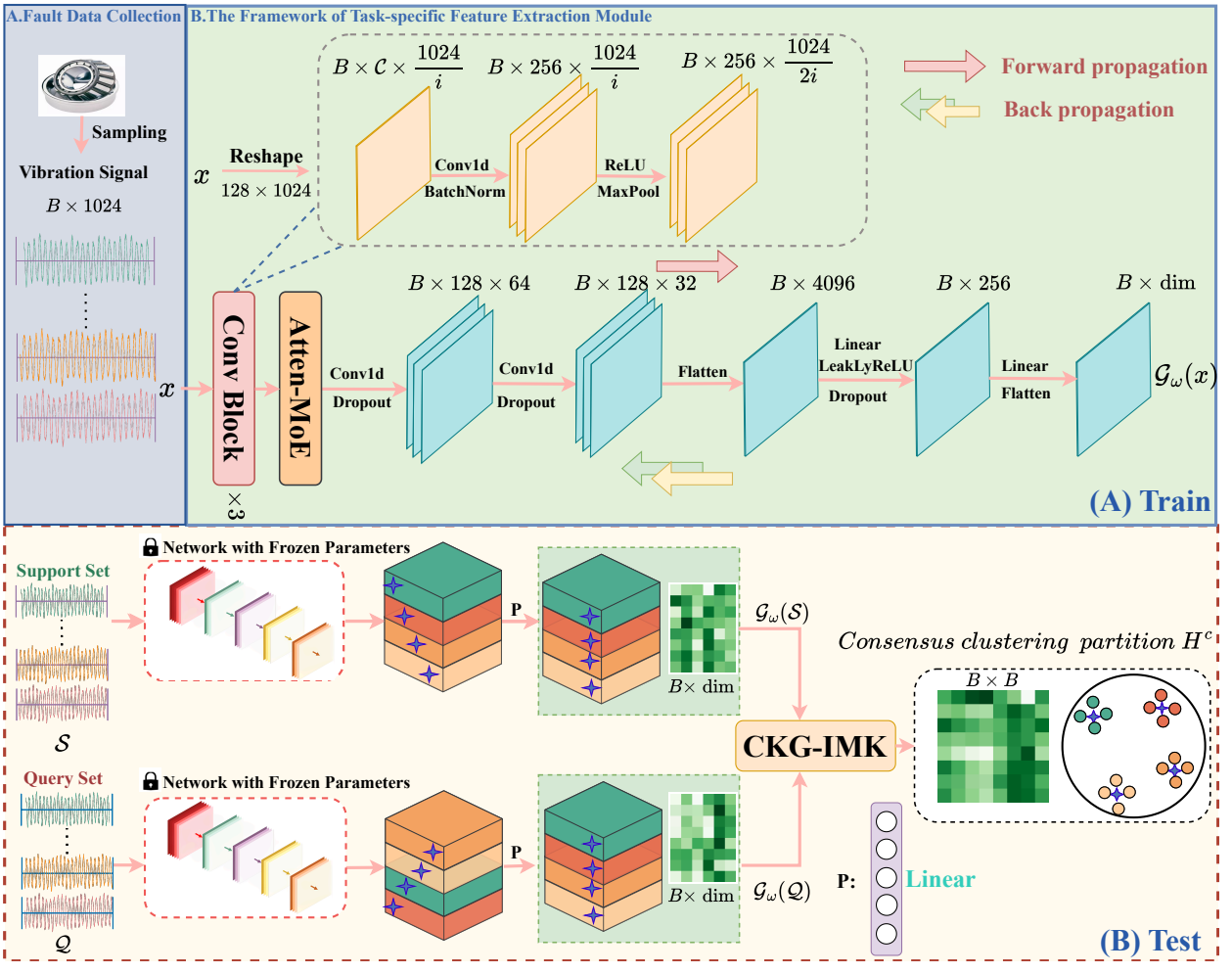
Graphical Abstract

Category knowledge-guided few-shot bearing fault diagnosis

Lingkai Hu, Feng Zhan, Wenkai Huang, Yikai Dong, Hao He, Guanjun Wu



Highlights

Category knowledge-guided few-shot bearing fault diagnosis

Lingkai Hu, Feng Zhan, Wenkai Huang, Yikai Dong, Hao He, Guanjun Wu

- Analyzing and accurately locating pertinent features for novel tasks.
- The category knowledge-guided (CKG) framework is introduced in this study.
- Tackling cold-start issues in early fault diagnosis of rotating machinery.
- Attaining exceptional clustering performance with high sensitivity for unknown samples.

Category knowledge-guided few-shot bearing fault diagnosis

Lingkai Hu[a,*], Feng Zhan[a,*], Wenkai Huang[a,**], Yikai Dong[a], Hao He[b], Guanjun Wu[b]

[a]School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou, 510006, China
[b]School of Politics and International Relations, East China Normal University, Shanghai, 200062, China

ARTICLE INFO

Keywords:
Bearing fault
Knowledge-guide
Few-shot learning
Early-stage fault diagnosis

ABSTRACT

Real-time bearing fault diagnosis plays a vital role in maintaining the safety and reliability of sophisticated industrial systems. However, the scarcity of labeled data in fault diagnosis, due to the difficulty of collecting fault samples and the high cost of labeling, poses a significant challenge in learning discriminative fault features from limited and complex monitoring signals. Few-shot learning (FSL) emerges as a potent method for extracting and accurately classifying features from severe fault signals. Nonetheless, challenges such as data scarcity and environmental noise significantly impede the efficacy of existing FSL methods in diagnosing incipient faults effectively. These limitations are primarily due to the inadequate consideration of inter-class correlations within noisy contexts by current FSL strategies, which restricts their ability to extrapolate familiar features to new classes. Consequently, there is a pressing demand for an FSL approach that can exploit inter-class correlations to address the hurdles of data insufficiency and environmental complexities, thereby facilitating the diagnosis of incipient faults in few-shot settings. This paper proposes a novel category-knowledge-guided model tailored for few-shot multi-task scenarios. By leveraging attribute data from base categories and the similarities across new class samples, our model efficiently establishes mapping relations for unencountered tasks, significantly enhancing its generalization capabilities for early-stage fault diagnosis and multi-task applications. This model ensures swift and precise FSL fault diagnosis under uncharted operational conditions. Comparative analyses utilizing the Case Western Reserve University bearing dataset and the Early Mild Fault Traction Motor bearing dataset demonstrate our model's superior performance against leading FSL and transfer learning approaches.

1. Introduction

Incidents of mechanical equipment failure, precipitated by progressive structural damage, malfunctions, and eventual loss of functionality, frequently result in significant human casualties and economic losses [1]. To mitigate these incidents, a variety of intelligent fault diagnosis techniques has been developed, including deep autoencoders (DAE) [2], convolutional neural networks (CNN) [3], recurrent neural networks (RNN), long short-term memory networks (LSTM) [4], generative adversarial networks (GAN) [5], and graph neural networks (GNN) [6]. Despite their advancements, these methods often require extensive manual parameter tuning and significant computational resources. With the industrial demand evolving from diagnosing basic components to encompassing large-scale machinery, the fault diagnosis landscape faces complex challenges: data scarcity, diverse operating conditions, poor training data quality, model efficiency, and the necessity for early fault diagnosis [7, 8, 9]. These complexities demand an approach that not only simplifies hyperparameter optimization and operates effectively with limited data but also maintains high sensitivity and robust generalization for prompt and accurate fault diagnosis. Few-shot learning (FSL) offers a viable path forward, addressing these critical needs [10].

FSL [11, 12] is a widely recognized fault diagnosis approach, enabling rapid adaptation to new tasks with minimal data. This approach alleviates the dependency on large datasets and expert insight, presenting a logical and promising direction for bearing fault diagnosis research [13, 14]. FSL, as an active fault diagnosis method, leverages feature vectors from a support set as auxiliary input to diagnostic models [15]. This approach enhances the diagnosis of specific fault manifestations, particularly when fault characteristics are subtle or system noise is substantial. Compared to passive fault diagnosis methods, FSL can perform fault diagnosis more effectively under these challenging conditions [16]. Nonetheless, FSL is not without its challenges. Data scarcity can precipitate model overfitting, undermining diagnostic precision [17, 18]. Furthermore, the inherent reliance on limited datasets may compromise the generalization

*Co-first author.
**Corresponding author.
E-mail address: smallkat@gzhu.edu.cn (Wenkai Huang)

ability of models, particularly in recognizing novel categories [19]. Compounding these issues, variable operational conditions (such as fluctuating loads and speeds) alter vibration signal distributions, creating disparities between source and target domains that can impede the effectiveness of pretrained models [20, 21]. In industrial settings, where operational conditions and environmental noise vary widely, these variations can cause identical fault characteristics to appear differently, challenging the direct application of pretrained models to cross-domain diagnostics [22]. FSL's core advantage, leveraging existing knowledge under conditions of data scarcity to classify new categories, underscores the need for innovative solutions to enhance its application in fault diagnosis.

Recent literature highlights that existing FSL approaches in fault diagnosis predominantly address permanent or severe faults, overlooking the intricacies of early-stage fault identification [9, 23, 24, 25]. A notable challenge arises when target domain fault labels diverge from those in the source domain, complicating knowledge transfer due to label inconsistency [26]. Moreover, optimization-based methods often fail to accurately capture category-specific attribute information, hindering the precise identification of new task-relevant features within the source task. This limitation significantly affects the diagnosis of early faults, which typically precede more severe bearing failures [27, 28]. The critical nature of early fault diagnosis for timely maintenance or replacement is thus underscored [29, 30]. However, the low magnitude of early faults renders them vulnerable to masking by prevalent background noise. Conventional FSL techniques have not sufficiently addressed the conveyance of class attribute information, further complicating the diagnosis of early faults [31, 32]. Additionally, the minimal differences between normal and early fault features challenge the effective extraction of diagnostic features for early-stage faults, presenting substantial hurdles in early fault diagnosis for rolling bearings.

The greatest challenges in diagnosing early bearing faults are weak vibration pulses, measurement noise, and interference from other vibration sources [33, 34, 35, 36]. Existing studies typically apply traditional signal-denoising methods to remove noise before using diagnostic algorithms to predict the type of failure [37]. For instance, Ahmadpour et al. [38] combined 3D finite element modeling with the Hilbert-Huang transform to address the fault diagnosis problem in industrial motors. Wang et al. [39] designed an on-rotor sensing system to enhance the signal-to-noise ratio in early fault vibration measurements. Additionally, researchers [40] introduced a signal processing method based on a deep binary mask signal separation model to extract features for early-stage fault diagnosis under variable speed conditions. These methods are primarily applied to stationary data, making it challenging to learn fault-sensitive features due to the weak nature of early faults [41]. This, in turn, affects the quality of fault isolation in the model to some extent [42]. In particular, the patterns exhibited by individual samples of early-stage faults are often inadequate for capturing key fault characteristics.

To analyze and precisely locate salient features relevant to new tasks [43, 44, 45] as well as to explicitly leverage inter-category relationships [46], encompassing commonalities for facilitating generalization across related categories [47] and uniqueness to reduce misclassification among similar categories, this paper introduces a novel category knowledge-guided (CKG) framework for cold-start cross-domain rotating machinery fault diagnosis and early fault diagnosis. Specifically, this research adopts an incomplete multiple kernel clustering matrix to quantify the inter-category correlations, thus generating a consensus clustering matrix to guide the joint learning for more robust tasks. This capability allows CKG to simultaneously handle multiple support categories, unlike the majority of existing methods, which require repeated meta-learning for each category. Further, the proposed framework is extended to more challenging tasks in the realm of FSL, including cold-start scenarios and early fault diagnosis, and its outstanding performance in data-scarce and complex environments is demonstrated. Finally, a theoretical analysis of the category knowledge-guided incomplete multiple kernel clustering (CKG-IMK) algorithm is provided, and comprehensive experiments and analyses are conducted to validate the stability and exceptional clustering performance of the proposed method. The primary contributions of this study are as follows:

1. A novel multiclass-based feature learning model is proposed for the analysis and localization of task-relevant features.
2. The introduced incomplete multiple kernel clustering method can effectively leverage the relative similarities among different categories in multi-task scenarios, achieving high sensitivity clustering performance for unknown samples. Additionally, this research theoretically investigates the effectiveness of the proposed CKG framework in terms of clustering generalization error.
3. Dependency on data volume and label quality is reduced, thus reducing the cost of data collection and processing while significantly enhancing the sensitivity and reliability of early diagnosis. Further, the model's robustness and interpretability are improved, enabling it to effectively address various challenges, including noise, anomalies, and uncertainties.

2. Theoretical foundation

Guided by category knowledge, this paper primarily explores the theories of FSL and incomplete multiple kernel clustering [48]. In this section, the notations for the multiple kernel clustering algorithm, along with the abbreviations used in this article, are introduced, followed by a brief overview of FSL methods [49] and the theoretical research on multiple kernel k-means [50].

2.1. Notations and abbreviations

The notations for the multiple kernel clustering algorithm, along with the abbreviations used in this article, are listed and described in the Appendix.

2.2. Few-shot learning

FSL [51] is a machine learning approach that leverages prior knowledge from multiple related tasks to enhance performance on new target tasks, with the support set denoted as $\mathcal{S}$ and the query set denoted as $\mathcal{Q}$. This approach allows the training of robust models that can classify a small sample of annotated data, recognize new classes, and improve the model's generalization and portability. Achieving this objective necessitates the effective utilization of prior knowledge and limited data to mitigate model bias and variance while avoiding overfitting and underfitting. Simultaneously, considerations must be made for factors like data quality, distribution, noise, and heterogeneity, as well as the complexity, diversity, and dynamics of tasks in meta-test scenarios.

In fault classification tasks, faults occurring in the same equipment typically exhibit a certain degree of similarity, and faults in the same category of equipment tend to share similar features [52]. Metric-based FSL methods capitalize on this characteristic by learning methods to represent the similarity between support and query samples in an embedding space to identify unknown samples. Specifically, this approach effectively measures the similarity between support and query samples in the embedding space. By learning similarity metrics in this embedding space, the model can accurately classify unknown samples into support categories that are similar to them [53]. This demonstrates the method's ability to handle various operating conditions or fault categories in practical applications, showcasing its feasibility in fault classification tasks.

Metric-based approaches utilize certain distance or similarity measures to compare samples in the test set with those in the support set (a limited amount of labeled data). However, metric-based FSL methods suffer from a drawback in that features unrelated to the classification task might mislead the model. Additionally, due to the limited number of samples in the support set, they often fail to identify the target features relevant to the task [54]. A model that could extract fault feature information more comprehensively from the data and explicitly learn significant features related to the task would be more reliable, reducing attention to task-irrelevant information and thereby enhancing the performance of fault classification tasks [55].

