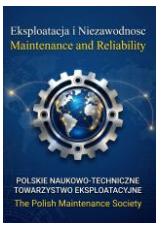


Eksploatacja i Niezawodnosc - Maintenance and Reliability

Volume 29 (2027), Issue 1

https://ein.org.pl

The Polish Maintenance Society

Physics-informed cross-domain fault diagnosis for rolling bearings under robustness and interpretability: a novel feature adaptation framework

Yi Xie[1], Ruyang Zheng[1], Xuepeng Guo[2], Shihan Tan [1], Dunxiang Zhu[1], Qiwei Hu[1,*]

[1] Shijiazhuang Campus, Army Engineering University of PLA, Hebei, P.R. China
[2] Command and Control Engineering College, Army Engineering University of PLA, Nanjing, China
* Corresponding Author: hu_q_w@163.com

Received: 14 February 2026
Revised: 16 July 2026
Accepted: 20 July 2026
Online: 27 July 2026

Abstract

Fault diagnosis of rolling bearings is crucial for operational safety. However, the scarcity of labeled data and significant domain shifts are two key challenges. Existing studies neglect robustness to physical disturbances and the interpretability of diagnostic decisions. To address these issues, this paper proposes a physics-informed cross-domain fault diagnosis framework. First, based on bearing fault modulation mechanisms, multi-domain features are extracted. A physics-regularized feature selection strategy combining Sparse Group Lasso with stability selection and Random Forest is used to retain fault-relevant features. Recognizing that velocity variations are a major source of domain shifts, order analysis is introduced as a physics-prior alignment method. Deep CORAL and pseudo-label self-training are combined for unsupervised knowledge transfer. Finally, multi-level interpretability analysis is embedded to quantify domain shift, track adaptation geometry and explain final decisions. Two case studies demonstrate its robustness, trustworthiness and engineering applicability.

Keywords: *fault diagnosis, robustness, unsupervised domain adaptation, interpretability, rolling bearings*

Article citation:
Xie Y, Zheng R, Guo X, Tan S, Zhu D, Hu Q, Physics-informed cross-domain fault diagnosis for rolling bearings under robustness and interpretability: a novel feature adaptation framework, Eksploatacja i Niezawodnosc – Maintenance and Reliability 2027: 29(1) http://doi.org/10.17531/ein/226160

Highlights

- A physics-informed and interpretable cross-domain fault diagnosis framework is developed.
- Robust, physically meaningful features are selected via SGL with stability selection.
- A robust diagnostic strategy is embedded by transfer learning.
- An end-to-end interpretable diagnostic pipeline is integrated into the framework.
- A Deep CORAL classifier coupled with multi-level post-hoc interpretability analysis.

1. Introduction

Rolling bearings are ubiquitous yet failure-prone components in rotating machinery, widely used in critical sectors such as aerospace, power systems, high-end manufacturing, and high-speed railway [1]. Under the demanding operating conditions of high-speed trains, rolling bearings are subjected to complex dynamic loads, strong vibrations, and high rotational speeds. Rolling bearings operate under these harsh and uncertain service conditions, which further increase system vulnerability and the likelihood of bearing-related failures. Unexpected bearing faults not only necessitate unplanned maintenance but can also propagate to adjacent components, degrading the reliability of the entire trainset. Hence, reliable and accurate fault diagnosis is crucial to ensuring railway safety and operational availability.

Conventional fault diagnosis methodologies typically follow three main steps: data acquisition, feature extraction, and fault classification [2]. Among various data acquisition techniques, vibration-based monitoring is prevalent due to its sensitivity to bearing defects [3]. Once signals are acquired, signal processing techniques like Fourier and wavelet transforms are employed to extract physics-based features from the time, frequency, or time-frequency domains [4,5]. These manually engineered features are then fed into shallow machine

learning models, such as the support vector machine (SVM), kernel extreme-learning machine (KELM), and artificial neural network (ANN) [6–8]. However, these methods heavily rely on expert knowledge for feature engineering and their diagnostic performance deteriorates significantly when confronted with large-scale, noisy data or when domain shifts occur between training and testing conditions [9,10].

In recent years, Deep Learning (DL) has emerged as a powerful tool for bearing fault diagnosis, owing to its exceptional ability to automatically learn hierarchical and complex features from massive datasets. DL models, particularly Convolutional Neural Networks (CNNs), can process raw vibration signals directly, bypassing the need for laborious manual feature engineering and achieving remarkable success in diagnostic tasks [11,12]. Lu and Xiao [13] proposed an improved DenseNet method for automobile bearing fault diagnosis, using recurrence plots to convert 1-D vibration signals into 2-D images, MobileViT attention and lightweight design to improve efficiency with reduced diagnosis time. Wen et al. [14] proposed a CNN-based fault diagnosis method for smart manufacturing, leveraging LeNet-5 to automatically extract features from raw data by converting signals into 2-D images. He and He [15] proposed a deep learning-based approach for bearing fault diagnosis, utilizing short-time Fourier transform (STFT) to preprocess sensor signals and an optimized LAMSTAR neural network for fault identification. Despite these significant advancements, there are still several obstacles in the deployment of DL from experiments to practical applications.

The first challenge is robustness, particularly the model's ability to adapt to varying operating conditions and noisy environments, which is referred to as robustness. The discrepancies of external factors between training and application, such as rotational speed, load, noise, and sensor placement, may cause a sharp decline in model performance. These discrepancies are known as domain shift [16,17]. To address this, Transfer Learning (TL), especially Unsupervised Domain Adaptation (UDA), has been widely adopted to transfer knowledge from a data-rich source domain to a label-scarce target domain [18,19]. Among them, Chen et al. [20] proposed a novel Joint Sliced Wasserstein Distance (JSWD) approach for unsupervised domain adaptation in mechanical fault diagnosis, addressing the issue of class conditional distribution. Zhong et al. [21] proposed an unsupervised deep convolutional dynamic joint distribution domain adaptive network for bearing fault diagnosis under variable conditions. Zhong et al. [22] proposed a deep feature construction and unsupervised domain adaptation strategy for rolling bearing fault diagnosis under varying working conditions. Zhang et al. [23] proposed an enhanced transfer joint matching (TJM) approach for cross-domain bearing fault diagnosis, addressing the issue of unbalanced sample distributions. Some other existing UDA methods often align feature distributions using statistical distance minimization (e.g., Maximum Mean Discrepancy (MMD), Correlation Alignment (CORAL)) or adversarial strategies. For example, Guo et al. [24] proposed a deep convolutional transfer learning network for intelligent fault diagnosis with unlabeled data, where the domain adaptation module learns domain-invariant features by maximizing domain recognition errors and minimizing the probability distribution distance. But many focus on global distribution alignment and overlook the preservation of task-specific decision boundaries, limiting their efficacy in fine-grained fault discrimination under complex conditions.

The second is the inherent lack of interpretability of DL models. As black box models, their opaque decision-making processes can undermine the credibility of their decisions when applied in cross-domain scenarios. Many methods have been proposed to elucidate model decisions and have made valuable progress, such as attention-based visualization of signal regions of interest, gradient-based attribution or activation-mapping for highlighting fault-relevant features, and physics-informed architectural embedding for enhanced intrinsic interpretability [25,26]. Zhou et al. [27] developed an interpretable model using residual and attention modules to extract deep features from system sequences, ensuring its explainability by visualizing the diagnostic process with attention weights and Class Activation Maps (CAM). Li et al. [28] proposed Multi-Layer Grad-CAM (MLG-CAM), a method that resolves feature resolution degradation and quantifies model interpretability in vibration-based fault diagnosis by fusing multi-resolution activation maps to highlight key features like cyclostationary impulses and fault frequencies. Chen et al. [29] developed an interpretable Variational Kernel Convolutional Neural Network (VKCNN)

for rolling bearing fault diagnosis. The model's architecture integrates variational kernels and approximate filters to extract physically meaningful Amplitude-Modulated and Frequency-Modulated (AM-FM) components from raw signals. Furthermore, it employs a channel attention mechanism with residual connections to dynamically assign weights to fault-related features across different frequency bands. This dual approach not only enhances the diagnostic performance of the model but also significantly improves the transparency of its decision-making process.

In summary, substantial research has contributed to enhancing either the robustness or the interpretability of diagnostic models. The former aims to maintain high accuracy in variable real-world environments, while the latter focuses on making model decisions transparent and trustworthy. However, the vast majority of current studies address these two critical aspects in isolation. Very few works have considered interpretability as a subsequent goal after achieving robustness against varying conditions and noise. In other words, these methods have largely overlooked the combined need for both robustness and interpretability, which is essential for deploying reliable and trustworthy fault diagnosis systems in safety-critical applications.

This paper proposes a novel Physics-informed Cross-Domain Fault Diagnosis Framework for rolling bearings, aiming to achieve effective fault diagnosis while ensuring model robustness and interpretability. The framework establishes a formal fault diagnosis pipeline that integrates robust feature extraction engineering, Random Forest (RF) classification, and a hybrid UDA strategy. By incorporating cross-domain feature collection and analysis of rolling bearings into the model training process, the proposed method ensures domain adaptability. Moreover, combined with the extensive interpretability analysis based on SHapley Additive exPlanations (SHAP), the interactive decision-making operations of the fault diagnosis process can be intuitively illustrated. The principal contributions of this work are as follows:

(1) A physics-informed and interpretable cross-domain fault diagnosis framework is developed for rolling bearings. By explicitly integrating physical knowledge, robust feature engineering, and unsupervised domain adaptation, the proposed framework enables reliable diagnostic knowledge transfer from data-rich laboratory settings to unlabeled real-world operating conditions.

(2) A robust diagnostic strategy is embedded into the proposed framework by jointly modeling bearing feature characteristics and domain shift. Specifically, a Sparse-Group Lasso (SGL) with stability selection and RF based selection strategy is first proposed to identify stable and physically meaningful features. Building on this, a hybrid domain adaptation scheme is developed to explicitly account for the effects of speed variation and environmental discrepancy, thereby strengthening cross-domain robustness. Furthermore, an integrated adaptation-and-decision procedure combining Deep CORAL with pseudo-label self-training is constructed, providing a unified interface that facilitates subsequent interpretability analysis.

(3) An end-to-end interpretable diagnostic pipeline is integrated into the framework through multi-level interpretability analysis. This includes target-domain permutation importance and SHAP-based global and category-wise attribution analysis.

The remainder of this paper is structured as follows: Section 2 reviews related work; Section 3 details the proposed methodology; Section 4 presents experimental results; and Section 5 concludes with limitations and future directions.

2. Related work

As critical components in many engineering applications, the fault diagnosis of rolling bearings has always been a research hotspot. In recent decades, the DL-based fault diagnosis has attracted widespread attention. While DL has significantly enhanced diagnostic capabilities, its practical deployment hinges on addressing key challenges related to data distribution discrepancies, operational robustness, and model transparency. Therefore, this section focuses on three critical research thrusts: cross-domain fault diagnosis, robustness enhancement strategies, and the integration of interpretability.

2.1. Cross-domain fault diagnosis of rolling bearings

A fundamental challenge in real-world rolling-bearing fault diagnosis is the domain shift phenomenon, where the statistical distribution of data collected during operation (target domain)

differs from the training data (source domain) due to variations in working conditions like speed, load, or environmental factors [30,31]. This discrepancy often leads to a severe degradation in the performance of models trained solely on source domain data. To address this, TL, and specifically UDA, has emerged as a dominant paradigm [32].

UDA techniques aim to leverage knowledge from a labeled source domain to perform diagnostics on an unlabeled target domain. These methods can be broadly classified into two main categories. The first category focuses on distribution alignment, which seeks to minimize a statistical distance metric between the feature representations of the source and target domains. Prominent methods in this class include aligning the mean and covariance of distributions using CORAL or minimizing the MMD [33]. The second category employs adversarial learning, where a domain discriminator network is trained to distinguish between source and target features, while the feature extractor is trained concurrently to produce domain-agnostic features that can fool this discriminator [34].

Numerous studies have demonstrated the efficacy of UDA. For instance, Wang et al. [35] proposed a dynamic collaborative adversarial domain adaptation network (DCADAN) for unsupervised fault diagnosis in rotating machinery, addressing the challenge of acquiring sufficient labeled fault data for new tasks. Xie et al. [36] proposed the target domain feature enhancement and feature-boundary alignment network (TDFE-FBAN) to improve unsupervised fault diagnosis in smart manufacturing, addressing issues with source domain priors and misalignment in high-dimensional features. Han et al. [37] proposed a deep transfer network (DTN) with joint distribution adaptation (JDA) for cross-domain fault diagnosis, reducing marginal and conditional distribution discrepancies between labeled source data and unlabeled target data to learn domain-invariant fault representations under varying operating conditions, fault severities, and fault types. Despite these advances, a common limitation persists: many UDA methods prioritize the alignment of global marginal and conditional distributions. This focus on coarse alignment can inadvertently compromise the fine-grained, class-specific decision boundaries essential for accurate fault discrimination in complex diagnostic tasks [36].

2.2. Fault diagnosis under robustness

In fault diagnosis, robustness denotes the capacity of a diagnostic model to retain reliable performance (e.g., classification accuracy, detection sensitivity) when confronted with realistic perturbations typical of industrial settings — such as sensor noise, changes in operating conditions, distribution shifts, or signal distortions — rather than only under clean, ideal, or lab controlled conditions [38]. While the cross-domain adaptation techniques discussed in Section 2.1 are a primary means of achieving robustness against distribution shifts, a holistic view of robustness also encompasses the intrinsic stability of features and the model's resilience to specific physical variations.

One key aspect of enhancing robustness is robust feature engineering. Instead of relying solely on a model to learn invariant representations, this approach aims to select or construct features that are inherently less sensitive to operational fluctuations and noise. Methodologies such as SGL and Stability Selection have been explored to identify a compact set of features that are not only highly discriminative but also consistently effective across different conditions, thereby improving the model's generalization capacity from the feature level. Zhao et al. [39] proposed the adaptive enhanced sparse period-group lasso (AdaESPGL) algorithm for bearing fault diagnosis, addressing the challenge of extracting impulses buried in heavy background noise.

