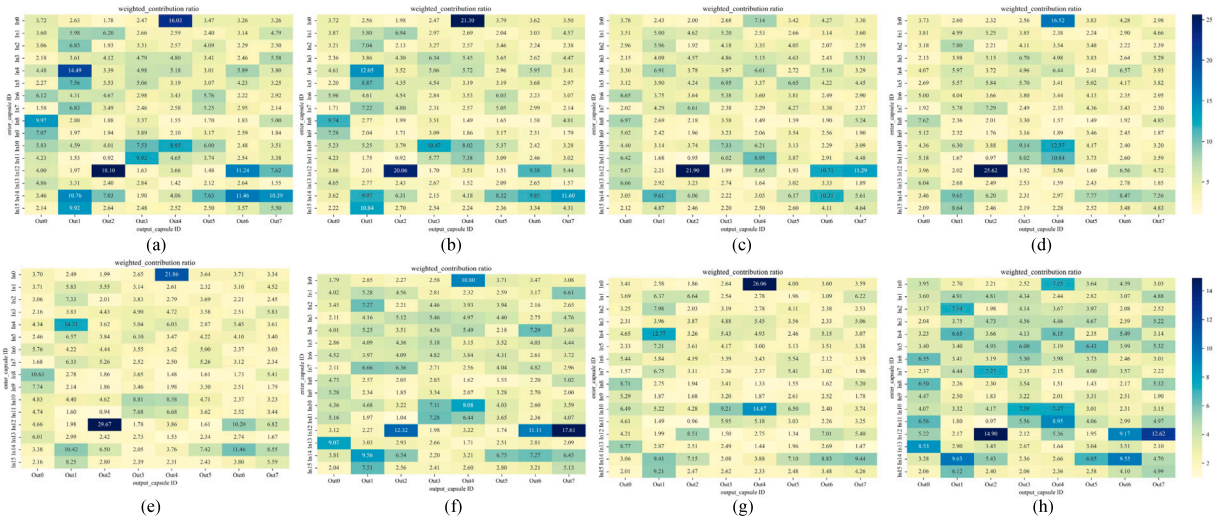

Fig. 15. Heatmap: mapping from lower-level capsules to higher-level capsules.

where ϵ is a small constant.

As illustrated in Fig. 15, the heatmap visually represents the contribution degrees of low-level capsules to high-level capsules. Although the contributions vary across different low-level capsules, they are positive, indicating that the low-level capsules, which integrate three physics-based information sources, make a positive contribution to fault localization.

6. Conclusion

Rail transit vehicle diagnosis is challenged by high rotational speeds, high loads, and limited samples. FLIP – an intelligent framework rooted in customized manufacturing and embodied intelligence – that systematically embeds human prior knowledge into smart diagnostic equipment. Building upon this paradigm, a multi-view signal processing priors embedded PIFCapsule is proposed. Firstly, a physics-inspired multi-view method, with wavelet transform, short-time Fourier transform, and blind deconvolution, is put forward to extract distinct physical properties and maximize the exploitation of complementary information across time, frequency and time-frequency. Secondly, a new non-iterative and computationally efficient attention fusion routing mechanism is proposed. AFR cannot only cope with a reduced number of capsules, improve intrinsic transformation capability of knowledge, and enhance generalization. Crucially, during the routing, AFR fully integrates the variations in complementary information, widen information space dimension and successfully captures implicit inter-view associations. Finally, a noise threshold amplitude ratio metric is proposed as a physics-informed regularization, which amplifies impact-related fault characteristics by weakening the learned noise.

However, several directions merit further investigation. Firstly, compared to some single-source physics-based information models, the parameter count and complexity of PIFCapsule are still relatively high. Further research is needed to achieve a more lightweight model while maintaining performance. Secondly, the features of PIFCapsule possess more physical interpretability and generalization, which hold potential for transfer learning across different machines. Thirdly, the first-layer paradigm can cover more types of signal processing and develop more signal processing-enabled convolution- or Transformer-based feature fusion modules. Finally, enhancing the performance of existing algorithms through Large language models (LLMs) [92] is an intriguing research direction. By leveraging LLM agents or assistants, researchers aim to further optimize the lightweight nature and performance of established algorithms. In the future, exploration can be conducted to deeply integrate the perception-action loop of

embodied intelligence with the reasoning capabilities of LLMs, so as to construct intelligent agents capable of active interaction, continuous learning, and autonomous decision-making. This will enable dynamic understanding, causal explanation, and adaptive diagnosis of faults in complex industrial systems.

