

Document Version

Final published version

Licence

CC BY

Citation (APA)

Xu, J., Yi, J., Wu, Z., Pang, Y., Liu, C., Cheng, G., & An, H. (2026). Prior Knowledge–Guided Adversarial Augmentation for Small-Sample Fault Diagnosis of Wind Turbine Drivetrain. *Structural Control and Health Monitoring*, 2026(1), Article 8394224. <https://doi.org/10.1155/stc/8394224>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

RESEARCH ARTICLE **OPEN ACCESS**

Prior Knowledge-Guided Adversarial Augmentation for Small-Sample Fault Diagnosis of Wind Turbine Drivetrain

Jin Xu¹  | Jiang Yi¹ | Zhaohong Wu¹ | Yusong Pang²  | Chang Liu³ | Gang Cheng¹  | Huang An⁴

¹School of Mechanical and Electrical Engineering, China University of Mining and Technology, Xuzhou, Jiangsu, China | ²Faculty of Mechanical Engineering, Delft University of Technology, Delft, South Holland, the Netherlands | ³School of Mechanical and Electrical Engineering, Xuzhou University of Technology, Xuzhou, Jiangsu, China | ⁴Dongfang Electric New Energy Technology (Chengdu) Co., Ltd, Chengdu, Sichuan, China

Correspondence: Gang Cheng (chg@cumt.edu.cn)

Received: 10 May 2026 | **Revised:** 7 July 2026 | **Accepted:** 8 July 2026

Guest Editor: Francesc Pozo

Keywords: data augmentation | fault diagnosis | rotating machinery | signal processing | small-sample learning

ABSTRACT

Wind turbine drivetrains operate under variable aerodynamic loading, fluctuating speed, and harsh field environments, making condition monitoring difficult when fault samples are scarce and imbalanced. To address this challenge, this paper proposes a small-sample data augmentation method that fuses variational mode decomposition–particle swarm optimization–symmetric dot pattern (VMD–PSO–SDP) feature optimization with a prior knowledge-guided Wasserstein conditional generative adversarial network (PK-WCGAN). First, VMD is used to decompose vibration signals into informative modal components. In combination with an improved PSO algorithm for optimizing SDP parameters, high-order fault feature maps are generated to enhance feature transferability. Then, real feature maps are used to pretrain the discriminator as prior knowledge, guiding the Wasserstein conditional generative adversarial network (WCGAN) to learn the target domain distribution. A gradient penalty mechanism is introduced to improve the quality of generated samples and training stability. Finally, generated samples are fused with real training samples to construct a hybrid dataset. Experimental results demonstrate that the generated data effectively reduce small-sample distribution mismatch. When the target data samples are expanded from 50 to 1000, the average recognition accuracy of the five models increases from 54.6% to 94.0%, and the risk of overfitting is reduced by up to 26.5 percentage points under the 10-sample setting. This method provides an effective solution for small-sample fault diagnosis in the wind turbine drivetrain.

1 | Introduction

As essential equipment of modern industry, rotating machinery is widely used in many fields such as aerospace and machinery manufacturing [1, 2]. Wind turbines are critical rotating assets in the energy industry, and their drivetrains operate continuously under varying aerodynamic loads, fluctuating speeds, and harsh field conditions. Gearboxes, shafts, bearings, and generators are the core load-transfer components, and localized damage in any of these components can lead to costly downtime or trigger secondary structural risks [3]. Therefore, accurate condition monitoring of the wind turbine drivetrain is of great significance

for the stable operation of the equipment and the prevention of accidents [4].

In the condition monitoring of rotating machinery, numerous experts have proposed a variety of fault diagnosis methods, which are broadly divided into signal processing and data-driven methods. From the perspectives of the time domain [5–7], frequency domain [8–10], and time-frequency domain [11–13], signal processing-based fault analysis methods extract fault information from vibration signals to identify fault characteristics. Shen et al. [14] proposed a bearing fault diagnosis method based on AGWO–particle swarm optimization (PSO)–VMD–TEFCG–AlexNet, optimized variational

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Copyright © 2026 Jin Xu et al. *Structural Control and Health Monitoring* published by John Wiley & Sons Ltd.

mode decomposition (VMD) parameters to improve signal decomposition, and combined time-frequency domain enhanced feature maps and deep learning to achieve efficient noise reduction and fault diagnosis in strong noise environments. Sun et al. [15] proposed a bearing fault diagnosis method based on EMD and improved Chebyshev distance. This method extracts intrinsic mode function (IMF) features via symmetric dot pattern (SDP) feature fusion and calculates the Chebyshev distance to achieve efficient fault diagnosis. With the rapid development of computer networks, data-driven fault diagnosis methods have gradually become a focal point in the field, primarily utilizing machine learning [16] and deep learning [17–19] techniques applied to image domains to learn deep latent features from large datasets for fault signal identification. Liu et al. [20] converted the time-domain signals to the frequency domain by the fast Fourier transform and constructed a k-nearest neighbor graph for graph node classification. This approach used meta-learning and a graph convolutional network (GCN) to achieve high accuracy in bearing fault classification with small sample sizes. Wu et al. [21] developed a bearing fault diagnosis model by integrating multiscale convolutional neural networks with long short-term memory networks. This model utilizes a CNN to extract spatial features and an LSTM to extract temporal features, enhancing robustness under variable loads and noise interference. Zhou et al. [22] introduced a hypergraph attention network with redundancy elimination and few-shot learning to model higher-order sample relationships under small-sample and noisy conditions. These studies improve feature representation and diagnostic discrimination, but their performance still depends strongly on the availability and representativeness of labeled fault samples.

However, the above methods suffer from data-dependent issues, particularly with respect to data size, distribution, and quality [23]. In small-sample data environments, the efficacy of these methodologies in ensuring model generalization is often compromised. This, in turn, results in suboptimal model performance in practical applications [24]. In practical industrial environments, fault samples are difficult to obtain because faults occur sporadically and maintenance records are often incomplete. For wind turbines, vibration acquisition is especially costly because sensors must operate on high, complex structures under harsh field conditions. Consequently, addressing small-sample limitations has emerged as a pivotal challenge in industrial fault diagnosis.

To address the challenge of small samples in industrial applications, researchers have primarily explored transfer learning, domain generalization, domain adaptation, and data augmentation. Transfer learning and domain-adaptation methods reduce distribution shift by reusing knowledge from source domains or aligning feature distributions. For instance, UDDGN [25] established domain-independent compact boundaries via class-wise adversarial training to handle universal diagnosis tasks in which target fault modes are unpredictable. OAIFAN [26] mitigated cross-speed discrepancies and noise interference by mapping multisource signals into a unified angular domain through equal-angle resampling. More recent studies have further focused on unseen or variable operating conditions. Gao et al. [27] proposed a domain feature decoupling network to separate domain-related and fault-related features for zero-shot diagnosis under unseen operating conditions. Li et al. [28]

developed DDNet, which combines domain augmentation and dual-invariance learning for small-sample domain generalization. Li et al. [29] proposed a joint collaborative adaptation network to address class imbalance and variable operating conditions through multisource domain adaptation and class-aware mechanisms.

Meanwhile, GANs have been demonstrated to be an effective data augmentation technique [30]. By generating synthetic fault samples, GAN-based methods can expand the training dataset, alleviate insufficient or imbalanced data, and improve downstream diagnostic models [31, 32]. Existing studies have improved GAN-based fault sample generation through multiscale residual structures, hybrid losses, cosine-similarity constraints, self-attention mechanisms, spectral normalization, and Wasserstein objectives [33–38]. Recent generative studies have moved beyond simple label-conditioned synthesis. Cheng et al. [39] proposed CRGAN-MDTR to generate fault samples during continuous degradation by retaining degradation trend information. Liu et al. [40] proposed ARC-GAN for zero-shot bearing fault diagnosis, using semantic auditing and attribute reconstruction to improve feature-semantic alignment under domain shift. These studies show that structural, temporal, or semantic constraints can improve generation quality and diagnostic generalization.

Although the above methods have achieved high accuracy in data augmentation and fault diagnosis, the following issues remain. First, existing feature extraction methods can extract time-frequency features, but their parameters are fixed and have not been optimized for cross-condition scenarios, resulting in limited transferability of features. When the target domain conditions (load and speed) change, the feature distribution is prone to shift, affecting the model's generalization ability; second, traditional GANs and their variants face issues such as training instability and gradient vanishing in small-sample scenarios, resulting in generated samples that are prone to noise or structural distortion. Additionally, they lack precise control over the target domain's conditional distribution, making it difficult to generate fault samples with consistent physical meaning; third, many methods generate samples either from raw signals or from unoptimized feature images, forcing GANs to learn the physical signal structure and fault-discriminative features simultaneously. This dual learning burden significantly degrades the quality of generation in data-scarce settings.

To address these limitations, this paper proposes a framework that combines physics-informed feature optimization with prior knowledge-guided adversarial generation. Specifically, an improved VMD-PSO-SDP procedure maps raw vibration signals onto an optimized and physically interpretable SDP feature manifold. These real target-domain feature maps are then used to pretrain the discriminator prior to adversarial training. After this auxiliary pretraining stage, the discriminator is converted into a Wasserstein critic, so the generator receives gradients from a critic that has already learned fault-related SDP structures. Therefore, the prior knowledge in prior knowledge-guided Wasserstein conditional generative adversarial network (PK-WCGAN) is not only provided as labels, input features, or a transferred classifier; it is also embedded in the discriminator representation, which guides the adversarial learning process. The main contributions are summarized as follows:

1. A VMD–PSO–SDP feature optimization mechanism is proposed. The SDP parameters are optimized after VMD decomposition, and a velocity-update penalty term is introduced into PSO to constrain redundant feature patterns and enhance cross-condition feature transferability.
2. A discriminator-prior initialization strategy is introduced for PK-WCGAN. The discriminator is first trained as an auxiliary fault classifier using real VMD–PSO–SDP feature maps; afterward, the auxiliary head is removed, and the learned convolutional representation is used to initialize the Wasserstein critic. This makes prior knowledge operate through adversarial gradients rather than only through labels, input features, handcrafted constraints, or a transferred diagnostic classifier.
3. A physically informed input strategy is adopted to reduce the GAN learning burden. Optimized VMD–PSO–SDP feature maps are used as generative targets, so the generator learns fault-discriminative distributions in an informative feature space rather than extracting signal structures from scratch. Experiments on cross-condition datasets show that the framework achieves over 90% accuracy when the training set is sufficiently augmented and reduces the risk of overfitting by up to 26.5% points in the 10-sample-per-class setting.

The remainder of this paper is organized as follows. Section 2 details the proposed data augmentation method. Sections 3 and 4 demonstrate the method's superior performance through experiments. Finally, Section 5 draws conclusions.

2 | Methodology

To address the problem of fault diagnosis for wind turbines with small sample sizes, this paper proposes a data augmentation framework that combines VMD–PSO–SDP feature optimization with PK-WCGAN. The overall process of the proposed research method is illustrated in Figure 1. First, in the feature optimization layer, the vibration signal is decomposed into modal components using VMD, and then the SDP parameters are optimized with a PSO algorithm that includes a penalty term to generate high-fidelity feature maps. In the data augmentation layer, the feature maps and class labels from real data serve as discriminator inputs for prior knowledge pretraining. During the primary training phase, the feature map class labels and random noise are input into the generator and discriminator. Finally, by inputting the labels of the data samples to be expanded and random noise, the model generates virtual samples similar to the target data, which are used to expand the dataset.

2.1 | Fusion of Feature Maps

This section proposes an improved PSO–SDP feature map fusion algorithm based on VMD. By introducing penalty terms during particle velocity updates, constraints are applied to particle positions or velocities to prevent convergence in unreasonable regions. Optimizing SDP transformation parameters enables VMD-generated feature maps to improve recognition accuracy.

2.1.1 | VMD-Based SDP

The SDP is a powerful visualization technique that converts one-dimensional time-domain signals into two-dimensional images

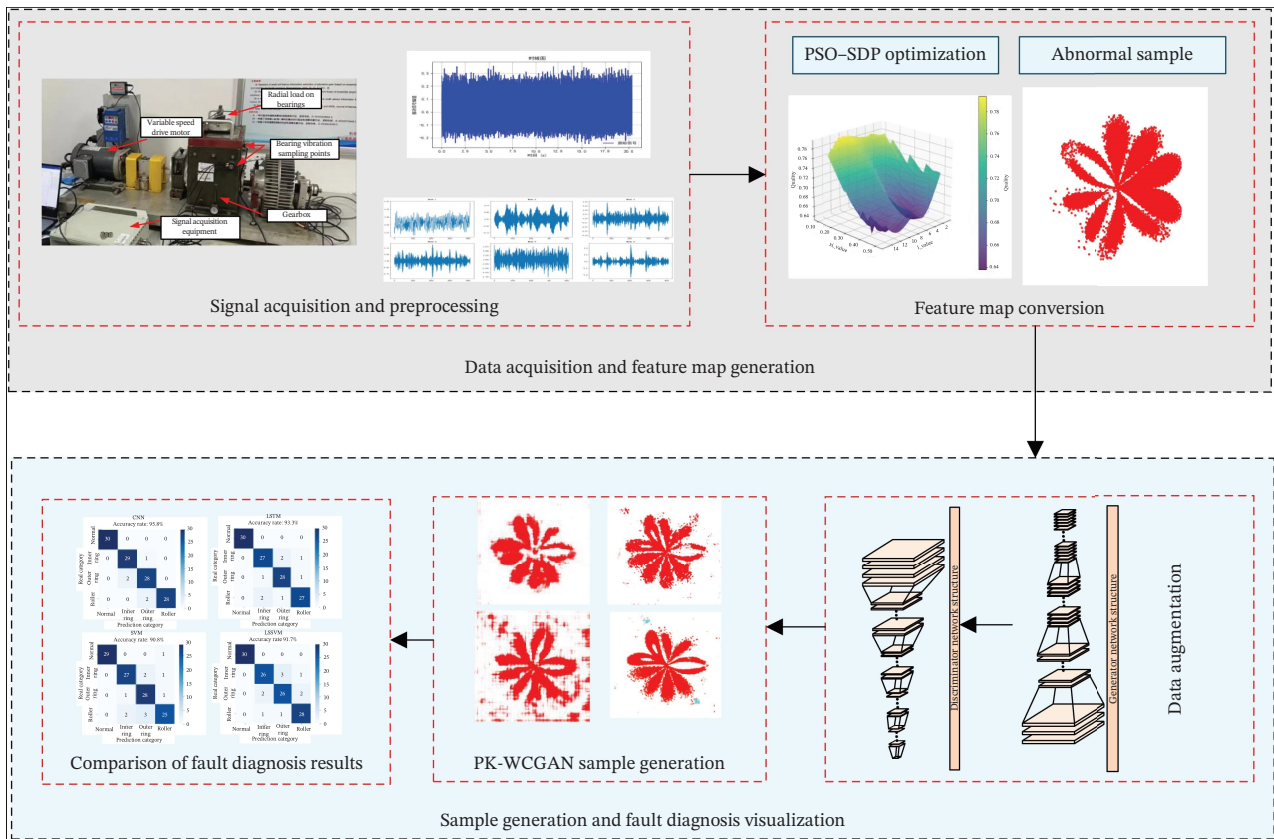


FIGURE 1 | The overall process of the proposed research method.

in polar coordinates, effectively uncovering the inherent periodicities and dynamic behaviors of the signal [41]. At its core, SDP maps the relationship between a given data point and its neighboring points onto a symmetric geometric pattern within a polar coordinate system. The conversion principle of SDP is to normalize the collected one-dimensional time-domain vibration signal and then convert it into symmetric polar coordinates using equations (1)–(3). The conversion principle diagram is shown in Figure 2.