Another critical strategy is the integration of physics-informed knowledge to directly counteract specific sources of variability. For instance, fluctuating rotational speed is a major challenge in machinery like high-speed trains, as it introduces non-stationarity into vibration signals and distorts fault signatures in the frequency domain [40]. Physics-based techniques like order analysis, which resamples signals from the time domain to the angular domain, can effectively normalize the effect of speed variations. By explicitly addressing known physical phenomena, these methods provide a more targeted and reliable path to robustness compared to purely data-driven adaptation approaches that treat all domain shifts as generic statistical discrepancies. A truly robust diagnostic system should therefore ideally combine the strengths of data-driven domain adaptation with the precision of physics-informed feature normalization and robust selection.

2.3. Fault diagnosis under interpretability

Beyond accuracy and robustness, the adoption of intelligent diagnostic systems in safety-critical sectors is critically dependent on their interpretability. The black-box nature of many deep learning models, where the decision-making logic is opaque to human users, creates a significant barrier to trust, verification, and deployment [41,42]. To address this, Explainable Artificial Intelligence (XAI) has become an indispensable research area in fault diagnosis.

XAI methodologies can be categorized into two main paradigms. The first is intrinsic interpretability (or interpretable-by-design models), where transparency is built into the model's architecture. This is often achieved by embedding physical principles or constraints into the model structure. Raissi et al. [43] proposed the Physics-Informed Neural Networks (PINNs) framework, which incorporates governing physical laws, initial conditions, and boundary conditions into neural network training through physics-informed loss functions, allowing the model predictions to remain consistent with known physical mechanisms. Melis et al. [44] designed the Self-Explaining Neural Network (SENN) framework, which incorporates structural constraints (regularization) and explicit explanation mechanisms (e.g., feature effects, additive explanations) during training to ensure a balance between model interpretability and performance. By ensuring that the learned features correspond to understandable physical phenomena, such models provide a more direct and inherent form of interpretability.

The second paradigm is post-hoc interpretability, which involves applying external techniques to explain the predictions of a pre-trained model. A popular approach is visualization through saliency maps, such as CAM and its derivatives. These methods highlight which parts of the input data were most influential in a model's decision. For example, Yang et al. [45] proposed an explainable intelligence fault diagnosis framework using CNNs to identify fault types in rotating machinery. The framework utilizes data from short-time Fourier transformation and employs a post hoc explanation method to visualize the features learned by the model. Brito et al. [46] proposed a new approach for fault detection and diagnosis in rotating machinery, addressing the challenges of unavailability of labeled historical data and the need for model explainability. The methodology consists of three parts: feature extraction (using time and frequency domain vibration features), fault detection (through unsupervised anomaly detection algorithms), and fault diagnosis (using SHAP for model interpretability).

Table 1. Comparison of recent methods on physics-informed robustness and interpretability.

Reference	Method	Year	Physics-informed Robustness	Cross-domain Interpretability	Observed Limitation
Wang et al. [37]	WJMMD-MDA	2023	✗	✗	Only multi-source alignment; Coarse global distribution alignment
Zhong et al. [22]	Deep Feature Construction + UDA	2023	✗	✗	Ignores physics-informed features and interpretability
Wang et al. [35]	DCADAN	2025	✗	✗	No exploitation of physical fault characteristics; Decisions in target domain are opaque
Xie et al. [36]	TDFE-FBAN	2025	✗	✗	Lacks explicit modeling of physical frequencies and harmonics
Zhong et al. [21]	DCDJAN	2024	Partial	✗	Lacks physically interpretable feature selection
Shao et al. [47]	LSFConvformer	2025	Partial	✗	Lacks explicit physics-informed feature modeling and cross-domain interpretability
Ma et al. [48]	TFD-Former	2025	✓	✗	Lacks explicit physics-informed feature modeling and cross-domain interpretability
Li et al. [28]	MLG-CAM	2023	Partial	Partial	Lacks explicit physical mechanism modeling across domains
Chen et al. [29]	VKCNN	2024	Partial	Partial	Cross-domain interpretability not fully addressed
Liu et al. [49]	AFAI-Transformer	2025	Partial	✓	Lacks explicit physics-informed feature modeling for cross-domain robustness
Zhang et al. [50]	WD-KANTF	2025	Partial	✓	No explicitly integrate physics-informed features for cross-domain adaptation.

To better highlight the limitations of existing approaches and their differences with respect to mechanism-informed robustness and interpretability, we provide a qualitative comparison of fault diagnosis methods published in recent years. Table 1 summarizes whether the methods adopted in recent literature explicitly incorporate physical information feature modeling to enhance robustness, and whether they provide interpretable decision outputs. Through this comparison, it

becomes clear that most of the latest methods either lack explicit physical guidance, do not provide interpretability across domains, or only partially mention both without connecting them. This has sparked a demand for the proposed physics-based interpretable framework. The information included in the table includes the author and citation, the method used, the year of publication, whether Physics-informed Robustness and Cross-domain Interpretability are adopted, and the limitations.

2.4. Research gaps

A systematic review of the literature above reveals that while significant strides have been made in the areas of cross-domain adaptation, robustness enhancement, and model interpretability, most of these research efforts have proceeded in isolation. The vast majority of studies focus on achieving high cross-domain accuracy or providing post-hoc explanations, but rarely both. However, the lack of robustness and interpretability in existing models substantially restricts their deployment in safety-critical components such as rolling bearings. To address this limitation, this paper proposes a unified framework that synergistically integrates robustness and interpretability for cross-domain fault diagnosis in rolling bearings.

3. Proposed methodology

In this section, the proposed cross-domain fault diagnosis framework is presented in detail. The schematic diagram of the proposed framework is shown in Figure 1, which is organized into three main components:

(1) Hybrid feature selection: SGL with stability selection is applied to generate and filter handcrafted features from the time, frequency, envelope, and order domains, yielding a compact subset of physically meaningful descriptors that are robust to noise and discriminative for bearing faults.

(2) Source-domain feature classification: a RF classifier is trained on the selected feature subset using only labeled source-domain data. The RF serves as an evaluation benchmark to verify the discriminative quality of the SGL-selected features.

(3) Unsupervised domain adaptation: the same selected feature subset is then fed into a Deep CORAL network with pseudo-label self-training, which aligns the source and target distributions in a latent space and refines the decision boundary for the target domain. Following the adaptation, a three-tier interpretability analysis is conducted, including ex-ante, adaptation process and post-hoc.

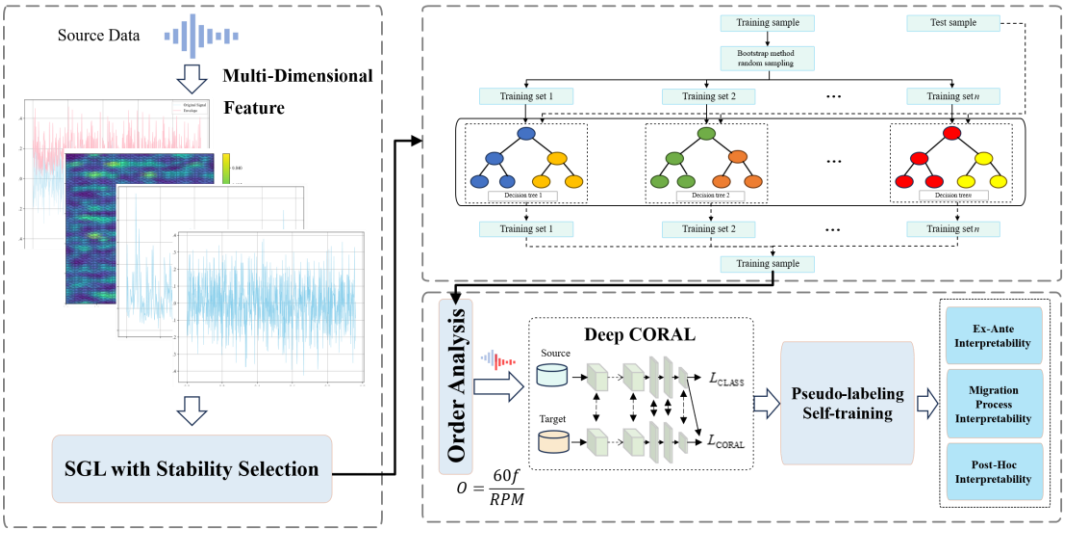


Figure 1. Structure of the Proposed Cross-domain Fault Diagnosis Framework.

3.1. Hybrid feature selection

3.1.1. Multi-dimensional feature extraction

The vibration characteristics of rolling bearings typically present three key properties: feature information is highly scattered, with a given fault type manifesting prominently only in a few frequency bands or time-domain indicators while remaining nearly indistinguishable in others; fault-related signatures tend to appear in clusters, as the characteristic frequency components, together with their harmonics and sidebands, vary synchronously; and the measured data are inevitably contaminated by substantial noise due to complex operating conditions, such that directly using all raw features can introduce considerable errors. To obtain a robust source-

domain representation, the vibration signals are first preprocessed and then described using multi-dimensional features guided by bearing failure mechanisms. In particular, time-domain statistical descriptors are extracted to quantify impact intensity, whereas frequency- and envelope-domain features are computed to emphasize periodic impulsive responses associated with localized defects.

1. Time domain feature statistics

Time-domain statistical features extract information directly from the time waveform of the vibration signal through a series of mathematical indicators, and their physical meanings are closely related to the actual operating status of the equipment. These statistics can intuitively reflect the overall energy level of vibration, the sharpness of signal distribution and the intensity of instantaneous impact. They are the most commonly used and direct basis for assessing the overall health status of rotating machinery and detecting abnormal impacts and early failures.

(1) Define $RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$ as RMS, which is equivalent to energy intensity. Commonly used RMS for vibration level assessment of rotating machinery

(2) Define $K_s = \frac{1}{N}\sum_{i=1}^{N} \frac{(x_i-\mu)^4}{\sigma^4}$ as kurtosis, which mainly reflects the tail thickness and peak height of the distribution. $\mu$is the mean of the signal.

(3) Define $P = \max_{1\le i\le N}|x_i|$ as the peak value (Peak), which refers to the instantaneous maximum amplitude.

2. Frequency domain statistics

Frequency-domain statistical features convert vibration signals from the time domain to the frequency domain for analysis, and their practical significance lies in revealing the frequency distribution structure of vibration energy. By analyzing the shape of the spectrum, components related to the fault characteristic frequency of bearings can be identified, different types of faults can be distinguished, and the main sources of system vibration and their stability can be evaluated, which is crucial for fault location and mechanism analysis.

(1) Define $H = -\sum_k p_k \ln p_k$ as spectral entropy, which is used to quantify the disorder of spectral energy. The P_k is the approximate power spectral density, defined as $P_k = \frac{\left|\sum_{i=1}^{N} x_i e^{-j2\pi(k-1)(i-1)/N}\right|^2}{N}$. The normalized spectral distribution is defined as $p_k = \frac{P_k}{\sum_{k=1}^{K} P_k}$. And $\sum_{k=1}^{K} p_k = 1$.

(2) Define $B = \sqrt{\sum_{k=1}^{K}(f_k - C)^2 p_k}$ as the spectral bandwidth, which is used to quantify the dispersion of energy relative to the spectrum center of mass. The C is the spectral centroid, which is the center of gravity position of the spectrum energy distribution, defined as $C = \sum_{k=1}^{K} f_k p_k$.

Before constructing fault-characteristic-frequency-related features, the geometry dependence of bearing characteristic frequencies is explicitly considered. For a rolling bearing with rolling-element number n_r, rolling-element diameter d_r, pitch diameter D_r, contact angle θ_r, and shaft rotational frequency f_r, the characteristic frequencies are calculated as:

$$\mathrm{BPFO} = \frac{n_r f_r}{2}\left(1 - \frac{d_r \cos\theta_r}{D_r}\right) \quad (1)$$

$$\mathrm{BPFI} = \frac{n_r f_r}{2}\left(1 + \frac{d_r \cos\theta_r}{D_r}\right) \quad (2)$$

$$\mathrm{BSF} = \frac{D_r f_r}{2 d_r}\left[1 - \left(\frac{d_r \cos\theta_r}{D_r}\right)^2\right] \quad (3)$$

$$FTF = \frac{f_r}{2}\left(1 - \frac{d_r \cos\theta_r}{D_r}\right) \quad (4)$$

These equations show that BPFO, BPFI, BSF, and FTF are determined jointly by bearing geometry and rotational speed. Therefore, fault-characteristic-frequency-related features should be constructed according to the available bearing parameters rather than treated as universal frequency templates.

3. Envelope spectrum alignment index

The envelope spectrum alignment index specifically targets the periodic impact modulation phenomenon that arises when components such as bearings sustain local damage. Its physical significance is to highlight the fault characteristic frequency and its harmonics that are overwhelmed by high-frequency resonance or noise by demodulating the vibration signal. These indicators can quantify the periodic intensity, harmonic structure and modulation sideband information of fault impact, and are the most effective and direct feature set for diagnosing local bearing damage.

(1) Define $\mathrm{Peak}(f_0) = max_{k\in I(f_0)}|E(f_k)|$ as the in-band peak (Peak), which refers to the maximum amplitude near the frequency. Define the in-band index as $I(f_0) = \{k: |f_k - f_0| \le \delta\}$, where δ denotes the half-bandwidth parameter that specifies the frequency tolerance around the target frequency f_0.

(2) Define $\mathrm{BandE}(f_0) = \sum_{k\in I(f_0)} E(f_k)^2 \Delta f$ as the in-band energy (BandE), which refers to the sum of the energy of all frequency components within a specific frequency bandwidth.

(3) Define $\mathrm{Eratio}(f_0) = \frac{\mathrm{BandE}(f_0)}{E_{\mathrm{total}}}$ as the energy ratio (Eratio), which refers to the normalization of local energy. The E_{total} is

the total energy, defined as $E_{\text{total}} = \sum_k |E(f_k)|^2 \Delta f$.

(4) Define $\text{HarmE}_M(f_0) = \sum_{m=1}^{M} BandE(mf_0)$ as the harmonic energy (HarmE), which refers to the sum of the energy of multiple vibration frequencies.

(5) Define $\text{HarmRatio}_M(f_0) = \frac{\text{HarmE}_M(f_0)}{E_{\text{total}}}$ as the harmonic ratio (HarmRatio), which refers to the ratio of frequency doubling energy to the total energy.

(6) Define $\text{SB}_Q(f_0) = \sum_{m=1}^{M} \sum_{q=1}^{Q} [\text{BandE}(mf_0 - qf_r) + \text{BandE}(mf_0 + qf_r)]$ as the sideband energy (SB), which is to count the energy sum of the first-order modulation sidebands on both sides of each octave frequency with the conversion frequency as the step distance.

