

Step-wise contrastive representation learning for diagnosing unknown defective categories in planetary gearboxes

Peng Chen ^{a,b}, Ruijin Zhang ^{a,1}, Shuai Fan ^c, Junyu Guo ^d, Xingkai Yang ^e

^a College of Engineering, Shantou University, Shantou, 515063, Guangdong, China

^b Key Laboratory of Intelligent Manufacturing Technology, Ministry of Education, Shantou, 515063, Guangdong, China

^c School of Mechanical and Electrical Engineering, Chengdu University of Technology, Chengdu, 610059, Sichuan, China

^d School of Mechatronic Engineering, Southwest Petroleum University, Chengdu, 610500, Sichuan, China

^e College of Mechanical and Vehicle Engineering, Hunan University, Changsha, 410082, Hunan, China

ARTICLE INFO

Keywords:

Planetary gearbox
Fault diagnosis
Signal processing
Contrastive learning
Semi-supervised learning

ABSTRACT

Although deep learning methods have demonstrated remarkable efficacy in the diagnostic analysis of planetary gearboxes, their practical application in industrial settings is considerably restrained by several inherent limitations. Among these critical challenges are the pronounced dependency on the quality of initial labeling, the constraints associated with loss function strategies employed in contrastive learning, the complexities involved in elucidating intra-cluster and inter-cluster dynamics, as well as the ongoing difficulties in recognizing unknown faults. In response to these challenges, this study introduces a novel methodology termed Step-wise Contrastive Representation Learning, which is specifically designed to address the issue of unidentified fault classes through the utilization of a constrained set of known fault categories. Furthermore, this research develops a comprehensive analytical framework that scrutinizes the relationships between intra-cluster and inter-cluster dynamics in both over-clustered and standard-clustered contexts, thereby facilitating enhanced optimization of model performance. In addition, a multi-step loss function strategy is proposed, encompassing stages from pre-training to advanced training phases, which significantly improves the model's adaptability to effectively capture the intricate characteristics of new or unrecognized instances. Ultimately, this framework aims to strengthen the development of a specialized classification network that enhances the accuracy and reliability of gear fault diagnosis, thereby yielding substantial improvements in generalization capabilities within dynamic industrial case studies.

1. Introduction

Planetary gearboxes are critical components utilized in a wide range of industrial applications, including wind turbines, gas turbines, electric motors, and hybrid vehicles [1–3]. The continuous operation under high-speed conditions, coupled with varying load scenarios, subjects these systems to significant mechanical stress. As a consequence, planetary gearboxes are particularly prone to a variety of damaging phenomena, such as broken or missing teeth, root cracks, and surface wear on the gear teeth. If these faults are not detected and addressed in a timely manner, they pose serious safety risks that can result in system failures or accidents. Therefore, the development and implementation of advanced fault diagnosis techniques have become not only essential but imperative. By enabling proactive maintenance interventions, these techniques help mitigate risks and contribute to the overall longevity

and efficiency of the machinery in which planetary gearboxes are integrated.

In recent years, data-driven fault diagnosis methods, particularly those incorporating deep learning techniques, have witnessed remarkable and rapid advancements [4–8]. These achievements can largely be attributed to the exceptional ability of deep learning to autonomously extract high-level representations from raw signals, thereby enabling highly accurate diagnostic predictions through end-to-end models. Over the past decade, various deep learning algorithms have gained significant traction, with notable examples including convolutional neural networks (CNNs) [9,10], generative adversarial networks (GANs) [11], and recurrent neural networks (RNNs) [12]. Each of these methods demonstrates impressive capabilities in different diagnostic applications. For instance, Han et al. [13] propose a deep transfer convolutional neural network (CNN) framework that improves diagnostic

* Corresponding author at: College of Engineering, Shantou University, Shantou, 515063, Guangdong, China.

E-mail address: pengchen@alu.uestc.edu.cn (P. Chen).