2.3. Multiple kernel k-means (MKKM)

As a non-linear technique, multiple kernel methods [56] can handle linearly inseparable data and achieve satisfactory clustering results in high-dimensional spaces [57]. However, in the context of multi-task settings, a single kernel function may not effectively handle heterogeneous data. To address this, the concept of multiple kernel subspace clustering (MKSC) [58] has been introduced. The principle behind MKSC involves extracting more information from the data using various kernel functions, thereby enhancing clustering performance. Given a set of observed data $\{x_i\}_{i=1}^n$ and the kernel mapping $\mathcal{F}(\cdot)$, the objective of MKKM is to partition the samples into k clusters by minimizing the sum of squares loss. This objective can be expressed as follows:

$$\begin{aligned} &\min_{\mathcal{W},\widehat{\mathcal{C}}} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{W}_{ij}\|\mathcal{F}(x_i)-\widehat{\mathcal{C}}_j\|_F^2 \\ &s.t.\ \sum_{j=1}^{m}\mathcal{W}_{ij}=1, \end{aligned} \tag{1}$$

where $\mathcal{W} \in \{0,1\}^{n\times k}$ represents the clustering assignments for each sample, and $\widehat{\mathcal{C}}_j$ denotes the centroid of the j-th cluster. In most cases, $\mathcal{F}(x_i) \in \mathbb{R}^d$, where $d \gg n$ or even infinite. Therefore, Eq.(1) cannot be directly optimized. Consequently, it is equivalently rewritten in matrix-vector form as follows:

$$\min_{\mathcal{W}} \mathrm{Tr}(\mathcal{K}_\xi) - \mathrm{Tr}(\xi^{1/2}\mathcal{W}^T\mathcal{K}_\xi, \mathcal{W}\xi^{1/2}), \tag{2}$$

where, $\mathcal{K}_\xi^{ij}(x_i, x_j) = \mathcal{F}(x_i)^T\mathcal{F}(x_j)$, $\xi = diag([n_1^{-1}, n_2^{-1}, \cdots, n_k^{-1}])$,$n_j = \sum_{i=1}^{n} \mathcal{W}_{ij}$, $\mathrm{Tr}(\cdot)$ denotes the trace norm, and ξ_j represents the fundamental kernel for the j-th weight. Discrete $\mathcal{W}$ makes Eq.(2) challenging to solve, and a common technique is to relax it, allowing for arbitrary values. MKKM can simultaneously learn ξ and the clustering assignment matrix H^C by defining $\mathrm{H}^C = \mathcal{W}\xi^{-1}$. The aforementioned problem can thus be transformed as follows:

$$\min_{\xi, \mathrm{H}^C} \mathrm{Tr}(\mathcal{K}_\xi(\mathrm{I}_n - \mathrm{H}^C(\mathrm{H}^C)^T))$$

$$s.t.\ \mathrm{H}^C \in \mathbb{R}^{n\times k}, (\mathrm{H}^C)^T\mathrm{H}^C = \mathrm{I}_k, \|\xi\| \geq 0. \tag{3}$$

Existing algorithms typically solve Eq.(3) through alternating optimization of H^C and ξ: (i) Fixing ξ to optimize H^C. For a specific kernel coefficient ξ, optimizing H^C in Eq.(3) is equivalent to the following Eq.(4):

$$\min_{\xi, \mathrm{H}^C} \mathrm{Tr}(\mathcal{K}_\xi(\mathrm{I}_n - \mathrm{H}^C(\mathrm{H}^C)^T))$$

$$s.t.\ \mathrm{H}^C \in \mathbb{R}^{n\times k}, (\mathrm{H}^C)^T\mathrm{H}^C = \mathrm{I}_k, \|\xi\| \geq 0. \tag{4}$$

Eq.(4) is a classic kernel k-means equation that can be easily optimized. An optimized kernel matrix $\mathcal{K}_\xi$ is parameterized in the following form: $\mathcal{K}_\xi = \sum_{j=1}^{k} \xi_j^2\mathcal{K}_\xi^j$, where $\{\mathcal{K}_\xi^j\}_{j=1}^k$ represents a set of precomputed kernel matrices. (ii) Fix H^C to optimize ξ. For a specific H^C, the optimization of ξ in Eq.(4) simplifies to the following:

$$\min_{\xi} \sum_{j=1}^{m} \xi_j^2 \mathrm{Tr}(\mathcal{K}_\xi(\mathrm{I}_n - \mathrm{H}^C(\mathrm{H}^C)^T))$$

$$s.t.\ \mathrm{H}^C \in \mathbb{R}^{n\times k}, (\mathrm{H}^C)^T\mathrm{H}^C = \mathrm{I}_k, \|\xi\| \geq 0. \tag{5}$$

Algorithm 1 presents a detailed optimization procedure for MKKM, where H^C and ξ are alternately optimized until convergence.

Algorithm 1 Multiple kernel k-means

Require: $\{\mathcal{K}_\xi^j\}_{j=1}^m, k, t = 1$
  Initialization $\xi = 1/\sqrt{m}$
  repeat
    Compute $(\mathrm{H}^C)^t$ in Eq. (2) with $\mathcal{K}_{\xi^t} = \sum_{j=1}^{m}(\xi_j^t)^2\mathcal{K}_\xi^j$
    Update ξ^t and $(\mathrm{H}^C)^t$ in Eq.(3)
    $t \leftarrow t + 1$
  until $|\xi^{(t+1)} - \xi^t| \leq e^{-4}$

3. Methodology

In this section, a category knowledge-guided incomplete multiple kernel clustering few-shot learning method is introduced, which is applicable to the problem of few-shot bearing fault diagnosis under various limited data conditions. This approach has demonstrated high sensitivity and accuracy in fault diagnosis performance on the Case Western Reserve University bearing dataset (CWRU) [59] for permanent and severe faults as well as on the Early Mild Fault Traction Motor bearing dataset (EMF-TM).

To ensure the continuity of bearing signal sequences, reduce computational time and space complexity, and maintain robustness and model fitting capability, a novel Attention Mixture of Experts (Atten-MoE) module is proposed. The Experts, functioning as lightweight subnetworks in the style of MoE modules, integrate a Localized Balancing Constraint during training, alleviating the burden of global knowledge distribution.

3.1. Problem definition

In contrast to traditional machine learning approaches, in FSL, training samples are treated as tasks or events rather than mere data instances. In the context of few-shot classification problems, this is typically formalized as an N-way K-shot classification problem. In this problem, the model is required to acquire K labeled fault samples from N different categories and accurately classify unlabeled faults [49, 60].

Specifically, given a dataset $D = \{(x_i, y_i), y_i \in \mathcal{L}\}_{i=1}^{I}$, D is divided into the meta-training set $D^{train} = \{(x_i, y_i), y_i \in \mathcal{L}^{train}\}_{i=1}^{I^{train}}$ and the meta-test set $D^{test} = \{(\widetilde{x}_i, \widetilde{y}_i), \widetilde{y}_i \in L^{test}\}_{i=1}^{I^{test}}$, where (x_i, y_i) represents the original features and label information of the i-th bearing sample in the meta-training set, and $D^{train} \cup D^{test} = D$, $L^{train} \cup L^{test} = L$. There is no intersection between the meta-training set and the meta-test set ($D^{train} \cap D^{test} = \emptyset$).

The FSL algorithm requires learning general meta-knowledge from multiple training sets to acquire new tasks. In accordance with the prior literature, we consider a meta-training set, denoted as T, comprising tasks $\mathfrak{T} = \{\mathfrak{T}^1, \mathfrak{T}^2, \cdots, \mathfrak{T}^T\}$. To construct each task $\mathfrak{T}^j$, we randomly select N categories from D^{train}, with each category containing M samples. Within each selected category, the M samples are further divided into two sets, each containing K and $M - K$ bearing fault samples, respectively. These sets are referred to as the support set $S^j = \{(x_i^j, y_i^j), y_i^j \in L^{train}\}_{i=1}^{N\times K}$ and the query set $Q^j = \{(\widehat{x_i^j}, \widehat{y_i^j}), \widehat{y_i^j} \in L^{train}\}_{i=1}^{N\times(M-k)}$. Similarly, D^{test} is divided into the labeled support set $S^\zeta = \{(x_i^\zeta, y_i^\zeta), y_i^\zeta \in L^{test}\}_{i=1}^{N\times K}$ and the unlabeled query set $Q^\zeta = (\widehat{x_j^\zeta})_{j=1}^{N\times(M-k)}$, with no intersection between these datasets($S^\zeta \cap Q^\zeta = \emptyset$).

During the training phase of FSL, the model is initially trained using the meta-training set, and the performance of the meta-trained model is evaluated using the meta-test set. The advantage of this task training strategy is that it enables the model to demonstrate good generalization performance on entirely new class samples. Consequently, our model can better adapt to various tasks and data, thereby enhancing its applicability and performance in practical applications.

3.2. Data preprocessing

To strike a balance between training efficiency and accuracy while showcasing the excellent performance of the FSL network and its adaptability to real industrial production environments, this study applied a random segmentation operation to the original signal sequences, dividing them into multiple data segments, each containing 1024 sampling points. There were only 500 samples per class. Data augmentation techniques were not employed in this research to increase the dataset size, as doing so might introduce signal redundancy between different data segments and consequently affect the reliability of experimental results [3].

It is worth noting that, before feeding the data into the model, data standardization was conducted. Specifically, a global standardization method, known as the z-score, was used to eliminate scale differences in the data, thereby aligning the bearing signal sequences with a standard Gaussian distribution.

3.3. Model architecture of category knowledge-guided

The proposed model consists of three modules, including a task-related feature extraction module, a category knowledge-guided incomplete multiple kernel clustering(CKG-IMK) algorithm, and the algorithm's extension.

A novel task-related feature extraction model has been proposed in this study (Fig. 1), aimed at extracting complex and salient fault-related features while addressing the issue of overfitting. This objective is accomplished through strategic enhancements to the model architecture and optimizations within the network layers, which increase the model's depth while simultaneously reducing the overall number of parameters. Fig. 1 illustrates the workflow of the proposed CKG framework.

The model, which is shown in Fig. 1, is employed for the processing of preprocessed fault signals, denoted as $x_i \in \mathbb{R}^{C\times 1024}$, where C represents the channels of input features. The feature calibration process begins with the application of a basic 1×1 Convolutional Block feature extractor. The design of this block is aimed at efficiently capturing complex patterns within the preprocessed fault signals while preserving crucial features associated with local faults. The Batch normalization modules are integrated to mitigate gradient vanishing or explosion issues, concurrently enhancing model robustness and generalization capabilities. Moreover, the ReLU activation function ensures the introduction of crucial non-linearity, which is essential for capturing fault-related patterns effectively. To maintain critical information while reducing spatial dimensions, Max-pooling is strategically employed, and dropout layers with a 0.5 dropout rate are strategically placed to mitigate overfitting by discouraging reliance on specific neurons.

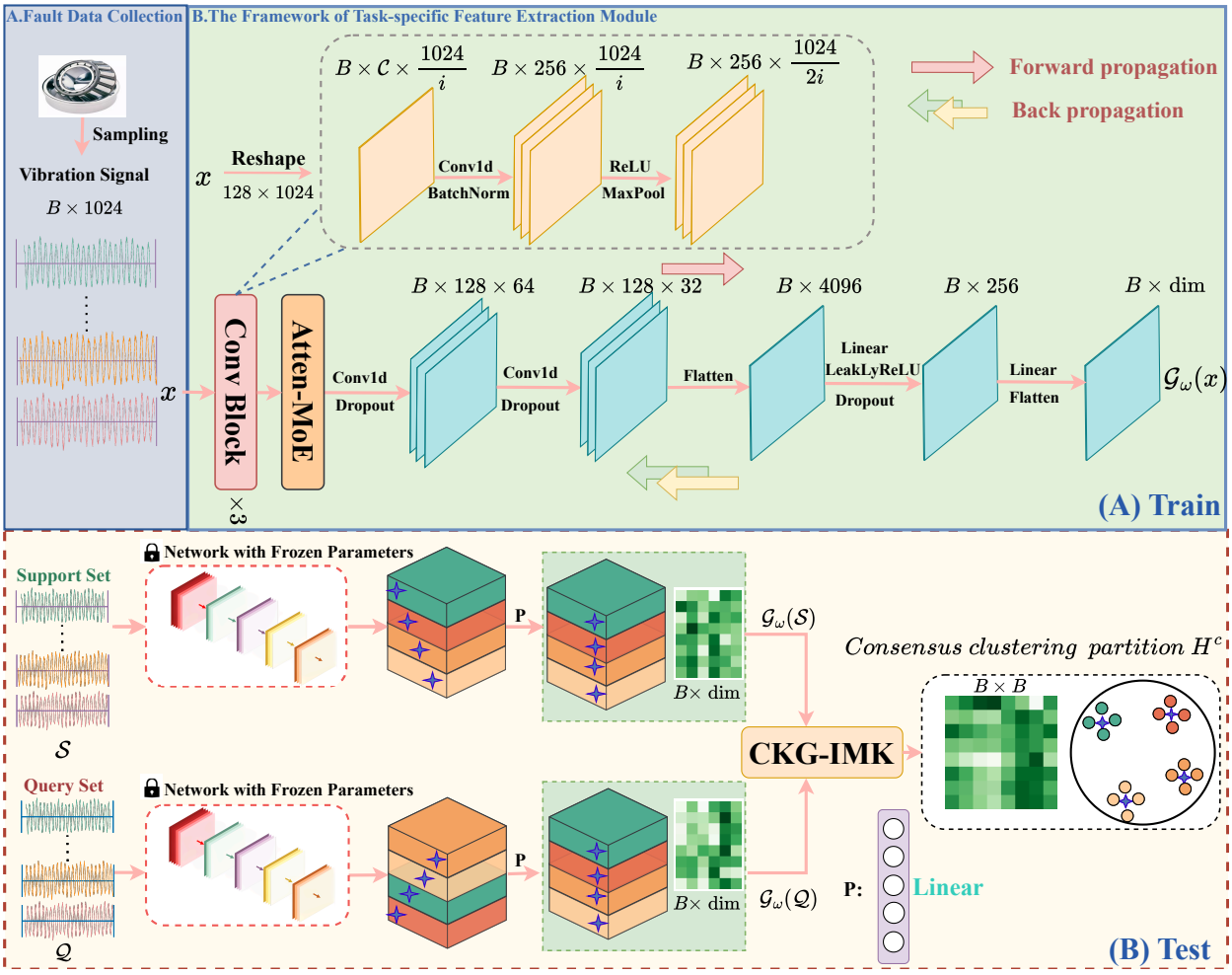


Fig. 1. Structure diagram of the CKG framework.

Subsequently, the Atten-MoE facilitates the learning of local features in multiclassification tasks and addresses imbalances in small-sample categories. A series of 1-D convolutional layers follow the Atten-MoE layer to extract in-depth features. Flattened feature maps from convolutional layers are then fed into fully connected layers, where features are abstracted and refined. The integration of Leaky-ReLU activation further enhances convergence speed. Afterward, the linear layer maps feature to output classes for fault diagnosis, effectively capturing both local and higher-level features. This approach achieves a balance between depth, parameter efficiency, and feature representation, thereby enhancing fault diagnosis performance significantly.

3.3.1. Attention Mixture of Experts

Merely employing the learned task embedding to encapsulate task-specific information tends to bias the task-shared experts towards overfitting the training data distribution. Experts, acting as lightweight subnetworks akin to an MoE-style plugin, introduce a Localized Balancing Constraint during training to mitigate the burden of global knowledge. In the realm of experts, opting for Dense-MoE [61] would notably escalate computational costs, as each input token is processed by every expert rather than a single one. Conversely, the use of Soft-MoE [62] would interrupt the sequence's continuity, requiring iterative computations and leading to decreased computational efficiency.