Polar radius:

$$R_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}. \quad (1)$$

Counterclockwise angle:

$$\theta_i = \theta + \frac{x_{i+l} - x_{\min}}{x_{\max} - x_{\min}} \xi. \quad (2)$$

Clockwise angle:

$$\theta_i = \theta - \frac{x_{i+l} - x_{\min}}{x_{\max} - x_{\min}} \xi, \quad (3)$$

where $x_{\max}$ means the maximum amplitude of the vibration signal; $x_{\min}$ means the minimum amplitude; x_i is the i th sample point of the vibration signal; l is the time interval parameter; θ is the mirror alignment angle; and ξ is the angle amplification factor. The subscripts 1 and 2 distinguish the counterclockwise and clockwise SDP branches.

However, the quality of the original SDP image is highly susceptible to noise and complex signal components inherent in raw vibration signals. To address this issue, this paper introduces VMD as a preprocessing step. VMD adaptively decomposes the complex raw signal into a discrete number of band-limited IMFs, effectively separating the fault-related impulse components from background noise and other interferences [42]. Applying the SDP transformation to informative IMFs, typically those containing dominant fault frequencies, can generate cleaner and more discriminative feature maps that are more conducive to cross-condition transfer learning.

As shown in Figure 3, the VMD–SDP decomposes raw signals into multiple time-frequency localized components.

The VMD algorithm decomposes the time-domain signal $x(t)$ into K predefined IMFs $u_k(t)$, each with a center frequency ω_k . The decomposition is achieved by solving the following constrained variational problem:

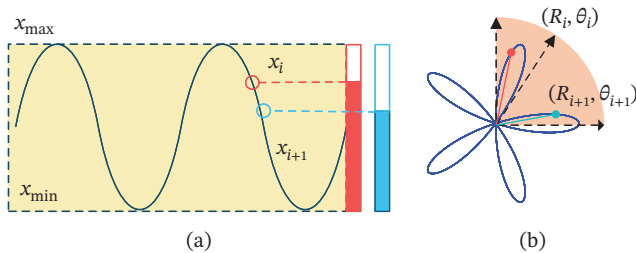


FIGURE 2 | Illustration of the SDP method's principle. (a) Time series. (b) SDP image.

$$\min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_{k=1}^K \left\| \partial_t [(\delta(t) + j/\pi t) * u_k(t)] e^{-j\omega_k t} \right\|_2^2 \right\}, \quad (4)$$

$$\text{s.t. } \sum_{k=1}^K u_k = f,$$

where $u_k(t)$ denotes the k th mode function, and it is obtained by minimizing the cost function; ω_k is the center frequency of the k th IMF; $\delta(t)$ is the Dirac delta function; ∂_t is the partial derivative with respect to time t ; and $*$ is the convolution operator.

Each mode function contains t time-domain points. These points are combined with SDP to construct feature maps using equations (5)–(7), as shown in the flowchart below.

Polar radius:

$$R_k[t] = \frac{u_k(t) - u_k(t)_{\min}}{u_k(t)_{\max} - u_k(t)_{\min}}. \quad (5)$$

Counterclockwise angle:

$$\theta_{1k}[t] = \theta + \frac{u_k(t)_{n+l} - u_k(t)_{\min}}{u_k(t)_{\max} - u_k(t)_{\min}} \xi. \quad (6)$$

Clockwise angle:

$$\theta_{2k}[t] = \theta - \frac{u_k(t)_{n+l} - u_k(t)_{\min}}{u_k(t)_{\max} - u_k(t)_{\min}} \xi, \quad (7)$$

where θ is the mirror alignment angle and ξ is the angle magnification factor.

In this mapping, the central point x_i determines the radial position, while the lagged point x_{i+l} influences the symmetric angular deflection. A periodic fault signature produces a structured, repetitive pattern, while noise results in chaotic scatter. The role of VMD is to purify the input signal $x(t)$ into IMFs $u_k(t)$, ensuring that the subsequent SDP transformation is applied to a signal component where the fault periodicity is prominent and unobscured by noise. This synergy between VMD's denoising and decomposition capabilities and SDP's power in geometric representation forms the foundation for generating highly transferable feature maps that are robust to variations in operating conditions.

2.1.2 | PSO–SDP for Optimizing Feature Map Parameters

Advanced optimization techniques play a crucial role in enhancing the front-end configuration and feature extraction of fault diagnosis systems. For example, PSRONet [43] formulated the optimization of sensor configurations as a Markov decision process, utilizing deep reinforcement learning to balance diagnostic accuracy and acquisition costs. Inspired by these optimization advancements, this section proposes an improved PSO–SDP feature map fusion algorithm. The proposed method combines PSO with SDP to optimize feature map parameters. By leveraging PSO's global optimization capability, the method adjusts SDP feature map generation parameters to more effectively characterize signal features and improve subsequent classification and fault diagnosis performance. The PSO–SDP optimization algorithm is illustrated in Figure 4.

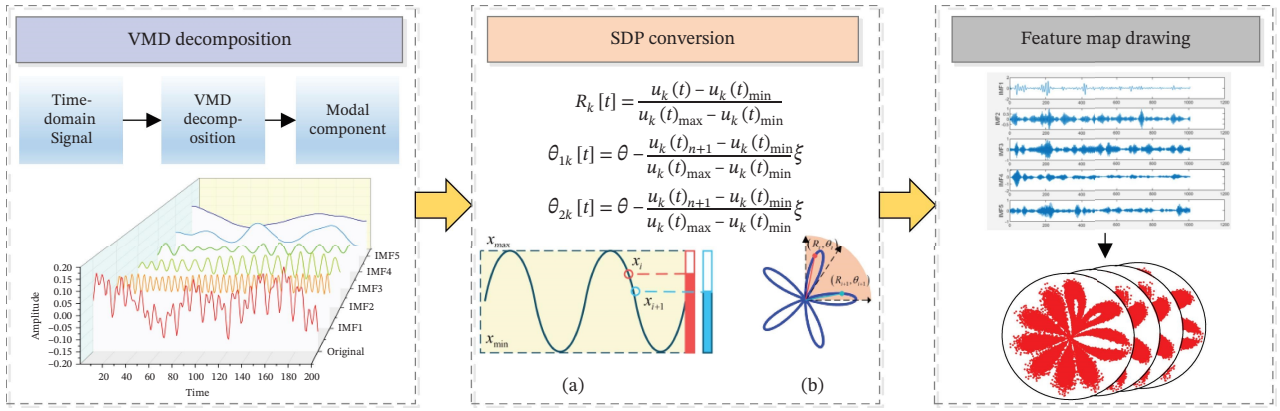


FIGURE 3 | VMD-SDP image plotting flowchart.

1. Each particle's position represents a parameter combination for SDP feature map generation. In this study, the particle is two-dimensional, defined by $\chi_i = (l, \xi)$, where l denotes the time interval parameter and ξ denotes the angle magnification factor. These two parameters were selected for optimization because they directly and significantly influence the geometric structure and, consequently, the discriminative power and transferability of the resulting SDP image. The initial positions of the particles are randomly distributed within predefined reasonable ranges: $l \in [0, 15]$ and $\xi \in [0.1, 0.628]$. These ranges were determined through preliminary experiments to ensure the generation of well-structured, nondegenerate SDP patterns that effectively capture the underlying characteristics of the signal.
2. Objective function: To comprehensively evaluate the quality of generated samples and prevent feature redundancy, the optimization objective combines distributional divergence and structural similarity (SSIM). The objective function $\text{Objective}(l, \xi)$ aims to minimize the Jensen-Shannon (JS) divergence between the generated sample distribution C_t and the real sample distribution C_s , while penalizing low SSIM:

$$\text{Objective}(l, \xi) = \min[F_{js}(C_s, C_t) + \lambda \cdot (1 - \text{SSIM}(C_s, C_t))], \quad (8)$$

where $F_{js}(C_i, C_j)$ is the JS divergence between the generated sample and the real sample distributions and λ is the penalty coefficient balancing the two metrics.

3. Update the velocity and position of the particles to make them move in the optimized direction. The optimized objective function $f(x_i)$ is the migration evaluation function between the generated sample distribution and the real sample distribution. The particle's velocity is updated in each iteration based on the individual optimal position P_i and the global optimal position G . The velocity update equation is as follows:

$$v_{in}(t+1) = \omega \cdot v_{in}(t) + c_1 \cdot r_1 \cdot (P_{in} - x_{in}(t)) + c_2 \cdot r_2 \cdot (G_n - x_{in}(t)). \quad (9)$$

To prevent particles from converging to unreasonable regions and causing feature redundancy, a penalty term is introduced during velocity updates:

$$v_{in}(t+1) = \omega \cdot v_{in}(t) + c_1 \cdot r_1 \cdot (P_{in} - x_{in}(t)) + c_2 \cdot r_2 \cdot (G_n - x_{in}(t)) - \lambda \cdot P(x_{in}(t)), \quad (10)$$

where $v_{in}(t)$ represents the velocity of particle i in the n th dimension at the t th iteration; ω is the inertia weight; c_1 and c_2 are learning factors; r_1 and r_2 are random numbers between $[0, 1]$; P_{in} represents the individual historical optimal position of the particle in the n th dimension; and G_n represents the global optimal position of the particle swarm in the n th dimension. λ is the penalty coefficient used to control the influence intensity of the penalty term. The function of the penalty term is as follows:

$$P(x_{in}(t)) = 1 - \text{SSIM}(x_{in}(t)), \quad (11)$$

where $\text{SSIM}(x_{in}(t))$ is the SSIM between the generated sample and real sample distributions, evaluating the degree of deviation under specific constraints. This effectively guides particles to capture domain relationships during optimization.

2.2 | PK-WCGAN Data Augmentation Framework

The PK-WCGAN data augmentation method is structured as shown in Figures 5 and 6. It comprises two main components: prior knowledge integration and Wasserstein distance-based conditional GAN.

Prior knowledge refers to accumulated domain expertise, research findings, and expert insights that guide model learning to improve performance and generalization. Wasserstein distance is mainly used to measure the difference between two probability distributions. In conditional generative adversarial networks (CGANs), these two probability distributions are the generated data distribution and the real data distribution, respectively, and are defined as in (12). When Wasserstein distance is used as the adversarial objective, the original cross-entropy loss is replaced by a critic-based distribution distance, so the generator is optimized to produce target-domain fault samples that satisfy the conditional labels while approaching the real-data distribution.

$$W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|], \quad (12)$$

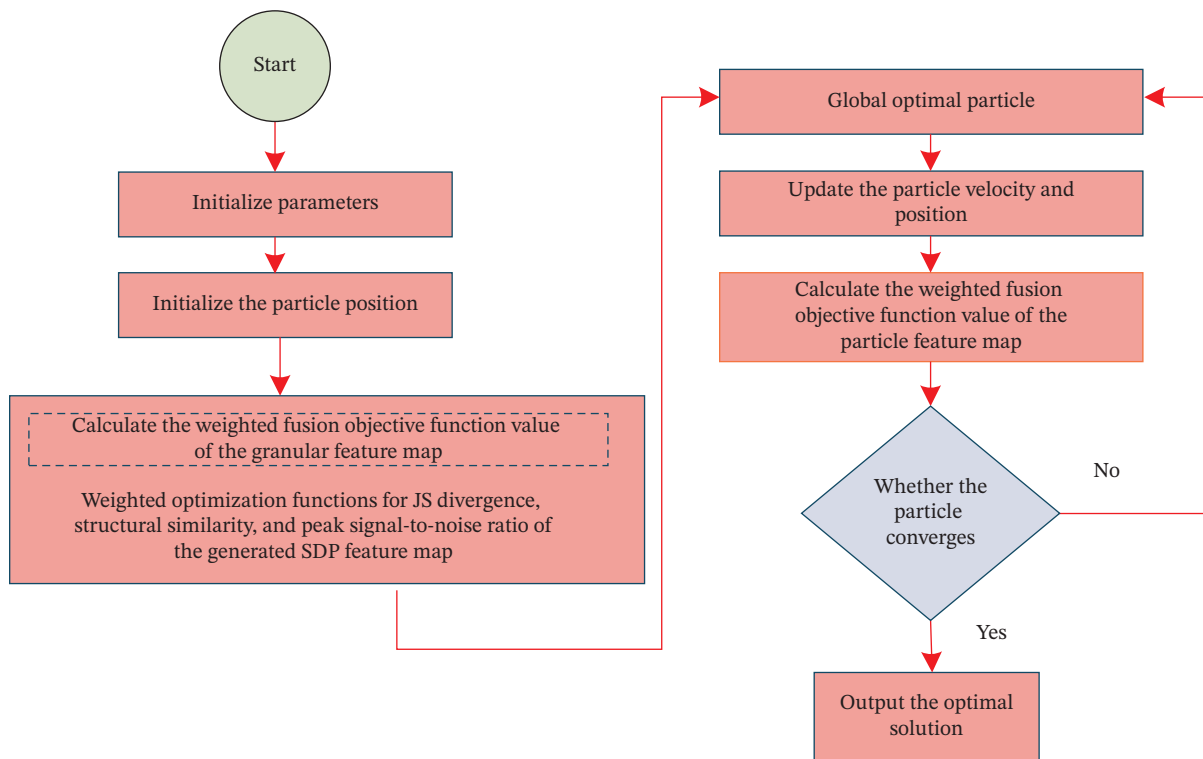


FIGURE 4 | PSO-SDP feature map parameter optimization method flowchart.