¹ These authors contributed equally to this work.

performance under conditions with limited labeled data by leveraging transfer learning and employing global average pooling (GAP) to reduce trainable parameters and mitigate the risk of overfitting. Similarly, Li et al. [14] present a novel fault diagnosis approach that integrates transfer learning with a dynamic model, enabling the extraction of domain-invariant features and achieving accurate fault classification for planetary gearboxes using both simulated and real-world data. Zhang et al. [15] introduce a nearly end-to-end deep learning method that combines Empirical Mode Decomposition (EMD) with a one-dimensional CNN to enhance fault diagnosis accuracy for wind turbine gearboxes, particularly under nonstationary operating conditions, using vibration signals. Moreover, Raghav et al. [16] propose an advanced signal processing approach combined with a deep convolutional neural network (DCNN) for fault diagnosis in spur gearboxes operating at low speeds and low loads, comparing four different image-based methods and demonstrating the superior performance of the Continuous Wavelet Transformation (CWT) method. Besides, Amiri et al. [17] propose a novel classifier that combines support vector machines, probabilistic neural networks, and deep neural networks using fuzzy systems to monitor UAV status and detect insulator faults, achieving superior performance compared to established classifiers. Lastly, Chen et al. [18] propose a deep convolutional generative adversarial network (DCGAN) scheme for health condition monitoring of wind turbine generator bearings. This innovative approach introduces a self-setting threshold mechanism, allowing the automatic definition of thresholds and overcoming the limitations of traditional threshold-setting methods, thereby enhancing diagnostic accuracy without the need for human intervention. These diverse methods highlight the growing potential of deep learning-based models in advancing fault diagnosis across various industrial applications.

In the field of semi-supervised learning applied to fault diagnosis within gearbox transmission systems, this area has garnered significant attention from researchers. Zhou et al. [19] present a novel fault detection and diagnosis framework specifically designed for gear systems, which employs a deep convolutional generative adversarial network (DCGAN) following a semi-supervised learning approach. By integrating unlabeled data into the training process, this framework significantly enhances diagnostic accuracy and capability, particularly regarding previously unseen faults, all while requiring limited labeled data. The efficacy of this approach is substantiated through systematic case studies conducted on experimental datasets. Similarly, Zhang et al. [20] propose a semi-supervised fault diagnosis method for gearboxes that effectively combines a feature pre-extraction mechanism utilizing wavelet transform with an improved generative adversarial network (IGAN). Their findings demonstrate not only enhanced diagnostic accuracy but also increased robustness against noise in gearbox fault datasets characterized by limited labeled samples and high levels of environmental noise. Furthermore, Zhao et al. [21] introduce a two-stage hybrid semi-supervised learning methodology aimed at improving fault diagnosis accuracy in rotating machinery. This approach incorporates pseudo-labeling and a novel consistency regularization mechanism to overcome challenges associated with initial model accuracy and one-dimensional data augmentation. Their results indicate a remarkable achievement of nearly 100% accuracy, even in scenarios where labeled samples are limited. Moreover, Luo et al. [22] offer an innovative Contrastive Vibration-Current (CVC) framework, which enhances fault diagnosis performance in electromechanical systems by leveraging synchronization information derived from both vibration and current signals. Through the application of advanced preprocessing and semi-supervised learning techniques, this framework significantly enhances the diagnostic capabilities of single-modality models, particularly improving the performance of current models that rely solely on single-modal signals. Fu et al. [23] introduce a semi-supervised prototype network known as TWSCE-SSPN, which proficiently diagnoses main bearing faults in tunnel boring machines under few-shot learning conditions. This method employs a novel two-stream wavelet

scattering convolutional encoder designed to enhance feature mapping, ultimately achieving significantly higher accuracy and robustness compared to existing methodologies, especially in noisy environments. Collectively, these studies underscore the potential of semi-supervised learning approaches in enhancing fault diagnosis accuracy and robustness in various mechanical systems. In short, these studies highlight the transformative potential of semi-supervised learning techniques in advancing fault diagnosis across various mechanical systems.