3.1.2. SGL with stability selection

To extract features that genuinely aid fault diagnosis and avoid the risk of overfitting, an extraction method combining SGL with Stability Selection is adopted. This method can not only automatically select a few key variables from a large number of time domain and frequency domain indicators, but also can identify these group related features. At the same time, it can also remain robust under small samples and strong noise.

Sparse-group sparse feature selection combines two classic regularization techniques. Sparse uses Lasso (L1 regularization) to screen out a few truly important indicators from numerous time domain and frequency domain features, and directly compress the coefficients of other irrelevant features to 0. Group Sparse uses Group Lasso to select or discard features belonging to the same fault frequency family as a whole, and then compresses all coefficients of the entire irrelevant group to 0. Combining the two achieves sparsity at both the group level and the individual feature level. Stability selection calculates the probability of each feature being selected through repeated subsampling and parameter perturbation to ensure that the features selected in a small sample and noisy environment are more reliable and reproducible.

In the case of multi-class classification, it is only necessary to replace the binary logistic regression of the SGL model with the multi-class soft-max logistic regression. The idea of sparse-group sparse and stability selection remains unchanged. Define discrete vibration signals as x_i. Suppose there are K categories, denoted by y_i ($y_i \in \{1, \dots, K\}$). Define the sample as (x_i, y_i). Establish a weight vector and intercept for each category k ($k \in \{1, \dots, K\}$), and define the linear predictor for sample i and category k as:

$$\eta_{i,k} = x_i^T w_k + b_k \tag{5}$$

where w_k is the weight vector and b_k is intercept.

The soft-max probability is:

$$\pi_{i,k} = \frac{\exp(\eta_{i,k})}{\sum_{l=1}^{K} \exp(\eta_{i,l})} \tag{6}$$

Define the groups of features as $\{G_1, G_2, \dots, G_G\}$. The SGL objective function for multi-category classification is formulated as:

$$\min_{\{w_k, b_k\}_{k=1}^{K}} \frac{1}{n} \sum_{i=1}^{n} [-\log \pi_{i,y_i}] + \lambda_1 \sum_{k=1}^{K} \|w_k\|_1 + \lambda_2 \sum_{g=1}^{G} \sqrt{|G_g|} \left\| \left(w_{1,G_g}, w_{2,G_g}, \dots, w_{K,G_g} \right) \right\|_2 \tag{7}$$

where λ_1 controls the sparsity of a single feature in each category, λ_2 controls the sparsity of the group, $\|\cdot\|_1$ and $\|\cdot\|_2$ are L1 and L2 norm. The grouped L2 penalty is computed as:

$$\left\| \left(w_{1,G_g}, \dots, w_{K,G_g} \right) \right\|_2 = \sqrt{\sum_{k=1}^{K} \left\| w_{k,G_g} \right\|_2^2} \tag{8}$$

At each iteration, we first perform a gradient descent step on the smooth cross-entropy loss to obtain an intermediate matrix U. Subsequently, the element-wise L1 proximal operator (soft-thresholding) is applied to each group sub-matrix:

$$\widetilde{\boldsymbol{W}}_{G_g} = S\left(U_{G_g}, \alpha \lambda_1 \right) \tag{9}$$

where $S(U_{G_g}, \alpha\lambda_1) = \text{sign}(U_{G_g}) max(\left| U_{G_g} \right| - \alpha\lambda_1, 0)$, α denotes the proximal-gradient step size.

Then the group-wise proximal operator is applied to scale the entire group:

$$\boldsymbol{W}_{G_g} \leftarrow max\left(0,1 - \frac{\alpha \lambda_2 \sqrt{|G_g|}}{\left\| \widetilde{\boldsymbol{W}}_{G_g} \right\|_F} \right) \widetilde{\boldsymbol{W}}_{G_g} \tag{10}$$

To enhance robustness under small sample sizes and high noise conditions, stability selection is integrated with SGL. For each subsample ratio $s \in (0,1)$, fit the model repeatedly on B subsampling iterations and L perturbation grids. In the multiclass setting, feature selection is defined at the individual-feature level by aggregating the coefficients of the same feature across all K categories. The j-th feature is regarded as selected in this run if its cross-class coefficient vector has a nonzero L2 norm. Accordingly, the stability selection probability is defined as:

$$\hat{\pi}_j = \frac{1}{BL}\sum_{b=1}^{B}\sum_{l=1}^{L}\mathbf{1}\left\{\left\|w_j^{(b,l)}\right\|_2 > 0\right\} \quad (11)$$

where $b \in \{1,2,\cdots,B\}$ is the number of subsampling iterations, and $l \in \{1,2,\cdots,L\}$ is the regular grid of each perturbation. $w_j^{(b,l)}$ the coefficient vector of the j-th feature across all K categories under the b-th subsample and l-th perturbation.

According to the model of Meinshausen and Bühlmann [51], when $\hat{\pi}_j \geq \tau_{SS}(0.5 < \tau_{SS} < 1)$, it was determined that this feature was stable and effective. Stability selection gives an approximate upper bound on the expected number of mis-selections:

$$E[V] \lesssim \frac{q^2}{(2\tau_{SS}-1)p} \quad (12)$$

where q is the average number of selected features for each sub-model, and p is t the total number of candidate variables. This bound guarantees that the final selected features maintain reliable diagnostic performance despite the presence of strong noise.

3.2. Source domain feature classification

SGL with stability selection yields a compact feature subset that is both physically interpretable and robust to noise. However, the sparsity-induced selection process itself does not directly indicate how well the retained features support fault discrimination. It is therefore necessary to evaluate the discriminative quality of the selected features before feeding them into the downstream domain-adaptation pipeline. To this end, an RF classifier is trained on the labeled source-domain data using the selected feature subset as input. RF is adopted for three reasons: (1) it naturally handles heterogeneous features, including time-domain statistics, spectral descriptors, envelope-domain indicators, and order-band measures, without requiring feature scaling or normality assumptions; (2) its ensemble structure improves robustness to noise and makes it less prone to overfitting than a single decision tree, especially when the number of features is moderate; and (3) it provides built-in out-of-bag (OOB) estimation, which offers an internal estimate of classification performance without requiring an additional validation set. The architecture of the trained RF model is illustrated in Figure 2.

It is important to emphasize that the RF classifier serves only as a source-domain evaluation benchmark and does not directly participate in the subsequent unsupervised domain adaptation stage. Once the discriminative quality of the selected features is confirmed on the source domain, the same feature subset, is passed to the Deep CORAL module. The Deep CORAL network independently learns a latent embedding from this shared feature input through a multilayer perceptron and is optimized jointly for source-domain discrimination, cross-domain covariance alignment and pseudo-label-based target adaptation. Therefore, the RF benchmark provides a transparent baseline against which the transferability gain achieved by Deep CORAL can be fairly assessed, without introducing parameter coupling or information leakage between the two models.

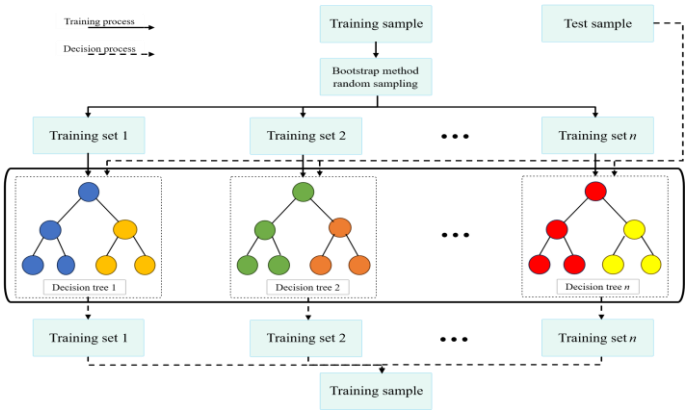


Figure 2. Training diagram of the RF algorithm.

3.3. Unsupervised domain adaptation

Since the conditions for data collection in different domains are usually different, cross-domain feature drift results. This article mainly discusses the situation where the rotation speed cannot be coupled. In order to ensure physical conservation, when performing feature processing, this article proposes an order-spectrum normalization framework to match the speed difference between the two domains to achieve normalized characterization of fault characteristics across speed conditions.

3.3.1. Order analysis

Since the source domain data and the target domain data are collected under different operating conditions, especially their rotation speeds are not identical. Consequently, the same bearing fault characteristic may appear at different absolute frequencies in the frequency spectrum or envelope spectrum. To reduce this cross-domain frequency-location mismatch, this study adopts order-domain normalization under the constant-

speed approximation.

$$f_r = \frac{\mathrm{RPM}}{60} \tag{13}$$

where RPM is the nominal rotational speed of a sample, which measured in revolutions per minute. and f_r is measured in Hz.

For a spectral frequency component f_s, its corresponding order is defined as:

$$O = \frac{f_s}{f_r} = \frac{60 f_s}{\mathrm{RPM}} \tag{14}$$

where O is dimensionless and represents the frequency component normalized by the shaft rotational frequency. Therefore, the frequency spectrum or envelope spectrum can be transformed into an order-domain representation by replacing the horizontal axis f_s with O. Under nominally constant-speed conditions, this operation is equivalent to a linear rescaling of the frequency axis.

Specifically, the whole vibration signal is first preprocessed by detrending and band-pass filtering. The envelope signal is then obtained using the Hilbert transform, and the envelope spectrum is calculated by FFT. After that, the frequency axis is normalized by the shaft rotational frequency to obtain the order spectrum $(O, |X(O)|)$. All spectrum-related features, including band energy, peak amplitude, spectral centroid, spectral entropy, spectral kurtosis, dominant order, harmonic energy, and sideband symmetry, are extracted directly in the order domain. The order-domain representation reduces the influence of rotational-speed variation by expressing spectral components as multiples of the shaft rotational frequency, thereby improving the comparability of fault-related modulation components across operating conditions. However, order normalization does not eliminate the influence of bearing geometry, because BPFO, BPFI, BSF, and FTF depend on the rolling-element number, rolling-element diameter, pitch diameter, and contact angle. Therefore, characteristic fault orders should not be assumed to be identical across different bearings.

In the proposed framework, when bearing geometry is available, characteristic fault orders are calculated from the corresponding bearing parameters. When complete bearing geometry is unavailable, exact characteristic orders from another bearing are not imposed directly. Instead, broad fault-sensitive order bands are constructed around physically meaningful fault-order regions, and feature operators such as peak amplitude, band energy, energy ratio, harmonic energy, harmonic ratio, and sideband energy are extracted within these bands. In this way, the extracted order-domain features retain their connection with bearing fault modulation mechanisms while avoiding the assumption that different bearings share identical BPFO, BPFI, BSF, or FTF values.

3.3.2. Domain adaptation without supervision

After the feature spaces of the source domain and the target domain are aligned, there is still a certain distribution shift problem. Define the source domain dataset as D_S, and the expression is:

$$D_S = \{(x_i^S, y_i^S)\}_{i=1}^{n_S} \tag{15}$$

where $x_i^S \in R^d$ is the extracted dimensional signal feature, $y_i^S \in \{1, \dots, K\}$ is the corresponding label category, and n_S is the number of samples in the source domain. Define $D_T = \{x_j^T\}_{j=1}^{n_T}$ is the target domain dataset. Among them, $x_j^T \in R^d$ is all unlabeled samples, and n_T is the number of samples in the target domain. The feature distributions of the source domain and the target domain are recorded as $P_S(x)$ and $P_T(x)$, respectively, and there is a significant distribution shift between them, this means $P_S(x) \neq P_T(x)$.

The UDA aims to learn a feature mapping $G(\cdot)$ such that the feature distributions become close:

$$P_S[G(x)] \approx P_T[G(x)] \tag{16}$$

Learn $G(\cdot)$ and classifier $f_c: R^d \to \{1, \dots, K\}$ through source domain supervised loss and cross-domain distribution alignment loss to improve target domain prediction performance.

3.3.3. Deep CORAL

Deep CORAL is adopted in this study as a domain adaptation strategy to mitigate the distribution discrepancy between the source domain and the target domain and thus improve diagnostic generalization under cross-domain operating conditions. In the proposed framework, the source-domain and target-domain samples are first represented by the same handcrafted feature vector extracted from the vibration signals. These feature vectors are then fed into a shared feature extractor implemented as a multilayer perceptron. Specifically, the Deep CORAL backbone consists of three fully connected hidden layers, each with multiple neurons. Each hidden layer is

followed by a ReLU activation and a Dropout operation. The output of the last hidden layer is taken as the latent representation for domain alignment, and a final linear classifier is used to predict the fault category. The core idea is to align domain distributions by minimizing the difference in second-order statistics of deep features of the source and target domains. Let $F_S \in \mathbb{R}^{n_S \times d}$ and $F_T \in \mathbb{R}^{n_T \times d}$ denote the mini-batch latent feature matrices of the source and target domains. The empirical covariance matrices are computed as:

$$C_S = \frac{1}{n_S - 1}\left(F_S^T F_S - \frac{1}{n_S}\left(\mathbf{1}_S{}^T F_S\right)^T \left(\mathbf{1}_S{}^T F_S\right)\right) \tag{17}$$

$$C_T = \frac{1}{n_T - 1}\left(f_T^T f_T - \frac{1}{n_T}\left(1_T{}^T f_T\right)^T \left(1_T{}^T f_T\right)\right) \tag{18}$$

where n_S and n_T are the source and target batch sizes and d is the latent feature dimension, $\mathbf{1}_S \in \mathbb{R}^{n_S \times 1}$ and $1_T \in \mathbb{R}^{n_T \times 1}$column vectors with all elements equal to 1.

Then, defining the CORAL loss as L_{CORAL}, and calculating the difference between the two covariance matrices.

$$L_{\text{CORAL}} = \frac{1}{4d^2}\|C_S - C_T\|_F^2 \tag{19}$$

where d is the dimension of the feature, $\frac{1}{4d^2}$ and is the normalization constant.

Finally, the total loss function is established:

$$L_{\text{TOTAL}} = L_{\text{CLASS}} + \lambda L_{\text{CORAL}} \tag{20}$$

where L_{CLASS} denoted as classification loss, λ is the factor used to weigh CORAL loss.