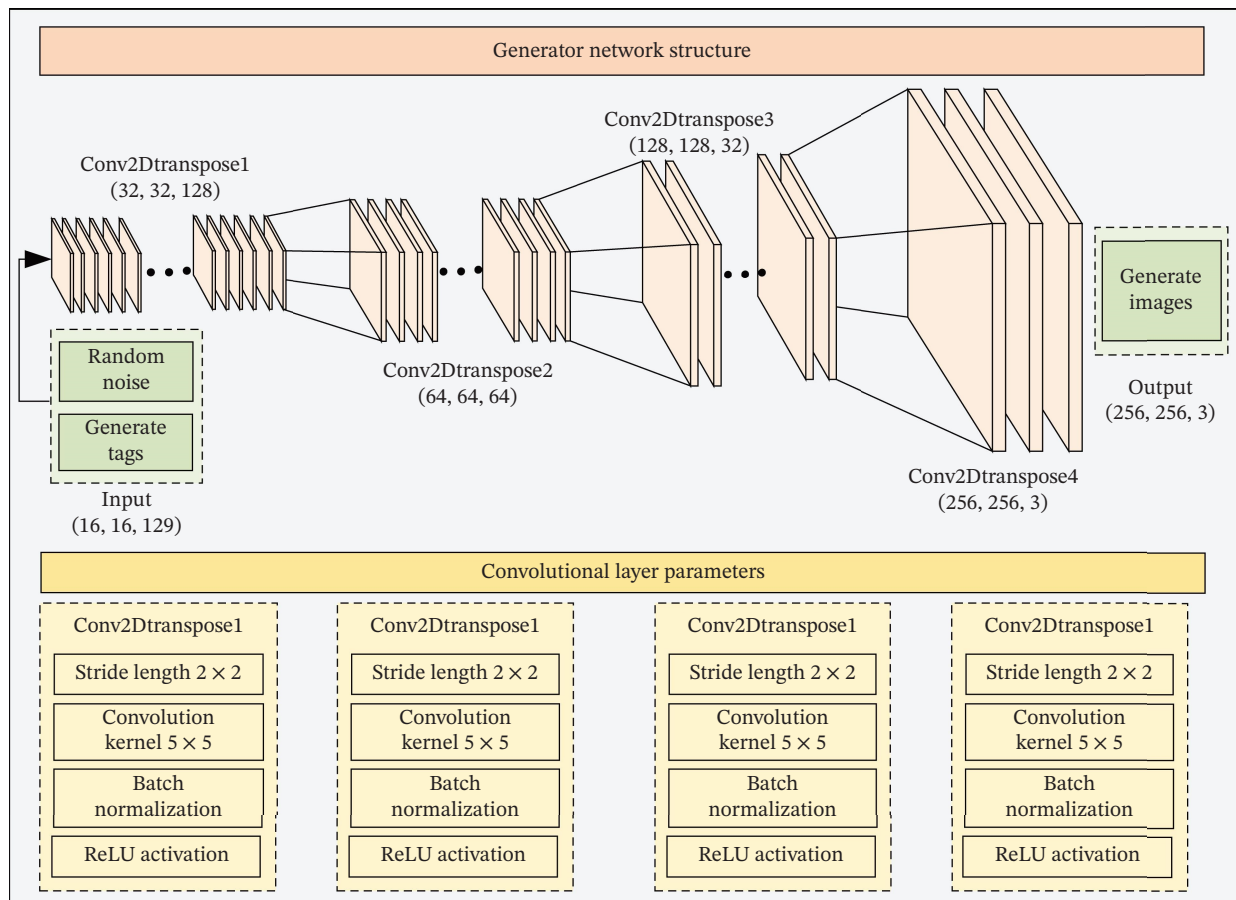


FIGURE 5 | Generator network architecture diagram.

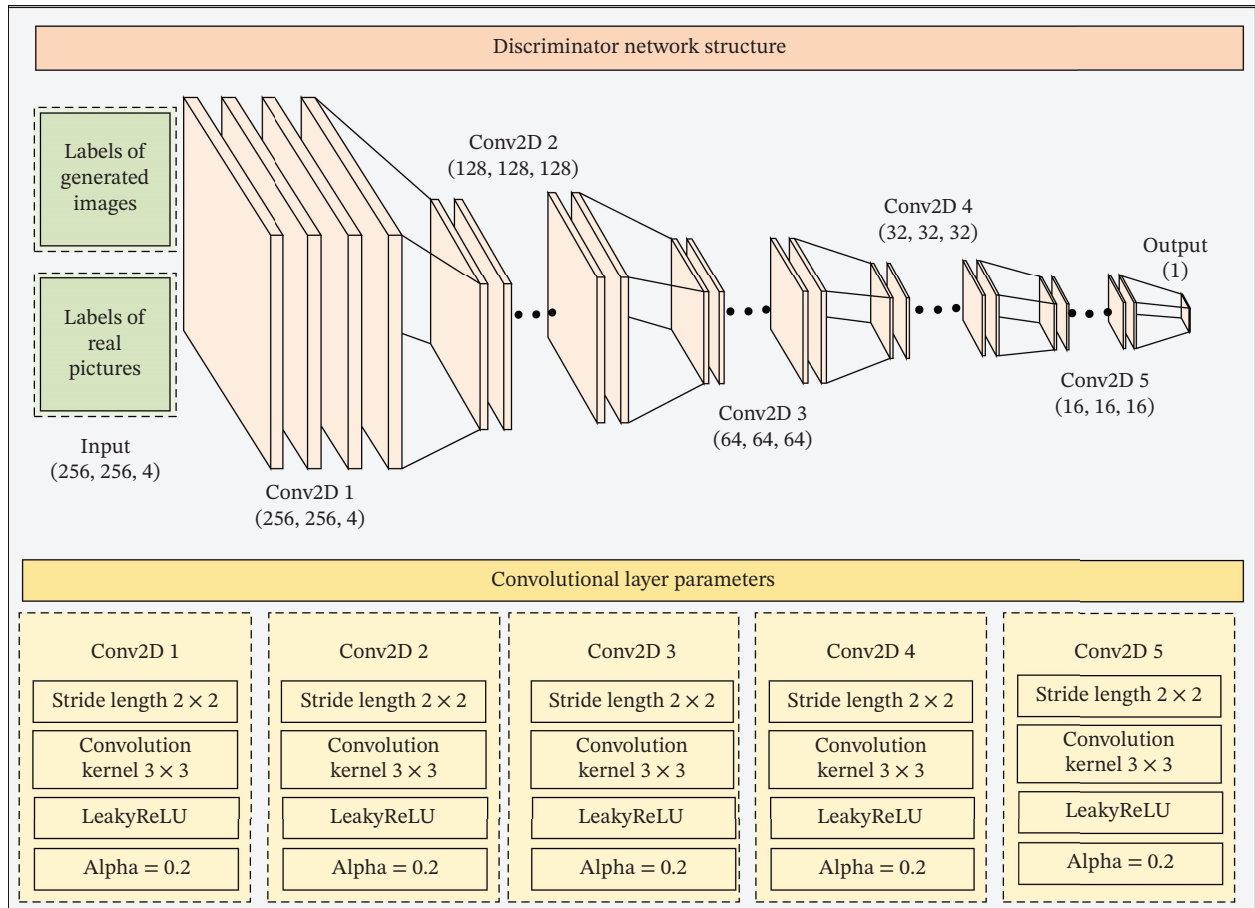


FIGURE 6 | Discriminator network architecture diagram.

where $\Pi(P_r, P_g)$ is the set of joint distributions and P_r and P_g denote the real-sample and generated-sample distributions, respectively.

To make the Wasserstein critic valid, the 1-Lipschitz constraint must be satisfied at the same time, as shown in (13). In this way, the scoring function can maintain sufficient smoothness and avoid the common problems of gradient vanishing or explosion in traditional GAN, thereby effectively improving the stability of model training and the quality of generated images.

$$|D(x_1) - D(x_2)| \leq \|x_1 - x_2\|, \quad (13)$$

where $D(x_1) - D(x_2)$ represents the score difference between two data points calculated by the degree penalty function, and this difference measures the response amplitude of the input data; $\|x_1 - x_2\|$ represents the distance between input samples.

Wasserstein conditional generative adversarial network (WCGAN) is a model extended from traditional CGANs. CGANs introduce conditional information to control the generation process, relying not only on random noise as the input to the generator but also incorporating additional information, such as category labels, text descriptions, or other features, into both the generator and discriminator. This improvement enables CGANs to generate samples that meet specific requirements while enhancing the diversity and accuracy of the generated output. WCGAN further optimizes the training stability and generation performance of CGAN by introducing the Wasserstein distance, thereby alleviating gradient vanishing and improving the model's

generation quality under conditional information control. Therefore, compared with traditional CGANs, WCGANs have significant advantages in the stability of the generation process and the realism of generated samples and are suitable for scenarios with scarce data and conditional constraints. In terms of structure, WCGANs inherit the basic framework of CGANs, consisting of a generator and a discriminator but differ in their specific implementations and optimization objectives. The task of the generator is to generate samples that are as close as possible to the true distribution based on the input random noise and conditional information, while the discriminator no longer outputs a binary classification result but instead outputs a real number score to measure the difference between the generated samples and the real samples. The objective function of the generator is to minimize the discriminator's score on the generated samples, thereby reducing the Wasserstein distance between the distributions of the generated samples and the real samples, as shown in the following equation:

$$\min_G \mathbb{E}_{z \sim p_z, c \sim p_c} [D(G(z, c))], \quad (14)$$

where $G(z, c)$ is the sample generated by the generator according to random noise z and conditions c ; $D(G(z, c))$ is the discriminator's score for the generated sample; p_z is the distribution of random noise; and p_c is the distribution of conditional information.

The objective of the discriminator is to maximize the Wasserstein distance, i.e., maximize the difference between the real data and

the generated data, as illustrated in (15). Through adversarial training between the generator and the discriminator, the ultimate goal of the generator is achieved, which is to make the scores of the generated samples as close as possible to those of the real samples. The generator is continuously optimized to minimize the Wasserstein distance between the distribution of the generated samples and that of the real data, thereby generating realistic samples that meet the condition information requirements.

$$\max_D [E_{x \sim p_r} [D(x)] - E_{x \sim p_{z,c} \sim p_c} [D(G(z, c))]], \quad (15)$$

where $E_{x \sim p_r} [D(x)]$ represents the discriminator's score for real samples and $E_{z \sim p_z, c \sim p_c} [D(G(z, c))]$ represents the discriminator's score for generated samples.

The generator utilizes four transposed convolutional layers, each with a 5×5 kernel and a 2×2 stride. This kernel size is widely adopted in GAN research [24, 31] due to its effective balance between receptive field coverage and computational efficiency. The 2×2 stride enables progressive doubling of the spatial dimensions at each stage ($16 \times 16 \rightarrow 32 \times 32 \rightarrow 64 \times 64 \rightarrow 128 \times 128 \rightarrow 256 \times 256$), facilitating a coherent and gradual image reconstruction. Such hierarchical upsampling supports the learning of multiscale features, capturing both global structural patterns and fine-grained textural details inherent in SDP images.

The discriminator receives a 256×256 VMD-PSO-SDP feature map together with the corresponding class condition. It is designed as a mirror image of the generator, comprising four consecutive convolutional blocks, each with a 3×3 kernel and a 2×2 stride. This architecture progressively halves the spatial dimensions ($256 \times 256 \rightarrow 128 \times 128 \rightarrow 64 \times 64 \rightarrow 32 \times 32 \rightarrow 16 \times 16$) while increasing the representation level from low-level SDP edges and textures to high-level fault-related patterns. The final critic layer maps the extracted representation to a single scalar output, which serves as the critic score for the computation of the Wasserstein distance. The layer-level discriminator structure and training status are summarized in Table 1.

The prior-knowledge pretraining phase is performed before adversarial training. In this study, prior knowledge specifically refers to the structural distribution of real target-domain VMD-PSO-SDP feature maps, including SDP textures, radial distributions, and class-discriminative fault patterns. During this phase, the discriminator backbone is temporarily connected to the auxiliary classification head shown in Table 1 and is trained only with real VMD-PSO-SDP feature maps and their fault labels using cross-entropy loss. The purpose is not to build the final classifier, but to initialize the convolutional filters so that they respond to physically meaningful SDP structures. After pre-training, the auxiliary head is removed, and the discriminator is converted back to a Wasserstein critic. Conv blocks 1-2, which encode edges and local SDP textures, are frozen to retain the learned structural prior knowledge, whereas Conv blocks 3-4 and the final critic layer are optimized during PK-WCGAN adversarial training to estimate the Wasserstein distance. Therefore, the prior knowledge is embedded in the discriminator parameters before adversarial training, rather than being supplied only as conditional labels, input features, or external loss constraints.

TABLE 1 | Discriminator structure used in the PK-WCGAN.

Stage	Operation	Output size	Training status
Input	VMD-PSO-SDP feature map with class condition	256×256	Used in pretraining and adversarial training
Conv block 1	Conv., 3×3 , stride 2×2	128×128	Frozen after pretraining
Conv block 2	Conv., 3×3 , stride 2×2	64×64	Frozen after pretraining
Conv block 3	Conv., 3×3 , stride 2×2	32×32	Fine-tuned
Conv block 4	Conv., 3×3 , stride 2×2	16×16	Fine-tuned
Auxiliary head	Classification head with cross-entropy loss	Fault-class logits	Pretraining only
Critic layer	Scalar output layer	Critic score	Adversarial training

Meanwhile, the prior knowledge pretraining alters the hypothesis space of the discriminator D . In a standard WCGAN, D is initialized randomly, and computing the Wasserstein distance $W(P_r, P_g)$ in a high-dimensional space with extremely small sample sizes often leads to high variance and uninformative gradients. By pretraining D with real physical feature maps, the network weights θ_D are optimized toward a fault structure-aware manifold $\mathcal{M}_{\text{prior}}$ that captures the structural characteristics of SDP features. Consequently, during the adversarial phase, the 1-Lipschitz constraint $|D(x_1) - D(x_2)| \leq \|x_1 - x_2\|$ is enforced not in a random, unconstrained pixel space, but on a physically meaningful feature manifold. This prior-constrained gradient penalty provides more informative adversarial gradients, guides the generator G to map the noise space toward the real fault feature distribution, and reduces mode collapse under data scarcity. As a result, the generated feature maps are expected to preserve more class-related physical structures, thereby improving generation quality and supporting better diagnostic generalization when they are combined with scarce real training samples.

The data augmentation algorithm model for Wasserstein-conditioned generative adversarial networks based on prior knowledge consists of three main parts: data preprocessing, model training, and virtual sample generation. The detailed process of each part is introduced as shown in Figure 7:

1. Data preprocessing: The collected original data are segmented into 4096 sampling points for each sample. Using the optimal parameters of the VMD-SDP feature map determined in the previous section, the segmented data are converted into corresponding feature maps. Ten feature maps are selected from each data class as training samples for the real data. These feature maps represent the features of each data class and provide a high-quality basis for subsequent model training, thereby improving the model's generalization and robustness.
2. Model training: The training process is divided into a prior knowledge pretraining stage and an adversarial training stage. In the prior knowledge pretraining stage, the discriminator acts as an auxiliary classifier. Real fault feature maps are used to train the discriminator using cross-entropy loss, enabling the network to extract high-level representations of specific fault types. During the subsequent WCGAN adversarial training stage, the pretrained weights of the discriminator's shallow convolutional layers are frozen to preserve the learned structural prior knowledge, while the deep layers are fine-tuned to measure the Wasserstein distance.
3. Virtual sample generation: After model training is complete, the trained PK-WCGAN model is invoked and saved, and the virtual sample generation process begins. By inputting the labels of the data samples to be expanded, along with random noise, the model generates virtual samples similar to the target data, expanding the dataset and addressing the problem of insufficient samples.

This design differs from transfer learning diagnosis methods because no source-domain diagnostic classifier is transferred to the final fault classifier. The pretrained discriminator is used only inside the generative model to shape the critic space and provide

more informative adversarial gradients for sample generation. It also differs from knowledge-informed GAN designs that inject prior knowledge only through conditional labels, external loss terms, or input representations because the prior fault-structure information is stored in the discriminator parameters before adversarial training.

3 | Experimental Data and Test Bench

3.1 | The Dynamic Diagnosis System (DDS) Experimental Equipment

The DDS test bench of China University of Mining and Technology is retained as a benchmark dataset for verifying the general feature optimization and adversarial augmentation capability. The DDS is shown in Figure 8. During the experiment, four types of bearing vibration signals were acquired by replacing bearings with different fault conditions, as illustrated in Figure 9, including normal bearings, outer ring cracks, inner ring cracks, and rolling element faults. The corresponding fault dimensions were 0.23, 0.3, and 0.1 cm, respectively. The sampling frequency was set to 5120 Hz. All the methods in the experiment are based on Python 3.10 and PyTorch 2.0.1 frameworks. The experimental environment is as follows: Intel 13600KF and NVIDIA RTX 3070 GPU.