Contrastive learning has notably emerged as a vital area of study, offering enhanced robustness and diagnostic accuracy by revealing the intrinsic structural features of data. This is particularly valuable in unsupervised or weakly supervised scenarios, where the availability of labeled datasets is often limited due to the difficulty and expense associated with their acquisition. Notably, techniques such as self-supervised learning (SS-Learning) and semi-supervised learning have gained widespread adoption and have achieved continuous advancements across various applications. For instance, Wang et al. [24] propose a self-supervised learning (SS-Learning) framework tailored specifically for machinery fault diagnosis. This framework excels in directly learning representative features from unlabeled signals, thereby improving diagnostic performance even when labeled data are scarce. It has been demonstrated to significantly outperform traditional convolutional neural networks (CNNs) across a spectrum of real-world datasets. In a similar vein, Zhu et al. [25] introduce a self-supervised fault diagnosis approach that integrates temporal predictive and similarity contrast learning (TPSCL) with a self-attention mechanism. This method is particularly effective for health monitoring of wind turbine gearboxes, as it adeptly extracts latent fault features from unlabeled vibration signals. Consequently, it enhances diagnostic accuracy even under conditions with limited labeled data and variable operating conditions. In another significant development, Zhou et al. [19] propose a deep convolutional generative adversarial network (DCGAN) framework for the semi-supervised fault detection and diagnosis of gear systems. By leveraging a substantial amount of unlabeled data, this method not only improves diagnostic accuracy but also extends the model's ability to identify previously unseen faults that were not present in the training dataset. Additionally, Cheng et al. [26] introduce a semi-supervised fault diagnosis approach utilizing a hybrid classification network and weighted pseudo-labeling (HCN-WPL). This strategy enhances diagnostic precision by combining autoencoder-driven feature extraction with a confidence-based pseudo-labeling mechanism, thus optimizing model performance in scenarios with limited labeled data. Furthermore, Liang et al. [27] propose a different semi-supervised learning method based on reactive power signals, employing a GAF-CNN-MTDL fault diagnosis model. By transforming reactive power data into two-dimensional images, this method effectively highlights critical gear fault features, thereby enabling highly accurate fault diagnosis with minimal dependence on labeled data. As a whole, these advancements vividly illustrate the substantial potential of contrastive learning techniques. They are particularly effective in overcoming the challenges posed by limited labeled data, significantly enhancing fault diagnosis accuracy across a variety of complex industrial applications.

While the aforementioned deep learning methods have shown impressive performance in various diagnostic tasks, they also face notable challenges for practical implementation in real-world industrial applications due to certain shortcomings:

1. Dependency on Initial Labeling: Although semi-supervised learning aims to reduce the reliance on labeled data, the quality and scope of the initial labeled dataset can significantly impact model performance. If the initial labeling is biased or incomplete, the effectiveness of the semi-supervised approach may be compromised, leading to suboptimal model outcomes.
2. Limitations of Loss Function Strategies in Contrastive Learning: The loss function employed in contrastive representation learning typically operates on a one-step basis rather than employing

a multi-step framework. This characteristic creates challenges in identifying and characterizing unknown categories, as the direct approach may hinder the model's ability to adequately capture and represent the intricacies associated with novel or unrecognized category instances.

3. Challenges in Understanding Intra-cluster and Inter-cluster Dynamics: Understanding the relationships between intra-cluster and inter-cluster structures is challenging in both over-clustering and standard scenarios. This is crucial for optimizing clustering performance but requires advanced knowledge of clustering methodologies. Additionally, results can be sensitive to the assumptions made, leading to misinterpretations or over-generalizations that undermine the validity of conclusions.
4. Inability to Detect Unknown Faults: Despite the significant advancements made in fault detection methodologies, there remains a persistent challenge in recognizing entirely new fault types that were not represented in the training dataset. This limitation restricts the applicability of these approaches, particularly in dynamic environments characterized by rapidly evolving conditions, where the emergence of new fault types can occur frequently and unexpectedly.