For the sake of maintaining the continuity of information in the sequence, achieving lower time and space complexity, while also preserving robustness and fitting ability, a novel Attention Mixture of Experts module is proposed, as illustrated in Fig. 2. This approach would cater to the precise computation of gradients and end-to-end learning. Given that a sequence $X \in \mathbb{R}^{L\times D}$ is the input to Atten-MoE. Atten-MoE initially normalizes and calculates,

for each input token, using projection matrices W^A to obtain slots. These slots are then passed through routers to generate attention scores A, which are subsequently used to perform an element-wise product with the output of the experts, denoted as V_i. Then we compute the output O_i by leveraging the attention scores A and the value matrix V:

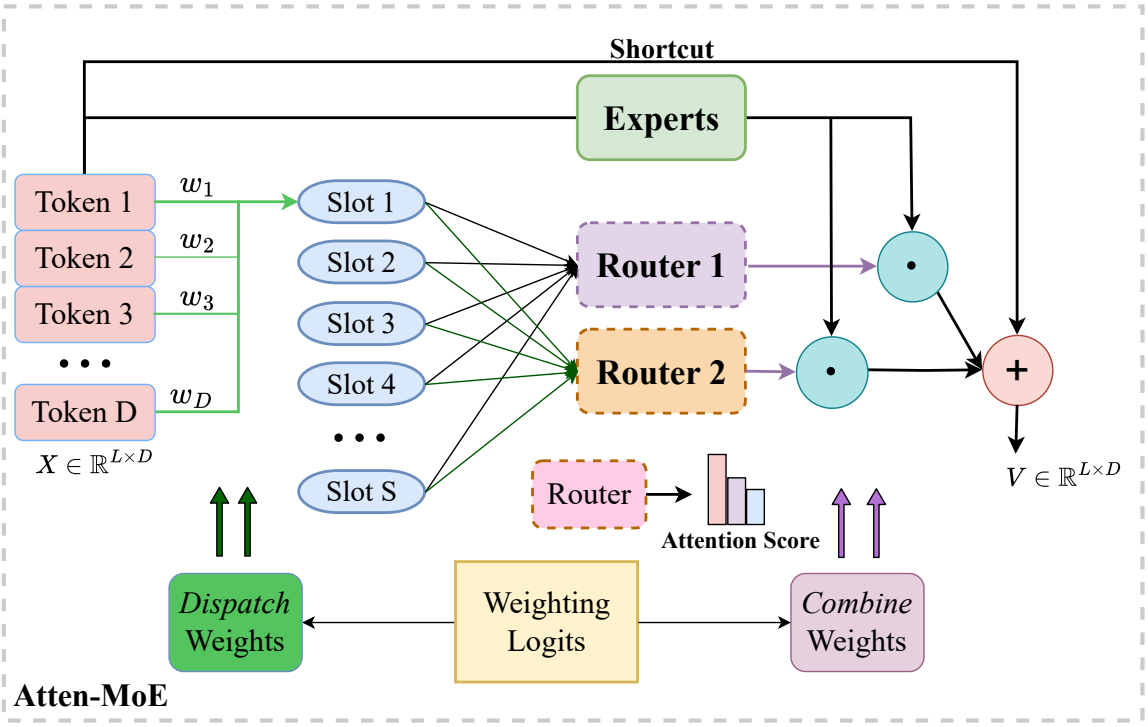


Fig. 2. The Atten-MoE routing algorithm.

$$A_i = R(W^A X) = Exp(W^A X_i) \tag{6}$$

$$V = E(W^V X) \tag{7}$$

$$O_i = \frac{\sum_{j=0}^{N-1} A_{i-j} \odot V_{i-j}}{\sum_{j=0}^{N-1} A_{i-j}} \in \mathbb{R}^{L\times D}. \tag{8}$$

The weights, $W^A A \in \mathbb{R}^{D\times D}$, $W^V \in \mathbb{R}^{D\times D}$, are learnable matrices, where N denotes the number of experts. The router, denoted as $R(\cdot)$ is a dense, fully-connected layer followed by the Softmax function. In Eq.(6), Attention A_i reflects the score of each token, denoted as $A_1, A_2, \cdots, A_i$, are multiplied with the experts' values V at each time step to calculate the corresponding output value O_i, weighted sum operation of some consecutive tokens in V. The symbol '$\odot$' represents the element-wise product between vectors (or matrices), which enhances the efficiency of computations. This approach allows for the development of diverse capabilities and efficient handling of different types of tasks.

Remark 1. The proposed feature extraction module strikes a balance between module depth, parameter efficiency, and feature representation. The combined application of convolutional layers, batch normalization, ReLU activation, Atten-MoE, and dropout regularization enables the module to capture complex fault-related patterns while preventing overfitting. This innovative module exhibits significant potential for enhancing the field of fault diagnosis.

3.3.2. Category knowledge-guided incomplete multiple kernel clustering algorithm

To jointly optimize the optimal kernel, maximum-margin hyperplane, and optimal clustering labels, a CKG-IMK algorithm is proposed to construct a consensus partition. The detailed specifics of this algorithm are presented below.

Assuming that the meta-training support set$\{(x_i, y_i), y_i \in L^{train}\}_{i=1}^{n}$ comprises a collection of n samples, and $\{\mathcal{G}(\cdot)\}_{o=1}^{m} : x_i \in X_S^{test} \Rrightarrow \{\mathcal{H}\}_{o=1}^{m}$ represents an encoder that maps different inputs x_i to Hilbert spaces $\{\mathcal{H}\}_{o=1}^{m}$,

yielding m multiple kernel observations $[(z_i)_1, (z_i)_2, \cdots, (z_i)_m]_{i=1}^n \in \mathbb{R}^n$, where $(z_i)_\varrho$ denotes the ϱ-th base kernel for the i-th sample. These m base kernels are obtained through the encoder $\{\mathcal{G}(\cdot)\}_{\varrho=1}^m$. These concepts can be precisely formulated mathematically as follows:

$$\{\mathcal{G}(x_i)\}_{\varrho=1}^m = [\xi_1(z_i)_1, \xi_2(z_i)_2, \cdots, \xi_m(z_i)_m], \tag{9}$$

where $\xi = [\xi_1, \xi_2, \cdots, \xi_m]$ represents a matrix containing normalization coefficients. These coefficients are adaptively optimized during the training process. ξ is capable of normalizing the multiple kernel observations $(z_i)_\varrho$, enabling local kernel alignment.

Hence, it can be assumed that the m basic kernels $(z_i)_\varrho$ share a latent proxy $h_i \in \mathbb{R}^k$ to represent each kernel sample z_i in the latent embedding space. Specifically, the m kernel samples $\{(z_i)_\varrho\}_{\varrho=1}^m$ can be represented through the latent proxies h_j and the corresponding mapping matrices $\{\ddot{\mathcal{P}}_\varrho\}_{\varrho=1}^m \in \mathbb{R}^{n\times k}$. These concepts can be precisely formulated mathematically as follows:

$$\begin{aligned} &\min_{\ddot{\mathcal{P}}_\varrho, h_i, \xi} \sum_{i=1}^n \sum_{\varrho=1}^m \|\xi_\varrho(z_i)_\varrho \ddot{\mathcal{P}}_\varrho - h_i\|_F^2 \\ &s.t.\ \|\xi\|_\varrho \geq 0, \ddot{\mathcal{P}}_\varrho^T \ddot{\mathcal{P}}_\varrho = \mathrm{I}_k. \end{aligned} \tag{10}$$

Based on the definition of the base kernel $(z_i)_\varrho$ and the fact that the samples x_i can be transformed into $\{\mathcal{G}(x_i)\}_{\varrho=1}^m$ through the encoder $\{\mathcal{G}(\cdot)\}_{\varrho=1}^m$, the kernel function can be expressed as follows:

$$\kappa_{\mathcal{G}}(x_i, x_i) = \{\mathcal{G}(x_i)\}_{\varrho=1}^m (\{\mathcal{G}(x_i)\}_{\varrho=1}^m)^{\mathrm{T}} = \sum_{\varrho=1}^m \xi_\varrho^2 \kappa_\varrho((z_i)_\varrho, (z_i)_\varrho). \tag{11}$$

Subsequently, the kernel matrix $\mathcal{K}_{\mathcal{G}}^i$ is computed by employing the defined kernel function $\kappa_{\mathcal{G}}(x_i, x_i)$. $\mathcal{K}_{\mathcal{G}}^i$ not only ensures the existence of potential partitions within a low-rank space but also enables the integration of complementary information among multiple base kernels, resulting in a consensus clustering partition, denoted as H^C. The aforementioned concept can be realized as follows:

$$\begin{aligned} &\min_{\ddot{\mathcal{P}}_\varrho, \mathrm{H}^C, \xi} \sum_{\varrho=1}^m \|\xi_\varrho(\mathcal{K}_{\mathcal{G}}^i)\ddot{\mathcal{P}}_\varrho - \mathrm{H}^C\|_F^2 \\ &s.t.\ \|\xi\|_\varrho \geq 0, \ddot{\mathcal{P}}_\varrho^T \ddot{\mathcal{P}}_\varrho = \mathrm{I}_k. \end{aligned} \tag{12}$$

By solving Eq.(12), one can infer a latent consensus partition, denoted as H^C, which fundamentally characterizes the data and uncovers the underlying structures shared by different kernels. Initially, a consensus partition matrix H^C is derived from the feature vectors $\{\mathrm{H}_\varrho\}_{\varrho=1}^m$, and subsequently, the partially overlapping partitions are computed using the learned consensus matrix H^C. In this manner, these two learning processes can seamlessly intertwine, allowing them to mutually negotiate and achieve enhanced clustering. The aforementioned concept can be implemented as follows:

$$\begin{aligned} &\max_{\mathrm{H}^C, \{\mathrm{H}_\varrho, \ddot{\mathcal{O}}_\varrho\}_\varrho^m} \mathrm{Tr}[(\mathrm{H}^C)^{\mathrm{T}}(\sum_{\varrho=1}^m \mathrm{H}_\varrho \ddot{\mathcal{O}}_\varrho)] \\ &s.t.\ \mathrm{H}^C \in \mathbb{R}^{n\times k}, (\mathrm{H}^C)\mathrm{H}^C = \mathrm{I}_k, \\ &\ddot{\mathcal{O}}_\varrho \in \mathbb{R}^{k\times k}, \ddot{\mathcal{O}}_\varrho^{\mathrm{T}} \ddot{\mathcal{O}}_\varrho = \mathrm{I}_k, \\ &\mathrm{H}_\varrho \in \mathbb{R}^{k\times k}, \mathrm{H}_\varrho^{\mathrm{T}} \mathrm{H}_\varrho = \mathrm{I}_k, \end{aligned} \tag{13}$$

where H^C and H_ϱ represent the consensus clustering matrix and the ϱ-th base clustering matrix, respectively, with k denoting the number of clusters and $\ddot{\mathcal{O}}_\varrho$ denoting the transposition matrix for the ϱ-th base, which aids in better alignment between H^C and H_ϱ. This necessitates the imputation of all incomplete elements and the deliberate

decomposition of the entire inferred similarity to facilitate clustering. This enhances the model's robustness throughout the optimization process. Ultimately, m partially incomplete base kernels $\{(z_i)_\varrho\}_{\varrho=1}^m$ are obtained, along with the clustering indicator matrix H^C. Let $\widehat{C} = [\widehat{C}_1, \widehat{C}_2, \cdots, \widehat{C}_k]$, where $\widehat{C}_k$ represents the centroids of each cluster, to reduce redundancy and enhance the diversity of the selected base kernels. Finally, k-means is employed to minimize the reconstruction loss:

$$\mathbb{E}[\min_{y\in\{e_1,e_2,\cdots,e_k\}} \|\{\mathcal{G}(x_i)\}_{\varrho=1}^m - \widehat{C}_y\|_F^2], \tag{14}$$

where $\{e_1, e_2, \cdots, e_k\}$ form the orthogonal bases of $\mathbb{R}^k$.

During the query phase of the meta-testing, a query set is constructed using samples $\{(x_i^S, y_i^S), x_i^S \in \mathcal{X}_S^{test}, y_i^S \in \mathcal{L}^{test}\}_{i=1}^n$ from k classes, where $x_i^S \in \mathcal{X}_S^{test}$ represents the labeled data. Additionally, unlabeled data $\{(x_i^Q), x_i^Q \in \mathcal{X}_Q^{test}\}_{i=1}^{N\times K}$ are employed as samples within the query set. Consistent with the previous approach, the kernel function is computed using unobserved elements $\{x_j^Q\}_{j=1}^{N\times(M-K)}$:

$$\kappa_{\mathcal{G}}(x_i^S, x_j^Q) = \{\mathcal{G}(x_i^S)\}_{\varrho=1}^m(\{\mathcal{G}(x_j^Q)\}_{\varrho=1}^m)^{\mathrm{T}} = \sum_{\varrho=1}^m \xi_\varrho^2 \kappa_\varrho(x_i^S, x_j^Q). \tag{15}$$

The encoder $\{\mathcal{G}(\cdot)\}_{\varrho=1}^m$ is employed to encode x_i^S and x_j^Q, resulting in encoding matrices $\widehat{\Theta}_i \in \mathbb{R}^{n\times c}$ and $\widehat{\delta}_j \in \mathbb{R}^{n\times c}$, respectively. $\mathcal{K}_{\mathcal{G}}^i$ represents the consensus clustering matrix used to measure the correlation between $\mathcal{X}_S^{test}$ and $\mathcal{X}_Q^{test}$, where alignment is only required for the similar samples of each data point with its nearest neighbors. The relevance aggregation algorithm is a critical step in cross-class computation for the meta-aggregation model, where the alignment of similarity between query features and support set categories is aggregated, and this can be achieved using the following equation:

$$\begin{aligned}
&\min_{\widehat{\Theta}_i, \ddot{\mathcal{P}}_\varrho, \mathrm{H}^C, \widehat{\delta}, \mathcal{B}, \xi} \sum_{\varrho=1}^m \|\xi_\varrho \mathcal{K}_{\mathcal{G}}^i \ddot{\mathcal{P}}_\varrho - \widehat{\Theta}_i \mathcal{B}\|_F^2 + \|\mathrm{H}^C - \widehat{\delta}_j \mathcal{B}\|_F^2 \\
&s.t.\ \widehat{\delta}_j \widehat{\delta}_j^T = \mathrm{I}_k, \ddot{\mathcal{P}}_\varrho^T \ddot{\mathcal{P}}_\varrho = \mathrm{I}_k, \widehat{\Theta}_i \widehat{\Theta}_i^{\mathrm{T}} = \mathrm{I}_n, (\mathrm{H}^C)\mathrm{H}^C = \mathrm{I}_k.
\end{aligned} \tag{16}$$

Eq.(12) and (16) are combined to yield:

$$\begin{aligned}
&\min_{\widehat{\Theta}_i, \ddot{\mathcal{P}}_\varrho, \mathrm{H}^C, \widehat{\delta}, \mathcal{B}, \xi} \sum_{\varrho=1}^m \|\xi_\varrho \mathcal{K}_{\mathcal{G}}^i \ddot{\mathcal{P}}_\varrho - \mathrm{H}^C\|_F^2 + \varpi \left(\sum_{\varrho=1}^m \|\xi_\varrho \mathcal{K}_{\mathcal{G}}^i \ddot{\mathcal{P}}_\varrho - \widehat{\Theta}_i \mathcal{B}\|_F^2 + \|\mathrm{H}^C - \widehat{\delta}_j \mathcal{B})\|_F^2 \right) \\
&s.t.\ \widehat{\delta}_j \widehat{\delta}_j^T = \mathrm{I}_k, \ddot{\mathcal{P}}_\varrho^T \ddot{\mathcal{P}}_\varrho = \mathrm{I}_k, \widehat{\Theta}_i \widehat{\Theta}_i^{\mathrm{T}} = \mathrm{I}_n, (\mathrm{H}^C)\mathrm{H}^C = \mathrm{I}_k,
\end{aligned} \tag{17}$$

where ϖ governs the consistency of cluster centers, dimension k controls the partitioning of latent dimensions, and $\mathcal{B}$ represents the consensus clustering center matrix. Notably, Eq.(17) utilizes the consensus clustering center $\mathcal{B}$ to connect the incomplete consensus partition matrix H^C with embedded cluster representations, characterizing this model as CKG. Furthermore, to ensure the aggregation of similar data within the clustering center matrix $\mathcal{B}$ and the separation of dissimilar data, guidance is drawn from global distribution information. The kernel function $\kappa_{\mathcal{F}}$ is employed to capture cross-category correlations, allowing the proposed algorithm to effectively utilize intra-cluster variations among samples and leverage inter-feature relationships for aggregated representations, thereby reducing misclassification and enhancing the model's generalization capabilities.