CRediT authorship contribution statement

Chao He: Writing – original draft, Software, Validation, Visualization, Methodology, Investigation, Proofreading, Funding acquisition. **Hongmei Shi:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jing-Xiao Liao:** Methodology. **Bin Liu:** Writing – review & editing, Methodology. **Qiuhai Liu:** Writing – review & editing. **Jianbo Li:** Supervision. **Zujun Yu:** Supervision, Project administration.

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.

Acknowledgments

The authors are grateful for the supports of the Fundamental Research Funds for the Central Universities (No. 2024QYBS027), the National Natural Science Foundation of China (No. 52272429), and State Key Laboratory of Advanced Rail Autonomous Operation (Contract No. RAO2023ZZ003).

Data availability

The data that has been used is confidential.

References

[1] J. Tan, J. Shi, L. Wu, B. Chen, H. Tang, C. Zhang, W. Zhang, S. Wang, J. Wan, Embodied intelligence empowering customized manufacturing: Architecture, opportunities, and challenges, IEEE Access 13 (2025) 92740–92755.

[2] W. Hu, G. Xin, J. Wu, G. An, Y. Li, K. Feng, J. Antoni, Vibration-based bearing fault diagnosis of high-speed trains: A literature review, High-Speed Railw. 1 (4) (2023) 219–223.

[3] Y. Huang, W. Huang, X. Hu, Z. Liu, J. Huo, UDDGN: Domain-independent compact boundary learning method for universal diagnosis domain generation, IEEE Trans. Instrum. Meas. 74 (2025) http://dx.doi.org/10.1109/TIM.2025.3554906.

[4] X. Hu, X. Zhang, F. Chen, Z. Liu, J. Liu, L. Tan, T. Tang, Simultaneous fault diagnosis for sensor and railway point machine for autonomous rail system, in: 2024 IEEE 27th International Conference on Intelligent Transportation Systems, ITSC, 2024, pp. 1011–1016, http://dx.doi.org/10.1109/ITSC58415.2024.10920166.

[5] Y. Ye, H. Li, Q. Wang, F. Li, C. Yi, X. Peng, C. Huang, J. Zeng, Fault diagnosis of railway wheelsets: A review, Measurement 242 (2025) 116169.

[6] C. He, H. Shi, J. Li, IDSN: A one-stage interpretable and differentiable STFT domain adaptation network for traction motor of high-speed trains cross-machine diagnosis, Mech. Syst. Signal Process. 205 (2023) 110846.

[7] C. He, H. Shi, J. Si, J. Li, Physics-informed interpretable wavelet weight initialization and balanced dynamic adaptive threshold for intelligent fault diagnosis of rolling bearings, J. Manuf. Syst. 70 (2023) 579–592.

[8] J. Shang, D. Xu, P. Liang, C. Jiang, H. Qiu, L. Gao, Out-of-domain generalization for remaining useful life prediction of rotating machinery from a single source: An adversarial contrastive learning approach, Mech. Syst. Signal Process. 236 (2025) 112965.

[9] C. He, H. Shi, X. Liu, J. Li, Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis, Knowl.-Based Syst. 288 (2024) 111499.

[10] M. Abtane, K. Dahi, H. Martinez, M. Sedki, H. El Kimi, C. Dahhassi, C. Hlib, L.F. Borges, Axle bearing fault diagnosis for high-speed trains: A comprehensive review of methodologies, technologies, challenges and emerging trends, Measurement 251 (2025) 117098.

[11] J. Shang, D. Xu, H. Qiu, C. Jiang, L. Gao, Domain generalization for rotating machinery real-time remaining useful life prediction via multi-domain orthogonal degradation feature exploration, Mech. Syst. Signal Process. 223 (2025) 111924.