To address the aforementioned challenges, a novel approach known as step-wise contrastive representation learning is proposed. This innovative method is specifically designed to confront the pervasive issue of unknown fault classes, effectively leveraging a limited but accessible set of known fault classes. Within the proposed framework, the dynamics of intra-cluster and inter-cluster relationships pertaining to both over-clustered and standard-clustered fault classes are thoroughly investigated. By engaging in this comprehensive exploration, the feature extraction capabilities at each stage of the model can be refined, thereby significantly enhancing the model's convergence speed and its overall ability to generalize to new data. Furthermore, this process ensures that the model learns in a productive direction, ultimately leading to the development of a classification network that is specifically trained for accurate and reliable gear fault detection.

The primary contributions of this paper can be summarized as follows:

1. Introduction of Step-wise Contrastive Representation Learning: This research proposes a novel approach called step-wise contrastive representation learning, which specifically targets the challenge of unidentified fault classes. By utilizing a limited yet accessible selection of known fault classes, this method effectively addresses unknown fault detection.
2. Comprehensive Framework for Analyzing Cluster Dynamics: A structured framework is introduced to comprehensively analyze the dynamics of intra-cluster and inter-cluster relationships, considering both over-clustered and standard-clustered fault classifications. This exploration enhances our understanding of fault relationships and aids in optimizing model performance.
3. Designation of Multi-Step Loss Function Strategies in Contrastive Learning: A multi-step framework for loss functions spanning from pre-training Stage 0 to Stage 3 is designed. This strategy aims to overcome the limitations of conventional direct methods that may hinder a model's ability to capture the complexities of novel or unrecognized category instances. By implementing a tiered loss function approach, we seek to enhance the model's adaptability and improve its performance in recognizing a wider range of data distributions.
4. Development of a Tailored Classification Network for Fault Detection: By implementing this approach, the model is guided toward more productive learning pathways, ultimately resulting in a classification network that is specifically calibrated for accurate and trustworthy gear fault detection.

This paper is systematically organized into four distinct sections, each of which plays a critical role in fostering a comprehensive understanding of the research topic at hand. In Section 2, the relevant theoretical frameworks are revisited, thereby establishing both the conceptual foundations and contextual backdrop for this investigation. Section 3 introduces and elaborates on the proposed step-wise contrastive learning model, providing an in-depth examination of its architecture and operational mechanisms. Following this, Section 4 is dedicated to the presentation and analysis of the experimental results obtained from evaluating the model across two case studies, which facilitates a critical assessment of the model's performance. Finally, in Section 5, the principal findings of the research are succinctly summarized, and the paper concludes by discussing the implications of these results, as well as suggesting potential avenues for future research.

2. Related works

The study of Novel Class Discovery (NCD) operates under the open-world assumption, which posits that data can be categorized into known and unknown types. Specifically, the complete dataset \mathcal{D} is partitioned into a labeled dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^n$ and an unlabeled dataset $\mathcal{D}_u = \{(x_i)\}_{i=1}^m$. The set of classes in the labeled dataset \mathcal{D}_l is denoted as C_l , while the set of classes in the unlabeled dataset \mathcal{D}_u is represented as C_u . Importantly, it is assumed that the intersection of these two sets of classes is non-empty, i.e., $C_l \cap C_u \neq \emptyset$, while they remain distinct, such that $C_l \neq C_u$. Within this framework, the classes corresponding to the intersection, $C_s = C_l \cap C_u$, are classified as known categories, whereas the classes represented by the difference $C_u - C_s$ are identified as new categories. Thus, the NCD method leverages a limited amount of labeled data from \mathcal{D}_l and treats all remaining data as unlabeled, thereby enabling the training of a model on \mathcal{D}_l and \mathcal{D}_u to cluster all samples. This approach allows for the classification of the entire dataset without requiring labels for the new class data found in $C_u - C_s$.