To validate the noise robustness of the proposed method, bearing fault signals were collected from the DDS test bench and then fused with Gaussian white noise at a signal-to-noise ratio (SNR) of 5 dB. After moving average (MA) filtering, sliding window segmentation is applied to each original signal record with a step size of 256. Each segmented sample group contains 4096 points. To avoid information leakage from overlapping sliding-window samples, the dataset was partitioned at the original signal-record level rather than at the window-sample level. Specifically, the original vibration records of each health state were first divided into training and test subsets using a stratified ratio of 80% and 20%, respectively. Sliding-window segmentation was then independently applied within each subset, with a window length of 4096 sampling points and a step size of 256. Therefore, samples generated from the same original signal record were assigned exclusively to one subset and could not appear simultaneously in the training or test sets. The comparison plots of the signals before and after noise addition are illustrated in Figure 10. The composition of the DDS dataset is shown in Table 2.

3.2 | Wind Turbine Field Dataset

The wind turbine field dataset was collected from the drivetrain condition monitoring system at the Mazongshan wind farm in Jiuquan City, Gansu Province. The dataset contains bearing inner-race and outer-race faults, as well as gear tooth breakage and wear faults observed during long-term operation. These fault types were selected because bearings are high-risk drivetrain components and because localized bearing defects, gear cracks, and tooth breakage are common causes of vibration changes in wind turbine transmission systems. The records cover multiple wind turbine drivetrain components, including the gearbox high-speed shaft, gearbox parallel stage, and generator non-drive-end bearing. These measuring points represent different load-transfer positions from the gearbox to the generator side and therefore provide a practical basis for evaluating whether the proposed

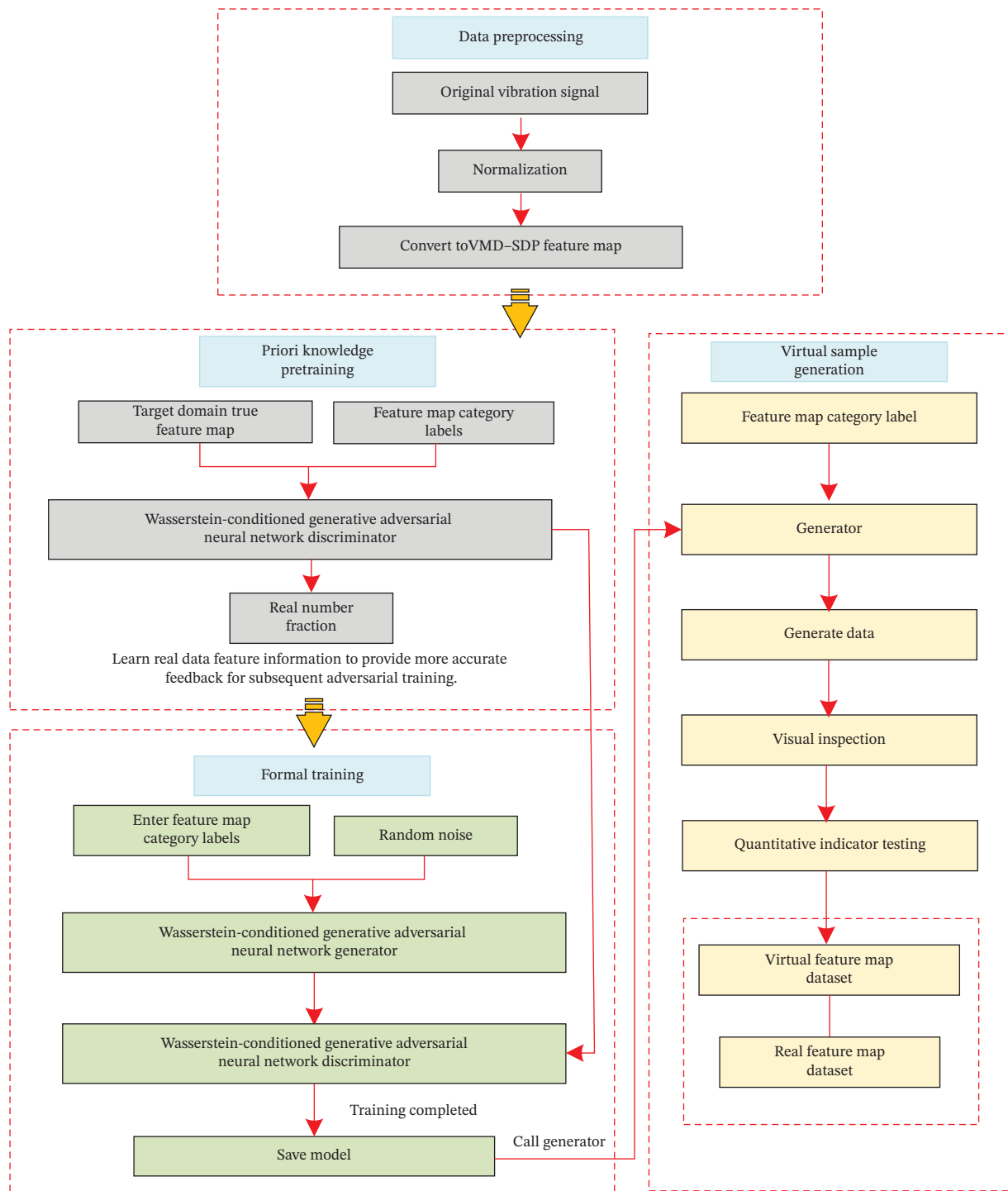


FIGURE 7 | Data augmentation process based on prior knowledge in WCGAN.

augmentation strategy can support transmission-system condition monitoring rather than a single isolated bearing case. The schematic diagram of the wind turbine drivetrain is shown in Figure 11. Figure 12 shows a specific fault image of the wind turbine drivetrain. Each sampling period is 1 s, and the sampling frequency is 8192 Hz. The sample composition is detailed in Table 3.

The experiment mainly focuses on optimizing the PSO-SDP feature map fusion method, conducting data augmentation experiments with traditional CGANs and the improved

Wasserstein generative adversarial network method, and evaluating the recognition effectiveness of generated datasets in DDS and wind turbine field diagnosis models.

4 | Experimental Results and Analysis

4.1 | Optimization Experiment of PSO-SDP Feature Map Fusion Method

In this section of the experiment, the PSO algorithm is combined with the symmetric point pattern method to optimize the feature

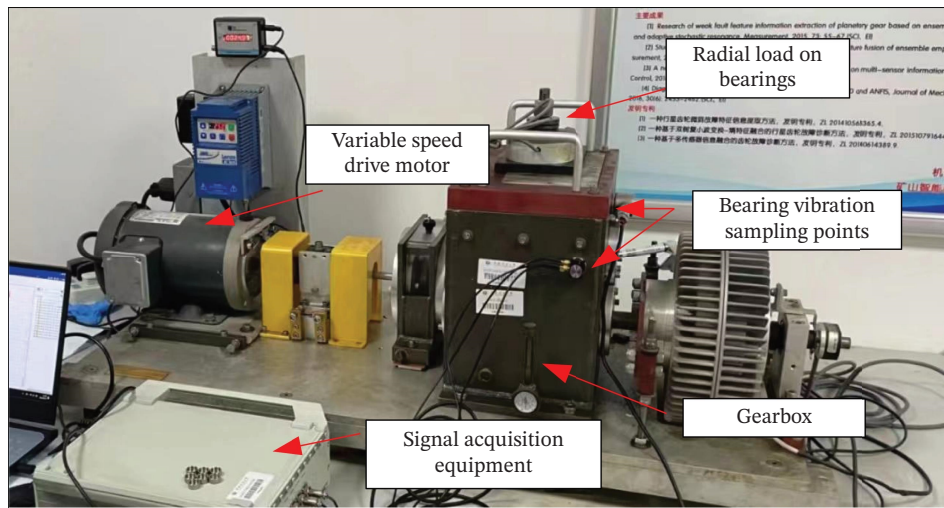


FIGURE 8 | DDS power transmission fault diagnosis comprehensive test bench.

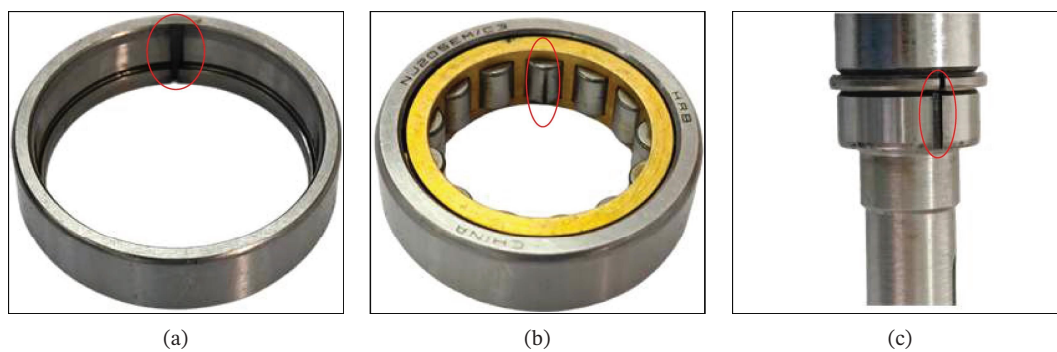


FIGURE 9 | The image of the DDS fault sample: (a) outer ring crack; (b) rolling element fault; (c) inner ring crack.

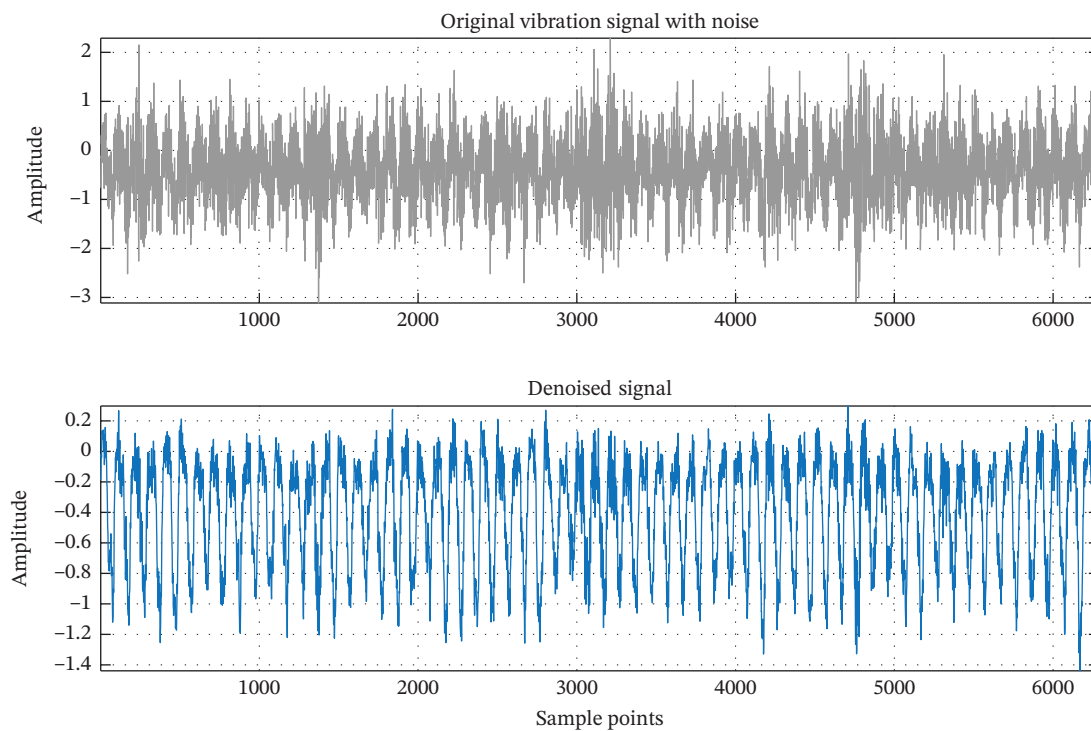


FIGURE 10 | Comparison of time-domain waveforms before and after denoising.

TABLE 2 | The experimental dataset of DDS.

Dataset	Motor load (%)	Motor speed (rpm)	Normal	Out ring	Inner ring	Roller
A	20	1500	1500	1500	1500	1500
B	40	1800	1500	1500	1500	1500

map fusion process and improve the effectiveness of feature extraction. In this subsection, the experiments utilize DDS data.

To ensure reproducibility of our experiments and provide a transparent overview of the model configurations, Table 4 summarizes the hyperparameter settings for the proposed framework. This includes the VMD decomposition configuration, the PSO optimization parameters, and the PK-WCGAN training settings. These hyperparameters were systematically selected and empirically tuned to balance the efficiency of feature extraction with the stability of the adversarial training process.

The experiment begins by performing normalization and denoising on the collected original one-dimensional time-domain vibration signals to ensure consistent data scales. Frequency-domain features are extracted using VMD, and these features are mapped to symmetric point patterns in polar coordinates with the SDP method to reveal the data's hidden structure and features. To further enhance feature map quality and the data augmentation model's performance, the PSO algorithm is employed to optimize the parameters of the generated SDP. In each iteration, particles evaluate their quality based on their fitness function, gradually approaching the global optimal solution. The results of the optimization process are displayed in Figure 13.

It can be seen from Figure 13 that when different values are selected for the time interval parameter and the angle magnification factor, the SSIM between the generated sample and the real sample varies, with a difference of about 0.2. This indicates that the choice of the time interval and the angle magnification factor significantly affects the quality of the generated samples.

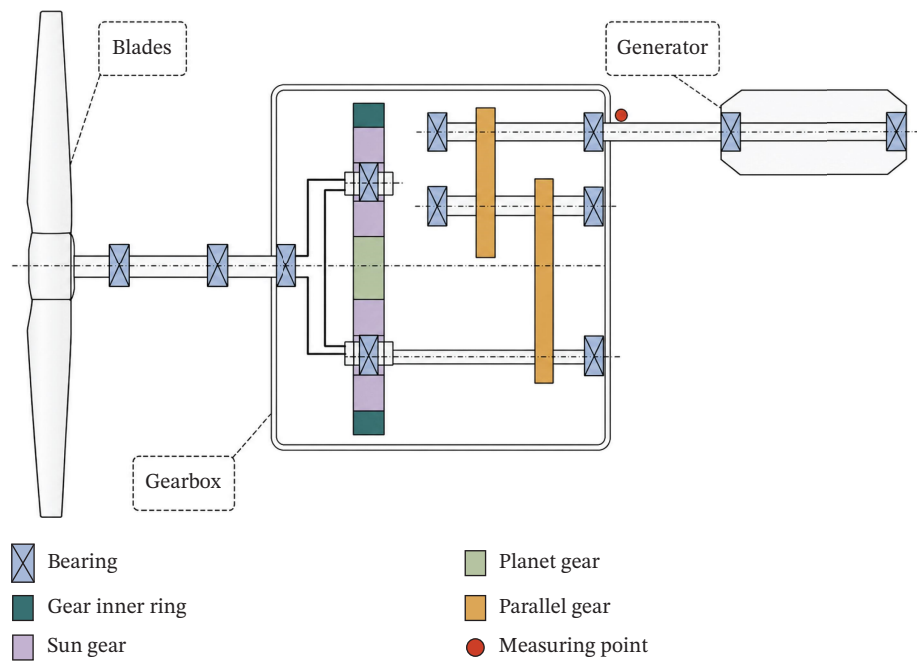


FIGURE 11 | Schematic diagram of wind turbine drivetrain.