[12] X. Li, Y. Wang, J. Xing, Y. Wang, Causal graph inference with adaptive dynamic structure learning for mechanism-oriented fault diagnosis in dynamic industrial systems, Reliab. Eng. Syst. Saf. 266 (2025) 111865.

[13] J. Cação, J. Santos, M. Antunes, Explainable AI for industrial fault diagnosis: A systematic review, J. Ind. Inf. Integr. 47 (2025) 100905.

[14] D. Chen, H.J. Yoon, Z. Wan, N. Alluru, S.W. Lee, R. He, T.J. Moore, F.F. Nelson, S. Yoon, H. Lim, D.D. Kim, J.-H. Cho, Advancing human-machine teaming: Concepts, challenges, and applications, 2025, arXiv preprint arXiv:2503.16518.

[15] T. Li, Z. Zhao, C. Sun, L. Cheng, X. Chen, R. Yan, R.X. Gao, WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis, IEEE Trans. Syst. Man Cybern.: Syst. 52 (4) (2021) 2302–2312.

[16] R. Liu, X. Ding, Q. Wu, Q. He, Y. Shao, An interpretable multiplication-convolution network for equipment intelligent edge diagnosis, IEEE Trans. Syst. Man Cybern.: Syst. 54 (6) (2024) 3284–3295.

[17] D. Wang, Y. Chen, C. Shen, J. Zhong, Z. Peng, C. Li, Fully interpretable neural network for locating resonance frequency bands for machine condition monitoring, Mech. Syst. Signal Process. 168 (2022) 108673.

[18] C. Xu, D. Tao, C. Xu, Multi-view learning with incomplete views, IEEE Trans. Image Process. 24 (12) (2015) 5812–5825.

[19] Z. Xu, K. Zhao, W. Zhang, W. Miao, K. Sun, J. Wang, M. Bashir, Collaborative and trustworthy fault diagnosis for mechanical systems based on probabilistic neural network with decision-level information fusion, J. Ind. Inf. Integr. 46 (2025) 100854.

[20] C. Li, W. Xie, B. Zheng, Q. Yi, L. Yang, B. Hu, C. Deng, An enhanced CLKAN-RF framework for robust anomaly detection in unmanned aerial vehicle sensor data, Knowl.-Based Syst. 319 (2025) 113690.

[21] S. Sabour, N. Frosst, G.E. Hinton, Dynamic routing between capsules, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 30, 2017, pp. 3856–3866.

[22] J. Choi, H. Seo, S. Im, M. Kang, Attention routing between capsules, in: 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019, IEEE, 2019, pp. 1981–1989.

[23] V. Mazzia, F. Salvetti, M. Chiaberge, Efficient-capsnet: Capsule network with self-attention routing, Sci. Rep. 11 (1) (2021) 14634.

[24] K. Duarte, B. Chen, N. Shvetsova, A. Rouditchenko, S. Thomas, A. Liu, D. Harwath, J. Glass, H. Kuehne, M. Shah, Routing with self-attention for multimodal capsule networks, 2021, arXiv preprint arXiv:2112.00775.

[25] Y. Ruqiang, S. Zuogang, W. Zhiying, X. Wengang, Z. Zhibin, W. Shibin, C. Xuefeng, Challenges and opportunities of XAI in industrial intelligent diagnosis: Priori-empowered, J. Mech. Eng. 60 (12) (2024) 1–20.

[26] D. Weikun, K.T. Nguyen, K. Medjaher, G. Christian, J. Morio, Physics-informed machine learning in prognostics and health management: State of the art and challenges, Appl. Math. Model. 124 (2023) 325–352.

[27] W. Ying, Y. Li, K. Noman, J. Zheng, D. Wang, K. Feng, Z. Li, Stockwell transform spectral amplitude modulation method for rotating machinery fault diagnosis, Mech. Syst. Signal Process. 223 (2025) 111884.

[28] Z. Wang, Y. Guo, W. Kang, X. Chen, A novel fault feature extraction method for planet-bearing based on generalized total variation model, IEEE Sens. J. 25 (11) (2025) 20870–20879.