Contrastive Learning, a prominent unsupervised learning technique, fundamentally aims to enhance the distinction between dissimilar samples while simultaneously drawing similar samples closer together in the feature space. In this context, samples that exhibit similarities to a specified reference sample are termed positive pairs, while those that differ are designated as negative samples. During training, contrastive learning algorithms work to learn effective feature representations by maximizing the similarity among positive samples and minimizing the similarity among negative samples. A primary advantage of contrastive learning lies in its ability to train models without the necessity of labeled data, thus rendering it particularly useful in scenarios where acquiring samples is straightforward, but labeling remains challenging. This method has found applications across a spectrum of tasks in computer vision and natural language processing, including image classification, object detection, image generation, and text representation learning.

To establish positive and negative sample pairs, the seminal work in contrastive learning, SimCLR [28], introduces a straightforward procedure. Specifically, SimCLR generates two augmented versions of the same sample through random data augmentation, and these augmented versions, denoted as x_i and x_j , are recognized as a positive sample pair. During training, SimCLR randomly selects N samples to form a batch; each original sample undergoes two random augmentations, resulting in a total of $2N$ data points. Notably, SimCLR does not engage in explicit negative sampling; for any given sample, only the augmented counterpart from the same original image is considered a positive pair, while the remaining $2N - 2$ samples are treated as negative samples. Consequently, the final loss function is formulated as follows:

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^N (l_{2k,2k-1} + l_{2k-1,2k}) \quad (1)$$

$$l_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp(\text{sim}(z_i, z_j)/\tau)} \quad (2)$$

$$\text{sim}(z_i, z_j) = \frac{z_i^T z_j}{\tau \|z_i\| \|z_j\|} \quad (3)$$

where $\mathbb{I}_{[k \neq i]}$ denotes the indicator function, which yields a value of 1 only when k is not equal to i . The parameter τ represents the temperature coefficient, while z_i and z_j refer to the outputs, as depicted in the figure above.

Since its introduction, the clustering approach employed by SimCLR has yielded promising results. However, it is worth noting that its performance is relatively limited when compared to recent advancements in the field, primarily due to its separate treatment of labeled and unlabeled data. One notable methodology, DTC [29], addresses this limitation through the application of transfer learning. In this approach, a base model is first learned from the labeled data and subsequently adapted to recognize samples from new classes. Moreover, ORCA [30, 31] presents a similar problem formulation aligned with the open-world assumptions outlined in this study. Its primary innovations involve the integration of adaptive supervised learning, the implementation of pairwise objectives, and the application of model regularization techniques. Although the problem setting in IIC [32] differs somewhat from that presented in this paper—wherein D_l and D_u are defined as two disjoint categories—it is possible to adapt the latter framework by considering D_l as a subset of D_u during testing. IIC primarily utilizes symmetrical Kullback–Leibler (KL) divergence to evaluate inter-class and intra-class similarities, optimizing a loss function based on this metric. However, it is essential to note that the scope of inter-class similarity control defined by IIC tends to be broad, focusing primarily on amplifying the distinction between known and new classes without specifically addressing the disparities among samples within each individual class. Another significant advancement, OpenNCD [33], innovatively employs a two-layer contrastive learning framework. This methodology strategically utilizes a greater number of prototypes than the actual number of classes in order to perform classification and incrementally clusters these prototypes throughout the training phase. Consequently, both prototype-level and prototype-group level similarities are delineated for improved performance.