Moreover, for enhanced scalability of the method, balancing the redundancy information among kernels and kernel details, the final dimensions of the input central matrix are set to be the same, allowing for matrix multiplication after transposition, enabling the method to handle inputs of different sizes. This approach enhances the method's value in industrial applications. This concept can be implemented as follows:

$$\begin{aligned}
&\mathcal{L}'(\omega; \mathcal{K}_{\mathcal{G}}^i, \mathcal{A}) = \left(\frac{\mathcal{K}_{\mathcal{G}}^i + |\mathcal{K}_{\mathcal{G}}^i|(\mathcal{J}_n - \mathcal{A}) + \mathcal{A}^2 - 2\mathcal{A}}{\mathcal{J}_n - \mathcal{A}} \right)^2 + \psi\Omega(\omega) \\
&s.t.\ \mathcal{L}^{test} \in \mathbb{R}^{n\times k}, \mathcal{L}^{test}(\mathcal{L}^{test})^T = \mathcal{A} \in \mathbb{R}^{n\times n},
\end{aligned} \tag{18}$$

where $\mathcal{L}^{test}$ represents a one-hot encoded matrix of size $n \times k$, $\mathcal{J}_n$ is an all-ones matrix,ω represents the training parameters of the model, and k denotes the number of categories. The regularization coefficient ψ is set to 0.005, and $\Omega(\omega)$ is a penalty term used to penalize model complexity. This is because model complexity is positively correlated with the number of coefficients, and the more coefficients there are, the more complex the model becomes. To control model complexity, it is possible to reduce the number of coefficients, which means limiting the number of non-zero elements in the vector. This can be achieved by introducing constraints into the optimization problem:

$$s.t.\ \|\omega\|_2 \leq \mathbb{C}, \tag{19}$$

where $\mathcal{A}_{ij} \in \{0, 1\}$, and thus the output expression of Eq.(15) is as follows:

$$\mathcal{L}'(\omega; \mathcal{K}^i_{\mathcal{G}}, \mathcal{A}) = \begin{cases} \mathcal{K}^i_{\mathcal{G}} + \psi\Omega(\omega) & if\ \mathrm{I}_k = 1 \\ \mathcal{A} + \psi\Omega(\omega) & if\ \mathrm{I}_k = 0. \end{cases} \tag{20}$$

The loss in training the network is computed by utilizing the mean squared loss function, which measures the discrepancy between the model's clustering output and the actual labels. The network is trained by minimizing the loss, represented as ℓ:

$$\ell = \|\mathcal{L}'(\omega; \mathcal{K}^i_{\mathcal{G}} - \mathcal{A}) - \mathcal{A}\|_F. \tag{21}$$

During the model training process, the training loss ℓ is propagated through the network using the backpropagation algorithm. Gradients are computed to update the network's parameters, ω, aiming to minimize the loss ℓ within the network. After each training epoch, the network is tested using data from unforeseen classes to evaluate its generalization ability. Visual results from t-SNE [63] visualization experiments demonstrate a significant enhancement in the distinctiveness and reliability of class separations, reducing misclassification, and strengthening the model's generalization capabilities.

Remark 2. The CKG framework enhances fault diagnosis by incorporating domain knowledge into an incomplete multiple kernel clustering matrix $\{\mathrm{H}_\varrho\}_{\varrho=1}^m$, which quantifies inter-category relationships for generalized learning and differentiates early-stage faults from normal operations. Category-specific feature learning is optimized by focusing on the most salient features, capturing both commonalities shared across categories and unique characteristics specific to certain fault types for accurate diagnostics. By guiding the learning process through a consensus clustering matrix H^C, the framework integrates domain-specific insights into how different faults evolve and interact over time, enabling the simultaneous handling of multiple fault categories and early diagnosis in challenging scenarios. The framework efficiently integrates prior knowledge of failure modes to improve fault diagnosis, even when labeled data is limited or incomplete. Furthermore, domain expertise aids in navigating complex environments by addressing noise and uncertainties, making the CKG approach particularly effective in industrial settings.

The following pseudocode outlines the key steps of the above algorithm. The final consensus clustering matrix H^C obtained from model training can effectively classify new samples.

Algorithm 2 Category knowledge-guided incomplete multiple kernel algorithm

Require: $\{\mathrm{H}_\varrho\}_{\varrho=1}^m$, k, $\mathcal{G}_\theta(\mathcal{S})$, $\mathcal{G}_\theta(\mathcal{Q})$, $\ddot{\mathcal{P}}_\varrho = \mathrm{I}_k$

Ensure: Consensus clustering matrix H^C

Initialize $\{\mathrm{H}_\varrho\}_{\varrho=1}^m = \mathrm{I}_{m\times k}$, $\{\ddot{\mathcal{P}}_\varrho\}_{\varrho=1}^m = \mathrm{I}_k$, $m = k$, $\xi = 1/\sqrt{m}$

repeat

Update $\mathcal{G}$ using the joint Eq.(11) and Eq.(12);

Update H^C using Eq.(12);

Update $\ddot{\mathcal{P}}_\varrho$ using Eq.(13);

Update $\mathcal{K}^i_{\mathcal{G}}$ using Eq.(14);

Update $\mathcal{B}$ using Eq.(16);

Update $\hat{\Theta}$, ξ using Eq.(17);

Update ω using Eq.(18);

until Eq.(21) is reached for convergence

3.3.3. Extension of CKG-IMK algorithm

This study introduces a CKG-IMK algorithm that estimates incomplete base clustering matrices $\{\mathrm{H}_{\varrho}\}_{\varrho=1}^{m}$ from multiple few-shot tasks by enhancing task-relevant features during the clustering process [64]. In the matching phase, the algorithm accurately distinguishes samples. Specifically, the algorithm defines the obtained $\mathrm{H}^{C} \in \mathbb{R}^{n\times k}$ as the incomplete multiple kernel clustering matrix, where m represents the number of clusters for each clustering task, n is the number of samples in each input, and k is the embedded dimension. To strike a balance between computational complexity and information retention in experiments, m is set equal to k.

In the CKG-IMK algorithm, the process begins with the computation of an incomplete similarity matrix $\mathcal{K}_{\mathcal{G}}$ to perform clustering and generate a set of incomplete base clustering matrices $\{\mathrm{H}_{\varrho}\}_{\varrho=1}^{m}$ for each base kernel. These obtained base clustering matrices are then used to learn the consensus clustering matrix H^{C}, which is subsequently employed to estimate each incomplete base clustering matrix. These two steps are iteratively performed until convergence is achieved. The underlying concept is to maximize alignment between the consensus clustering matrix and adaptively weighted base clustering matrices with the optimal arrangement. This prior knowledge is capable of integrating features from different categories [65], facilitating the learning of the consensus clustering matrix and thereby enhancing the performance and efficiency of clustering.

4. Experimental results and analysis

4.1. Experimental setup

In the experiment, the model's training iterations were set to 100, with a batch size of 128, and the Adam optimizer with a learning rate of 1e-4 was employed. Additionally, the learning rate underwent logarithmic decay based on the number of iterations. During the meta-training phase, the training dataset was utilized for supervised learning. In the meta-testing phase, the model with the highest accuracy was loaded and used for clustering the data in the query set. This study adhered to the standard principles of small-sample classification, incorporating three query modes: 1-shot, 3-shot, and 5-shot. To ensure fairness and consistency in our experiments, all competing models were tested under the same running environment, which included an Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz, NVIDIA GeForce GTX3090 GPU, CUDA 11.0, and PyTorch 1.12.

Regarding the evaluation of the results, various common validation methods were employed to comprehensively assess the fault diagnosis outcomes. Given the presence of different types of fault data in practical applications, including permanent faults and early-stage minor faults, the CWRU dataset [59] and the EMF-TM bearing dataset were chosen as the basis for two case studies. In the first case, the CWRU dataset containing vibration signals from 10 different bearing conditions was utilized. In the second case, fault diagnosis across different interdisciplinary scenarios was explored, encompassing rolling bearings under various loads and different fault severities. Multiple experiments were conducted to thoroughly validate the robust generalization capabilities and outstanding diagnostic performance of our proposed CKG method, highlighting its applicability across different fault contexts.

4.2. Case1: CWRU dataset

1) Description of the CWRU Dataset: The CWRU dataset, which is provided by CWRU Bearing, has become a widely used benchmark for diagnosing faults in bearings. The data are collected using accelerometers from fan- and drive-end deep groove ball bearings, which are sampled at the frequency of 48 kHz. The dataset consists of vibration signals from three predesigned bearing faults obtained through electrical discharge machining: ball fault (BF), inner race fault (IF), and outer race fault (OF). BF occurs in the rolling elements, leading to irregular periodic vibrations. IF arises in the rotating inner ring, generating high-frequency vibration signals. OF, on the other hand, occurs in the stationary outer ring, resulting in low-frequency periodic vibrations. Each bearing fault includes three fault sizes: 0.007 inches, 0.014 inches, and 0.021 inches. The signals were collected at a sampling frequency of 48 kHz under different loads (1HP, 2HP, and 3HP). Thus, for each load condition, there are 10 bearing states (one normal state and nine fault states) concerned with the specific information presented in Table 1.

In practical industrial settings, the partitioning of data is crucial, especially when dealing with different fault sizes and types. To better simulate the variations in rolling bearing conditions encountered in real-world usage, three distinct fault types and healthy conditions from the 10 available categories were selected as the meta-testing set, while the remaining six fault types were used as the meta-training set. This data partitioning approach aimed to reflect the model's robustness and generalization ability in classifying unknown non-stationary faults during actual healthy operational

conditions. Given the significant disparities in data distribution, the model needs to adapt to various fault types and sizes, thereby better accommodating the diversity and complexity of real-world conditions.

Table 1
Details of the CWRU dataset.

Label	Fault Type	Defect Size (inch)	Accelerometer	Load (hp)
(a)	Normal	None	Drive end	1
(b)	BF	0.007	Drive end	1
(c)	BF	0.014	Drive end	1
(d)	BF	0.021	Drive end	1
(e)	IF	0.007	Drive end	1
(f)	IF	0.014	Drive end	1
(g)	IF	0.021	Drive end	1
(h)	OF	0.007	Drive end	1
(i)	OF	0.014	Drive end	1
(j)	OF	0.021	Drive end	1

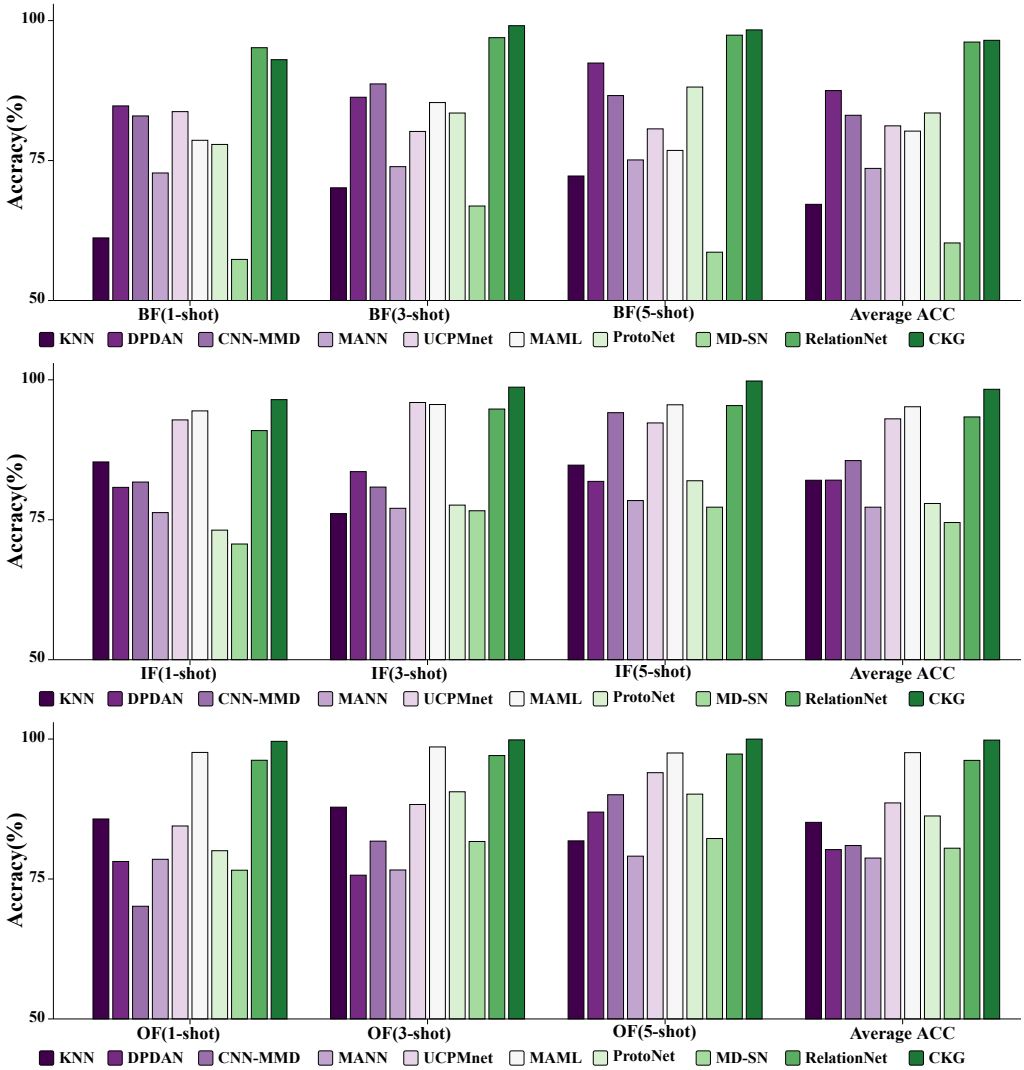


Fig. 3. Performance of various methods on three types of bearing operation data, with healthy data included as the test set: (a) ball fault (BF); (b) inner race fault (IF); (c) outer race fault (OF).

2) Performance Comparison with Existing Methods: Nine recently published fault diagnosis methods were employed in this experiment to validate the performance of the proposed CKG method on the CWRU dataset. These methods include K-nearest neighbor algorithms (KNN) [66], DPDAN [67], CNN-MMD [68], MANN [69], UCPMnet [70], MAML [71], ProtoNet [72], MD-SN [73], and RelationNet [74]. To comprehensively demonstrate the effectiveness of the CKG method, clustering methods like KNN and two domain adaptation methods were also included, as transfer learning methods like DPDAN and CNN-MMD are widely applied in real-world scenarios. In the experiment, three different fault types were merged with the normal state to construct three meta-test sets, each consisting of four categories. Multiple repeated experiments were conducted, including 1-shot, 3-shot, and 5-shot tasks. To ensure fairness, all methods were trained using the same backbone architecture and hyperparameters, and each method underwent 10 experiments to determine the average accuracy, minimizing the impact of uncertainties arising from random network initialization and neural network training. The experimental results are presented in Fig. 3 and Table 2. For ease of reading, the optimal and suboptimal accuracy results are indicated with bold and underlined formatting in Table 2.