3.3.4. Pseudo-labeling self-training

A high-confidence pseudo-labeling strategy is adopted to further utilize unlabeled data in the target domain. Define the prediction distribution of the initial classifier in the target domain as:

$$p_\theta(y|x^T) = g_\theta(x^T) \tag{21}$$

Select the sample set S_{PL} that satisfies:

$$S_{\text{PL}} = \left\{x^T \in D_T \,\middle|\, \max_k p_\theta(k|x^T) \geq \tau\right\} \tag{22}$$

where$\tau \in (0,1)$ is the confidence threshold. For each $x^T \in S_{\text{PL}}$, assign a pseudo-label:

$$\hat{y}^T = \arg\max_k p_\theta(k|x^T) \tag{23}$$

Merge it with the aligned source domain samples:

$$D_{\text{aug}} = D_S \cup \{(x^T, \hat{y}^T)|x^T \in S_{PL}\} \tag{24}$$

Minimize cross entropy again on the augmented data set:

$$\min_{\theta'} \frac{1}{|D_{\text{aug}}|}\sum_{(x,y)\in D_{\text{aug}}} l(g_{\theta'}(x), y) \tag{25}$$

Obtain the fine-tuned classifier $g_{\theta'}$.

Define target entropy as the performance of the model's prediction confidence for target domain samples after self-training:

$$H(x_t) = -\sum_{k=1}^{K} p_\theta(y_k|x_t)\, log\, p_\theta\,(y_k|x_t) \tag{26}$$

where $p(y_k|x_t)$ is the predicted distribution. The lower the entropy, the more confident the model output is and the higher the confidence level.

Algorithm: Overall process of Deep CORAL-based domain adaptation

Input: Source-domain features x_i^S and labels y_i^S, target-domain features x_i^T, shared Deep CORAL network f with parameters θ, Stage-I epochs, Stage-II rounds, Stage-II epochs, learning rates, batch size, trade-off coefficient λ, confidence threshold τ.

For each epoch in Stage-I **do**
  Sample source-domain mini-batch (x_i^S, y_i^S).
  Sample target-domain mini-batch x_i^T.
  Input x_i^S and x_i^T into f.
  Obtain source logits and latent features.
  Obtain target latent features.
  Evaluate L_{CLASS} by source-domain labels y_i^S.
  Evaluate L_{CORAL} by the covariance discrepancy between source and target latent features.
  Evaluate total loss $L_{\text{TOTAL}} = L_{\text{CLASS}} + \lambda L_{\text{CORAL}}$.
  Update the network parameters θ by back-propagation.
End For

x_i^T into the trained network and obtain the target prediction probabilities.
For each round **do**
  Select high-confidence target samples according to $\max_k p_\theta(k|x^T) \geq \tau$.
  Generate pseudo-labels for the selected target samples.
  Merge the pseudo-labeled target samples with the labeled source-domain samples.
  For each epoch in Stage-II **do**
    Evaluate the cross-entropy loss on the augmented dataset.
    Update the network parameters θ by minimizing the cross-entropy loss on D_{aug}.
  End For
  Recompute the prediction probabilities of the target-domain samples.
End For

Output: the final classifier f_c and the final target-domain prediction $\hat{y}^T$.

3.3.5. Interpretability analysis

Model interpretability is essential for cross-domain fault diagnosis, as domain adaptation models are often perceived as "black boxes." Improving transparency helps users understand why adaptation is needed, how alignment changes representations, and whether predictions are reliable—thereby reducing both blind distrust and over-trust.

(1) Ex-ante Explainability

The core of ex-ante explainability is to reveal domain differences before adaptation, correlate bearing failure mechanisms, explore the necessity and key directions of adaptation, and avoid blind adaptation. This is achieved mainly by quantifying cross-domain feature drift to reveal domain gaps and guide adaptation. Specifically, standardized differences in mean and variance are used to generate a drift report, and the most shifted features are highlighted via visualization. These results are further interpreted using bearing fault mechanisms, which identifies the key variables and provides a physically meaningful direction for subsequent adaptation.

(2) Adaptation Process Interpretability

During the adaptation process, this paper uses the Deep CORAL alignment method to achieve linear covariance alignment. The principal components of tension or compression are tracked and geometric changes are quantified to verify correlation with the bearing mechanism. If matched with the direction of spectral energy accumulation, it means that CORAL stretches along the fault-sensitive subspace, and the alignment direction is the same as the direction of the shock modulation mechanism, demonstrating that the domain adaptation process prioritizes subspaces associated with physical mechanisms, making the transfer process easier to interpret.

(3) Post-hoc Explainability

To demonstrate the reliability of the model's decisions, this paper adopts a multi-view explanation framework for post-hoc interpretation. First, a rank-importance approach is used to check whether the most important features match the failure mechanism. This method mainly quantifies the contribution of each feature to the model's decision-making by randomly varying individual features in the target domain data and observing the decrease in model prediction consistency. When the feature with the highest contribution aligns with physical mechanisms, it indicates that the model's logic is physically consistent.

SHAP is then employed to provide global and local explanations, feature contributions are linked to fault categories through SHAP ranking and swarm visualization, and key features are ultimately identified. SHAP analysis can clearly point out the main basis for a sample to be judged as a specific fault category, enabling traceable reasoning from statistical decision-making to physical fault diagrams. The SHAP analysis provides more than a ranking of feature importance, and it also offers a physically interpretable bridge between the diagnosis model and the failure mechanism of the bearing system. In this study, the features with large SHAP contributions are mainly associated with fault-characteristic frequencies, harmonic-energy descriptors, and order-related indicators. These quantities are closely related to the mechanical response induced by localized defects on the inner ring, outer ring, or rolling elements. Therefore, when such features consistently contribute to the predicted fault category, the SHAP results support the view that the model is not relying on arbitrary statistical fluctuations, but is instead capturing physically meaningful degradation signatures. For maintenance practice, this interpretability is valuable because it helps identify which fault-related indicators are driving the diagnosis result. In practical monitoring, engineers are often less concerned with a black-box class label alone than with which measurable symptom is becoming dominant. SHAP-based explanations can therefore assist maintenance personnel in prioritizing inspection targets, distinguishing likely fault locations, and evaluating whether the observed change is consistent with known failure evolution. This is particularly useful for condition-based maintenance, where early intervention depends not only on detection accuracy but also on confidence in the physical plausibility of the diagnosis. From an engineering perspective, the SHAP results also improve the deployability of the proposed framework in cross-domain scenarios. Since the diagnosis model is transferred across different operating conditions or devices, interpretability helps verify whether the transferred model still focuses on physically meaningful fault descriptors rather than spurious domain-specific patterns. In this sense, SHAP analysis enhances model transparency, supports engineering trust, and provides additional evidence that the

proposed framework remains aligned with bearing fault physics after domain adaptation.

4. Ilustrative cases

4.1. Feasibility verification under simple working conditions

This study leverages two distinct datasets to validate the proposed transfer learning model: the Case Western Reserve University (CWRU) bearing dataset [52] as the source domain and vibration data from an operational high-speed train as the target domain.

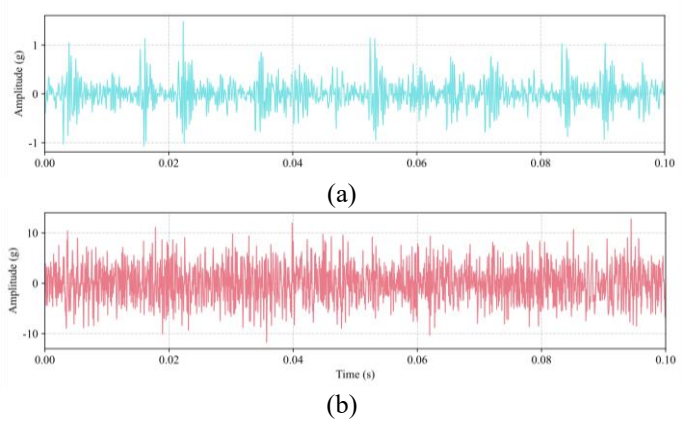
4.1.1. Experimental settings

The source-domain dataset used in this study was constructed from the extracted CWRU bearing data. Specifically, 161 labeled files were selected from four subsets, namely 12 kHz Drive End (DE) data, 12 kHz Fan End (FE) data, 48 kHz DE data, and 48 kHz normal-condition data. In the CWRU source domain, DE, FE, and BA denote the drive-end, fan-end, and base accelerometer channels of the motor test rig, respectively. The raw signal length is not consistent across files and contains a range of sample points ranging from 96,000 to 384,000. The dataset covers four health conditions: normal (N), ball fault (B), inner-race fault (IR), and outer-race fault (OR). Fault diameters include 0.007, 0.014, 0.021, and 0.028 inches, and the OR samples further involve fault locations at 3, 6, and 12 o'clock. In total, the source dataset contains 4 normal files, 40 B files, 40 IR files, and 77 OR files. These samples span four motor load conditions 0-3 hp, with shaft speeds approximately ranging from 1718 to 1798 rpm. In the experiments, each raw vibration file was treated as an independent signal and converted into a set of overlapping windows for subsequent feature extraction. A window length of 20 shaft revolutions was used throughout the study. For the CWRU source domain, the sliding step was set to 10 revolutions, producing 2887 source windows in total, including 1346 OR windows, 734 B windows, 727 IR windows, and 80 N windows. Detailed information is summarized in Table 2.

The target domain consists of 16 vibration files collected from an in-service high-speed train. Each file contains 256000 sample points. The data were collected at a fixed train speed of approximately 90 km/h, corresponding to about 600 rpm, with a 32 kHz sampling frequency. The target-domain vibration signal was collected by a single accelerometer mounted on the axlebox. For the target domain, a smaller sliding step of 4 revolutions was adopted to preserve more local fault information under the limited amount of available data. To keep the evaluation set balanced across classes, 16 windows were extracted from each target file, yielding 256 target windows in total, with 64 windows per class. During the domain adaptation process, these target-domain samples are treated as completely unlabeled and no fault labels are used for training or adaptation. The ground-truth labels are reserved exclusively for model evaluation. This setup provides an appropriate testbed for evaluating the proposed unsupervised domain adaptation framework while enabling quantitative performance assessment. As illustrated in Figure 3, a clear domain shift exists between the two domains: the source-domain signals exhibit relatively distinct periodic impulsive patterns, whereas the target-domain signals are characterized by stronger background noise and weaker fault-related impulses.

Table 2. Detailed description of the source domain dataset (CWRU).

Category	Health Condition	Fault Diameters (inches)	Motor Load (HP)	Number of Windows
0	Ball Fault	0.007, 0.014, 0.021, 0.028	0, 1, 2, 3	734
1	Inner Race Fault	0.007, 0.014, 0.021, 0.028	0, 1, 2, 3	727
2	Outer Race Fault	0.007, 0.014, 0.021	0, 1, 2, 3	1346
3	Normal	-	0, 1, 2, 3	80
Total	-		-	2887



(a)

(b)

Figure 3. Comparison of raw vibration signals: (a) source domain (CWRU, inner race fault); (b) target domain (high-speed train).

For the CWRU source domain, bearing characteristic frequencies were calculated using the published geometric

parameters of the SKF6205 and SKF6203 bearings. For the high-speed-train target domain, the complete geometric parameters of the in-service axlebox bearing are not publicly available. Therefore, the target-domain BPFO, BPFI, BSF, and FTF were not assumed to be identical to those of the CWRU bearings. Following the geometry-aware order-band strategy, target-domain features were extracted from fault-sensitive order bands of the single axlebox sensor signal and interpreted as order-band descriptors rather than exact target-bearing characteristic-frequency components.

Before feature computation, each raw signal segment was linearly detrended to remove the DC component and slow baseline drift. A fourth-order Butterworth band-pass filter was then applied using zero-phase forward-backward filtering with the Gustafsson method. The lower cut-off frequency was fixed at 500 Hz to suppress low-frequency components mainly associated with shaft rotation, structural vibration, and operating-condition drift. The upper cut-off frequency was set to $min(16000, 0.49f_o)$, where f_o denotes the original sampling frequency of each file. This setting preserves high-frequency resonance components related to bearing impacts while avoiding components too close to the Nyquist frequency. After filtering, signals were resampled to a unified sampling frequency of 32 kHz for subsequent windowing and feature extraction. Envelope-domain features were extracted from the filtered signal using the Hilbert transform. Specifically, the analytic signal was first obtained, and the envelope was computed as the modulus of the analytic signal. The mean value of the envelope was then removed to suppress the DC component and avoid zero-order spectral leakage. The envelope spectrum was calculated using the real-valued FFT, and the corresponding frequency axis was obtained according to the sampling frequency.

The retained metadata included shaft speed, rotational frequency, and the number of shaft revolutions. Based on these preprocessed windows, time-domain, spectral, envelope-domain, per-revolution, and order-domain features were then extracted. For the CWRU source domain, fault-characteristic-frequency-related features were constructed according to the bearing geometry of the corresponding CWRU bearings. The components associated with FTF, BPFO, BPFI, and BSF were quantified together with their first five harmonics and modulation sidebands. For the high-speed-train target domain, the envelope spectrum was converted into the order domain by normalizing frequency with respect to the shaft rotational frequency. Fault-sensitive order-band descriptors were then extracted from the single axlebox sensor signal and interpreted as order-band descriptors rather than exact target-bearing characteristic-frequency components. These descriptors include peak amplitude, band energy, energy ratio, harmonic energy, harmonic ratio, and sideband energy within the predefined order intervals. For order-domain feature extraction, the envelope-spectrum frequency axis was normalized by the nominal shaft rotational frequency. The order-domain half-bandwidth was set to 0.1 orders. Harmonic-energy descriptors were computed using the first M = 5 harmonics, and sideband-energy descriptors were computed with Q = 3 sidebands. These parameters were fixed for both source and target domains.

For source-domain feature screening, SGL with stability selection was applied to 187 candidate features grouped into five feature groups, namely BSF-related descriptors, BPFI-related descriptors, BPFO-related descriptors, FTF-related descriptors, and the remaining statistical descriptors. The perturbation amplitude was set to 0.4, the stability threshold was set to 0.6. The maximum number of iterations was 400, the bootstrap repetition number was 20, and the random forest classifier used 300 trees with the balanced-subsample class-weight strategy. The key symbolic hyperparameters used in feature screening, domain adaptation, and pseudo-label refinement are summarized in Table 3.

Table 3. Key symbolic hyperparameters used in the proposed framework.