In addition to these methodologies, active learning constitutes a pivotal branch of semi-supervised learning that grapples with the challenge of a vast number of unlabeled samples while requiring labels for only a limited number. Active learning focuses on selecting the most informative samples for labeling, thereby minimizing the number of labeled samples needed while preserving model performance. The core principle of active learning is to prioritize labeling samples characterized by the highest uncertainty. These uncertain samples are typically positioned near decision boundaries in the feature space, and the supervision signal derived from these samples offers more informative insights compared to randomly selected samples. Such samples are often referred to as hard samples, and the concept parallels the difficulties encountered by humans during their own learning processes. Addressing challenging problems often necessitates enhanced logical reasoning capabilities, thereby facilitating deeper knowledge acquisition; similarly, this principle applies to machine learning frameworks.

3. The proposed step-wise contrastive representation learning

This section elucidates the proposed step-wise contrastive representation learning methodology, seen in Fig. 1, which is designed to unearth latent categories in the domain of gearbox fault diagnosis. The model designs a step-wise two-level contrast learning approach that operates on multiple prototypes. By leveraging a limited set of known class faults within signal samples, this method facilitates the identification of both known and unknown class faults in real-world signal samples. Consequently, this approach addresses the challenging problem of unknown class fault identification in open-world signal

data, which has hitherto been a significant obstacle in the field of fault diagnosis. By employing this advanced contrastive learning technique, the model effectively learns to distinguish between various fault categories, even those not explicitly presented during the training phase. This is achieved through the careful construction of a feature space where similar faults cluster together, while dissimilar faults are pushed apart. As a result, the model can generalize its learned representations to novel fault classes, thereby significantly improving its utility in real-world applications.

The proposed model's overall architecture comprises an encoder E and a prototype system, which consists of two trainable tensors, denoted as G_{coarse} and G_{fine} . It is important to note that the number of columns in each tensor corresponds to the dimensionality of the encoder's output. The two tensors serve distinct yet complementary purposes: G_{coarse} is employed to facilitate the grouping of samples into broad classes, while G_{fine} is utilized for more granular classification into smaller, more specific classes.

To elaborate further, G_{coarse} , which is responsible for coarse-grained classification, is structured as a tensor with dimensions (n, dim) . In this context, n represents the initial number of classes defined by the model, effectively determining the number of categories into which the data will be assigned. The parameter dim corresponds to the dimensionality of the feature vector produced by encoder E . Each row within G_{coarse} can be conceptualized as a prototype, essentially serving as a representative model for the feature vector extracted from a particular class of samples after processing through the encoder. Consequently, the final classification process involves first extracting features via the encoder, followed by computing the similarity between the sample's feature vector and each prototype. This computation yields an n -dimensional vector, which effectively represents the similarity (or assignment probability) between the sample and each prototype. The classification outcome is then determined by identifying the prototype with the highest similarity score. To enhance the normalization of computational outcomes and ensure consistency in subsequent processes, this research introduces slight modifications to the similarity (assignment probability) computation. Nonetheless, it is important to note that these adjustments do not alter the underlying principle, which remains firmly based on the cosine similarity methodology. Specifically, the assignment probability p_{coarse} is computed as follows:

$$p_{coarse} = \frac{\exp\left(\frac{1}{\tau} feature \cdot G_{coarse}^T\right)}{\sum \exp\left(\frac{1}{\tau} feature \cdot G_{coarse}^T\right)} \quad (4)$$

where $feature$ denotes the feature vector extracted from the sample by the encoder, and τ represents the temperature coefficient, which modulates the softness of the probability distribution.