Table 2
Performance of CKG and nine existing methods.

Method	BF (1-shot)	BF (3-shot)	BF (5-shot)	Average ACC
KNN	61.18%	70.13%	72.24%	67.85%
DPDAN	84.76%	86.31%	92.42%	87.50%
CNN-MMD	82.97%	88.69%	86.61%	83.09%
MANN	72.78%	73.91%	75.11%	73.60%
UCPMnet	83.74%	80.20%	80.66%	81.20%
MAML	78.62%	85.36%	76.81%	80.26%
ProtoNet	77.88%	83.49%	88.13%	83.50%
MD-SN	57.34%	66.88%	58.63%	60.28%
RelationNet	**95.16**%	96.95%	97.40%	96.17%
CKG	93.02%	**99.09**%	**98.34**%	**96.48**%
Method	IF (1-shot)	IF (3-shot)	IF (5-shot)	Average ACC
KNN	85.33%	76.10%	84.76%	82.06%
DPDAN	80.79%	83.60%	81.86%	82.08%
CNN-MMD	81.74%	80.84%	94.13%	85.57%
MANN	76.28%	77.05%	78.43%	77.25%
UCPMnet	92.84%	95.95%	92.30%	93.03%
MAML	94.45%	95.59%	95.54%	95.19%
ProtoNet	73.15%	77.62%	81.97%	77.91%
MD-SN	70.67%	76.61%	77.25%	74.51%
RelationNet	90.93%	94.78%	95.40%	93.37%
CKG	**96.46**%	**98.69**%	**99.79**%	**98.31**%
Method	OF (1-shot)	OF (3-shot)	OF (5-shot)	Average ACC
KNN	85.72%	87.83%	81.81%	85.12%
DPDAN	78.13%	75.68%	86.95%	80.25%
CNN-MMD	70.13%	81.76%	90.05%	80.98%
MANN	78.52%	76.61%	79.08%	78.74%
UCPMnet	84.47%	88.31%	93.99%	88.59%
MAML	97.61%	98.59%	97.51%	97.57%
ProtoNet	80.05%	90.58%	90.16%	86.26%
MD-SN	76.57%	81.70%	82.23%	80.50%
RelationNet	96.21%	97.05%	97.32%	96.19%
CKG	**99.59**%	**99.85**%	**99.99**%	**99.81**%

Based on the experimental data presented in Table 2, it can be observed that the accuracy of KNN fluctuates significantly under different operating conditions, indicating its subpar clustering performance in the presence of shifting data distributions. Meanwhile, domain adaptation methods, such as DPDAN and CNN-MMD, show a declining trend in performance when confronted with knowledge transfer across different labels. This decline can be attributed to their high reliance on data distribution similarity between the source and target domains as well as the similarity of learning tasks between the two domains. Additionally, the representative FSL method, MD-SN, has some limitations. Its complex structure is prone to overfitting, and it struggles to capture inter-task correlations, leading to poorer performance on the meta-test set. Compared to CKG, MD-SN exhibits a significant performance gap of up to 35.68 percentage points. It is worth noting that classical methods like MANN require an ample amount of labeled data for training the diagnostic model. Limited meta-training samples may cause model instability and hinder its generalizability to target fault diagnosis tasks.

For the three types of bearing operation, the CKG method has significantly improved ball fault diagnostic accuracy compared to other methods. Specifically, the CKG method achieved an improvement of 8.26% - 31.84% in 1-shot ball fault diagnosis, 2.14% - 32.21% in 3-shot ball fault diagnosis, and 1.94% - 39.71% in 5-shot ball fault diagnosis. Furthermore, the CKG method also demonstrated improved accuracy in diagnosing inner race faults, with an improvement of 2.74% - 22.59% in 1-shot diagnosis, 2.74% - 22.54% in 3-shot diagnosis, and 4.25% - 22.54% in 5-shot diagnosis. Similarly, for outer race faults, the CKG method achieved an improvement of 3.38% - 29.46% in 1-shot diagnosis, 1.26% - 24.17% in 3-shot diagnosis, and 2.67% - 20.91% in 5-shot diagnosis.

Overall, the proposed CKG method demonstrates greater robustness and superior performance in these nine experiments when compared to the recently published nine other fault diagnosis methods. These experimental results comprehensively showcase the robustness and superiority of the CKG method across diverse working conditions.

4.3. Case2: EMF-TM dataset

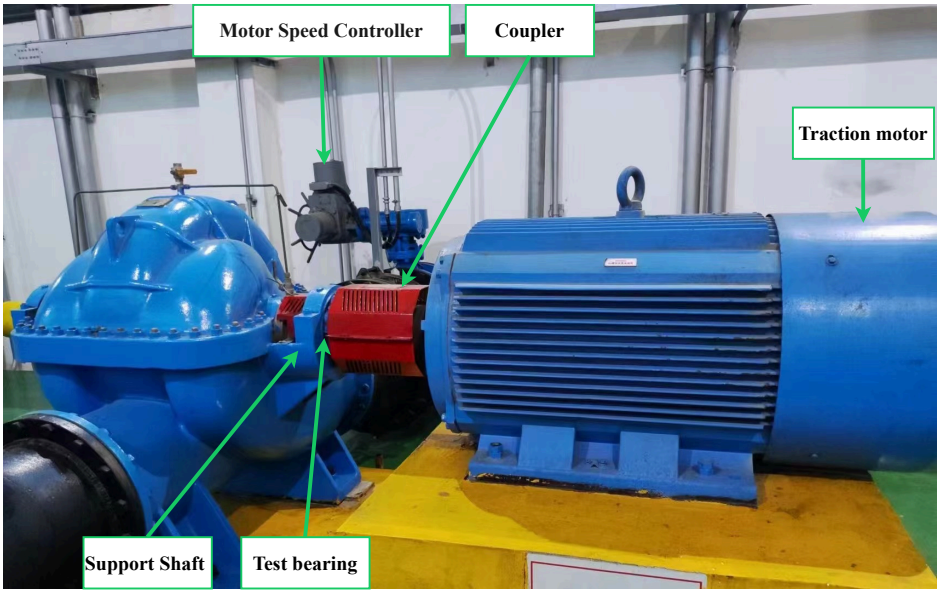


Fig. 4. Case 2 - EMF-TM bearing test bench.

1) Description of the EMF-TM Dataset: The dataset was acquired from motors with broken rotor bars operating at different load currents (0.2A, 0.5A, and 0.8A, with the device capable of a maximum load of 1A) and varying levels of fault severity. The load currents of the motors fall into three categories: 0.2A (light load [LL]), 0.5A (moderate load [ML]), and 0.8A (heavy load [HL]). Additionally, three levels of fault severities were considered, namely, 0.005 (trivial fault), 0.200 (moderate fault), and 1.000 (severe fault), resulting in a total of 10 bearing conditions, which also include the healthy state (H). The data were sampled at a frequency of 2.4 KHz. The experimental platform used for data collection is illustrated in Fig. 4, while a comprehensive dataset description can be found in Table 3, with a visual diagram illustrated in Fig. 5.

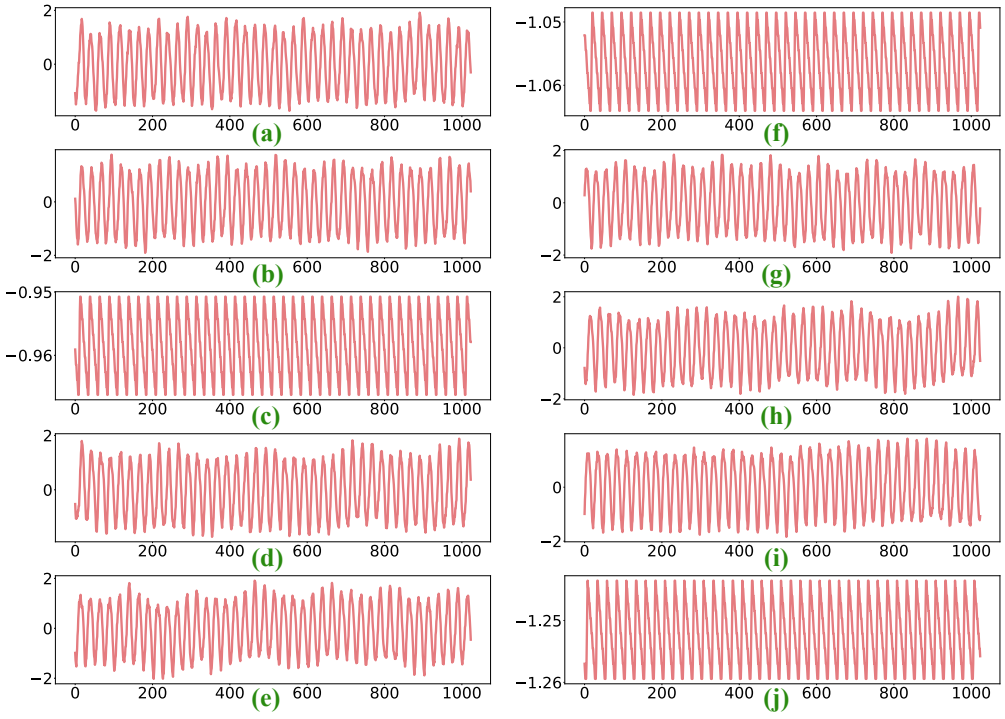


Fig. 5. Raw vibration signals for 10 bearing states of the EMF-TM dataset.

Table 3
Details of the EMF-TM dataset.

Label	Load Current	Degree of Fault	Accelerometer	Temperature(°C)
(a)	-	None	Drive end	45
(b)	LL	0.005	Drive end	45
(c)	LL	0.200	Drive end	45
(d)	LL	1.000	Drive end	45
(e)	ML	0.005	Drive end	45
(f)	ML	0.200	Drive end	45
(g)	ML	1.000	Drive end	45
(h)	HL	0.005	Drive end	45
(i)	HL	0.200	Drive end	45
(j)	HL	1.000	Drive end	45

2) Performance Comparison with Existing Methods: In this experiment, nine recently published fault diagnosis methods were employed to evaluate the performance of the proposed CKG method on the EMF-TM dataset. These methods include KNN, DPDAN, CNN-MMD, MANN, UCPMnet, MAML, ProtoNet, MD-SN, and RelationNet. In the experiment, three different load types were merged with the normal state to construct three meta-test sets, each consisting of four categories, to assess the model's ability to sensitively classify minor faults. Multiple repeated experiments were conducted, including 1-shot, 3-shot, and 5-shot tasks. To prevent overfitting, an early stopping strategy was employed during training.

In the meta-training phase, the model with the highest accuracy was saved to perform meta-test tasks. To ensure fairness, all methods were trained using the same backbone architecture and hyperparameters, and each method underwent 10 repeated experiments to determine the average accuracy, minimizing the impact of uncertainties arising from random network initialization and neural network training. The experimental results are presented in Fig. 6 and Table 4. For ease of reading, the optimal and suboptimal accuracy results are indicated with bold and underlined formatting in Table 4.

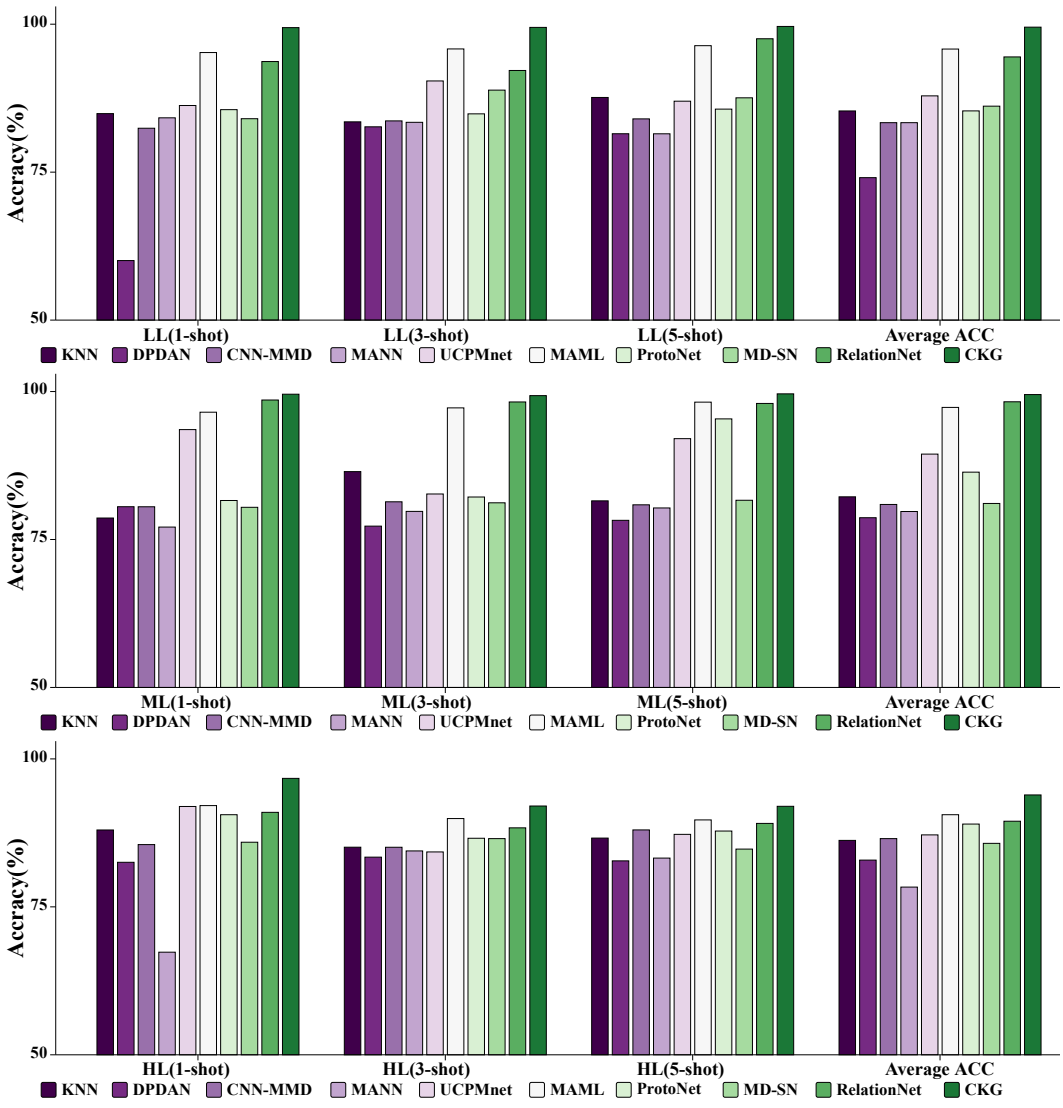


Fig. 6. Performance of various methods on three types of bearing operation data, with healthy data included as the test set:(a) light load data;(b) moderate load data;(c) heavy load data.