Parameter	Value
λ_1	0.06
λ_2	0.25
subsample ratio	0.6
bootstrap repetitions	20
regularization perturbations	3
stability threshold	0.6
convergence tolerance	10^{-6}
maximum number of iterations	400
number of trees	300
λ	10.0
τ	0.9

Following Section 3, the adaptation backbone was implemented as a Deep CORAL network with two fully connected hidden layers of 64 and 32 neurons, each followed by ReLU activation and dropout with a rate of 0.3. A final linear

classifier was used for fault-category prediction. In Stage I, the network was optimized using Adam with a learning rate of 0.001, a weight decay of 10^{-5}, a batch size of 64, and 500 training epochs. To address the imbalance in the source-domain windows, class-weighted cross-entropy was used, where the weight of each class was computed as the total number of samples divided by the product of the number of classes and the corresponding class frequency. In this stage, the source-domain classification loss and the covariance alignment loss between D_S and D_T are jointly minimized. The CORAL loss weight was set to 10.0. Before optimization, the source and target domain features were standardized using StandardScaler with statistics computed from their respective domains. In Stage II, pseudo-label-based fine-tuning was conducted for three rounds. In each round, target windows were selected only when their maximum softmax probability exceeded the round-dependent threshold τ and the probability margin between the top-1 and top-2 classes was no smaller than 0.10. To alleviate pseudo-label imbalance, the number of selected windows in each predicted class was limited to at most 100 according to the confidence ranking. The selected pseudo-labeled target windows were merged with the original source windows to form the augmented training set D_{aug}, and the classifier was further updated using cross-entropy loss. Adam was adopted with a learning rate of 0.0002, a weight decay of 10^{-5}, a batch size of 64, and 60 fine-tuning epochs in each round.

4.1.2. Source domain data feature analysis and extraction

Figure 4 uses the RMS and kurtosis of the BA channel as two-dimensional coordinates to visualize the distribution of different health conditions in the source domain. Different colors and marker shapes represent different fault categories. The normal sample is located near the low-RMS and low-kurtosis region, showing a clear difference from most fault samples. Among the fault categories, however, the distributions are not completely separated. The B samples are mainly concentrated in the low-to-moderate kurtosis region with a relatively wide RMS range, while the IR samples mostly appear in the low-RMS and moderate-kurtosis region. The OR samples show a broader distribution and include several high-kurtosis regions, indicating stronger impulsive characteristics for part of the outer-race fault samples. Meanwhile, noticeable overlap exists among B, IR, and OR in the low-to-moderate RMS and kurtosis ranges. Therefore, the RMS-kurtosis feature pair of the BA channel provides useful but limited discriminative information. It can reflect differences in vibration energy and impulsiveness among fault types, but it is insufficient to fully separate all categories by itself. This observation motivates the use of multi-channel and multi-domain feature extraction for subsequent fault classification.

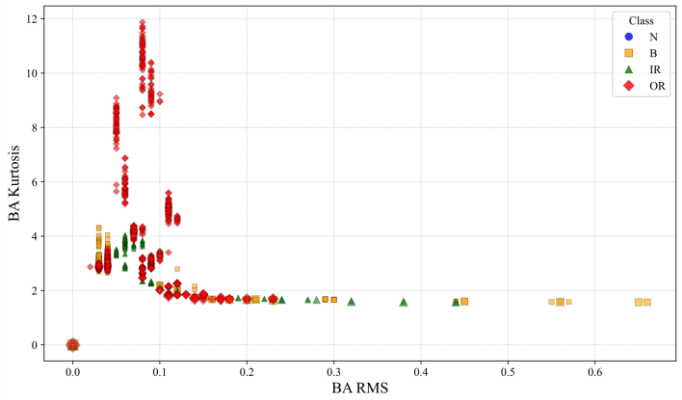


Figure 4. RMS-kurtosis scatter plot of BA channel under different health conditions.

Figure 5 shows the Pearson correlation heat map of RMS and kurtosis features extracted from the DE, FE, and BA channels in the source domain. The diagonal elements are all 1.00, indicating self-correlation. Most off-diagonal correlation coefficients are relatively small, suggesting that the RMS and kurtosis features from different channels and feature types provide largely complementary information. Specifically, the correlation between DE_rms and DE_kurtosis is only 0.02, indicating that vibration energy and impulsiveness at the DE channel are almost independent in this feature space. The correlations between DE_rms and the other channel features are also weak or slightly negative, including FE_rms -0.11, FE_kurtosis -0.19, BA_rms -0.28, and BA_kurtosis -0.11. Relatively higher positive correlations are observed between DE_kurtosis and BA_kurtosis 0.48, between FE_rms and FE_kurtosis 0.34, and between FE_kurtosis and BA_rms 0.30. These results indicate that some impulsiveness-related features across channels share partial common information, while the overall correlation level remains moderate or weak. Therefore, fusing RMS and kurtosis features from multiple measurement channels can provide complementary descriptions of vibration energy and impact characteristics, which is beneficial for subsequent fault feature construction and classification.

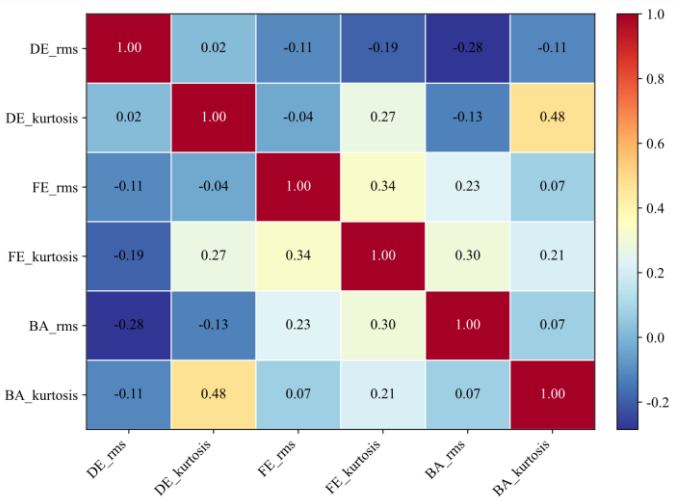


Figure 5. Block correlation heat map of RMS and kurtosis at DE, FE and BA channels in source domain.

Following the proposed methodology, 187 candidate features were initially constructed from fault-characteristic order-band descriptors and general statistical descriptors. SGL with stability selection was then applied to identify stable and discriminative features, and an RF classifier was trained on the selected feature subset. Figure 6 shows the importance scores of the selected Top-10 features. The most important feature is DE_BPFI_SB_Q3_ord, with an importance score of 0.1378, followed by DE_BSF_harmE_M5_ord 0.1240, DE_BPFI_harmRatio_M5_ord 0.1111, and DE_BPFI_Eratio_ord 0.1049. These leading features are mainly associated with BPFI- and BSF-related order-band descriptors, indicating that the source-domain RF classifier relies strongly on inner-race-sensitive sideband statistics and rolling-element-sensitive harmonic or energy-ratio information. In contrast, no BPFO-related descriptor appears among the Top-10 features, suggesting that BPFO-related information is not a dominant discriminative factor in this selected feature subset. Among the CWRU source-domain channels, the DE-channel features contribute most prominently, indicating that the drive-end signal contains strong bearing-fault-related information in the CWRU dataset.

Four commonly used multi-class classification models, RF, Logistic Regression (LR), Support Vector Classification (SVC), and Multilayer Perceptron (MLP), were selected for source-domain model evaluation. The evaluation indicators are accuracy, precision, recall, and F1 score. The F1 score is a widely used metric for classification performance evaluation, especially for imbalanced datasets, because it balances precision and recall. It is defined as $\mathrm{F1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\mathrm{precision} + \mathrm{recall})}$.

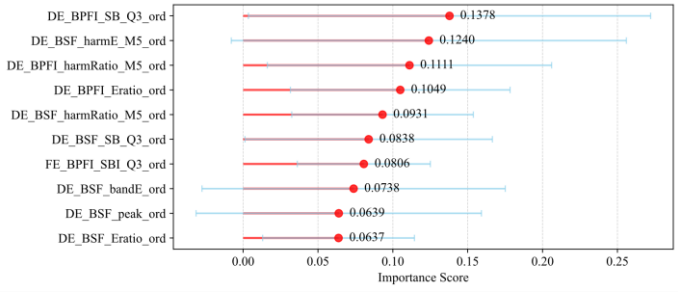


Figure 6. The top-10 feature importance scores from the trained RF model.

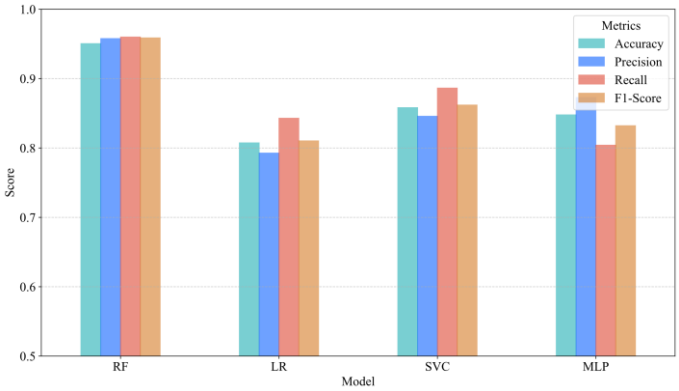


Figure 7. Performance scores of various models on the source domain.

As shown in Figure 7, RF achieves the best overall performance among the four models, with all four metrics exceeding 0.95. Specifically, its accuracy, precision, recall, and F1-score are approximately 0.950, 0.958, 0.960, and 0.959, respectively. SVC and MLP also show relatively strong performance, with all metrics generally above 0.80. SVC obtains approximately 0.858 accuracy, 0.846 precision, 0.886 recall, and 0.862 F1-score, while MLP achieves approximately 0.848 accuracy, 0.873 precision, 0.804 recall, and 0.832 F1-score. LR performs comparatively weaker, with accuracy and F1-score around 0.81 and precision slightly below 0.80. Overall, the results indicate that the selected features provide effective discriminative information for source-domain fault diagnosis, among which RF demonstrates the most stable and superior classification performance.

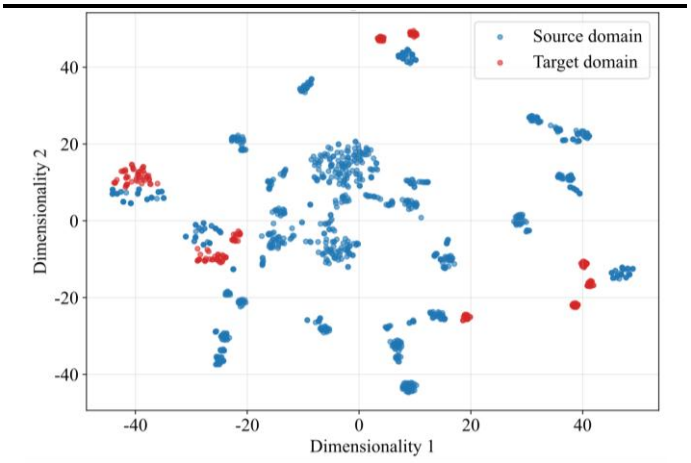
4.1.3. Target domain transfer diagnosis

Table 4 shows the iteration loss statistics during Deep CORAL training. As training proceeds from 10 to 500 iterations, the classification loss decreases from 1.152 to 0.139, while the total

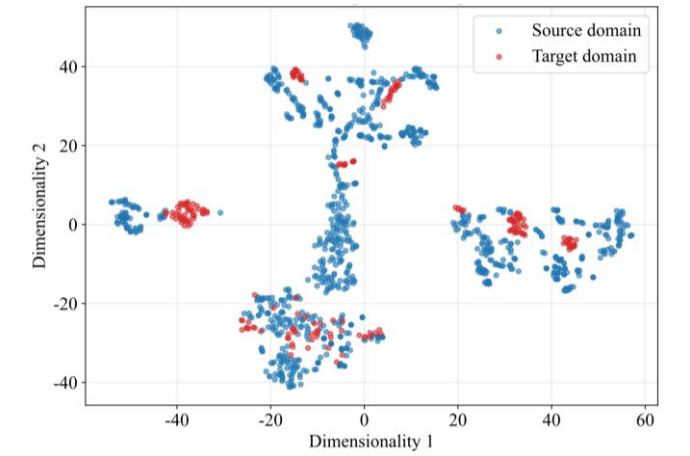
loss decreases from 1.159 to 0.166. The CORAL loss remains within a relatively small range of 0.007 ~ 0.055 and gradually stabilizes after the early training stage. These results indicate that the model achieves stable joint optimization of source-domain classification and source-target covariance alignment. The decreasing total loss suggests that Deep CORAL effectively reduces the distribution discrepancy between the source and target domains while maintaining discriminative classification ability.

Table 4. Iteration loss statistics during deep CORAL training.

Iteration	Classification Loss	CORAL Loss	Total Loss
10	1.152	0.007	1.159
50	0.420	0.055	0.475
100	0.287	0.050	0.337
200	0.179	0.031	0.210
300	0.181	0.027	0.208
400	0.166	0.024	0.190
500	0.139	0.027	0.166



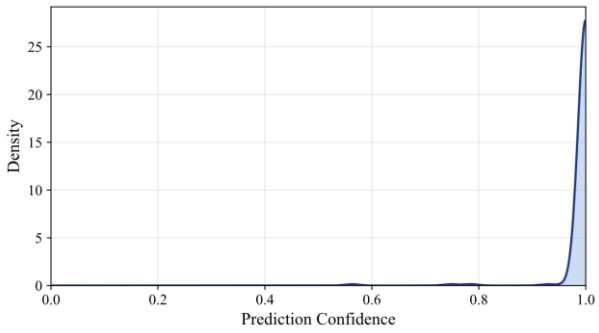
(a)



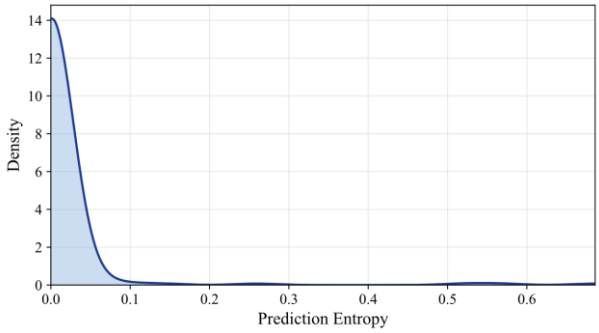
(b)

Figure 8. The t-SNE visualization of source-domain and target-domain feature distributions: (a) before alignment; (b) after alignment.