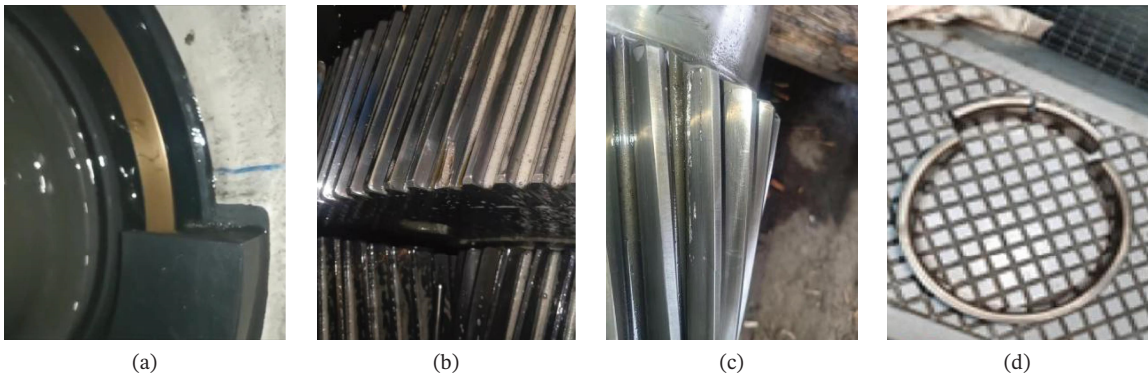


FIGURE 12 | Specific faults of the wind turbine drivetrain. (a) Bearing outer race faults. (b) Partial tooth breakage faults. (c) High-speed shaft gear cracks. (d) Bearing inner race faults.

TABLE 3 | Wind turbine drivetrain field dataset.

Health state	Measuring point	Speed range (rpm)	Samples
Generator non-drive-end bearing inner race fault	Gearbox high-speed shaft	904–1229	142
Motor side bearing outer ring	Gearbox high-speed shaft	902–1224	185
Intermediate shaft gear broken	Gearbox parallel stage	—	73
High-speed shaft gear crack	Gearbox high-speed shaft	915–1253	127

TABLE 4 | Detailed hyperparameter configurations for the proposed framework.

Module	Parameter	Value
VMD	Number of modes (K)	5
	Penalty factor (α)	2000
	Tolerance (τ)	$1e - 7$
PSO-SDP	Population size	20
	Learning factors (c_1, c_2)	[1.5, 1.5]
	Inertia weight (ω)	[0.4–0.9]
	Maximum iterations	80
	Time interval parameter (l)	0–15
	Angle magnification factor (ξ)	0.1–0.628
PK-WCGAN	Optimization objective	Structural similarity between generated samples and real samples
	Optimizer	Adam ($\beta_1 = 0.5, \beta_2 = 0.9$)
	Learning rate	0.0001
	Batch size	64
	Gradient penalty coefficient (λ)	10
	Critic iterations (n_c)	5
	Training epochs	2500

After multiple experimental analyses, the optimal parameters are $l_{\text{value}} = 10$ and $\xi_{\text{value}} = 0.3085$. Under this condition, the global optimal SSIM between the generated sample and the real sample reaches 0.796. This result indicates the important role of parameter optimization in improving the quality and SSIM of generated images and also verifies the effectiveness and robustness of the model-generated samples.

To further analyze the performance of the PSO-SDP method, a genetic optimization algorithm is used for comparison. The experimental parameter settings are shown in Table 5, and the

experimental optimization process results are shown in Figure 14.

Figure 14 shows that the global optimal SSIM between the sample generated by the genetic optimization algorithm optimizing SDP and the real sample reaches 0.783. By comparison, PSO-SDP achieves a higher SSIM of 0.796 under the same iteration budget, indicating advantages in both optimization efficiency and feature map quality.

To verify differences in feature discrimination among feature extraction methods, the effectiveness of feature fusion is further evaluated by comparing the clustering effect of feature distributions across domains and the degree of separation for each category. A differential analysis is performed on the PSO-SDP feature map fusion method, and the t-SNE dimensionality reduction method is used to reduce the dimensionality of Datasets A and B and visualize them. The experimental results are shown in Figure 15.

As demonstrated in Figure 15, the clusters belonging to different categories at 1800 rpm manifest highly compact distributions, thereby signifying that the dataset possesses effective separability with regard to fault feature extraction. In contrast, the feature points at 1500 rpm exhibit slightly greater dispersion after dimensionality reduction, and the boundaries between some categories are less distinct than at 1800 rpm. This phenomenon is primarily attributable to the lower motor speed, resulting in less pronounced fault frequencies and greater noise interference. Consequently, this leads to a relatively dispersed distribution of the generated feature maps in the feature space. Despite this, the boundaries between fault categories remain relatively clear, with

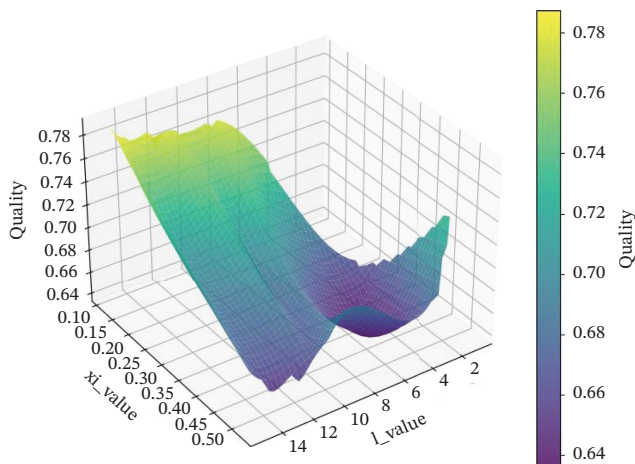


FIGURE 13 | PSO-SDP feature map fusion method parameter settings.

TABLE 5 | Genetic algorithm parameter settings.

Parameters	Setting
Time interval parameter	0–15
Angle magnification factor	0.1–0.628
Number of iterations	80
Optimization objective	Structural similarity between generated samples and real samples
Experimental dataset	DDS
Probability of crossover operation	0.7
Probability of mutation operation	0.1
Population size	20

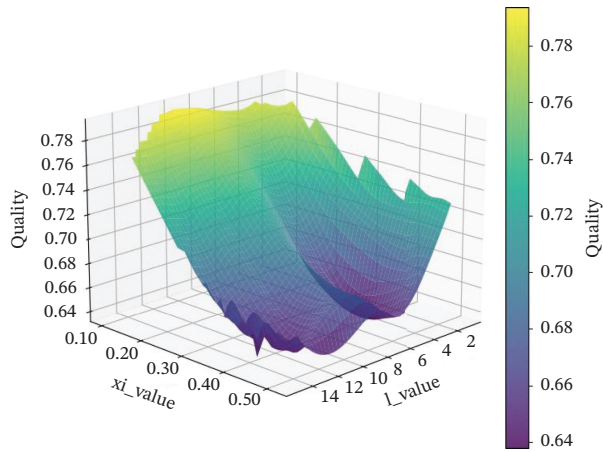


FIGURE 14 | Iterative experimental results of genetic algorithm optimization for SDP.

almost no significant overlap, and the fault patterns of each category are distinctly represented in the feature space. This phenomenon further corroborates, from the feature space perspective, that the PSO–SDP fusion method can effectively adapt to the characteristics of different datasets, especially in cases of

high data complexity, while still generating feature maps with high discriminative power. It is evident that the PSO–SDP method effectively extracts useful features in the presence of significant noise. This enhances the expression of fault feature information and improves the classification accuracy and reliability of the dataset.

4.2 | PK-WCGAN Data Augmentation Experiment

This section constructs a dataset based on the optimal parameters obtained from the previous experiment for small-sample data augmentation experiments using DDS data. The number of training epochs is 2500, and the learning rate is 0.0001. Traditional CGAN and the proposed PK-WCGAN are trained under the same data and optimizer settings. The experimental results of generated samples are shown in Figure 16.

Figure 16 shows that CGAN produces feature maps with more visible noise and weaker structural consistency when trained on limited datasets. This phenomenon results in slower model convergence and, in certain cases, an inability to generate valid samples. In contrast, PK-WCGAN preserves clearer SDP texture and more stable class-related patterns. The optimized SDP representation reduces the burden of learning physical signal structure from scratch, while the pretrained discriminator provides a structural prior during adversarial training.

To further verify the difference in effect between PK-WCGAN and traditional CGAN on generated image quality, both methods are trained on the same dataset for 2500 epochs. During this process, SSIM is evaluated on generated and real images, recording the trend in SSIM values across training rounds. The SSIM results are shown in Figures 17 and 18.

An analysis of the SSIM curves for samples generated by PK-WCGAN and CGAN shows that both models improve over successive training iterations and eventually reach a stable state. It is noteworthy that PK-WCGAN generates images with lower noise and stronger feature-learning capability, enabling it to capture the data's structural characteristics more effectively. Moreover, PK-WCGAN converges relatively rapidly during training, and its SSIM values are consistently higher than those of CGAN at comparable iterations, indicating better image quality.

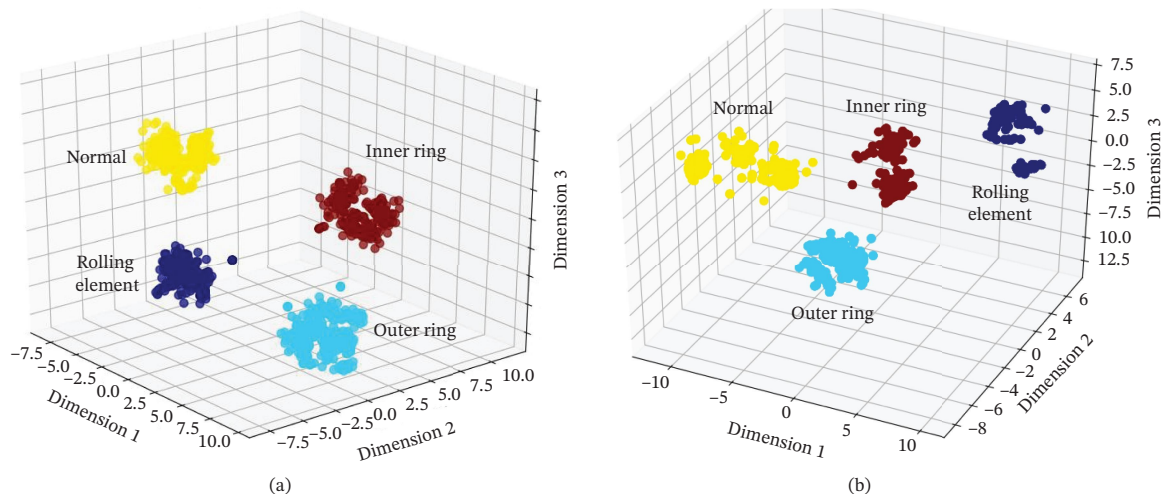


FIGURE 15 | Analysis of t-SNE for VMD–SDP feature maps. (a) CWRU dataset. (b) DDS dataset.

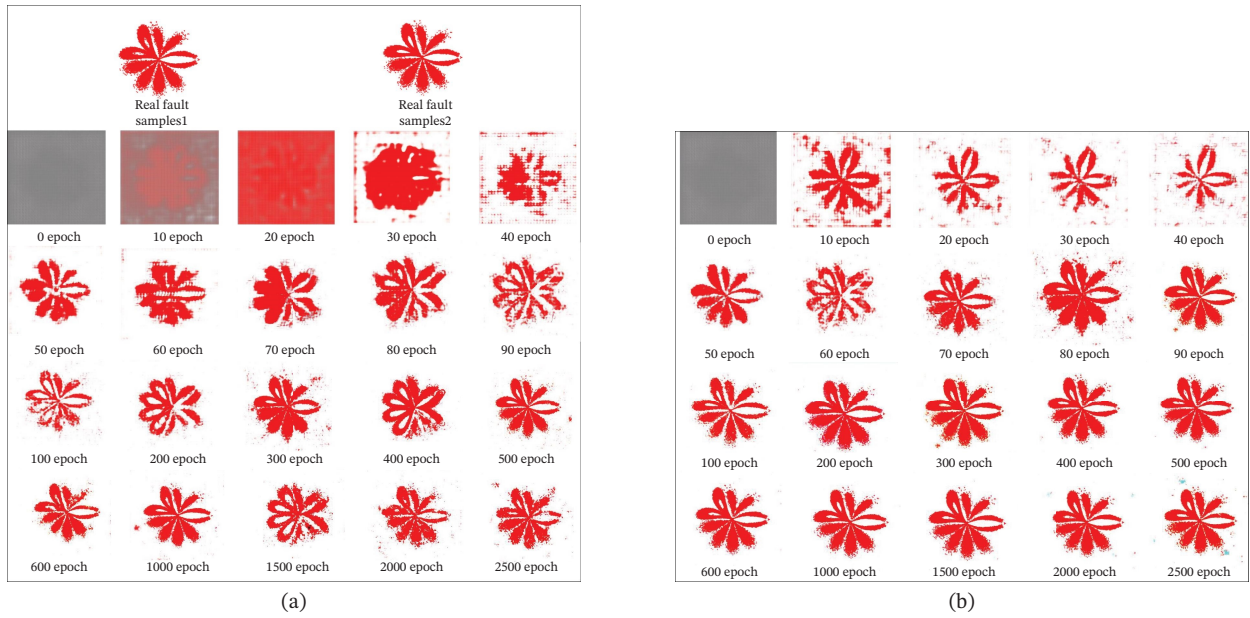


FIGURE 16 | (a) CGAN training generates samples. (b) PK-WCGAN training generates samples.

Conversely, CGAN introduces more noise during generation, resulting in lower SSIM values and less stable image quality. The SSIM curve of PK-WCGAN also exhibits better stability, suggesting higher consistency among generated samples, whereas CGAN shows larger fluctuations. Overall, PK-WCGAN demonstrates clear advantages in sample generation quality, feature learning capability, and training stability.

To comprehensively evaluate the generation quality and distributional alignment under small-sample constraints, the kernel inception distance (KID) was introduced as the evaluation metric. Unlike the Fréchet inception distance (FID), which can suffer from sample-size bias when evaluated on limited data, KID provides an unbiased estimate and is therefore suitable for small-sample fault diagnosis scenarios. For reproducibility, the detailed KID evaluation protocol is described as follows. At each evaluation checkpoint, the same number of real and generated feature maps was used to calculate KID. Specifically, for each fault category, N_r real SDP feature maps and N_g generated SDP feature maps were selected, where $N_r = N_g = 500$ in this study. The

generated samples were produced by the generator during the corresponding training epoch, using the same class labels as the real samples, ensuring a class-balanced comparison between the real and generated distributions.