[29] S.S. Haykin, Adaptive Filter Theory, Pearson Education India, 2002.

[30] Z. Wang, X. Yu, Y. Guo, W. Kang, X. Chen, A noise-enhanced feature extraction method combined with tunable Q-factor wavelet transform and its application to planet-bearing fault diagnosis, Appl. Acoust. 239 (2025) 110845.

[31] T. Lin, Z. Ren, L. Zhu, Y. Zhu, K. Feng, W. Ding, K. Yan, M. Beer, A systematic review of multi-sensor information fusion for equipment fault diagnosis, IEEE Trans. Instrum. Meas. 74 (2025) 1, http://dx.doi.org/10.1109/TIM.2025.3529577.

[32] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: Recent progress and new challenges, Inf. Fusion 38 (2017) 43–54.

[33] Y. Xiao, H. Shao, J. Wang, B. Cai, B. Liu, From deterministic to Bayesian: Adapting pre-trained models for human-computer collaborative fault diagnosis via post-hoc uncertainty, J. Ind. Inf. Integr. 47 (2025) 100921.

[34] H. Shao, Y. Xiao, J. Leng, X. Zhao, B. Liu, Collaborative human-computer fault diagnosis via calibrated confidence estimation, Adv. Eng. Inform. 65 (2025) 103349.

[35] S. Wen, Y. Chen, X. Pan, W. Zhuang, X. Li, Enhancing fault troubleshooting through human-machine collaboration: A multi-stage reasoning approach, in: 2024 IEEE 20th International Conference on Automation Science and Engineering, CASE, 2024, pp. 460–467.

[36] G. Li, Y. Li, S. Sun, H. Wang, J. Zhao, B. Sun, J. Shi, Self-improving few-shot fault diagnosis for nuclear power plant based on man-machine collaboration, Nucl. Eng. Des. 420 (2024) 113051.

[37] B. Kim, Interactive and Interpretable Machine Learning Models for Human Machine Collaboration (Ph.D. thesis), Massachusetts Institute of Technology, 2015.

[38] G. Han, J. Chen, L. Liu, Z. Wang, F. Zhang, Y. Abudurexiti, An interpretable CNN with wavelet group policy embedded for intelligent fault diagnosis, IEEE Trans. Instrum. Meas. 73 (2024) 1–15, http://dx.doi.org/10.1109/TIM.2024.3368479.

[39] Q. Li, H. Li, W. Hu, S. Sun, Z. Qin, F. Chu, Transparent operator network: A fully interpretable network incorporating learnable wavelet operator for intelligent fault diagnosis, IEEE Trans. Ind. Inform. 20 (6) (2024) 8628–8638.

[40] F. Abid, M. Salmi, A. Brahim, Robust interpretable deep learning for intelligent fault diagnosis of induction motors, IEEE Trans. Instrum. Meas. 69 (6) (2019) 3506–3515.

[41] M. Sadooghi, C. Hu, Physics-based convolutional neural network for fault diagnosis of rolling element bearings, IEEE Sens. J. 19 (11) (2019) 4181–4192.

[42] C. Gao, Z. Wang, Y. Guo, H. Wang, H. Yi, MPINet: Multiscale physics-informed network for bearing fault diagnosis with small samples, IEEE Trans. Ind. Inform. 20 (12) (2024) 14371–14380.

[43] J. Zhong, Z. Yan, C. Ruan, et al., M-IPISincNet: An explainable multi-source physics-informed neural network based on improved SincNet for rolling bearings fault diagnosis, Inf. Fusion 115 (2025) 102761.

[44] H. Lu, V.P. Nemani, V. Barzegar, et al., A physics-informed feature weighting method for bearing fault diagnostics, Mech. Syst. Signal Process. 191 (2023) 110171.

[45] R. Liu, X. Ding, Y. Shao, Prior-knowledge-guided mode filtering network for interpretable equipment intelligent diagnosis under varying speed conditions, Adv. Eng. Inform. 61 (2024) 102493.

[46] R. Liu, X. Ding, S. Liu, et al., Knowledge-informed FIR-based cross-category filtering framework for interpretable machinery fault diagnosis under small samples, Reliab. Eng. Syst. Saf. 254 (2025) 110610.