Figure 8 shows the t-SNE visualization of the source-domain and target-domain feature distributions before and after Deep CORAL alignment. In Figure 8(a), the source-domain samples and target-domain samples are projected into several separated regions, indicating a clear distribution discrepancy between the two domains when raw features are used. After Deep CORAL alignment, as shown in Figure 8(b), the target-domain samples move closer to the source-domain feature clusters and the overlap between the two domains is increased in the two-dimensional embedding space. This result suggests that Deep CORAL reduces the second-order statistical discrepancy between the source and target feature distributions and improves cross-domain feature alignment.



(a)



(b)

Figure 9. Distribution of target-domain pseudo-label filtering indicators: (a) Kernel density distribution of prediction confidence; (b) Kernel density distribution of prediction entropy.

Figure 9(a) presents the kernel density distribution of the maximum prediction probability for target-domain samples. The maximum prediction probability is used as the confidence score of each target-domain prediction. The distribution is strongly concentrated near 1.0, indicating that most target-domain samples are assigned highly confident predictions by the adapted model. Meanwhile, several minor density peaks

appear in the medium-to-high confidence range, suggesting that a small number of samples still exhibit relatively lower prediction confidence. Therefore, the confidence threshold used for pseudo-label selection is necessary to exclude uncertain target samples and reduce the risk of introducing unreliable pseudo-labels. Figure 9(b) shows the kernel density distribution of prediction entropy on the target domain. Prediction entropy measures the uncertainty of the model output; a smaller entropy value indicates a more deterministic class prediction. The entropy distribution is mainly concentrated near zero, which is consistent with the high-confidence pattern observed in Figure 9(a). However, a long-tail distribution can also be observed, indicating that a small proportion of target-domain samples still have higher uncertainty. These results suggest that the adapted model produces confident predictions for most target samples, while confidence-based and entropy-aware pseudo-label filtering remains important for preventing uncertain samples from affecting the subsequent self-training process.

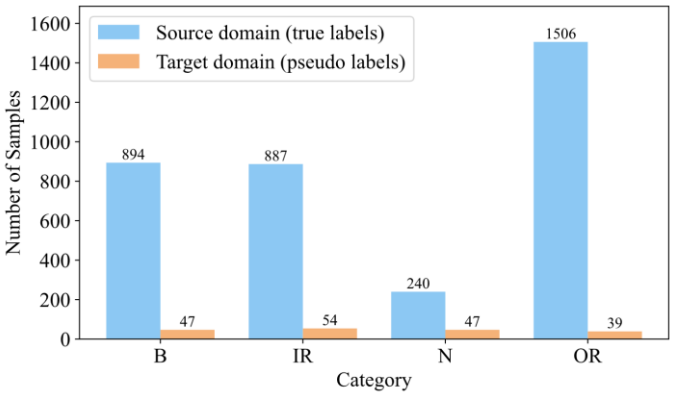


Figure 10. Source domain and target domain (pseudo-label) category distribution comparison.

Figure 10 compares the category distribution between the source-domain samples with true labels and the selected high-confidence target-domain pseudo-labels. In the source domain, the four categories are imbalanced, with 894 B samples, 887 IR samples, 240 N samples, and 1506 OR samples. In the target domain, 187 high-confidence pseudo-labeled samples are selected under the confidence threshold of 0.9, including 47 B samples, 54 IR samples, 47 N samples, and 39 OR samples. These results indicate that the model can identify a subset of target-domain samples with reliable prediction confidence for pseudo-label-based fine-tuning. However, the selected target pseudo-labels are still limited in number and show a mildly uneven class distribution; therefore, the pseudo-label distribution should be interpreted cautiously. Combining Figures 8–10, the results suggest that the Deep CORAL plus pseudo-labeling strategy improves domain alignment, produces confident predictions for part of the target domain, and provides a basis for further target-domain refinement.

To further evaluate the effectiveness of the proposed framework, comparative experiments were conducted with several representative methods, namely Fine KNN [53], unsupervised class incremental learning network (UCILN) [54] and dynamic collaborative adversarial domain adaptation network (DCADAN) [35], on the target domain bearing fault diagnosis task. Considering the potential distribution mismatch and class imbalance in cross-domain diagnosis, Raw Accuracy, Macro-F1, and AUC were adopted as the main evaluation metrics. For baseline implementation, Fine-KNN was implemented as CORAL whitening alignment followed by 1-nearest-neighbor classification. UCILN was implemented as CORAL whitening alignment followed by an RF classifier with 600 trees and class-weight strategy. DCADAN was trained with a learning rate of , weight decay of , dropout of 0.3 and hidden feature dimensions of 64 and 32. To obtain statistically reliable results, 100 replicate experiments were performed for each method and the mean ± standard deviation of all runs was reported. The corresponding results are summarized in Table 5.

Table 5. Performance comparison of different methods on the target domain ($p < 0.05$, t-test).

Method	Raw Accuracy (%)	Macro-F1 (%)	AUC (%)
Fine KNN	36.98 ± 0.44	30.69 ± 0.38	57.99 ± 0.29
UCILN	43.75 ± 0.66	31.26 ± 0.54	80.26 ± 0.94
DCADAN	90.21 ± 1.33	90.10 ± 1.38	99.05 ± 0.57
Proposed	**98.02 ± 1.21**	**98.01 ± 1.22**	**99.92 ± 0.12**

As shown in Table 5, the proposed method consistently outperforms all comparison methods across the three evaluation metrics. Specifically, the proposed method achieves a Raw Accuracy of 98.02% ± 1.21%, a Macro-F1 of 98.01% ± 1.22%, and an AUC of 99.92% ± 0.12%. These results are higher than those of Fine KNN, UCILN, and DCADAN, indicating that the proposed framework provides stronger target-domain diagnostic performance under the investigated cross-domain condition. Fine KNN achieves only 36.98% ± 0.44% Raw Accuracy and 57.99% ± 0.29% AUC, suggesting that the shallow distance-based classifier is highly sensitive to the source-target distribution shift. UCILN improves the AUC to

80.26% ± 0.94%, but its Raw Accuracy and Macro-F1 remain relatively low, indicating that incremental representation learning alone is insufficient for robust target-domain diagnosis. DCADAN achieves 90.21% ± 1.33% Raw Accuracy, 90.10% ± 1.38% Macro-F1, and 99.05% ± 0.57% AUC, showing strong performance but still remaining below the proposed method. Overall, the comparison demonstrates the effectiveness of combining feature selection, Deep CORAL alignment, and pseudo-label refinement. Nevertheless, because the target-domain evaluation is based on a limited transductive dataset, further validation on larger-scale datasets is still necessary.

To investigate the efficacy of each module in the proposed adaptation strategy, this paper conducted an ablation study. The diagnostic accuracies, verified against a manually labeled target domain subset, are detailed in Table 6.

Table 6. Comparison of diagnostic accuracy on the target domain.

Method	Description	Accuracy (%)
Source-only	train on source, no CORAL alignment, no pseudo-labels	88.44
+pseudo-labeling (without CORAL)	no CORAL term, pseudo-label self-training	91.56
+Deep CORAL (without pseudo-labeling)	Deep CORAL alignment, no pseudo-label self-training	92.60
Full Method	**Deep CORAL + pseudo-label self-training**	**98.02**

The results in Table 6 reveal a clear and progressive performance improvement as additional adaptation modules are incorporated. The source-only model, trained on the source domain without CORAL alignment or pseudo-labeling, achieves 88.44% accuracy on the target domain. When pseudo-label self-training is introduced without CORAL alignment, the accuracy increases to 91.56%, indicating that high-confidence target samples can help refine the decision boundary. When Deep CORAL alignment is used without pseudo-labeling, the accuracy further increases to 92.60%, showing that covariance alignment is effective in reducing the source-target distribution discrepancy. The full method, which combines Deep CORAL alignment with pseudo-label self-training, achieves the highest accuracy of 98.02%. Compared with the source-only model, the full method improves accuracy by 9.58 percentage points. These results confirm that Deep CORAL and pseudo-label refinement provide complementary benefits: CORAL reduces latent-space distribution mismatch, while pseudo-labeling further adapts the classifier to confident target-domain samples.

4.1.4. Interpretability analysis

In order to reveal the cross-domain differences before adaptation, cross-domain drift is quantified using standardized mean differences to identify the features most affected by working-condition and device changes. Figure 11 compares the distribution difference of each selected feature between the source and target domains. The horizontal axis represents the standardized mean difference, and the vertical axis lists the feature names. A larger value indicates a stronger mean shift between the two domains. The most significant domain discrepancy appears in t_shape, with a standardized mean difference of 1.84, indicating that the waveform shape factor is highly sensitive to the change from the CWRU test rig to the high-speed-train axlebox condition. Other features with relatively large shifts include h_spec_gini 0.70, h_sig_peak_frac 0.53, ord_entropy 0.45, ord_spread 0.37, and ord_centroid 0.36. These results suggest that the source-target discrepancy is mainly reflected in general time-domain waveform characteristics, spectral concentration, peak-related properties, and order-domain distribution descriptors, rather than being limited to a single fault-characteristic-frequency component. Therefore, the selected features capture both global signal-distribution differences and order-domain structural differences caused by changes in operating conditions and measurement systems.

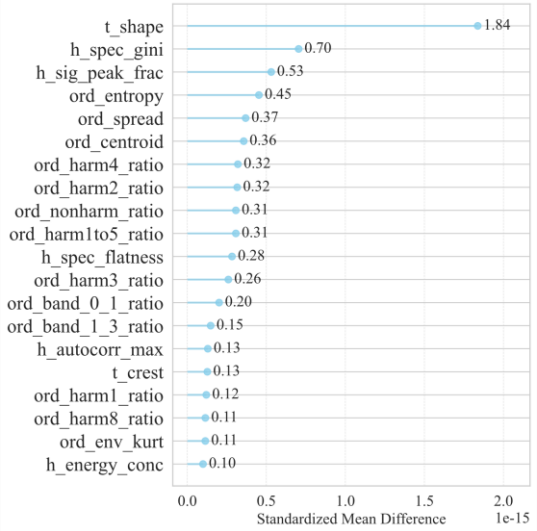


Figure 11. Cross-domain standardized mean differences of selected feature.

Figure 12 shows the zoom strength of Deep CORAL along the covariance principal axes during second-order distribution alignment. The horizontal axis denotes the covariance principal-axis number, and the vertical axis denotes the zoom strength applied to each axis. A value greater than 1 indicates expansion of the corresponding covariance direction, whereas a value smaller than 1 indicates compression. The first two principal axes are expanded, with zoom strengths of 1.16 and 1.28, respectively. In contrast, most remaining axes are compressed to different degrees, with representative values of 0.81, 0.71, 0.82, 0.85, 0.72, 0.78, and 0.83 for axes 3–9, and lower values ranging from 0.36 to 0.62 for axes 10–20. This pattern indicates that Deep CORAL does not uniformly transform all feature directions. Instead, it selectively expands a small number of dominant covariance directions while compressing many secondary directions, thereby reducing the second-order statistical discrepancy between the source and target domains.

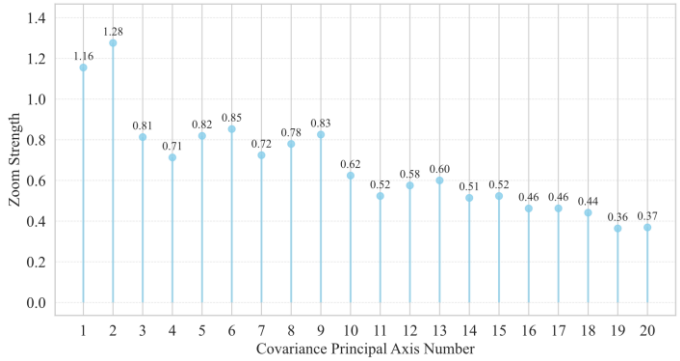


Figure 12. Deep CORAL aligned principal axis scaling intensity distribution.

In the post-hoc explanation stage, permutation importance was computed on the aligned target-domain samples to assess which features most strongly affect the stability of the model's target-domain decisions under feature-wise perturbation. Specifically, one feature was randomly shuffled at a time, and the resulting decrease in prediction consistency was recorded. A larger decrease indicates that the model is more sensitive to that feature in the target domain. Figure 13 summarizes the top 20 permutation importance scores, where the horizontal axis denotes the mean permutation importance with standard deviation, and the vertical axis lists the feature names. The most influential feature is ord_harm7_ratio, with a mean importance of 0.176, followed by ord_centroid 0.107, ord_harm9_ratio 0.092, ord_band_1_3_ratio 0.084, and ord_harm3_ratio 0.074. Other moderately important features include ord_band_0_1_ratio 0.057, h_spec_flatness 0.046, ord_harm5_ratio 0.045, t_crest 0.044, and ord_harm2_ratio 0.044. Overall, the model's target-domain decisions are mainly affected by order-domain harmonic-ratio and band-ratio descriptors, together with several time-domain and spectral-shape descriptors. This indicates that the adapted model relies primarily on order-domain distribution and harmonic-structure information.

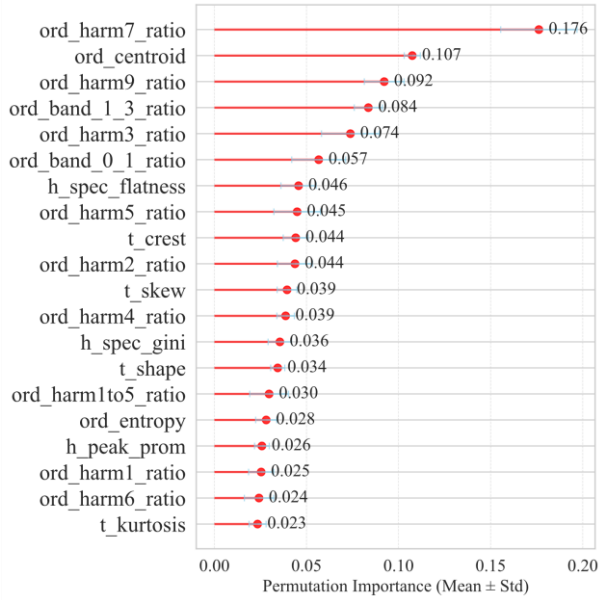


Figure 13. Permutation importance ranking of the top 20 features on the aligned target-domain.