Based on the experimentation, it was found that CNN-MMD has a faster training speed but performs poorly in terms of test accuracy, often struggling to converge. The DPDAN method performs inadequately under light loads, indicating that it learns specific features during training, lacking transferability. The MANN and MAML methods require meticulous feature engineering and higher computational resources, with fluctuating performance across different datasets. The ProtoNet method has limitations in capturing effective and information-rich fault features. Finally, the RelationNet method appears to be sensitive to the size and quality of the dataset, resulting in unstable accuracy levels.

The CKG method has shown significant improvement in light load diagnostic accuracy compared to other methods for the three types of load currents. Specifically, the CKG method achieved an improvement of 4.21% - 39.36% in 1-shot ball fault diagnosis, 3.64% - 16.80% in 3-shot ball fault diagnosis, and 3.84% - 25.55% in 5-shot ball fault diagnosis. Furthermore, the CKG method also demonstrated improved accuracy in diagnosing moderate load faults, with an improvement of 0.98% - 22.45% in 1-shot diagnosis, 1.07% - 22.05% in 3-shot diagnosis, and 1.40% - 21.40% in 5-shot diagnosis. Similarly, for heavy load faults, the CKG method achieved an improvement of 4.60% - 29.37% in 1-shot diagnosis, 2.12% - 8.64% in 3-shot diagnosis, and 2.30% - 9.22% in 5-shot diagnosis.

Table 4
Performance of CKG and nine existing methods.

Method	LL (1-shot)	LL (3-shot)	LL (5-shot)	Average ACC
KNN	84.89%	83.50%	87.62%	85.34%
DPDAN	60.06%	82.66%	81.49%	74.07%
CNN-MMD	82.42%	83.66%	84.00%	83.36%
MANN	84.18%	83.42%	81.48%	83.36%
UCPMnet	86.26%	90.41%	86.99%	87.89%
MAML	95.21%	95.82%	96.37%	95.80%
ProtoNet	85.56%	84.85%	85.65%	85.35%
MD-SN	84.03%	88.86%	87.56%	86.15%
RelationNet	93.69%	92.18%	97.54%	94.47%
CKG	**99.42**%	**99.46**%	**99.62**%	**99.50**%
Method	ML (1-shot)	ML (3-shot)	ML (5-shot)	Average ACC
KNN	78.63%	86.47%	81.53%	82.21%
DPDAN	80.54%	77.26%	78.24%	78.68%
CNN-MMD	80.53%	81.37%	80.86%	80.92%
MANN	77.10%	79.75%	80.33%	79.73%
UCPMnet	93.57%	82.69%	92.04%	89.43%
MAML	96.51%	97.24%	98.21%	97.32%
ProtoNet	81.59%	82.18%	95.38%	86.38%
MD-SN	80.44%	81.20%	81.63%	81.09%
RelationNet	98.57%	98.24%	97.99%	98.27%
CKG	**99.55**%	**99.31**%	**99.61**%	**99.49**%
Method	HL (1-shot)	HL (3-shot)	HL (5-shot)	Average ACC
KNN	87.99%	85.08%	86.62%	86.23%
DPDAN	82.53%	83.40%	82.77%	82.90%
CNN-MMD	85.52%	85.07%	88.00%	86.53%
MANN	67.34%	84.45%	83.24%	78.34%
UCPMnet	91.96%	84.28%	87.25%	87.16%
MAML	92.11%	89.92%	89.69%	90.57%
ProtoNet	90.57%	86.59%	87.81%	88.99%
MD-SN	85.92%	86.53%	84.76%	85.73%
RelationNet	90.97%	88.35%	89.10%	89.47%
CKG	**96.71**%	**92.04**%	**91.99**%	**93.91**%

The proposed CKG method efficiently clusters based on inter-task correlated features, achieving the highest diagnostic accuracy while demonstrating stability and superiority in multiple aspects. It not only exhibits high-speed operation and low computational memory usage but also attains high accuracy during the training process. Importantly, CKG can effectively distinguish minor early-stage faults from healthy states under the interference of strong background noise, even when there are slight variations in load conditions. Thus, CKG represents a valuable method for addressing complex, dynamic, and high-demand real-world applications, especially in the realm of early and rapid fault diagnosis.

3) Performance Comparison under Different Noise Conditions: In practical industrial applications, sensor-collected vibration signals often contain varying levels of background noise. Therefore, this article evaluates the performance of the proposed CKG method, along with nine other recently published methods, under varying noise conditions, with signal-to-noise ratios (SNR) ranging from -4 to 10 dB

$$\mathrm{SNR} = 10 \lg \left(\frac{p_{signal}}{p_{noise}} \right), \tag{22}$$

where p_{signal} represents the signal power, and p_{noise} represents the noise power. In this study, all experiments were conducted 10 times independently, with the structure and parameters of the CKG model and the other nine models kept unchanged. The performance of the CKG method, along with the highest accuracy achieved by the other methods under different noise conditions, is shown in Table 5. It is evident that, across three types of bearing operations, the proposed CKG method achieves higher diagnostic accuracy than the other nine methods under various noise conditions. This demonstrates the robust capability of the CKG method in effectively classifying different bearing states in noisy environments through end-to-end representation learning.

Table 5
Performance of different methods on EMF-TM dataset under different SNRs.

Method	**-4**	**-2**	**0**	**2**	**4**	**6**	**8**	**10**
Light Load (1-shot)								
KNN	84.05%	84.96%	83.20 %	85.87%	83.78%	83.70%	86.52%	83.22%
DPDAN	81.99%	81.76%	82.19%	81.83%	81.16%	83.54%	83.71%	81.75%
CNN-MMD	85.12%	82.59%	84.43%	83.36%	82.55%	85.60%	83.97%	86.29%
MANN	79.23%	78.18%	81.09%	81.17%	81.26%	83.27%	85.96%	83.01%
UCPMnet	94.88%	95.02%	93.18%	92.80%	87.84%	84.14%	84.09%	85.20%
MAML	93.86%	97.33%	95.32%	95.84%	97.39%	96.41%	98.41%	98.38%
ProtoNet	99.24%	98.57%	98.73%	98.93%	99.21%	98.94%	96.40%	98.88%
MD-SN	84.68%	85.67%	84.70%	85.45%	84.41%	83.06%	85.16%	83.52%
RelationNet	89.39%	86.81%	85.44%	86.07 %	87.11%	86.03%	87.20%	88.75%
CKG	**99.37%**	**99.58%**	**99.38%**	**99.37%**	**99.42%**	**99.34%**	**99.26%**	**99.34%**
Moderate Load (1-shot)								
KNN	84.78%	80.14%	85.18%	81.35%	82.99%	80.49%	82.65%	83.25%
DPDAN	80.09%	85.62%	80.19%	78.24%	79.60%	60.17%	79.30%	78.67%
CNN-MMD	79.09%	79.77%	79.79%	79.18%	81.79%	81.49%	80.52%	81.82%
MANN	82.23%	80.64%	81.79%	84.07%	83.89%	83.31%	85.20%	86.19%
UCPMnet	93.87%	93.48%	95.00%	80.76%	95.59%	93.48%	88.68%	93.79%
MAML	93.56%	97.39%	94.34%	97.34%	97.83%	96.69%	98.50%	96.37%
ProtoNet	87.90%	82.58%	87.12%	80.27%	84.89%	81.85%	82.84%	97.42%
MD-SN	79.26%	82.48%	79.58%	79.72%	80.80%	83.32%	88.02%	81.43%
RelationNet	83.41%	80.69%	82.90%	98.18%	82.68%	82.99%	82.07%	80.37%
CKG	**99.18%**	**99.66%**	**99.47%**	**99.84**	**99.20%**	**99.91%**	**99.61%**	**99.02%**
Heavy Load (1-shot)								
KNN	86.17%	89.34%	86.61%	87.38%	87.80%	88.31%	85.85%	87.11%
DPDAN	84.52%	78.25%	82.88%	83.62%	82.41%	83.61%	83.20%	83.56%
CNN-MMD	85.24%	81.79%	86.11%	84.25%	85.28%	85.18%	83.60%	84.07%
MANN	79.46%	85.74%	81.04%	82.55%	84.37%	85.50%	85.25%	84.73%
UCPMnet	88.46%	84.67%	87.50%	88.70%	91.23%	88.69%	83.65%	87.11%
MAML	89.05%	91.30%	89.69%	89.02%	87.62%	91.60%	88.86%	89.09%
ProtoNet	87.64%	85.13%	86.33%	90.36%	90.35%	87.39%	85.27%	87.48%
MD-SN	84.94%	**97.03%**	87.03%	87.17%	88.99%	90.57%	85.57%	87.06%
RelationNet	94.33%	90.24%	93.97%	91.19%	87.36%	87.94%	87.60%	91.06%
CKG	**94.59%**	96.34%	**94.50%**	**91.42%**	**93.17%**	**91.83%**	**92.89%**	**92.34%**

Moreover, to validate the efficiency of the proposed model in early fault diagnosis, the number of samples processed per second by each model was compared, as shown in Table 6. The results show that the proposed CKG model ranks second only to the simpler CNN-MMD structure in terms of efficiency, while significantly outperforming CNN-MMD in diagnostic performance, demonstrating the overall capability of the CKG model.

Table 6
Comparison of model training speeds in terms of samples processed per second.

Model	KNN	DPDAN	CNN-MMD	MANN	UCPMnet	MAML	ProtoNet	MD-SN	RelationNet	**CKG**
Samples per second ↑	751838.29	766099.38	**1088193.27**	836496.65	939220.25	785305.13	875162.06	954285.09	372763.40	1030634.14

4) Classification Performance of the CKG Method: To validate the model's ability to handle the variability and uncertainty of early fault data, four fault types were randomly selected from the ten available types for FSL testing across 10 independent experiments under an SNR of 8 and a 4-way, 1-shot setting—conditions designed to challenge the model's performance. Given the varying requirements for diagnostic methods in practical applications, it is essential to employ multiple evaluation metrics (i.e., accuracy, precision, recall, and F1-score) to assess the performance of the proposed CKG method. The calculation formulas are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{24}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{25}$$

$$\text{F1-score} = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{26}$$

where TP, FP, FN, and TN represent the number of true positives, false positives, false negatives, and true negatives, respectively. The accuracy, precision, recall, and F1-score all fall within the range of zero to one, with higher values indicating improved fault diagnosis performance.

Table 7
Performance evaluation of the CKG method with ten independent experiments.

Training number	Accuracy	Precision	Recall	F1-score
1	98.89%	98.87%	99.87%	99.85%
2	98.94%	98.90%	98.92%	98.92%
3	99.90%	99.90%	99.80%	99.86%
4	99.60%	99.68%	99.59%	99.64%
5	95.81%	95.97%	95.20%	95.15%
6	99.03%	98.12%	98.05%	98.05%
7	97.09%	96.49%	96.50%	96.48%
8	97.94%	97.81%	97.77%	97.77%
9	98.96%	97.86%	97.84%	97.84%
10	99.99%	99.90%	99.80%	99.85%
Average	98.62%	97.46%	98.33%	98.34%
Best	99.99%	99.90%	99.80%	99.86%
Worst	95.81%	95.97%	95.20%	95.15%

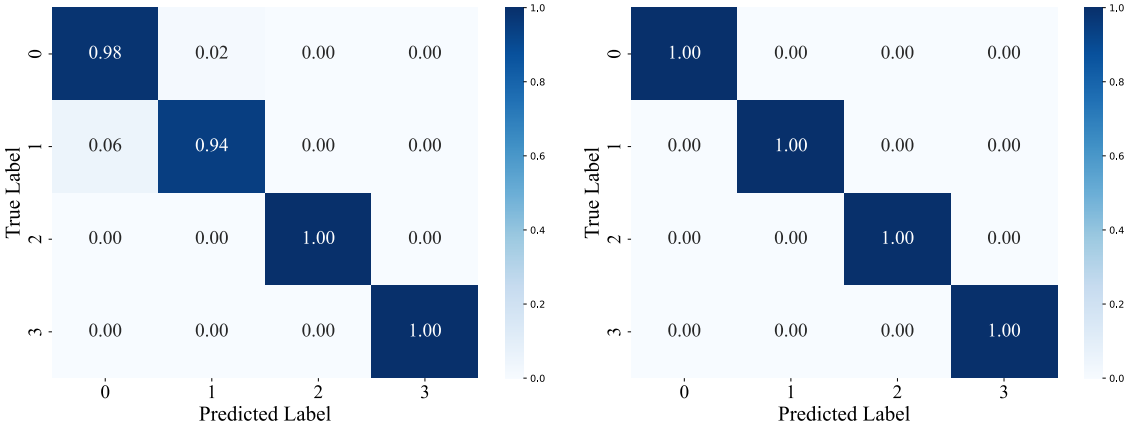


Fig. 7. The worst classification results are shown on the left, and the best on the right, visualized using confusion matrices.

Table 7 presents the testing results from ten trials, showing that the highest accuracy, precision, recall, and F1-score achieved are 99.99%, 99.90%, 99.80%, and 99.85%, respectively, indicating that most bearing states are correctly classified. Additionally, the average values of these metrics—98.62%, 97.46%, 98.33%, and 98.34%—demonstrate the stability of the proposed CKG method in early bearing fault diagnosis. Even the lowest values, 95.81%, 95.97%, 95.50%, and 95.15%, all exceed 95%, confirming that all verification metrics surpass this threshold. Overall, these results indicate that the proposed CKG method accurately and reliably diagnoses various roller-bearing faults, even when facing uncertainty and variability in early-stage fault diagnosis. Additionally, the best and worst classification results are visualized using confusion matrices in Fig. 7.

5) Performance Comparison of MoE Techniques for Accuracy and Computational Efficiency: To demonstrate the effectiveness of the proposed Atten-MoE module, a comparison of accuracy and computational efficiency using different MoE techniques, including Atten-MoE, Dense-MoE, and Soft-MoE, was conducted on the EMF-TM dataset. The performance comparison is presented in Table 8.

Table 8
Performance comparison of MoE models across different loads and resource usage.

Method	LL (1-shot)	ML (1-shot)	HL (1-shot)	Speed (M sample/s)	Memory Usage (M)
Dense-MoE	99.33%	99.31%	96.66%	0.83	13.23
Soft-MoE	96.21%	95.47%	89.83%	**1.09**	10.44
Atten-MoE	**99.42%**	**99.55%**	**96.71%**	1.03	**10.22**

As shown in Table 8, the proposed Atten-MoE model outperforms the other two algorithms in terms of accuracy across all load conditions, while maintaining higher computational speed and lower memory usage. This demonstrates the ability of Atten-MoE to reduce computational costs while preserving sequence continuity.

6) Advanced Feature Visualization: The feature extraction capabilities of the aforementioned methods were visualized using the t-SNE method, and the results are depicted in Fig. 8. In comparison to the nine recently published methods, the proposed CKG method exhibits well-separated and compact data clusters. Nearly all samples are clustered within their respective regions, and there are reasonable inter-class differences among all four states. It is noteworthy that under the interference of strong background noise, the subtle distinctions between the healthy state and early fault features pose a significant challenge for early fault diagnosis in rolling bearings. Other models tend to confuse early minor faults with healthy operational conditions, which can be fatal in the context of real industrial vulnerabilities.