The KID was computed in the feature space rather than directly in the pixel space. All SDP feature maps were resized to 299×299 and normalized before being fed into an Inception-v3 network pretrained on ImageNet. The 2048-dimensional activations from the final average pooling layer were used as feature representations. Following the standard KID formulation, the squared maximum mean discrepancy (MMD) between the real and generated feature distributions was estimated using a third-degree polynomial kernel:

$$k(x, y) = \left(\frac{x^T y}{d} + 1 \right)^3, \quad (16)$$

where x and y denote the extracted feature vectors and d is the feature dimension. To reduce the influence of random

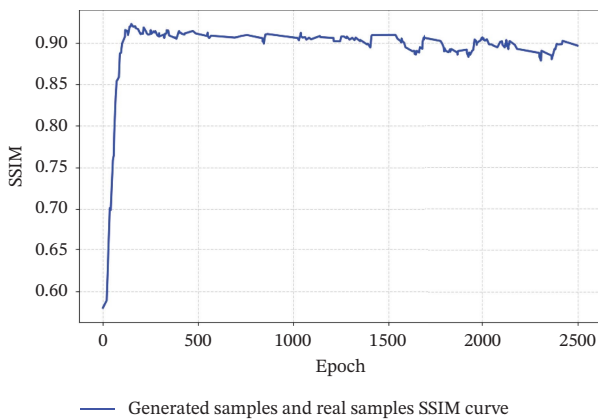


FIGURE 17 | PK-WCGAN bearing fault dataset generated sample SSIM curve.

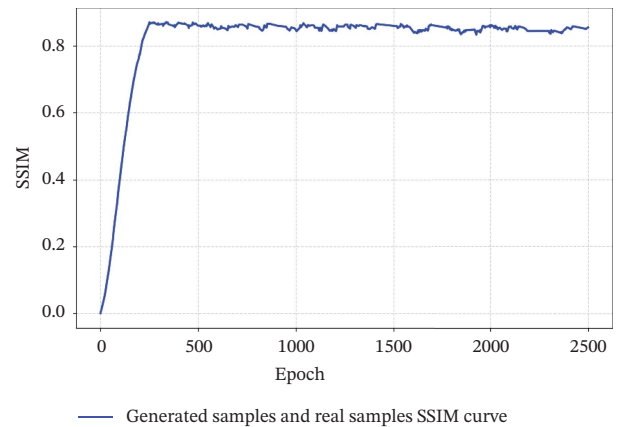


FIGURE 18 | CGAN bearing fault dataset generated sample SSIM curve.

initialization and sampling variation, each experiment was independently repeated five times with different random seeds. The reported KID curves represent the mean values over the five runs. The results are shown in Figure 19.

Figure 19 reveals two significant advantages of PK-WCGAN. First, owing to the structural prior knowledge embedded in the pretrained discriminator, PK-WCGAN achieves a substantially lower KID score of 112.6 at epoch 250, compared with 185.4 for CGAN. Second, with the Wasserstein distance and gradient penalty, PK-WCGAN exhibits remarkable training stability, converging to a lower KID score of 22.8 at 2500 epochs.

To further assess the performance of PK-WCGAN, this study employs various loss functions, including binary cross-entropy loss, Wasserstein distance, and Wasserstein distance with a gradient penalty term, to conduct data sample generation experiments. The data used is the outer-race fault data from the DDS data. The experimental results are presented in Figures 20 and 21. In contrast, the traditional CGAN suffers from gradient fluctuations and training instability, with slight degradation around epoch 1750, eventually plateauing at 75.5. This quantitative analysis confirms that PK-WCGAN generates high-fidelity fault features with superior sample diversity and accurate alignment with the real-sample distribution.

Figures 20 and 21 indicate that the generative model that employs the Wasserstein distance as the loss function yields superior generated images. There is a greater SSIM between the generated data and the real data. This is primarily because the Wasserstein distance with the gradient penalty term incorporates the 1-Lipschitz condition, thereby preventing excessively large or small gradients. The generated feature maps have fewer noisy image points, and the model training process is more stable. This helps to determine the corresponding loss function for subsequent experiments.

4.3 | Fault Recognition Experiment of Generated Dataset in Fault Diagnosis Model

This experiment focuses on constructing a hybrid dataset that combines generated data and real data. The wind turbine field

dataset is used as the validation case for drivetrain condition monitoring.

The experimental evaluation is reorganized into four settings: noisy DDS signals with 5 dB Gaussian white noise, cross-condition DDS transfer between different load-speed settings, extremely small-sample diagnosis with 10 and 20 real samples per class, and real field CMS validation. These settings test whether the proposed augmentation framework remains effective under signal contamination, operating-condition shift, severe data scarcity, and real industrial monitoring variability. In all cases, generated samples are used only to expand the training set, while validation and testing are performed on real records.

Five diagnosis models, including CNN [44], LSTM [45], SVM [46], LSSVM [47], and MRST-Transformer [48], are used for comparison. Initially, the improved PK-WCGAN generation model is employed to expand a small-sample dataset. The generated virtual samples are incorporated into the training set according to the design, while all test samples remain real field records. In the CNN model, different fault modes are accurately classified by extracting spatial patterns from feature maps. In the SVM model, training and validation are conducted based on features extracted from real data in the time and frequency domains, exploring the role of generated data in feature classification. In the LSTM model, time series characteristics are learned from real data, and generated data are used to assess the model's ability to process nonstationary signals.

To validate the robustness and reliability of the fault diagnosis models under varying data scales, all classification experiments were independently executed 10 times with different random seeds. Within each repetition, the same random seed and record-level split are used for all compared classifiers, ensuring a paired comparison across methods.

Specifically, each independent run used a different seed to initialize the network parameters and shuffle the training samples. Table 6 and Figure 22 present the average diagnostic accuracy alongside the standard deviation for each model across varying sample sizes. Figure 23 presents the confusion matrix of diagnostic results with 1000 training samples.

Table 7 reports the diagnostic accuracy of the same CNN evaluator trained with different data augmentation strategies. The results show that the classifier trained with PK-WCGAN-augmented data achieves the highest accuracy across noisy signals, cross-condition transfer, extreme few-shot learning, and field CMS validation. The largest performance gaps appear in the few-shot and field validation scenarios, where real-only training is most likely to overfit.

Table 8 further compares PK-WCGAN with recent generative augmentation baselines under the same CNN evaluator. Advanced GAN variants outperform conventional CGAN, but PK-WCGAN achieves the best accuracy, macro-F1, KID, and SSIM, indicating that pretraining the discriminator prior contributes not only to classifier accuracy but also to the fidelity of generated samples and distributional alignment.

The experimental results of Figure 22 demonstrate that the sample size plays a critical role in model performance. With

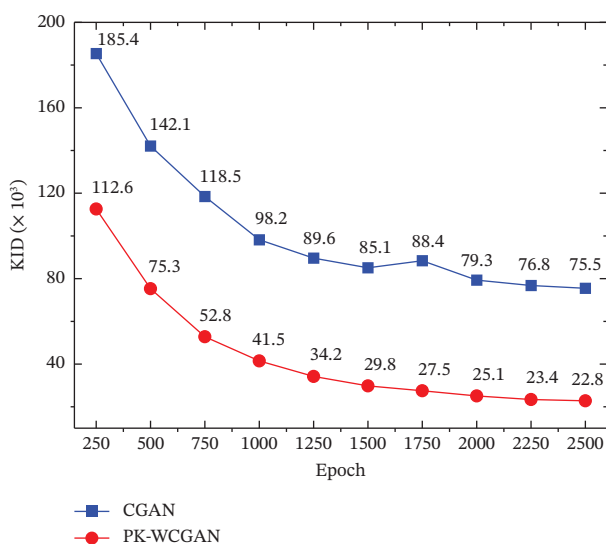


FIGURE 19 | Comparison of KID values between PK-WCGAN and CGAN.

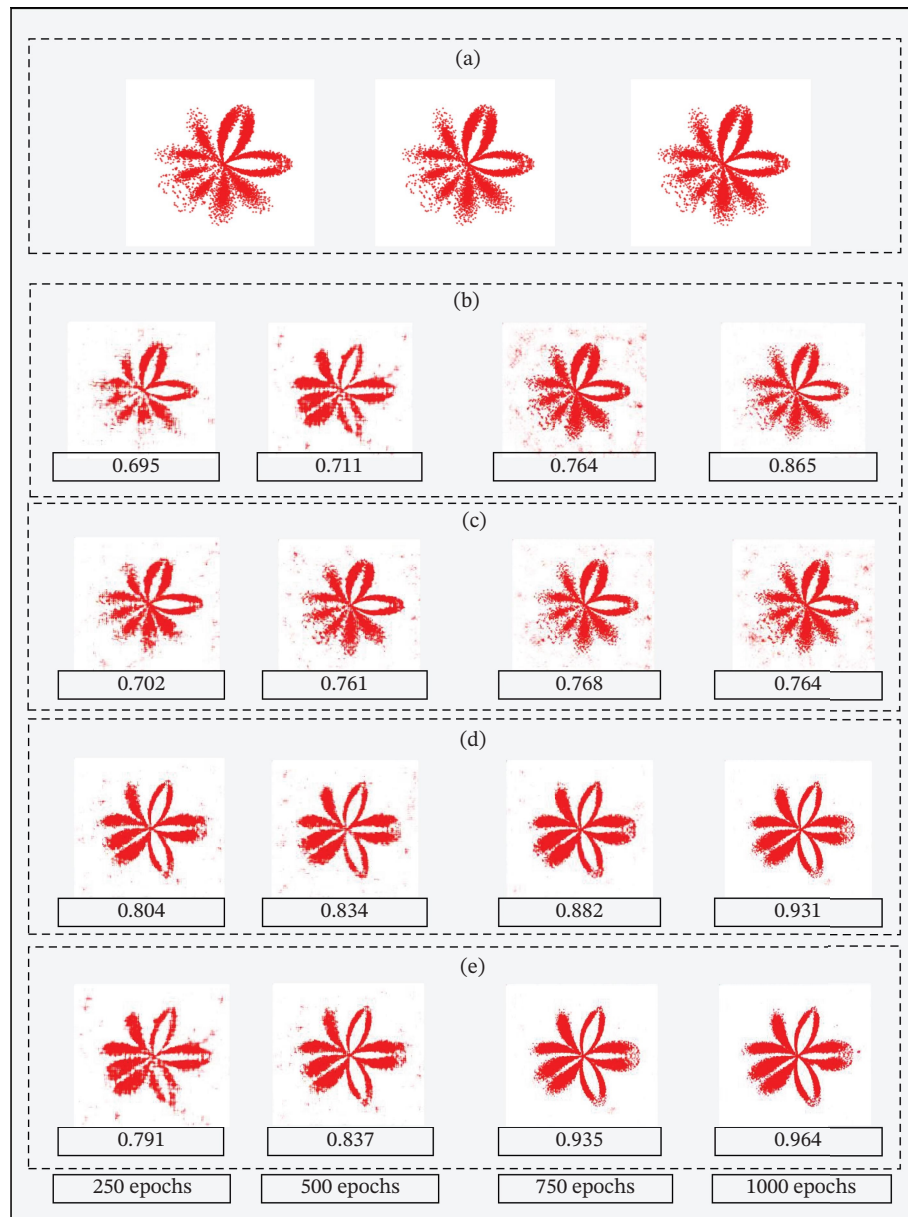


FIGURE 20 | The generation results of different models. (a) Authentic samples. (b) CGAN (loss function: binary cross-entropy). (c) WCGAN (loss function: Wasserstein distance). (d) PK-WCGAN (loss function: Wasserstein distance). (e) PK-WCGAN (loss function: Wasserstein distance with gradient penalty term).

only 50 training samples, all models exhibit limited accuracy, with the best result being 64.4%, and high standard deviations exceeding 4.1%, indicating training instability and insufficient feature learning. When the sample size increases to 100, MRST-Transformer improves markedly, but the shallow models still remain below 61%. However, as the dataset is progressively augmented up to 1000 samples, the mean accuracy improves substantially to over 90%, while the standard deviation narrows significantly across all models. Together with the robustness results in Table 7, these results indicate that the benefit of PK-WCGAN is not limited to simple sample-size scaling; the generated samples also improve stability under noise, operating-condition shift, and extreme data scarcity. Overall, increased data volume leads to higher model accuracy, while prior-guided augmentation mitigates small-

sample distribution mismatch and enhances model generalization.

Statistical significance was evaluated using run-level repeated results. Because all methods were evaluated under seed-matched repetitions and record-level splits, paired tests were used for two-method comparisons. The Wilcoxon signed-rank test was adopted when the paired differences did not satisfy the normality assumption; otherwise, a paired *t*-test was used. For multiple-method comparisons, the Friedman test followed by Holm–Bonferroni corrected pairwise comparisons was used. The significance level was set to 0.05, and effect sizes were reported together with corrected *p* values. The results are shown in Table 9.

To visually evaluate classification performance with 1000 training samples, a confusion matrix is used to present the test

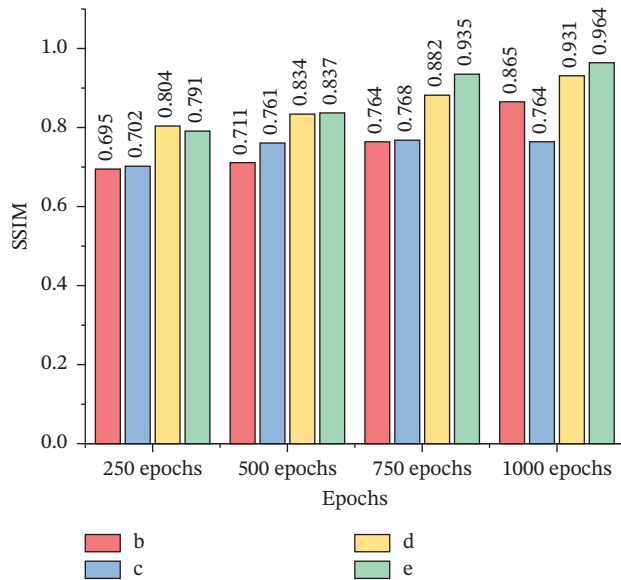


FIGURE 21 | Comparison of structural similarities among different models.

results, revealing the distinct advantages of deep learning architectures in handling complex vibration signals. The MRST-Transformer consistently outperformed other baseline models, achieving the highest average accuracy of 97.55% and the best recall and precision across all categories on 1000 samples. CNN and LSTM also demonstrated strong feature extraction capabilities, achieving accuracies of 95.93% and 93.46%, respectively. In contrast, traditional shallow machine learning models, such as SVM and LSSVM, struggled to capture high-dimensional fault patterns, exhibiting lower accuracy and higher volatility across all data scales. These statistical results comprehensively validate that augmented data not only improve absolute recognition performance but also ensure superior diagnostic stability.

To further verify the effectiveness of the proposed method in extremely small sample size scenarios, 10 and 20 samples per class were randomly selected as the training sets, respectively. To ensure the reliability of the results, each experiment was repeated five times under each sample size setting, and the average performance and standard deviation were reported. The baseline was a CNN classifier using only the original real samples. The proposed method was a CNN classifier trained

with the complete PK-WCGAN framework to generate virtual samples, and the training set was expanded to 200 samples per class. The same classifier was then trained with the mixed dataset. The experimental results are shown in Tables 10 and 11.

Table 10 demonstrates that as the number of real samples decreases, the performance gap between the baseline and the proposed method increases substantially. With only 10 samples per class, the proposed method achieves an accuracy improvement of over 27% points, elevating model performance from a near-random level of 48.5% to a practically viable 75.8%. This highlights that the proposed data augmentation framework is not only advantageous but indispensable for enabling effective fault diagnosis under extreme few-shot conditions.