To further analyze the decision logic of the adapted model, SHAP values were computed for target-domain samples and summarized separately for each category. DeepSHAP gradient-based explainer was applied to the trained Deep CORAL network. The explainer was wrapped so that the explained model output corresponded to the per-class softmax probability. The background reference distribution consisted of 100 target-domain windows randomly sampled without replacement, and SHAP values were computed for 500 randomly sampled target-domain windows, or for all available target-domain windows when fewer than 500 samples were available. All random sampling procedures used a fixed random seed. For multi-class interpretation, class-specific SHAP values were obtained for each output class. Per-class mean absolute SHAP values were then aggregated and averaged across classes to rank the top-20 important features. Figure 14 shows the category-wise SHAP summary plots, where the horizontal axis denotes the SHAP value, the vertical axis denotes the feature name, and the color

indicates the feature value magnitude. A larger absolute SHAP value indicates a stronger contribution to the model output. The results show that the adapted model mainly relies on a combination of time-domain, spectral, and order-domain features. For the Ball Fault category, important features include h_spec_gini, ord_nonharm_ratio, ord_band_3_6_ratio, ord_harm1to5_ratio, and h_autocorr_max. For the Inner Race Fault category, the most influential features are t_skew, ord_centroid, ord_harm7_ratio, t_impulse, and h_autocorr_max. For the Outer Race Fault category, the dominant features are mainly order-band descriptors, such as ord_band_3_6_ratio, ord_band_0_1_ratio, ord_band_1_3_ratio, ord_band_6_10_ratio, and ord_centroid. For the Normal category, the model is mainly affected by t_skew, h_spec_gini, ord_harm7_ratio, h_autocorr_max, t_crest, and ord_centroid. These results indicate that the model forms its target-domain decisions by jointly considering waveform asymmetry, impulsiveness, spectral concentration, autocorrelation structure, and order-domain energy distribution. The category-wise attribution patterns are consistent with the physical characteristics of bearing vibration signals, confirming that the adapted model learns interpretable and fault-discriminative representations in the target domain.

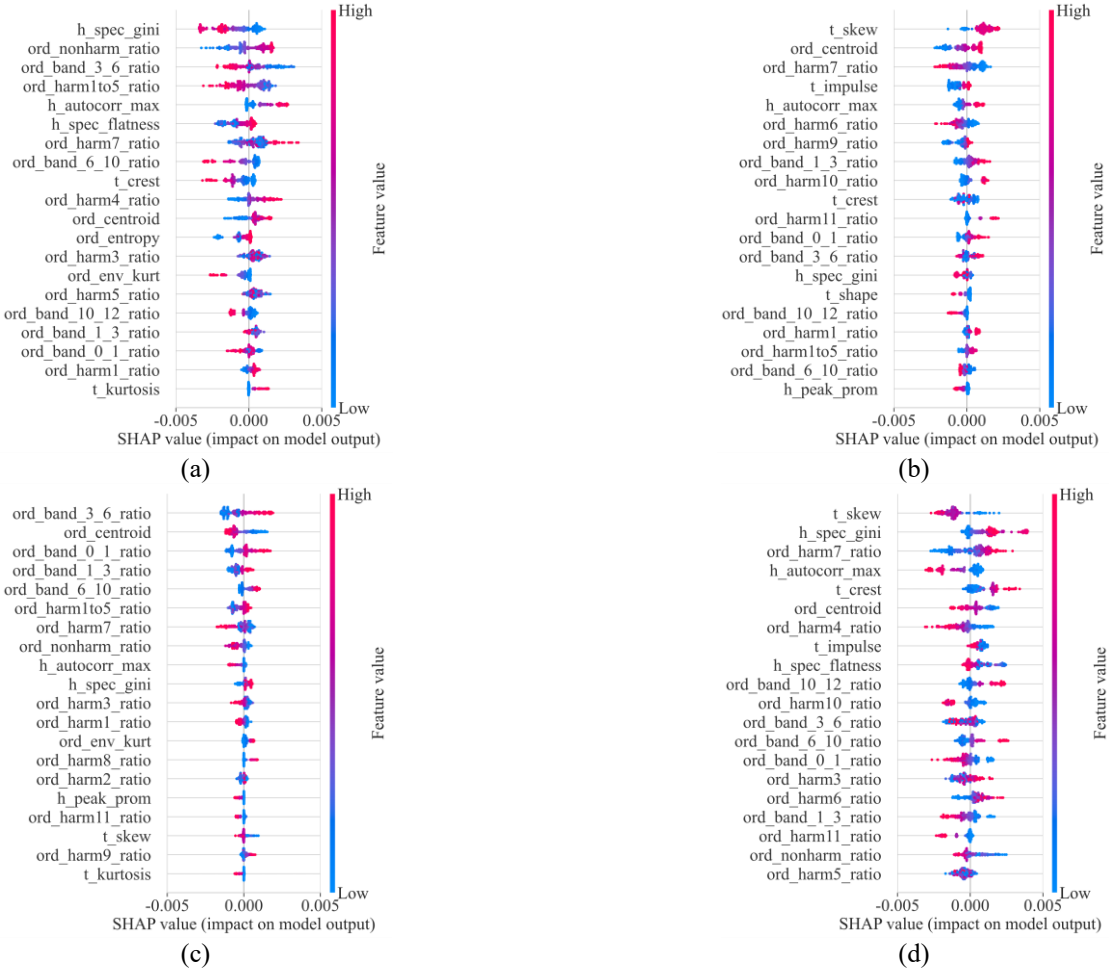


Figure 14. Category-wise distributions of SHAP Values for the main features in the target domain: (a) ball fault; (b) inner race fault; (c) outer race fault; (d) normal.

Figure 15 presents the relative contribution of the global mean absolute SHAP values across all target-domain samples. The top contributor is h_spec_gini, followed closely by ord_harm7_ratio, ord_band_3_6_ratio, ord_centroid and t_skew. These five features together account for approximately 35.5% of the total absolute attribution, indicating that the model's global decisions are distributed across spectral sparsity measures, high-order harmonic ratios, band-energy descriptors, spectral centroids, and time-domain asymmetry indicators. This broad distribution suggests that after domain adaptation, the model integrates complementary information from spectral flatness, order-band ratios, time-domain statistical moments, and harmonic structures to achieve robust and generalizable fault diagnosis in the target domain. Such diversity in feature

usage also enhances the model's resilience to sensor-channel mismatch and varying fault signatures across domains.

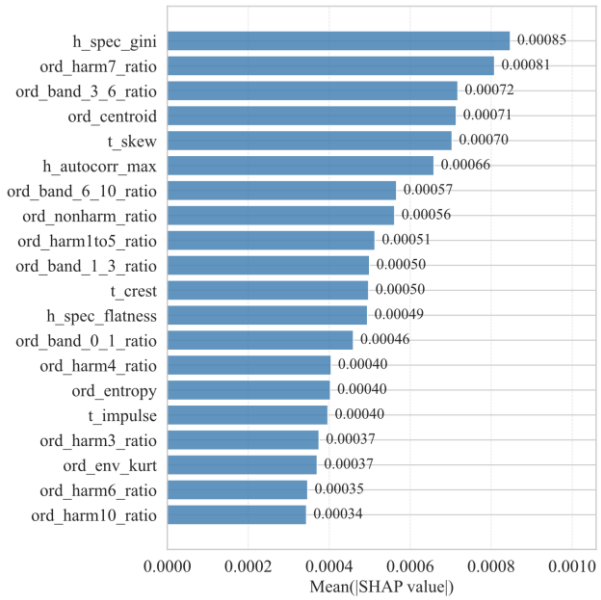


Figure 15. Contribution percentages of global mean absolute SHAP values on the target domain.

4.2. External generalization verification under complex working conditions

To further validate the generalizability of the proposed framework under more challenging working conditions, including cross-speed operation and strong ambient noise, additional cross-domain experiments were conducted using an in-house bearing fault diagnosis dataset. The dataset was acquired from a centrifugal fan test rig equipped with a 3.7 kW induction motor and two roller bearings, namely N205 and NU205. Vibration signals were collected using a single accelerometer at a sampling frequency of 50 kHz under two shaft speeds: 600 r/min and 1000 r/min. Four health conditions were considered: normal condition, inner-race fault, outer-race fault, and ball fault. The faults were artificially introduced by wire electrical discharge machining. Following the same sliding-window strategy, each raw signal was segmented into overlapping windows corresponding to 10 shaft revolutions, with a step size of 2 revolutions. The 600 r/min subset was used as the labelled source domain, whereas the 1000 r/min subset was used as the target domain. Target-domain labels were not used during training and were reserved only for evaluation. Because the source and target signals are acquired on the same test rig and bearing, the domain shift here arises predominantly from the change in rotational speed, which alters the characteristic fault frequencies, the modulation of fault-induced impulses, and the overall spectral distribution. A total of 84 candidate features were constructed, including time-domain statistics, spectral descriptors, envelope indicators, and fault-characteristic order-band descriptors related to BPFO, BPFI, BSF, and FTF, calculated in both the frequency and order domains. Sparse Group Lasso with stability selection was then applied to the source-domain data to identify stable and discriminative features. The features were grouped by fault family, including BSF, BPFI, BPFO, and FTF, together with a residual group. The top 24 features were retained for all subsequent tasks. Table 7 reports the fault-type settings for the training and test splits of the in-house testbed dataset, while Table 8 summarizes the key symbolic hyperparameters used in the proposed framework.

Table 7. Setting of fault types in the training and test splits of the in-house testbed dataset.

Fault type	Source-domain	Target-domain
Ball Fault	46	79
Inner Race Fault	46	79
Outer Race Fault	46	79
Normal	146	246

Table 8. Key symbolic hyperparameters used in the proposed framework.

Parameter	Value
λ_1	0.06
λ_2	0.25
subsample ratio	0.6
bootstrap repetitions	20
regularization perturbations	3
stability threshold	0.6
Deep CORAL hidden layers	[64, 32]
Deep CORAL learning rate	0.001
CORAL loss weight	0.5
Pseudo label confidence threshold	0.95
Pseudo label margin threshold	0.08
convergence tolerance	10^{-6}
maximum number of iterations	400
number of trees	100
λ	0.5
τ	0.95

To provide a comprehensive evaluation, accuracy, balanced accuracy, macro-F1, and AUC were jointly adopted as the main performance metrics. Table 9 summarizes the diagnostic results for the transfer task. The proposed method achieves an average accuracy of 96.76%, together with a balanced accuracy of 96.55%, a macro-F1 of 95.85%, and an AUC of 99.43%. The high balanced accuracy and macro-F1 indicate effective recognition across all four fault categories, while the AUC value

confirms the strong discriminative ability of the proposed method under severe distribution shift.

Table 9. External generalization performance of the proposed method.

Accuracy (%)	Balanced Accuracy (%)	Macro-F1 (%)	AUC (%)
96.76 ± 0.32	96.55 ± 0.70	95.85 ± 0.61	99.43 ± 0.29

To assess the practical resilience of the proposed framework, additive white Gaussian noise (AWGN) was injected into the raw target-domain signals at different signal-to-noise ratios, namely 20 dB, 10 dB, 5 dB, 0 dB, and -5 dB. The target windows were then re-extracted from the noisy signals, and the same pre-trained model was evaluated without retraining. This setting simulates real-world scenarios in which test-time signal quality deteriorates due to environmental noise. The results are reported in Table 10.

Table 10. Noise robustness results of the proposed method under different SNR conditions.

SNR (dB)	Accuracy	Balanced Accuracy	Macro-F1	AUC
20	96.55 ± 0.48	96.23 ± 0.76	95.53 ± 0.75	99.37 ± 0.31
10	96.07 ± 0.72	95.28 ± 1.12	94.61 ± 1.13	98.97 ± 0.46
5	95.38 ± 0.43	93.86 ± 0.85	93.26 ± 0.83	98.56 ± 0.72
0	93.31 ± 1.69	89.91 ± 2.63	89.25 ± 2.92	98.44 ± 0.55
-5	88.89 ± 1.14	83.02 ± 1.74	80.82 ± 2.52	97.77 ± 0.04

The noise robustness results show a gradual performance degradation as the SNR decreases from 20 dB to -5 dB. This trend is expected, since stronger noise progressively masks fault-related impulses and weakens discriminative signal features. Nevertheless, even under the severe noise condition of -5 dB, the proposed method still achieves an accuracy of 88.89% and an AUC of 97.77%. These results indicate that the proposed framework maintains reliable diagnostic capability under strong noise interference and demonstrates favorable robustness in complex operating environments.

5. Conclusion

This paper proposed a physics-informed and interpretable cross-domain fault diagnosis framework for rolling bearings under domain shift and limited target-domain labels. The framework integrates multi-domain feature extraction, Sparse-Group Lasso with stability selection, source-domain feature evaluation, order-domain normalization, Deep CORAL-based distribution alignment, pseudo-label refinement, and multi-level interpretability analysis. By combining mechanism-guided feature construction with unsupervised domain adaptation, the proposed method improves both diagnostic transferability and decision transparency. Experimental results on the CWRU-to-high-speed-train transfer task and the in-house cross-speed testbed demonstrate the effectiveness of the proposed framework. Compared with representative baseline methods, the proposed method achieves higher target-domain accuracy, macro-F1, and AUC. The ablation study further confirms that Deep CORAL alignment and pseudo-label self-training provide complementary benefits: covariance alignment reduces source-target distribution discrepancy, while high-confidence pseudo-labels further refine the target-domain decision boundary. The additional noisy-condition experiments also show that the proposed framework maintains favorable robustness under strong interference. The interpretability analysis indicates that the model decisions are mainly associated with physically meaningful descriptors, including order-domain harmonic ratios, order-band energy distributions, spectral concentration measures, and time-domain impulsiveness indicators. The permutation-importance and SHAP results suggest that the adapted model does not rely solely on arbitrary statistical correlations, but instead focuses on features that are closely related to bearing fault modulation mechanisms. This improves the transparency of the cross-domain diagnostic process and provides useful evidence for engineering trust.

Several promising research directions remain for future work. First, the proposed framework can be extended to more complex fault evolution scenarios and larger-scale multi-component systems, where the curse of dimensionality may necessitate more efficient feature selection or representation learning strategies. Additionally, more applications that fuse multi-sensor data, such as vibration, temperature, acoustic emission, to respond to more comprehensive and resilient diagnostic systems can be explored. Finally, joint optimization of fault diagnosis confidence and maintenance decision-making under resource constraints represents an important and practical direction for future investigation, especially in safety-critical systems such as high-speed railways.