[47] R. Liu, X. Ding, Q. Wu, et al., An interpretable multiplication-convolution network for equipment intelligent edge diagnosis, IEEE Trans. Syst. Man Cybern.: Syst. 54 (6) (2024) 3284–3295.

[48] Q. Chen, X. Dong, G. Tu, et al., TFN: An interpretable neural network with time-frequency transform embedded for intelligent fault diagnosis, Mech. Syst. Signal Process. 207 (2024) 110952.

[49] Q. Qian, J. Zhang, J. Luo, Y. Qin, Integrated-dispersion manifold distance: A new distribution discrepancy metric for machine fault transfer diagnosis under time-varying conditions, IEEE Trans. Cybern. (2025) 1–13, http://dx.doi.org/10.1109/TCYB.2025.3630879.

[50] Z. Zhao, R. Zhao, Y. Jiao, Random convolution layer: an auxiliary method to improve fault diagnosis performance, J. Intell. Manuf. 35 (6) (2024) 2937–2949, http://dx.doi.org/10.1007/s10845-024-02458-4.

[51] F. Lu, Q. Tong, X. Jiang, S. Du, J. Xu, J. Huo, Z. Zhang, Envelope spectrum neural network with adaptive domain weight harmonization for intelligent bearing fault diagnosis under cross-machine scenarios, Adv. Eng. Inform. 62 (2024) 102787.

[52] Y. Dalian, Z. Junjun, L. Hui, Capsule networks for intelligent fault diagnosis: a roadmap of recent advancements and challenges, Expert Syst. Appl. 296 (2025) 128814, http://dx.doi.org/10.1016/j.eswa.2025.128814.

[53] H. Sun, D. He, Z. Lao, Z. Jin, Z. Wei, J. Wu, MF-MSRNet: a fault diagnosis method for train bogie bearing based on multi-source data fusion and multi-scale residual network, Nondestruct. Test. Eval. (2025) 1–38, http://dx.doi.org/10.1080/10589759.2025.2542388.

[54] Y. Xu, H. Tao, W. Li, Y. Zhong, CapsFormer: a novel bearing intelligent fault diagnosis framework with negligible speed change under small-sample conditions, IEEE Trans. Instrum. Meas. 72 (2023) 1–11, http://dx.doi.org/10.1109/TIM.2023.3318693.

[55] X. Wang, H. Chen, J. Zhao, C. Song, Y. Zhang, Z.-X. Yang, P.K. Wong, Wind turbine fault diagnosis for class-imbalance and small-size data based on stacked capsule autoencoder, IEEE Trans. Ind. Inform. 20 (11) (2024) 12694–12704.

[56] D. Zhao, S. Liu, T. Zhang, H. Zhang, Z. Miao, Subdomain adaptation capsule network for unsupervised mechanical fault diagnosis, Inform. Sci. 611 (2022) 301–316.

[57] R. Huang, J. Li, Y. Liao, J. Chen, Z. Wang, W. Li, Deep adversarial capsule network for compound fault diagnosis of machinery toward multidomain generalization task, IEEE Trans. Instrum. Meas. 70 (2021) 1–11, http://dx.doi.org/10.1109/TIM.2020.3042300.

[58] H. Xing, X. Jiang, Q. Song, Q. Wang, J. Liu, Z. Zhu, Comprehensive feature integrated capsule network for machinery fault diagnosis, Expert Syst. Appl. 260 (2025) 125450.

[59] Y. Xiong, Z. Liu, J. Tan, L. Hao, Multi-scale adaptive-routing capsule contrastive network-based intelligent fault diagnosis method for rotating machinery under noisy environment and labels, Adv. Eng. Inform. 62 (2024) 102712.

[60] H. Tang, J. Xia, Y. Bai, C. Chen, Y. Leng, EFLightCaps: an efficient feature-focused lightweight capsule network with frequency-domain regularization for rotating machinery fault diagnosis, Meas. Sci. Technol. 36 (3) (2025) 036130.