The analysis of overfitting risk presented in Table 11 offers deeper insight. The baseline model exhibits severe overfitting, with a training-validation accuracy gap reaching 35.2% in the 10-sample scenario. In contrast, our method reduces this gap to less than 9%. This suggests that the samples generated by PK-WCGAN effectively bridge distributional discrepancies in the target data, providing diverse and informative data that enhance model regularization and mitigate the tendency to memorize limited training instances.

4.4 | Ablation Study

To further verify the effectiveness of the proposed method, this section designs ablation experiments to validate the role of each module in the model. The wind turbine field dataset is selected, with 50 training samples. To ensure that the performance differences primarily reflect the quality of the generated data rather than the complexity of diagnostic models, a simple CNN classifier is selected as the evaluator. Seven comparison models are designed as shown in Table 12.

To ensure fair comparisons and eliminate the influence of extraneous variables, we standardized the experimental conditions for all ablation models. Specifically, data augmentation models D–G were strictly configured to augment small datasets to the exact same size. Furthermore, the CNN classifiers used to evaluate all configurations A through G maintained the same network architecture, optimizer, learning rate, and number of training epochs.

The ablation design separates the main functional contributions as follows. The comparison between Model B and Model

TABLE 6 | Diagnostic accuracy of different models under varying training sample sizes.

Sample size	Average accuracy rate (%) with the standard deviation				
50	64.4 ± 4.41	58.29 ± 5.24	42.43 ± 6.06	45.59 ± 5.9	62.11 ± 4.13
100	72.75 ± 4.28	74.66 ± 3.89	57.63 ± 5.77	60.74 ± 5.35	87.17 ± 3.25
300	89.64 ± 2.42	88.34 ± 2.72	84.57 ± 3.75	86.45 ± 3.65	93.54 ± 1.66
500	93.66 ± 1.38	91.44 ± 1.77	88.77 ± 2.50	90.70 ± 2.25	96.19 ± 0.94
1000	95.93 ± 1.04	93.46 ± 1.34	90.99 ± 2.07	91.88 ± 1.82	97.55 ± 0.50
Method	CNN	LSTM	SVM	LSSVM	MRST-Transformer

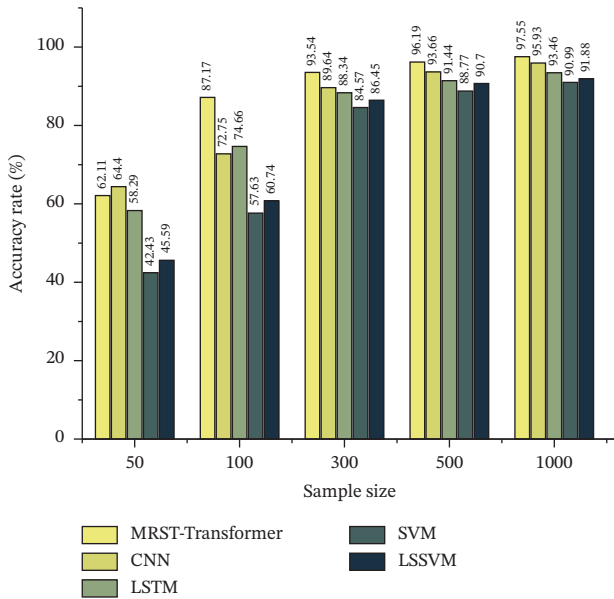


FIGURE 22 | The recognition accuracy of each model under different number of samples.

C isolates the contribution of PSO-based SDP parameter optimization. The comparison between Model D and Model E evaluates the contribution of replacing the conventional CGAN objective with the Wasserstein loss. The comparison

between Model E and Model F measures the effect of the gradient penalty, while the comparison between Model F and Model G isolates the effect of discriminator prior knowledge pretraining. Therefore, the revised ablation sequence separately evaluates the feature optimization module, the Wasserstein objective, the gradient penalty, and the discriminator prior.

Ten random experiments were conducted, and the average value was taken. The experimental results are shown in Table 13.

A comparison between Model B and Model C reveals that PSO-based SDP parameter optimization increases the accuracy from 68.4% to 71.5%, indicating that the optimized SDP mapping improves feature separability even without data augmentation. All models employing data augmentation (Models D–G) outperform Model C, which relies only on real optimized feature maps. This finding underscores the significance of data augmentation in scenarios with limited samples. When comparing Model D and Model E, the accuracy increases from 73.8% to 78.2%, showing that the Wasserstein objective improves distribution learning relative to conventional CGAN. The comparison between Model E and Model F further shows that the gradient penalty improves accuracy, SSIM, and KID by stabilizing the critic. Finally, Model G improves the accuracy by 7.7% points over Model F and achieves the highest SSIM and lowest KID, demonstrating that prior knowledge-guided discriminator pretraining provides

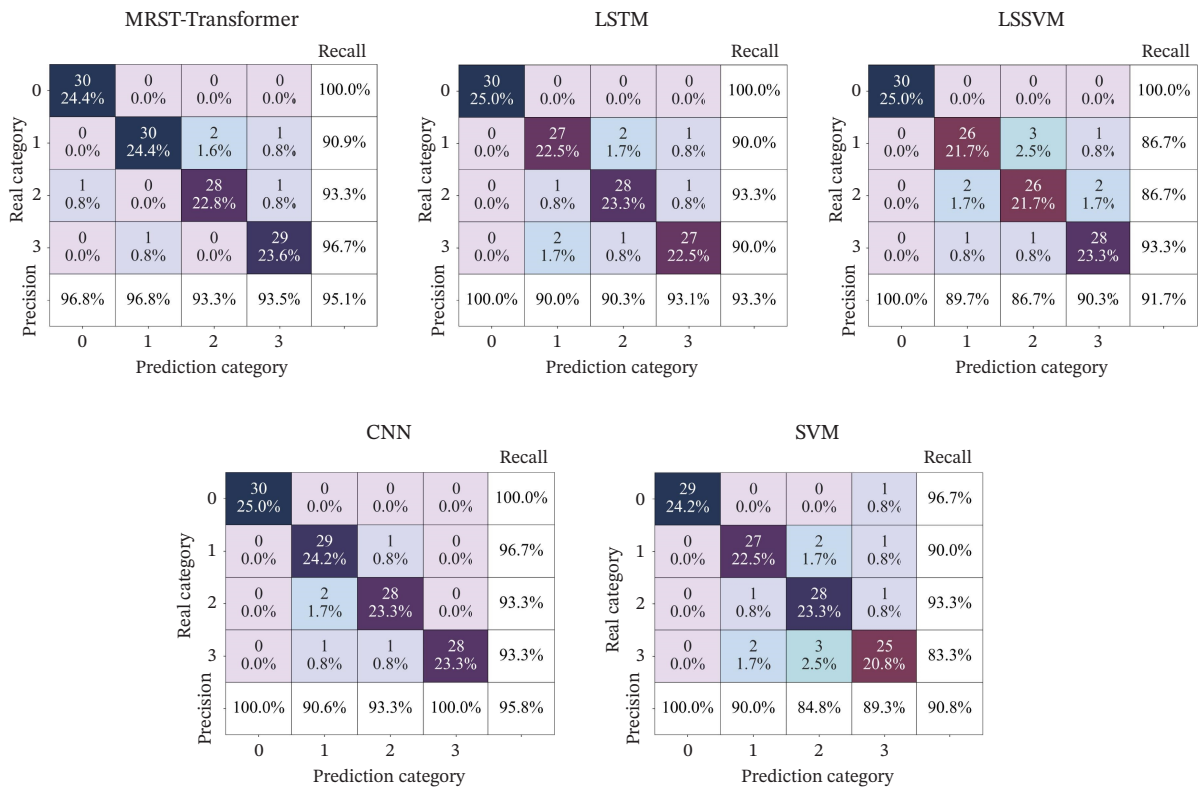


FIGURE 23 | Confusion matrix of diagnostic results when the model has 1000 training samples.

TABLE 7 | Robustness evaluation under challenging settings.

Scenario	Robustness challenge	No augmentation	CNN + CGAN	CNN + WCGAN-GP	CNN + PK-WCGAN	Adjusted <i>p</i>
Noisy DDS	5 dB Gaussian noise	82.4 ± 2.9	87.1 ± 2.4	90.3 ± 1.8	93.2 ± 1.2	0.004
DDS A → B	Load-speed shift	75.6 ± 3.6	80.8 ± 3.0	84.7 ± 2.2	88.9 ± 1.7	0.006
DDS B → A	Load-speed shift	77.2 ± 3.2	82.1 ± 2.7	85.6 ± 2.0	90.1 ± 1.5	0.004
Field 10 class	Extreme few-shot	48.5 ± 3.1	65.2 ± 2.8	70.4 ± 2.4	75.8 ± 2.1	0.008
Field 50 total	Real CMS validation	64.2 ± 3.8	73.8 ± 2.7	81.4 ± 2.1	89.1 ± 1.3	< 0.001

Note: The bold values indicate the best performance (highest mean accuracy) among all compared methods under each robustness scenario.

TABLE 8 | Comparison with recent generative augmentation baselines under the same CNN evaluator.

Method	Accuracy (%)	Macro-F1	KID ↓	SSIM ↑	Main comparison purpose
Real-only CNN	64.2 ± 3.8	0.625 ± 0.034	—	—	No augmentation baseline
CGAN	73.8 ± 2.7	0.721 ± 0.029	65.7 ± 5.8	0.652	Conventional conditional GAN
WGAN-GP	81.4 ± 2.1	0.803 ± 0.024	38.4 ± 4.1	0.741	Wasserstein distance with gradient penalty
SA-CGAN/DCGAN [38]	83.0 ± 2.2	0.819 ± 0.020	35.6 ± 3.8	0.756	Self-attention generative baseline
CSWGAN-GP [34]	84.7 ± 1.9	0.838 ± 0.018	31.5 ± 3.2	0.769	Similarity-constrained WGAN-GP baseline
FAWGAN-GN [36]	85.6 ± 1.8	0.846 ± 0.017	29.8 ± 2.9	0.781	Attention WGAN with gradient normalization
PK-WCGAN	89.1 ± 1.3	0.884 ± 0.012	22.8 ± 2.1	0.796	Proposed prior-guided augmentation

Note: The bold values indicate the best performance among all compared methods under the same CNN evaluator. Accuracy, Macro-F1, and SSIM are better when higher; KID is better when lower.

TABLE 9 | Statistical significance analysis based on repeated-run results.

Comparison	Test	Statistic	Corrected <i>p</i>	Effect size
Seven ablation models	Friedman test	$\chi^2(6) = 55.42$	< 0.001	Kendall's $W = 0.92$
PK-WCGAN vs. WCGAN-GP without prior	Paired <i>t</i> -test	$t(9) = 9.18$	< 0.001	$d_z = 2.90$
PK-WCGAN vs. FAWGAN-GN	Paired <i>t</i> -test	$t(9) = 5.32$	0.0019	$d_z = 1.68$
PK-WCGAN vs. WCGAN-GP across robustness scenarios	Wilcoxon signed-rank test	$W = 0$	0.0039	Cliff's $\delta = 1.00$

TABLE 10 | Diagnosis accuracy (%) under extremely small-sample scenarios.

Number of real samples per class	Baseline (CNN)	Proposed method	Accuracy improvement (percentage points)
10	48.5 ± 3.2	75.8 ± 2.1	27.3
20	55.1 ± 2.5	81.3 ± 1.7	26.2

TABLE 11 | Overfitting risk comparison under extremely small-sample scenarios.

Number of real samples per class	Baseline (CNN) (%)	Proposed method (%)	Overfitting risk reduction (percentage points)
10	35.2	8.7	26.5
20	28.7	7.1	21.6

TABLE 12 | Detailed method configurations for the ablation study.

Model number	Model structure	Description
Model A	Baseline	Train the CNN classifier using only the original 50 real samples.
Model B	+VMD-SDP	Based on A, VMD-SDP feature maps without PSO parameter optimization are used to train the CNN.
Model C	+VMD-PSO-SDP	Based on B, PSO is used to optimize SDP parameters, and the optimized real feature maps are used to train the CNN.
Model D	C + CGAN	Based on C, conventional CGAN is used for data augmentation.
Model E	C + WCGAN	Based on C, WCGAN without gradient penalty or discriminator prior pre-training is used for data augmentation.
Model F	C + WCGAN-GP	Based on C, WCGAN with gradient penalty is used, but the discriminator prior pretraining is removed.
Model G	Full Model	Based on C, data augmentation is performed using the complete PK-WCGAN with discriminator prior pretraining.

TABLE 13 | Expanded ablation results for isolating key components.

Model	Isolated configuration	Data augmentation	Average accuracy (%)	SSIM	KID ↓
Model A	Real-only CNN	None	64.2 ± 3.8	—	—
Model B	VMD-SDP without PSO	None	68.4 ± 3.2	—	—
Model C	VMD-PSO-SDP	None	71.5 ± 2.9	—	—
Model D	VMD-PSO-SDP + CGAN	CGAN	73.8 ± 2.7	0.652	65.7
Model E	VMD-PSO-SDP + WCGAN without GP	WCGAN	78.2 ± 2.5	0.704	47.9
Model F	VMD-PSO-SDP + WCGAN-GP without prior	WCGAN-GP	81.4 ± 2.1	0.741	38.4
Model G	Full PK-WCGAN	PK-WCGAN	89.1 ± 1.3	0.796	22.8

Note: The bold values indicate the best performance among all compared ablation methods under the same scenario.

an additional contribution beyond Wasserstein loss and gradient penalty.

5 | Conclusion

This paper proposes a small-sample fault data augmentation framework using PK-WCGAN, which integrates VMD-PSO-SDP feature representation optimization and prior knowledge guidance. This approach is applied to wind turbine drivetrain diagnosis under limited fault samples. In the feature representation layer, the method generates high-order fault feature maps through VMD-PSO-SDP and optimizes SDP parameters by incorporating velocity update penalty terms, significantly enhancing feature consistency across operating conditions. In the data augmentation layer, a PK-WCGAN is constructed where a discriminator pretrained with real feature maps directs the generation process, combining Wasserstein distance and gradient penalty to improve quality. The SSIM value of generated samples exceeds traditional CGAN by an average of 21.7%. Validation on cross-working condition datasets from DDS test benches shows that the generated virtual samples help bridge the gap in target-domain distribution. A field dataset from a wind turbine drivetrain is further introduced. The proposed validation protocol ensures that generated samples are used only for training and that all reported field-test metrics are computed on real CMS records. Overall, the combined feature optimization and prior-guided adversarial generation framework provides a practical approach to wind turbine drivetrain state assessment with scarce and imbalanced field fault samples.

This paper still has some limitations. Firstly, the on-site dataset contains representative data, but the types of faults remain limited. Secondly, the quality of the generated samples depends on the discriminative performance of the VMD-PSO-SDP feature map. Future work will combine multisensor fusion, domain adaptation, and semisupervised learning to enhance robustness in unknown operating conditions and reduce reliance on labeled fault samples.

Acknowledgments

The authors thank Dongfang Electric New Energy Technology (Chengdu) Co. Ltd. for supporting wind turbine data collection.