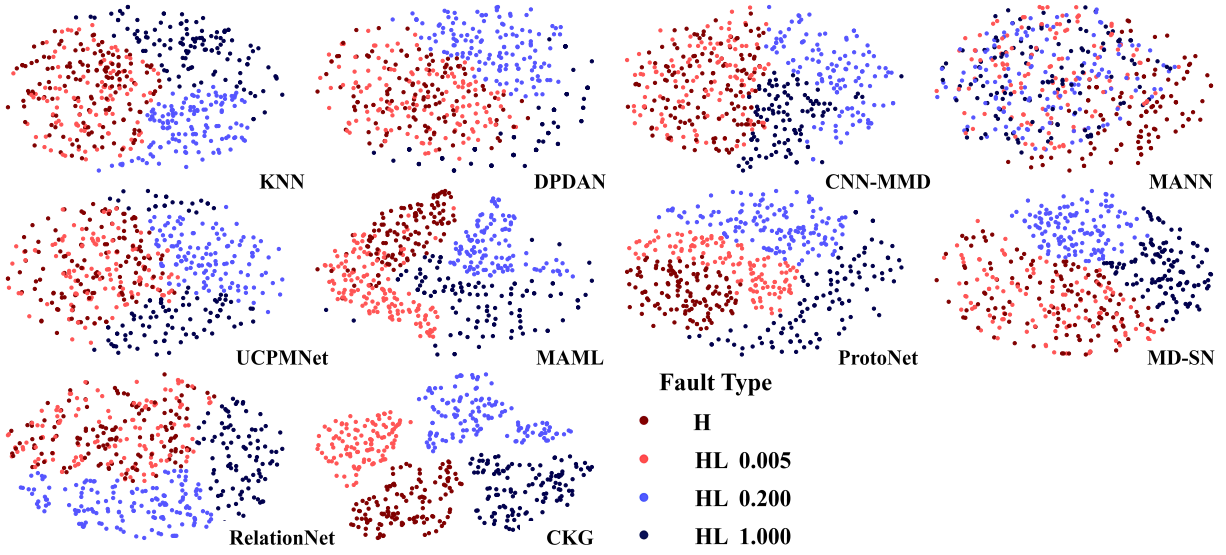


Fig. 8. t-SNE-based visualization of extracted features by different methods. The labels H, HL 0.005, HL 0.200, and HL 1.000 correspond to healthy state, trivial fault, moderate fault, and severe fault, respectively.

The traditionally used KNN clustering method yielded dispersed learned features and had a poor ability to learn inter-task correlated features. In addition, it exhibited low convergence accuracy. Transfer learning methods like DPDAN and CNN-MMD incurred high tuning costs in the target domain. ProtoNet, while capable of distinguishing between the four categories, lacked clarity and was prone to misclassification. Visual results indicate that the CKG method effectively extracts discriminative features from vibration signals in fault diagnosis. It maximizes classification boundaries and achieves optimal clustering results, surpassing the performance of the nine recently published methods.

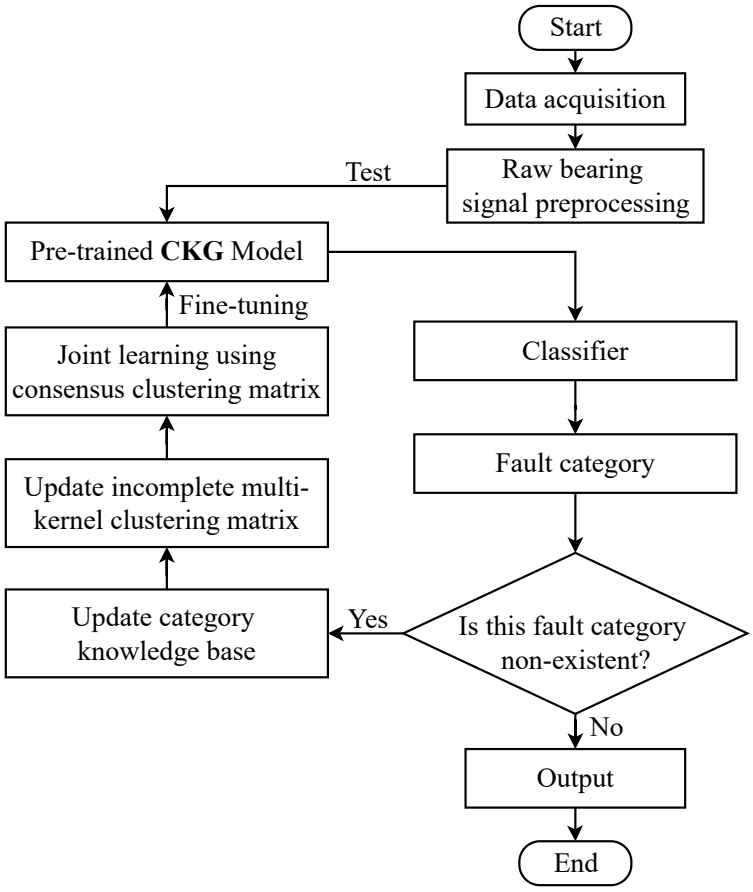


Fig. 9. Implementation details of the category knowledge-guided model in various bearing fault diagnosis scenarios.

7) Practical Application of CKG Method: As shown in Fig. 9, the implementation of the knowledge-guided few-shot learning approach begins with data collection and preprocessing, where fault-related data is gathered and refined through noise reduction, outlier detection, and standardization. During the application phase, the model employs a pretrained CKG framework, leveraging knowledge-guided feature selection to improve efficiency by focusing on task-relevant features. Following this, the classifier (linear layer) generates diagnostic results. If the fault belongs to a previously known category, the diagnosis is provided directly; otherwise, the process includes an update of the category knowledge base, followed by an update to the incomplete multiple kernel clustering matrix. This step quantifies inter-category relationships, enhancing the model's ability to generalize across different fault categories. Simultaneously, a consensus clustering matrix guides the joint learning process, allowing the model to be fine-tuned to handle multiple support categories, even in data-scarce environments.

The CKG-based methodology is currently being applied in practical production environments for the early diagnosis of faults in bearings. This diagnostic application utilizes sensor data collected from a wind turbine gearbox, as shown in Fig. 10, with further details provided in Fig. 9. The fault signals are processed by a pretrained CKG model specifically designed for fault diagnosis, which promptly generates alerts when anomalies are diagnosed. In future work, fault-tolerant control strategies will be implemented to enhance equipment reliability further.

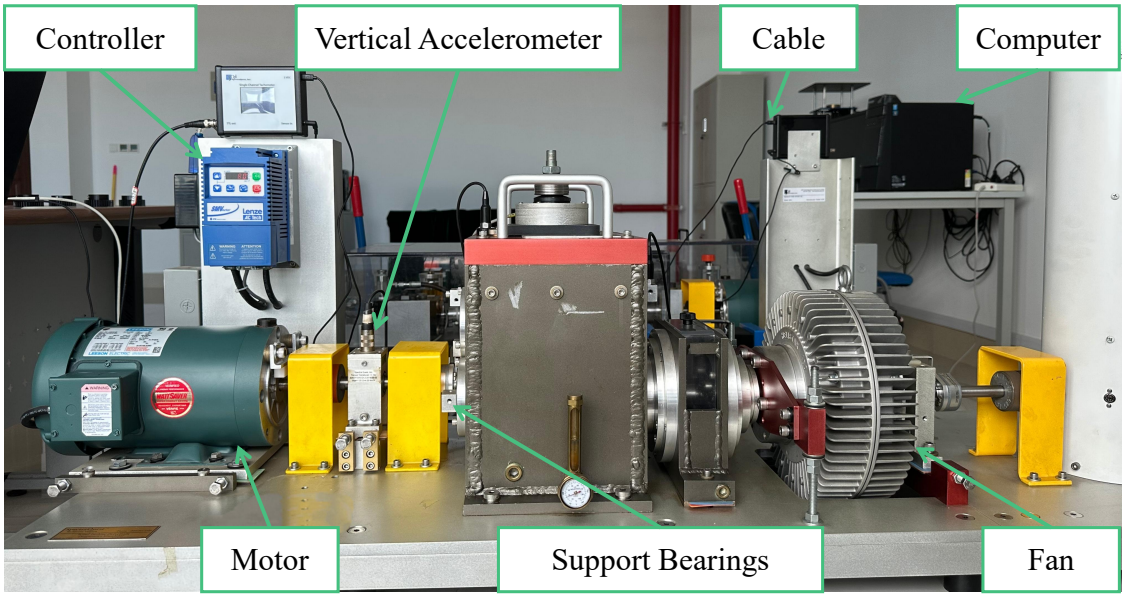


Fig. 10. Case study of fault diagnosis for wind turbine gearbox bearings.

5. Conclusions and future works

In order to address the more challenging cross-domain cold-start tasks and early fault diagnosis tasks within FSL and to reduce the complexity of FSL, this study introduced an innovative CKG method. This method effectively utilizes an incomplete multiple kernel clustering algorithm to capture category information among data. Unlike traditional inductive approaches, this method is capable of classifying unlabeled query instances in a single pass, thus overcoming the challenges of small-sample category knowledge imbalances. Further, the proposed CKG method was compared with nine other cross-domain fault diagnosis methods. Multiple experiments were conducted on the CWRU bearing dataset and the traction motor dataset, assessing the performance of these methods across classic 4-way, 1-shot, 3-shot, and 5-shot tasks. The experimental results demonstrate that, even with limited samples in both support and query sets, the proposed category-guided incomplete clustering method outperforms the other nine fault diagnosis methods.

Furthermore, the visual results also demonstrate that the proposed CKG jointly optimizes the best kernel, maximum-margin hyperplane, and optimal clustering labels while effectively distinguishing minor faults from normal states. This highlights the high fault diagnosis accuracy of the method. Notably, the CKG method was tested and applied in an actual industrial production context, offering a more reliable and efficient solution for bearing fault diagnosis. In summary, this study provides a viable solution for the field of small-sample bearing fault diagnosis and offers valuable insights for future research endeavors.

A primary limitation of this study is the complexity involved in selecting hyperparameters, which can significantly affect the model's generalization ability when applied to datasets with varying sampling rates or rotational speeds. To address this, future work will focus on designing algorithms that adaptively determine key hyperparameters. This approach will enhance efficiency, reduce the need for manual intervention, and improve generalization across diverse cases. Furthermore, enhancing the timeliness of early fault diagnosis in motor bearings remains a crucial direction. After fault diagnosis, fault-tolerant control mechanisms should be implemented to ensure stable motor operation. The proposed method can also be adapted and optimized for other mechanical fault diagnosis challenges, such as in gearboxes, motors, and engines [75]. Additionally, the findings of this research can be extended to broader data classification tasks in various industrial applications.

CRediT authorship contribution statement

Lingkai Hu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft. **Feng Zhan:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft. **Wenkai Huang:** Formal analysis, Resources, Writing – review

& editing, Supervision, Project administration, Funding acquisition. **Yikai Dong:** Conceptualization, Software, Validation, Visualization, Writing – review & editing. **Hao He:** Writing – review & editing, Funding acquisition. **Guanjun Wu:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to express their gratitude to Guangzhou Sanki Automotive Gasket Co., Ltd. for providing test data samples and verifying the algorithm studied in this article during actual production work. This work was supported by Guangzhou Youth Science and Technology Education Project under Grant KP2024403, as well as the East China Normal University Innovation Team for Artificial Intelligence Governance (2024QKT001).

Appendix

Table 9
Nomenclature

Notations	Interpretation	Notations	Interpretation
$\mathcal{W}_{ij}$	Clustering assignment for the i-th data sample point and the j-th cluster centroid	$\mathcal{F}(\cdot)$	Kernel mapping
$\mathcal{K}_{\xi}$	Kernel function	$\mathrm{Tr}(\cdot)$	Trace norm
$\{(x_i, y_i)\}_{i=1}^{n}$	n training samples pair	$\mathcal{G}(\cdot)$	Encoder function
$\{\mathcal{H}\}_{o=1}^{m}$	Hilbert spaces	$(z_i)_o$	Latent representation
ξ	Normalization coefficients	h_j	Latent proxies
$\{\ddot{\mathcal{P}}_o\}_{o=1}^{m}$	Mapping matrices	$\mathcal{K}_{\mathcal{G}}$	Kernel function
H^{C}	Cluster assignment	$\ddot{\mathcal{G}}_o$	Transposition matrix
$\widehat{C}_k$	Learned cluster centroids	e_k	Orthogonal base
(x_i^S, y_i^S)	Support set pair	x_i^Q	Query set input
ϖ	Learnable parameter	$\widehat{\Theta}_i$	Encoding matrices
$\mathcal{B}$	Consensus clustering center matrix	$\mathcal{A}$	Label matrix
$\mathcal{J}_n$	All-ones matrix	ψ	Regularization coefficient
$\Omega(\cdot)$	Penalty term	ℓ	Loss function

Table 10
Abbreviations

Abbreviations	Interpretation	Abbreviations	Interpretation
CKG	Category knowledge-guided	BF	Ball fault
FSL	Few-shot learning	IF	Inner race fault
CKG-IMK	Category knowledge-guided incomplete multiple kernel clustering	OF	Outer race fault
MKKM	Multiple kernel k-means	LL	Light load
CWRU	Case Western Reserve University bearing dataset	ML	Moderate load
EMF-TM	Early Mild Fault Traction Motor bearing dataset	HL	Heavy load
MKSC	Multiple kernel subspace clustering	TP	True positives
Atten-MoE	Attention mixture of experts	FP	False positives
H	Healthy state	FN	False negatives
SNR	Signal-to-noise ratios	TN	True negatives

References

[1] Fasikaw Kibrete, Dereje Engida Woldemichael, and Hailu Shimels Gebremedhen. Multi-sensor data fusion in intelligent fault diagnosis of rotating machines: A comprehensive review. *Measurement*, page 114658, 2024.

[2] Zheng Yang, Binbin Xu, Wei Luo, and Fei Chen. Autoencoder-based representation learning and its application in intelligent fault diagnosis: A review. *Measurement*, 189:110460, 2022.

[3] Yihao Xue, Rui Yang, Xiaohan Chen, Zhongbei Tian, and Zidong Wang. A novel local binary temporal convolutional neural network for bearing fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 2023.

[4] Kai Liu, Ningyun Lu, Feng Wu, Ridong Zhang, and Furong Gao. Model fusion and multiscale feature learning for fault diagnosis of industrial processes. *IEEE Transactions on Cybernetics*, 2022.

[5] Quan Sun, Fei Peng, Xianghai Yu, and Hongsheng Li. Data augmentation strategy for power inverter fault diagnosis based on wasserstein distance and auxiliary classification generative adversarial network. *Reliability Engineering & System Safety*, 237:109360, 2023.

[6] Mingkuan Shi, Chuancang Ding, Rui Wang, Changqing Shen, Weiguo Huang, and Zhongkui Zhu. Graph embedding deep broad learning system for data imbalance fault diagnosis of rotating machinery. *Reliability Engineering & System Safety*, page 109601, 2023.

[7] Kai Zhong, Jiayi Wang, Shuiqing Xu, Chao Cheng, and Hongtian Chen. Overview of fault prognosis for traction systems in high-speed trains: A deep learning perspective. *Engineering Applications of Artificial Intelligence*, 126:106845, 2023.

[8] Devendra Sahu, Ritesh Kumar Dewangan, and Surendra Pal Singh Matharu. An investigation of fault detection techniques in rolling element bearing. *Journal of Vibration Engineering & Technologies*, 12(4):5585–5608, 2024.

[9] Xiaohan Chen, Rui Yang, Yihao Xue, Mengjie Huang, Roberto Ferrero, and Zidong Wang. Deep transfer learning for bearing fault diagnosis: A systematic review since 2016. *IEEE Transactions on Instrumentation and Measurement*, 2023.