[61] J. Shang, D. Xu, H. Qiu, L. Gao, C. Jiang, P. Yi, A novel data augmentation framework for remaining useful life estimation with dense convolutional regression network, J. Manuf. Syst. 74 (2024) 30–40.

[62] Q. Qian, Q. Wen, R. Tang, Y. Qin, DG-Softmax: A new domain generalization intelligent fault diagnosis method for planetary gearboxes, Reliab. Eng. Syst. Saf. 260 (2025) 111057.

[63] Y. Xu, S. Kohtz, J. Boakye, P. Gardoni, P. Wang, Physics-informed machine learning for reliability and systems safety applications: State of the art and challenges, Reliab. Eng. Syst. Saf. 230 (2023) 108900.

[64] Y. Wu, B. Sicard, S.A. Gadsden, Physics-informed machine learning: A comprehensive review on applications in anomaly detection and condition monitoring, Expert Syst. Appl. 255 (2024) 124678.

[65] J. Tan, J. Wan, H. Cai, H. Shao, M. Safran, S.A. AlQahtani, Domain-knowledge-driven intelligent attribute definition for zero-shot fault diagnosis of bearings, IEEE Trans. Ind. Inform. 21 (7) (2025) 5286–5296.

[66] Y. Qin, H. Liu, Y. Wang, Y. Mao, Inverse physics–informed neural networks for digital twin–based bearing fault diagnosis under imbalanced samples, Knowl.-Based Syst. 292 (2024) 111641.

[67] J.-X. Liao, C. He, J. Li, J. Sun, S. Zhang, X. Zhang, Classifier-guided neural blind deconvolution: A physics-informed denoising module for bearing fault diagnosis under noisy conditions, Mech. Syst. Signal Process. 222 (2025) 111750.

[68] L.N. Garpelli, D.S. Alves, K.L. Cavalca, H.F. de Castro, Physics-guided neural networks applied in rotor unbalance problems, Struct. Health Monit. 22 (6) (2023) 4117–4130.

[69] N. Jia, W. Huang, C. Ding, J. Wang, Z. Zhu, Physics-informed unsupervised domain adaptation framework for cross-machine bearing fault diagnosis, Adv. Eng. Inform. 62 (2024) 102774.

[70] F. Kibrete, D.E. Woldemichael, H.S. Gebremedhen, Multi-sensor data fusion in intelligent fault diagnosis of rotating machines: A comprehensive review, Measurement 232 (2024) 114658.

[71] D. Sun, Y. Li, Z. Liu, S. Jia, K. Noman, Physics-inspired multimodal machine learning for adaptive correlation fusion based rotating machinery fault diagnosis, Inf. Fusion 108 (2024) 102394.

[72] W. Ying, L. Li, Y. Li, T. Wang, J. Zheng, K. Feng, Trustworthy multimodal feature-enhanced fusion network for non-contact rotating machinery fault diagnosis, Inf. Fusion 124 (2025) 103377.

[73] Y. Keshun, W. Puzhou, H. Peng, G. Yingkui, A sound-vibration physical-information fusion constraint-guided deep learning method for rolling bearing fault diagnosis, Reliab. Eng. Syst. Saf. 253 (2025) 110556.

[74] T. Yang, S. Wang, S. Jiang, H. Ma, L. Jiang, Q. Han, X. Wang, X. Li, MSPI-Net framework: a novel optimizer-powered multi-source physical information fusion approach for intelligent diagnosis and interpretability of bearings, Expert Syst. Appl. 296 (2025) 129279.

[75] L. Zhang, L. Zhao, C. Wang, Sparse representation by novel cascaded dictionary for bearing fault diagnosis using bi-damped wavelet, Int. J. Adv. Manuf. Technol. 124 (7) (2023) 2365–2381.

[76] J. Tan, J. Wan, B. Chen, M. Safran, S.A. AlQahtani, R. Zhang, Selective feature reinforcement network for robust remote fault diagnosis of wind turbine bearing under non-ideal sensor data, IEEE Trans. Instrum. Meas. 73 (2024) 1–11, http://dx.doi.org/10.1109/TIM.2024.3375958.