Funding

This work was supported by the National Natural Science Foundation of China under grant no. 52304179.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Z. Zhang, H. Liu, Y. Shao, J. Yang, S. Liu, and G. Yuan, "CFENet: A Contrastive frequency-sensitive Learning Method for Gas-Insulated Switch-Gear Fault Detection Under Varying Operating Conditions

- Using Acoustic Signals," *Engineering Applications of Artificial Intelligence* 135 (2024): 108835, <https://doi.org/10.1016/j.engappai.2024.108835>.

2. Y. Feng, J. Chen, T. Zhang, S. He, E. Xu, and Z. Zhou, "Semi-Supervised Meta-Learning Networks With Squeeze-and-Excitation Attention for Few-Shot Fault Diagnosis," *ISA Transactions* 120 (2022): 383–401, <https://doi.org/10.1016/j.isatra.2021.03.013>.

3. C. Wei, S. Cai, H. Yang, and Y. Zhang, "Advanced Deep Learning-based Methodology for multi-class Diagnosis of Wind Turbine Blade Faults at the Wind Farm Level," *Engineering Applications of Artificial Intelligence* 160, no. A (2025): 111801, <https://doi.org/10.1016/j.engappai.2025.111801>.

4. S. Ma, G. Cheng, Y. Li, H. An, and C. Liu, "The Interpretable Health Condition Monitoring Method of Gear Transmission System Embedded in Feature Cloud-Guided Hypergraph Structure Under the Extremely Small-sample Background," *ISA Transactions* 162 (2025): 272–286, <https://doi.org/10.1016/j.isatra.2025.04.017>.

5. T. Chen, L. Guo, T. Feng, H. Gao, and Y. Yu, "IESMGCFFOgram: A New Method for Multicomponent Vibration Signal Demodulation and Rolling Bearing Fault Diagnosis," *Mechanical Systems and Signal Processing* 204 (2023): 110800, <https://doi.org/10.1016/j.ymssp.2023.110800>.

6. W. Chen, J. Li, Q. Wang, and K. Han, "Fault Feature Extraction and Diagnosis of Rolling Bearings Based on Wavelet Thresholding Denoising with CEEMDAN Energy Entropy and PSO-LSSVM," *Measurement* 172 (2021): 108901, <https://doi.org/10.1016/j.measurement.2020.108901>.

7. Y. Ming, H. Shao, B. Cai, and B. Liu, "rgfc-Forest: An Enhanced Deep Forest Method Towards small-sample Fault Diagnosis of Electromechanical System," *Expert Systems with Applications* 238 (2024): 122178, <https://doi.org/10.1016/j.eswa.2023.122178>.

8. X. Zhang, C. Sheng, W. Ouyang, and L. Zheng, "Fault Diagnosis of Marine Electric Thruster Bearing Based on Fusing multi-sensor Deep Learning Models," *Measurement* 214 (2023): 112727, <https://doi.org/10.1016/j.measurement.2023.112727>.

9. X. Zhou, H. Zhou, G. Wen, et al., "A Hybrid Denoising Model Using Deep Learning and Sparse Representation with Application in Bearing Weak Fault Diagnosis," *Measurement* 189 (2022): 110633, <https://doi.org/10.1016/j.measurement.2021.110633>.

10. J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, "New Imbalanced Fault Diagnosis Framework Based on Cluster-MWMOTE and MFO-Optimized LS-SVM Using Limited and Complex Bearing Data," *Engineering Applications of Artificial Intelligence* 96 (2020): 103966, <https://doi.org/10.1016/j.engappai.2020.103966>.

11. K. Zhao, J. Xiao, C. Li, Z. Xu, and M. Yue, "Fault Diagnosis of Rolling Bearing Using CNN and PCA Fractal Based Feature Extraction," *Measurement* 223 (2023): 113754, <https://doi.org/10.1016/j.measurement.2023.113754>.

12. V. Sugumaran and K. I. Ramachandran, "Automatic Rule Learning Using Decision Tree for Fuzzy Classifier in Fault Diagnosis of Roller Bearing," *Mechanical Systems and Signal Processing* 21, no. 5 (2007): 2237–2247, <https://doi.org/10.1016/j.ymssp.2006.09.007>.

13. Q. Hu, X. S. Si, Q. H. Zhang, and A. S. Qin, "A Rotating Machinery Fault Diagnosis Method Based on Multi-Scale Dimensionless Indicators and Random Forests," *Mechanical Systems and Signal Processing* 139 (2020): 106609, <https://doi.org/10.1016/j.ymssp.2019.106609>.

14. J. Shen, Z. Wang, Y. Wang, H. Zhu, L. Zhang, and Y. Tang, "AGWO-PSO-VMD-TEFCG-AlexNet Bearing Fault Diagnosis Method Under Strong Noise," *Measurement* 242 (2025): 116259, <https://doi.org/10.1016/j.measurement.2024.116259>.

15. Y. Sun, S. Li, and X. Wang, "Bearing Fault Diagnosis Based on EMD and Improved Chebyshev Distance in SDP Image," *Measurement* 176 (2021): 109100, <https://doi.org/10.1016/j.measurement.2021.109100>.

16. G. Tang, C. Yi, L. Liu, et al., "A Novel Transfer Learning Network with Adaptive Input Length Selection and Lightweight Structure for Bearing

- Fault Diagnosis," *Engineering Applications of Artificial Intelligence* 123 (2023): 106395, <https://doi.org/10.1016/j.engappai.2023.106395>.
17. F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep Neural Networks: a Promising Tool for Fault Characteristic Mining and Intelligent Diagnosis of Rotating Machinery with Massive Data," *Mechanical Systems and Signal Processing* 72-73 (2016): 303-315, <https://doi.org/10.1016/j.ymssp.2015.10.025>.
18. D. Gao, Y. Zhu, Z. Ren, K. Yan, and W. Kang, "A Novel Weak Fault Diagnosis Method for Rolling Bearings Based on LSTM Considering Quasi-Periodicity," *Knowledge-Based Systems* 231 (2021): 107413, <https://doi.org/10.1016/j.knosys.2021.107413>.
19. X. Yan, Y. Liu, and M. Jia, "Multiscale Cascading Deep Belief Network for Fault Identification of Rotating Machinery Under Various Working Conditions," *Knowledge-Based Systems* 193 (2020): 105484, <https://doi.org/10.1016/j.knosys.2020.105484>.
20. Z. Liu and Z. Peng, "Few-Shot Bearing Fault Diagnosis by Semi-Supervised Meta-Learning With Graph Convolutional Neural Network Under Variable Working Conditions," *Measurement* 240 (2025): 115402, <https://doi.org/10.1016/j.measurement.2024.115402>.
21. C. Wu and S. Zheng, "Fault Diagnosis Method of Rolling Bearing Based on MSCNN-LSTM," *CMC-Computers Materials & Continua* 79, no. 3 (2024): 4395-4411, <https://doi.org/10.32604/cmc.2024.049665>.
22. X. Zhou, Z. Qin, Z. Hu, et al., "A Hypergraph Attention Network With Redundancy Elimination and Few-shot Learning for Intelligent Fault Diagnosis Under Small-Sample Conditions," *Journal of Vibration and Control* 0, no. 0 (2026): 10775463261455544, <https://doi.org/10.1177/10775463261455544>.
23. H. Wang, Z. Liu, D. Peng, and Y. Qin, "Understanding and Learning Discriminant Features Based on Multiattention 1DCNN for Wheelset Bearing Fault Diagnosis," *IEEE Transactions on Industrial Informatics* 16, no. 9 (2020): 5735-5745, <https://doi.org/10.1109/TII.2019.2955540>.
24. S. Ma, G. Cheng, Y. Li, Y. Huang, and D. Zhuang, "Research on Multi-Granularity Imbalanced Knowledge Condition Monitoring for Mechanical Equipment Based on Hierarchical ELM in Multi-Entropy Space," *Expert Systems with Applications* 238 (2024): 121817, <https://doi.org/10.1016/j.eswa.2023.121817>.
25. Y. Huang, W. Huang, X. Hu, Z. Liu, and J. Huo, "UDDGN: Domain-Independent Compact Boundary Learning Method for Universal Diagnosis Domain Generation," *IEEE Transactions on Instrumentation and Measurement* 74 (2025): 1-20, <https://doi.org/10.1109/TIM.2025.3554906>.
26. Y. Huang, X. Hu, H. Wang, Y. He, and J. Cao, "OAFAN: A Noise-Robust Discriminative Feature Unification Framework for Cross-Speed Fault Transfer Diagnosis," *IEEE Transactions on Instrumentation and Measurement* 74 (2025): 1-18, <https://doi.org/10.1109/TIM.2025.3577843>.
27. T. Gao, J. Yang, W. Wang, and X. Fan, "A Domain Feature Decoupling Network for Rotating Machinery Fault Diagnosis Under Unseen Operating Conditions," *Reliability Engineering & System Safety* 252 (2024): 110449, <https://doi.org/10.1016/j.res.2024.110449>.
28. D. Li, J. Yang, and J. He, "Ddnet: a Domain Augmentation and Dual Invariance Learning-Guided Small-Sample Domain Generalization Network for Bearing Fault Diagnosis," *Measurement Science and Technology* 37, no. 3 (2026): 036110, <https://doi.org/10.1088/1361-6501/ae31a4>.
29. Y. Li, J. Yang, W. Wang, and T. Gao, "A Joint Collaborative Adaptation Network for Fault Diagnosis of Rolling Bearing Under Class Imbalance and Variable Operating Conditions," *Advanced Engineering Informatics* 69 (2026): 103931, <https://doi.org/10.1016/j.aei.2025.103931>.
30. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Networks," *Communications of the ACM* 63, no. 11 (2020): 139-144, <https://doi.org/10.1145/3422622>.
31. H. Lv, J. Chen, T. Pan, T. Zhang, Y. Feng, and S. Liu, "Attention Mechanism in Intelligent Fault Diagnosis of Machinery: A Review of Technique and Application," *Measurement* 199 (2022): 111594, <https://doi.org/10.1016/j.measurement.2022.111594>.
32. G. Zhang, X. Kong, Q. Wang, et al., "Multi-Source Partial Domain Adaptation Method Based on Pseudo-Balanced Target Domain for Fault Diagnosis," *Knowledge-Based Systems* 284 (2024): 111255, <https://doi.org/10.1016/j.knosys.2023.111255>.
33. J. Liu, C. Zhang, and X. Jiang, "Imbalanced Fault Diagnosis of Rolling Bearing Using Improved MsR-GAN and Feature Enhancement-Driven CapsNet," *Mechanical Systems and Signal Processing* 168 (2022): 108664, <https://doi.org/10.1016/j.ymssp.2021.108664>.
34. X. Gu, Y. Yu, L. Guo, H. Gao, and M. Luo, "CSWGAN-GP: A New Method for Bearing Fault Diagnosis Under Imbalanced Condition," *Measurement* 217 (2023): 113014, <https://doi.org/10.1016/j.measurement.2023.113014>.
35. W. Wan, S. He, J. Chen, A. Li, and Y. Feng, "QSCGAN: An Un-Supervised Quick Self-Attention Convolutional GAN for LRE Bearing Fault Diagnosis Under Limited Label-Lacked Data," *IEEE Transactions on Instrumentation and Measurement* 70 (2021): 3527816-16, <https://doi.org/10.1109/TIM.2021.3125973>.
36. J. Fan, X. Yuan, Z. Miao, Z. Sun, X. Mei, and F. Zhou, "Full Attention Wasserstein GAN With Gradient Normalization for Fault Diagnosis Under Imbalanced Data," *IEEE Transactions on Instrumentation and Measurement* 71 (2022): 3517516-16, <https://doi.org/10.1109/TIM.2022.3190525>.
37. R. Wang, Z. Chen, S. Zhang, and W. Li, "Dual-Attention Generative Adversarial Networks for Fault Diagnosis Under the Class-Imbalanced Conditions," *IEEE Sensors Journal* 22, no. 2 (2022): 1474-1485, <https://doi.org/10.1109/JSEN.2021.3131166>.
38. J. Luo, J. Huang, J. Ma, and S. Liu, "Application of self-attention Conditional Deep Convolutional Generative Adversarial Networks in the Fault Diagnosis of Planetary Gearboxes," *Proceedings of the Institution of Mechanical Engineers-Part O: Journal of Risk and Reliability* 238, no. 2 (2024): 260-273, <https://doi.org/10.1177/1748006X221147784>.
39. Y. Cheng, P. Wang, H. Gu, J. Zeng, and J. Ma, "Fault Sample Generation Under Continuous Degradation Using CRGAN-MDTR: a Conditional Recurrent GAN with Maximum Degenerate Trend Retention," *Applied Soft Computing* 192 (2026): 114728, <https://doi.org/10.1016/j.asoc.2026.114728>.
40. C. Liu, L. Yuan, K. Lv, W. Xiao, G. Peng, and T. Ran, "ARC-GAN: an Auditing and Reconstructing Framework for zero-shot Bearing Fault Diagnosis," *Control Engineering Practice* 175 (2026): 107089, <https://doi.org/10.1016/j.conengprac.2026.107089>.
41. F. Huang, X. Zhang, G. Qin, et al., "Demagnetization Fault Diagnosis of Permanent Magnet Synchronous Motors Using Magnetic Leakage Signals," *IEEE Transactions on Industrial Informatics* 19, no. 4 (2023): 6105-6116, <https://doi.org/10.1109/TII.2022.3165283>.
42. K. Dragomiretskiy and D. Zosso, "Variational Mode Decomposition," *IEEE Transactions on Signal Processing* 62, no. 3 (2014): 531-544, <https://doi.org/10.1109/TSP.2013.2288675>.
43. X. Hu, H. Zhang, J. Cao, Y. Huang, H. Wang, and T. Tang, "PSRONet: A Deep Reinforcement Learning-Based Sensor Configuration Framework in Railway Point Machines Fault Diagnosis," *IEEE Transactions on Instrumentation and Measurement* 75 (2026): 1-13, <https://doi.org/10.1109/TIM.2026.3654732>.
44. R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep Learning and Its Applications to Machine Health Monitoring," *Mechanical Systems and Signal Processing* 115 (2019): 213-237, <https://doi.org/10.1016/j.ymssp.2018.05.050>.
45. X. Chen, B. Zhang, and D. Gao, "Bearing Fault Diagnosis Base on multi-scale CNN and LSTM Model," *Journal of Intelligent*

Manufacturing 32, no. 4 (2021): 971–987, <https://doi.org/10.1007/s10845-020-01600-2>.

46. A. Widodo and B. S. Yang, “Support Vector Machine in Machine Condition Monitoring and Fault Diagnosis,” *Mechanical Systems and Signal Processing* 21, no. 6 (2007): 2560–2574, <https://doi.org/10.1016/j.ymssp.2006.12.007>.

47. X. Yan, X. Hua, D. Jiang, and L. Xiang, “A Novel Robust Intelligent Fault Diagnosis Method for Rolling Bearings Based on SPAVMD and WOA-LSSVM Under Noisy Conditions,” *Measurement Science and Technology* 35, no. 5 (2024): 056121, <https://doi.org/10.1088/1361-6501/ad29e3>.

48. X. Wu, H. Peng, X. Cui, T. Guo, and Y. Zhang, “Multichannel Vibration Signal Fusion Based on Rolling Bearings and MRST-Transformer Fault Diagnosis Model,” *IEEE Sensors Journal* 24, no. 10 (2024): 16336–16346, <https://doi.org/10.1109/JSEN.2024.3380002>.