References

1. Randall R, Antoni J. Rolling element bearing diagnostics—A tutorial. *Mechanical Systems and Signal Processing* 2011; 25(2): 485-520,

https://doi.org/10.1016/j.ymssp.2010.07.017.
2. Hoang D, Kang H. A survey on deep learning based bearing fault diagnosis. *Neurocomputing* 2019; 335: 327-335, https://doi.org/10.1016/j.neucom.2018.06.078.
3. Cerrada M, Sánchez R, Li C, Pacheco F, Cabrera D, de Oliveira J, Vásquez R. A review on data-driven fault severity assessment in rolling bearings. *Mechanical Systems and Signal Processing* 2018; 99: 169-196. https://doi.org/10.1016/j.ymssp.2017.06.012.
4. Seryasat O, Honarvar F, Rahmani A. Multi-fault diagnosis of ball bearing using FFT, wavelet energy entropy mean and root mean square. In: Proceedings of the 2010 IEEE International Conference on Systems, Man and Cybernetics; Istanbul, Turkey. IEEE; 2010. p. 4295-4299. https://doi.org/10.1109/ICSMC.2010.5642389.
5. Cao W, Jia X, Liu Y, Hu Q. Selective maintenance optimization for fuzzy multi-state systems. *Journal of Intelligent & Fuzzy Systems* 2018; 34(1): 105-121. https://doi.org/10.3233/JIFS-17031.
6. Zhang W, Li C, Peng G, Chen Y, Zhang Z. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mechanical Systems and Signal Processing* 2018; 100: 439-453. https://doi.org/10.1016/j.ymssp.2017.06.022.
7. Lei Y, He Z, Zi Y. EEMD method and WNN for fault diagnosis of locomotive roller bearings. *Expert Systems with Applications* 2011; 38(6): 7334-7341. https://doi.org/10.1016/j.eswa.2010.12.095.
8. Lin L, Hongbing J. Signal feature extraction based on an improved EMD method. *Measurement* 2009; 42(5): 796-803. https://doi.org/10.1016/j.measurement.2009.01.001.
9. Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi A. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing* 2020; 138: 106587. https://doi.org/10.1016/j.ymssp.2019.106587.
10. Tan S, Hu Q, Guo C, Zhu D, Dong E, Zhang F. Operational readiness-oriented condition-based maintenance and spare parts optimization for multi-state system. *Reliability Engineering & System Safety* 2025; 264: 111367. https://doi.org/10.1016/j.ress.2025.111367.
11. Zhu Z, Lei Y, Qi G, Chai Y, Mazur N, Annamdas V. A review of the application of deep learning in intelligent fault diagnosis of rotating machinery. *Measurement* 2023; 206: 112346. https://doi.org/10.1016/j.measurement.2022.112346.
12. Wei L, Peng X, Cao Y. Multi-Scale Graph Transformer for rolling bearing fault diagnosis. *Eksploatacja i Niezawodność – Maintenance and Reliability* 2025; 27(2): 194779. https://doi.org/10.17531/ein/194779.
13. Lu X, Xiao Y. Fault diagnosis method of automobile rolling bearing based on transfer learning and improved DenseNet. *Eksploatacja i Niezawodność – Maintenance and Reliability* 2025; 27(2): 194675. https://doi.org/10.17531/ein/194675.
14. Wen L, Li X, Gao L, Zhang Y. A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Transactions on Industrial Electronics* 2018; 65(7): 5990-5998. https://doi.org/10.1109/TIE.2017.2774777.
15. He M, He D. Deep learning based approach for bearing fault diagnosis. *IEEE Transactions on Industry Applications* 2017; 53(3): 3057-3065. https://doi.org/10.1109/TIA.2017.2661250.
16. Ngaopitakkul A, Bunjongjit S. An application of a discrete wavelet transform and a back-propagation neural network algorithm for fault diagnosis on single-circuit transmission line. *International Journal of Systems Science* 2013; 44(9): 1745-1761. https://doi.org/10.1080/00207721.2012.670290.
17. Attoui I. Novel fast and automatic condition monitoring strategy based on small amount of labeled data. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 2022; 52(2): 1266-1275. https://doi.org/10.1109/TSMC.2020.3018102.
18. Do C, Ng A. Transfer learning for text classification. In: Advances in Neural Information Processing Systems 18. Cambridge, MA: MIT Press; 2005. p. 299-306.
19. Sun W, Chen J, Li J. Decision tree and PCA-based fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing* 2007; 21(3): 1300-1317. https://doi.org/10.1016/j.ymssp.2006.06.010.
20. Chen P, Zhao R, He T, Wei K, Yang Q. Unsupervised domain adaptation of bearing fault diagnosis based on join sliced Wasserstein distance. *ISA Transactions* 2022; 129: 504-519. https://doi.org/10.1016/j.isatra.2021.12.037.
21. Zhong J, Lin C, Gao Y, Zhong J, Zhong S. Fault diagnosis of rolling bearings under variable conditions based on unsupervised domain adaptation method. *Mechanical Systems and Signal Processing* 2024; 215: 111430. https://doi.org/10.1016/j.ymssp.2024.111430.
22. Zhong Z, Liu H, Mao W, Xie X, Cui Y. Rolling bearing fault diagnosis across operating conditions based on unsupervised domain

adaptation. *Lubricants* 2023; 11(9): 383. https://doi.org/10.3390/lubricants11090383.

23. Zhang Z, Chen H, Li S, An Z. Unsupervised domain adaptation via enhanced transfer joint matching for bearing fault diagnosis. *Measurement* 2020; 165: 108071. https://doi.org/10.1016/j.measurement.2020.108071.
24. Guo L, Lei Y, Xing S, Yan T, Li N. Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. *IEEE Transactions on Industrial Electronics* 2018; 66(9): 7316-7325. https://doi.org/10.1109/TIE.2018.2877090.
25. Cação J, Santos J, Antune M. Explainable AI for industrial fault diagnosis: A systematic review. *Journal of Industrial Information Integration* 2025; 9: 100905. https://doi.org/10.1016/j.jii.2025.100905.
26. Cao W, Jia X, Hu Q, Zhao J, Wu Y. A literature review on selective maintenance for multi-unit systems. *Quality and Reliability Engineering International* 2018; 34(5): 824-845. https://doi.org/10.1002/qre.2293.
27. Zhou D, Huang D, Tie M, Zhang X, Hao J, Wu Y, Shen Y, Wang Y. A fault diagnosis method based on interpretable machine learning model and decision visualization for HVs. *Applied Intelligence* 2025; 55(4): 270. https://doi.org/10.1007/s10489-024-06219-x.
28. Li S, Li T, Sun C, Yan R, Chen X. Multilayer Grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis. *Journal of Manufacturing Systems* 2023; 69: 20-30. https://doi.org/10.1016/j.jmsy.2023.05.027.
29. Chen G, Tang G, Zhu Z. VKCNN: An interpretable variational kernel convolutional neural network for rolling bearing fault diagnosis. *Advanced Engineering Informatics* 2024; 62: 102705. https://doi.org/10.1016/j.aei.2024.102705.
30. Xiao Y, Shao H, Yan S, Wang J, Peng Y, Liu B. Domain generalization for rotating machinery fault diagnosis: A survey. *Advanced Engineering Informatics* 2025; 64: 103063. https://doi.org/10.1016/j.aei.2024.103063.
31. Ding Y, Jia M, Zhuang J, Cao Y, Zhao X, Lee C. Deep imbalanced domain adaptation for transfer learning fault diagnosis of bearings under multiple working conditions. *Reliability Engineering & System Safety* 2023; 230: 108890. https://doi.org/10.1016/j.ress.2022.108890.
32. Lu B, Zhang Y, Liu Z, Wei H, Sun Q. A novel sample selection approach based universal unsupervised domain adaptation for fault diagnosis of rotating machinery. *Reliability Engineering & System Safety* 2023; 240: 109618. https://doi.org/10.1016/j.ress.2023.109618.
33. Sun B, Feng J, Saenko K. Correlation alignment for unsupervised domain adaptation. In: Csurka G, editor. Domain Adaptation in Computer Vision Applications. Cham: Springer International Publishing; 2017. p. 153-171. https://doi.org/10.1007/978-3-319-58347-1_8.
34. Cai G, Wang Y, He L, Zhou M. Unsupervised domain adaptation with adversarial residual transform networks. *IEEE Transactions on Neural Networks and Learning Systems* 2020; 31(8): 3073-3086. https://doi.org/10.1109/TNNLS.2019.2935384.
35. Wang X, Jiang H, Mu M, Dong Y. A dynamic collaborative adversarial domain adaptation network for unsupervised rotating machinery fault diagnosis. *Reliability Engineering & System Safety* 2025; 255: 110662. https://doi.org/10.1016/j.ress.2024.110662.
36. Xie M, Liu J, Li Y, Yi C. An unsupervised domain adaptation method for intelligent fault diagnosis based on target feature enhancement and feature-boundary alignment. *Journal of Intelligent Manufacturing* 2025. https://doi.org/10.1007/s10845-025-02586-5.
37. Han T, Liu C, Yang W, Jiang D. Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application. *ISA transactions*, 2020, 97: 269-281. https://doi.org/10.1016/j.isatra.2019.08.012.
38. Liu H, Wei C, Sun B. Quantitative evaluation for robustness of intelligent fault diagnosis algorithms based on self-attention mechanism. *Journal of Internet Technology* 2024; 25(6): 921-929. https://doi.org/10.70003/160792642024112506012.
39. Zhao Z, Wu S, Qiao B, Wang S, Chen X. Enhanced sparse period-group lasso for bearing fault diagnosis. *IEEE Transactions on Industrial Electronics* 2019; 66(3): 2143-2153. https://doi.org/10.1109/TIE.2018.2838070.
40. Zhang H, Che W, Cao Y, Guan Z, Zhu C. Condition monitoring and fault diagnosis of rotating machinery towards intelligent manufacturing: Review and prospect. *Iranian Journal of Science and Technology, Transactions of Mechanical Engineering* 2025; 49(1): 1-34. https://doi.org/10.1007/s40997-024-00783-w.
41. Qamar T, Bawany NZ. Understanding the black-box: towards interpretable and reliable deep learning models. *PeerJ Computer Science* 2023; 9: e1629. https://doi.org/10.7717/peerj-cs.1629.
42. Jiao XZ, Zhang JJ, Cao JH. A physics-guided transfer learning framework with consistency verification for cross-domain bearing fault diagnosis. *Eksploatacja i Niezawodność – Maintenance and Reliability* 2026; 28(2): 211797. https://doi.org/10.17531/ein/211797.
43. Raissi M, Perdikaris P, Karniadakis G. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics* 2019; 378: 686-707.

https://doi.org/10.1016/j.jcp.2018.10.045.

44. Alvarez-Melis D, Jaakkola TS. Towards robust interpretability with self-explaining neural networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems 31. Red Hook, NY: Curran Associates; 2018. p. 7775-7784.
45. Yang D, Karimi HR, Gelman L. An explainable intelligence fault diagnosis framework for rotating machinery. *Neurocomputing* 2023; 541: 126257. https://doi.org/10.1016/j.neucom.2023.126257.
46. Brito L, Susto G, Brito J, Duarte M. An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. *Mechanical Systems and Signal Processing* 2022; 163: 108105. https://doi.org/10.1016/j.ymssp.2021.108105.
47. Shao H, Lai Y, Liu H, Wang J, Liu B. LSFConvformer: A lightweight method for mechanical fault diagnosis under small samples and variable speeds with time-frequency fusion. *Mechanical Systems and Signal Processing* 2025; 236: 113016. https://doi.org/10.1016/j.ymssp.2025.113016.
48. Ma R, Chen L, Feng Y, Zhou Z, Xie J. TFD-former: Time-frequency domain fusion decoders for effective and robust fault diagnosis under time-varying speeds. *Knowledge-Based Systems* 2025; 316: 113410. https://doi.org/10.1016/j.knosys.2025.113410.
49. Liu KY, Li YF, Cui ZY, Qi G, Wang B. Adaptive frequency attention-based interpretable Transformer network for few-shot fault diagnosis of rolling bearings. *Reliability Engineering & System Safety* 2025; 263: 111271. https://doi.org/10.1016/j.ress.2025.111271.
50. Zhang Y, Zhao X, Peng Z, Xu R, Chen P. WD-KANTF: An interpretable intelligent fault diagnosis framework for rotating machinery under noise environments and small sample conditions. *Advanced Engineering Informatics* 2025; 66: 103452. https://doi.org/10.1016/j.aei.2025.103452.
51. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2010; 72(4): 417-473. https://doi.org/10.1111/j.1467-9868.2010.00740.x.
52. Case School of Engineering. Download a Data File. https://engineering.case.edu/bearingdatacenter/download-data-file; 2025 [accessed 1 December 2025].
53. Alonso-González M, Díaz VG, Pérez BL, G-Bustelo B, Anzola J. Bearing fault diagnosis with envelope analysis and machine learning approaches using CWRU dataset. *IEEE Access* 2023; 11: 57796-57805. https://doi.org/10.1109/ACCESS.2023.3283466.
54. Zhan Y, Yang R, Zhang Y, Wang Z. Mitigating catastrophic forgetting in cross-domain fault diagnosis: An unsupervised class incremental learning network approach. *IEEE Transactions on Instrumentation and Measurement* 2025; 74: 1-14. https://doi.org/10.1109/TIM.2024.3500047.

Symbol	Notation
x_i	i -th vibration sample
K	Number of fault categories
y_i	Category of sample i
w_k	Weight vector of category k
b_k	Intercept of category k
$\pi_{i,k}$	Probability of sample i belonging to class k
G_g	g -th feature group
λ_1	Sparsity penalty coefficient
λ_2	Group sparsity penalty coefficient
$\hat{\pi}_j$	Selection probability of feature j
τ_{SS}	Stability-selection threshold
f	Frequency of signal
O	Order of signal
$x(t)$	Time domain signal at time t
$\omega(t)$	Angular speed at time t
$\theta(t)$	Cumulative rotation angle at time t

$x_\theta(\theta)$	Angle-domain resampled signal
D_S	Source domain dataset
D_T	Target domain dataset
n_S,n_T	Number of source and target samples
C_S,C_T	Covariance matrices of source and target features
L_{CORAL}	CORAL loss
L_{CLASS}	Classification loss
L_{TOTAL}	Total loss
λ	The factor used to weigh classification loss and CORAL loss
S_{PL}	Pseudo-labeled target sample set
$\hat{y}^T$	Pseudo label of target sample
$H(x_t)$	Prediction entropy for target sample x_t