[77] C. He, H. Shi, R. Li, J. Li, Z. Yu, Interpretable modulated differentiable STFT and physics-informed balanced spectrum metric for freight train wheelset bearing cross-machine transfer fault diagnosis under speed fluctuations, Adv. Eng. Inform. 62 (2024) 102568.

[78] A. Ding, Y. Qin, B. Wang, L. Guo, L. Jia, X. Cheng, Evolvable graph neural network for system-level incremental fault diagnosis of train transmission systems, Mech. Syst. Signal Process. 210 (2024) 111175.

[79] T. Zhang, J. Chen, S. He, Z. Zhou, Prior knowledge-augmented self-supervised feature learning for few-shot intelligent fault diagnosis of machines, IEEE Trans. Ind. Electron. 69 (10) (2022) 10573–10584.

[80] R. Liu, F. Wang, B. Yang, S.J. Qin, Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions, IEEE Trans. Ind. Inform. 16 (6) (2020) 3797–3806.

[81] Z. Shang, Z. Zhao, R. Yan, Denoising fault-aware wavelet network: A signal processing informed neural network for fault diagnosis, Chin. J. Mech. Eng. 36 (1) (2023) 9.

[82] L. Jia, T.W. Chow, Y. Yuan, GTFE-Net: A gramian time frequency enhancement CNN for bearing fault diagnosis, Eng. Appl. Artif. Intell. 119 (2023) 105794.

[83] L. Zhang, J. Lin, Z. Yang, H. Shao, B. Liu, C. Li, Wave-ConvNeXt: An efficient and precise fault diagnosis method for IIoT leveraging tailored ConvNeXt and wavelet transform, IEEE Internet Things J. 11 (13) (2024) 23096–23109.

[84] Y. Feng, C. Zheng, J. Chen, T. Pan, J. Xie, S. He, H. Wang, Beyond deep features: Fast random wavelet kernel convolution for weak-fault feature extraction of rotating machinery, Mech. Syst. Signal Process. 224 (2025) 112057.

[85] Z. Chen, J. Chen, S. Liu, Y. Feng, S. He, E. Xu, Multi-channel calibrated transformer with shifted windows for few-shot fault diagnosis under sharp speed variation, ISA Trans. 131 (2022) 501–515.

[86] S. Han, H. Shao, J. Cheng, X. Yang, B. Cai, Convformer-NSE: A novel end-to-end gearbox fault diagnosis framework under heavy noise using joint global and local information, IEEE/ASME Trans. Mechatronics 28 (1) (2023) 340–349.

[87] H. Fang, J. Deng, Y. Bai, B. Feng, S. Li, S. Shao, D. Chen, CLFormer: A lightweight transformer based on convolutional embedding and linear self-attention with strong robustness for bearing fault diagnosis under limited sample conditions, IEEE Trans. Instrum. Meas. 71 (2022) 1–8, http://dx.doi.org/10.1109/TIM.2021.3132327.

[88] S. Yan, H. Shao, J. Wang, X. Zheng, B. Liu, LiConvFormer: A lightweight fault diagnosis framework using separable multiscale convolution and broadcast self-attention, Expert Syst. Appl. 237, Part A (2023) 121338.

[89] B. Wang, Y. Xiong, L. Tan, A high-precision aeroengine bearing fault diagnosis based on spatial enhancement convolution and vision transformer, IEEE Trans. Instrum. Meas. 74 (2025) 1–15, http://dx.doi.org/10.1109/TIM.2024.3502884.

[90] Y. Yu, H.R. Karimi, L. Gelman, A.E. Cetin, MSIFT: A novel end-to-end mechanical fault diagnosis framework under limited & imbalanced data using multi-source information fusion, Expert Syst. Appl. 274 (2025) 126947.
[91] Q. Li, B. Chen, Q. Chen, X. Li, Z. Qin, F. Chu, HSE: A plug-and-play module for unified fault diagnosis foundation models, Inf. Fusion 123 (2025) 103277.
[92] J. Wang, T. Li, Y. Yang, S. Chen, W. Zhai, DiagLLM: multimodal reasoning with large language model for explainable bearing fault diagnosis, Sci. China Inf. Sci. 68 (6) (2025) 160103.