[10] Chuanhang Qiu, Tang Tang, Tianyuan Yang, and Ming Chen. Learning to generalize with latent embedding optimization for few-and zero-shot cross domain fault diagnosis. *Expert Systems with Applications*, page 124280, 2024.

[11] Zhen Liu and Zhenrui Peng. Few-shot bearing fault diagnosis by semi-supervised meta-learning with graph convolutional neural network under variable working conditions. *Measurement*, page 115402, 2024.

[12] Haidong Shao, Xiangdong Zhou, Jian Lin, and Bin Liu. Few-shot cross-domain fault diagnosis of bearing driven by task-supervised anil. *IEEE Internet of Things Journal*, 2024.

[13] Liang Zeng, Junjie Jian, Xinyu Chang, and Shanshan Wang. A meta-learning method for few-shot bearing fault diagnosis under variable working conditions. *Measurement Science and Technology*, 35(5):056205, 2024.

[14] Ye Rong, Dongmei Guo, Qingyi Kong, Guanglong Wang, Zhixin Ren, and Zihao Tian. Fault diagnosis of rotating machinery bearings based on multi-scale attention feature fusion under few shot and complex working conditions. *Journal of Electrical Systems*, 20(3):13–27, 2024.

[15] Xingchen Fu, Jianfeng Tao, Keming Jiao, and Chengliang Liu. A novel semi-supervised prototype network with two-stream wavelet scattering convolutional encoder for tbm main bearing few-shot fault diagnosis. *Knowledge-Based Systems*, 286:111408, 2024.

[16] Manh-Hung Vu, Van-Quang Nguyen, Thi-Thao Tran, Van-Truong Pham, and Men-Tzung Lo. Few-shot bearing fault diagnosis via ensembling transformer-based model with mahalanobis distance metric learning from multiscale features. *IEEE Transactions on Instrumentation and Measurement*, 2024.

[17] Dengming Zhang, Kai Zheng, Yin Bai, Dengke Yao, Dewei Yang, and Shaowang Wang. Few-shot bearing fault diagnosis based on meta-learning with discriminant space optimization. *Measurement Science and Technology*, 33(11):115024, 2022.

[18] Likui Qiao, Yuxian Zhang, and Qisen Wang. Fault detection in wind turbine generators using a meta-learning-based convolutional neural network. *Mechanical Systems and Signal Processing*, 200:110528, 2023.

[19] Jianing Liu, Hongrui Cao, and Yang Luo. An information-induced fault diagnosis framework generalizing from stationary to unknown nonstationary working conditions. *Reliability Engineering & System Safety*, 237:109380, 2023.

[20] Yihao Xue, Rui Yang, Xiaohan Chen, Weibo Liu, Zidong Wang, and Xiaohui Liu. A review on transferability estimation in deep transfer learning. *IEEE Transactions on Artificial Intelligence*, 2024.

[21] Shengnan Tang, Jingtao Ma, Zhengqi Yan, Yong Zhu, and Boo Cheong Khoo. Deep transfer learning strategy in intelligent fault diagnosis of rotating machinery. *Engineering Applications of Artificial Intelligence*, 134:108678, 2024.

[22] Peng Ding, Xiaoli Zhao, Haidong Shao, and Minping Jia. Machinery cross domain degradation prognostics considering compound domain shifts. *Reliability Engineering & System Safety*, 239:109490, 2023.

[23] Bin Pang, Qiuhai Liu, Zhenli Xu, Zhenduo Sun, Ziyang Hao, and Ziqi Song. Fault vibration model driven fault-aware domain generalization framework for bearing fault diagnosis. *Advanced Engineering Informatics*, 62:102620, 2024.

[24] Mohammed Hakim, Abdoulhdi A Borhana Omran, Ali Najah Ahmed, Muhannad Al-Waily, and Abdallah Abdellatif. A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations. *Ain Shams Engineering Journal*, 14(4):101945, 2023.

[25] Ke Feng, JC Ji, Yongchao Zhang, Qing Ni, Zheng Liu, and Michael Beer. Digital twin-driven intelligent assessment of gear surface degradation. *Mechanical Systems and Signal Processing*, 186:109896, 2023.

[26] Sheng Li, Ke Feng, Yadong Xu, Yongbo Li, Qing Ni, Ke Zhang, Yulin Wang, and Weiping Ding. Cross-modal zero-sample diagnosis framework utilizing non-contact sensing data fusion. *Information Fusion*, 110:102453, 2024.

[27] Chunlin Zhang and Yuling Liu. A two-step denoising strategy for early-stage fault diagnosis of rolling bearings. *IEEE Transactions on Instrumentation and Measurement*, 69(9):6250–61, 2020.

[28] Sheng Li, Qiubo Jiang, Yadong Xu, Ke Feng, Zhiheng Zhao, Beibei Sun, and George Q Huang. Digital twin-assisted interpretable transfer learning: A novel wavelet-based framework for intelligent fault diagnostics from simulated domain to real industrial domain. *Advanced Engineering Informatics*, 62:102681, 2024.

[29] Ke Feng, JC Ji, Qing Ni, and Michael Beer. A review of vibration-based gear wear monitoring and prediction techniques. *Mechanical Systems and Signal Processing*, 182:109605, 2023.

[30] Ming Xie, Jianxin Liu, Yifan Li, Ke Feng, and Qing Ni. An ensemble domain adaptation network with high-quality pseudo labels for rolling bearing fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 2024.
[31] Shuangshan Hou, Jinde Zheng, Haiyang Pan, Ke Feng, Qingyun Liu, and Qing Ni. Multivariate multi-scale cross-fuzzy entropy and ssa-svm-based fault diagnosis method of gearbox. *Measurement Science and Technology*, 35(5):056102, 2024.
[32] Qing Ni, JC Ji, Ke Feng, Yongchao Zhang, Dongdong Lin, and Jinde Zheng. Data-driven bearing health management using a novel multi-scale fused feature and gated recurrent unit. *Reliability Engineering & System Safety*, 242:109753, 2024.
[33] M Pandiyan and T Narendiranath Babu. Systematic review on fault diagnosis on rolling-element bearing. *Journal of Vibration Engineering & Technologies*, pages 1–35, 2024.
[34] Qing Ni, JC Ji, Benjamin Halkon, Ke Feng, and Asoke K Nandi. Physics-informed residual network (piresnet) for rolling element bearing fault diagnostics. *Mechanical Systems and Signal Processing*, 200:110544, 2023.
[35] Hossein Shayeghi, Ali Ahmadpour, and Mir Mohsen Hosseini Khashe Heiran. Optimal operation of wind farm in presence of pumped-storage station as smart infrastructure and load estimation using artificial neural networks. In *2017 smart grid conference (SGC)*, pages 1–7. IEEE, 2017.
[36] Leila Bagherzadeh, Hossein Shahinzadeh, Hossein Shayeghi, Abdolmajid Dejamkhooy, Ramazan Bayindir, and Mohammadreza Iranpour. Integration of cloud computing and iot (cloudiot) in smart grids: Benefits, challenges, and solutions. In *2020 international conference on computational intelligence for smart power system and sustainable energy (CISPSSE)*, pages 1–8. IEEE, 2020.
[37] Sheng Li, JC Ji, Yadong Xu, Ke Feng, Ke Zhang, Jingchun Feng, Michael Beer, Qing Ni, and Yuling Wang. Dconformer: A denoising convolutional transformer with joint learning strategy for intelligent diagnosis of bearing faults. *Mechanical Systems and Signal Processing*, 210:111142, 2024.
[38] Ali Ahmadpour, Abdolmajid Dejamkhooy, and Hossein Shayeghi. Fault diagnosis of hts–slim based on 3d finite element method and hilbert–huang transform. *IEEE Access*, 10:35736–35749, 2022.
[39] Zuolu Wang, Dawei Shi, Yuandong Xu, Dong Zhen, Fengshou Gu, and Andrew D Ball. Early rolling bearing fault diagnosis in induction motors based on on-rotor sensing vibrations. *Measurement*, 222:113614, 2023.
[40] Yi Wang, Jiakai Ding, Haoran Sun, Yi Qin, and Baoping Tang. Deep signal separation for adaptive estimation of instantaneous phase from vibration signals. *Expert Systems with Applications*, 246:123187, 2024.
[41] Jiaxian Chen, Dongpeng Li, Ruyi Huang, Zhuyun Chen, and Weihua Li. Multi-scale dilated convolutional auto-encoder network for weak feature extraction and health condition detection. In *2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2024.
[42] Khadija Shaheen, Apoorva Chawla, Ferdinand Evert Uilhoorn, and Pierluigi Salvo Rossi. Sensor-fault detection, isolation and accommodation for natural-gas pipelines under transient flow. *IEEE Transactions on Signal and Information Processing over Networks*, 2024.
[43] Yifei Ding, Jichao Zhuang, Peng Ding, and Minping Jia. Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings. *Reliability Engineering & System Safety*, 218:108126, 2022.
[44] Zhe Wu, Li Su, and Qingming Huang. Decomposition and completion network for salient object detection. *IEEE transactions on image processing*, 30:6226–39, 2021.
[45] Linglan Zhao, Ge Liu, Dashan Guo, Wei Li, and Xiangzhong Fang. Boosting few-shot visual recognition via saliency-guided complementary attention. *Neurocomputing*, 507:412–27, 2022.
[46] Zun Li, Congyan Lang, Jun Hao Liew, Yidong Li, Qibin Hou, and Jiashi Feng. Cross-layer feature pyramid network for salient object detection. *IEEE Transactions on Image Processing*, 30:4587–98, 2021.
[47] Chaofan Chen, Xiaoshan Yang, Jinpeng Zhang, Bo Dong, and Changsheng Xu. Category knowledge-guided parameter calibration for few-shot object detection. *IEEE Transactions on Image Processing*, 32:1092–1107, 2023.
[48] Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1136–49, 2022.
[49] Shen Zhang, Fei Ye, Bingnan Wang, and Thomas G Habetler. Few-shot bearing fault diagnosis based on model-agnostic meta-learning. *IEEE Transactions on Industry Applications*, 57(5):4754–64, 2021.
[50] Xinwang Liu. Simplemkkm: Simple multiple kernel k-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5174–86, 2022.
[51] Huan Wang, Xindan Wang, Yizhuo Yang, Konstantinos Gryllias, and Zhiliang Liu. A few-shot machinery fault diagnosis framework based on self-supervised signal representation learning. *IEEE Transactions on Instrumentation and Measurement*, 2024.
[52] Chuanjiang Li, Shaobo Li, Yixiong Feng, Konstantinos Gryllias, Fengshou Gu, and Michael Pecht. Small data challenges for intelligent prognostics and health management: a review. *Artificial Intelligence Review*, 57(8):1–52, 2024.
[53] Lanjun Wan, Le Huang, Jiaen Ning, Changyun Li, and Keqin Li. A novel meta-transfer learning approach via convolutional multi-head self-attention network for few-shot fault diagnosis. *Knowledge-Based Systems*, page 112113, 2024.
[54] Xiao Zhang, Weiguo Huang, Chuancang Ding, Jun Wang, Changqing Shen, and Juanjuan Shi. Cross-supervised multisource prototypical network: A novel domain adaptation method for multi-source few-shot fault diagnosis. *Advanced Engineering Informatics*, 61:102538, 2024.
[55] Zhe Wang, Yi Ding, Te Han, Qiang Xu, Hong Yan, and Min Xie. Adaptive attention-driven few-shot learning for robust fault diagnosis. *IEEE Sensors Journal*, 2024.
[56] Ben Yang, Xuetao Zhang, Zhiping Lin, Feiping Nie, Badong Chen, and Fei Wang. Efficient and robust multiview clustering with anchor graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6200–13, 2022.
[57] Zhenglai Li, Chang Tang, Xiao Zheng, Xinwang Liu, Wei Zhang, and En Zhu. High-order correlation preserved incomplete multi-view subspace clustering. *IEEE Transactions on Image Processing*, 31:2067–80, 2022.
[58] Xiaoqian Zhang, Shuai Zhao, Jing Wang, Li Guo, Xiao Wang, and Huaijiang Sun. Purity-preserving kernel tensor low-rank learning for robust subspace clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[59] KA Loparo. Case western reserve university bearing data center. *Bearings Vibration Data Sets, Case Western Reserve University*, pages 22–8, 2012.
[60] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.
[61] Xiaonan Nie, Shijie Cao, Xupeng Miao, Lingxiao Ma, Jilong Xue, Youshan Miao, Zichao Yang, Zhi Yang, and CUI Bin. Dense-to-sparse gate for mixture-of-experts. 2021.
[62] Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023.
[63] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):5416, 2019.
[64] Miaomiao Li, Yi Zhang, Chuan Ma, Suyuan Liu, Zhe Liu, Jianping Yin, Xinwang Liu, and Qing Liao. Regularized simple multiple kernel k-means with kernel average alignment. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
[65] Xinwang Liu. Hyperparameter-free localized simple multiple kernel k-means with global optimum. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[66] Bing Song, Shuai Tan, Hongbo Shi, and Bo Zhao. Fault detection and diagnosis via standardized k nearest neighbor for multimode process. *Journal of the Taiwan Institute of Chemical Engineers*, 106:1–8, 2020.
[67] Huanjie Wang, Xiwei Bai, Jie Tan, and Jiechao Yang. Deep prototypical networks based domain adaptation for fault diagnosis. *Journal of Intelligent Manufacturing*, pages 1–11, 2022.
[68] Jinyang Jiao, Ming Zhao, Jing Lin, and Kaixuan Liang. Residual joint adaptation adversarial network for intelligent transfer fault diagnosis. *Mechanical Systems and Signal Processing*, 145:106962, 2020.
[69] Tianyuan Li, Xin Su, Wei Liu, Wei Liang, Meng-Yen Hsieh, Zhuhui Chen, XuChong Liu, and Hong Zhang. Memory-augmented meta-learning on meta-path for fast adaptation cold-start recommendation. *Connection Science*, 34(1):301–18, 2022.
[70] Tian Zhang, Jinyang Jiao, Jing Lin, Hao Li, Jiadong Hua, and Dong He. Uncertainty-based contrastive prototype-matching network towards cross-domain fault diagnosis with small data. *Knowledge-Based Systems*, 254:109651, 2022.
[71] Jian Lin, Haidong Shao, Xiangdong Zhou, Baoping Cai, and Bin Liu. Generalized maml for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals. *Expert Systems with Applications*, page 120696, 2023.
[72] Wenkang Zhou and Ning Li. Semi-supervised prototype network with cbam and data selector for few-shot bearing fault diagnosis. In *2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, pages 1–6. IEEE, 2022.
[73] Xiaosong Xing, Wei Guo, and Xuecheng Wan. An improved multidimensional distance siamese network for bearing fault diagnosis with few labelled data. In *2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing)*, pages 1–6. IEEE, 2021.
[74] Hulin Ruan, Yi Wang, Xiaomeng Li, Yi Qin, and Baoping Tang. An enhanced non-local weakly supervised fault diagnosis method for rotating machinery. *Measurement*, 189:110433, 2022.
[75] Afzal Ahmed Soomro, Masdi B Muhammad, Ainul Akmar Mokhtar, Mohamad Hanif Md Saad, Najeebullah Lashari, Muhammad Hussain, Umair Sarwar, and Abdul Sattar Palli. Insights into modern machine learning approaches for bearing fault classification: A systematic literature review. *Results in Engineering*, page 102700, 2024.