For G_{fine} , which is employed for fine-grained classification, it is observed that it is structured as a tensor with dimensions $(n * k, dim)$. Here, k signifies the over-clustering factor, allowing for a more nuanced categorization of samples. It is worth noting that the operational principles of G_{fine} closely mirror those of G_{coarse} . The feature vector, when processed through G_{fine} , also yields an assignment probability. The key distinction lies in the dimensionality of the resulting probability vector: while G_{coarse} produces an n -dimensional vector, G_{fine} generates an $n * k$ -dimensional vector, reflecting its capacity for more fine-grained classification. The calculation of the assignment probability p_{fine} follows a similar formula:

$$p_{fine} = \frac{\exp\left(\frac{1}{\tau} feature \cdot G_{fine}^T\right)}{\sum \exp\left(\frac{1}{\tau} feature \cdot G_{fine}^T\right)} \quad (5)$$

The learning process of the model is divided into four distinct phases, each representing a different level of contrastive learning. These phases are systematically designed to incrementally enhance the model's capacity to differentiate between various fault categories. It is important to highlight that the first phase utilizes a purely unsupervised learning approach, guided by a relatively simple loss function.

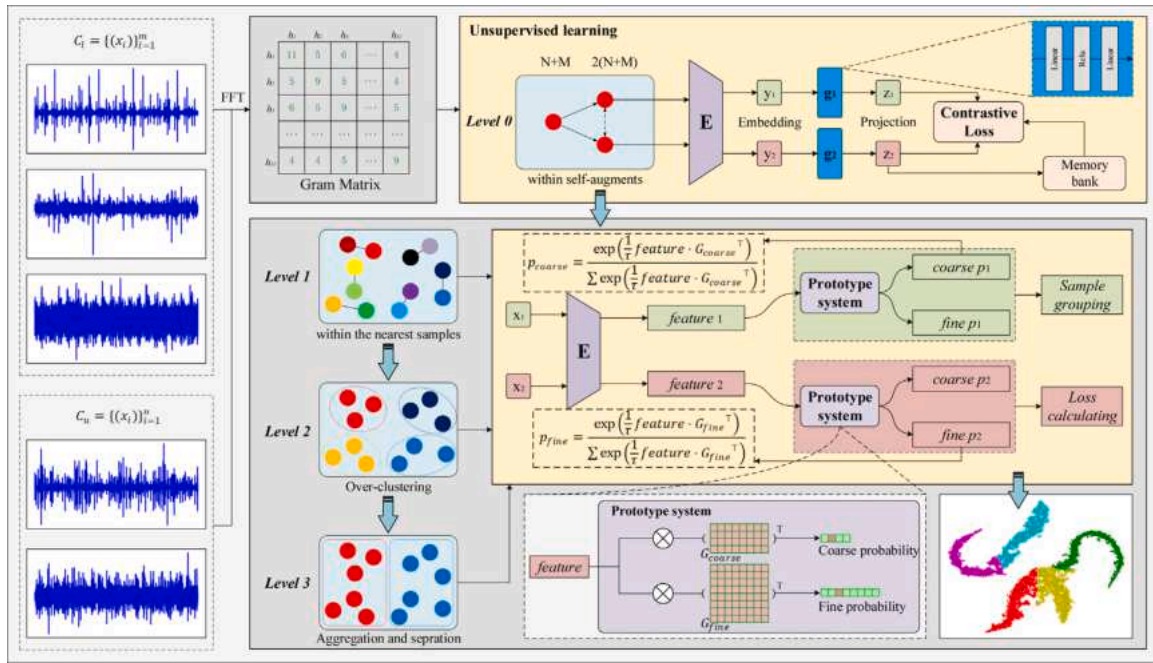


Fig. 1. The proposed architecture of step-wise contrastive representation learning.

However, the subsequent three phases adopt semi-supervised learning techniques, incorporating increasingly sophisticated optimization strategies to refine the model’s performance.

3.1. The learning process of pre-training stage 0: Unsupervised learning

The Stage 0, which is based on Level 0 contrastive learning, focuses on fundamental unsupervised learning. This stage essentially adopts the contrastive learning methodology of SimCLR for model pretraining. During this phase, only the encoder E is trained, leaving the subsequent prototype system untouched. The process involves generating two augmented images from each sample to create positive pairs. These augmented images are then fed into the encoder E to obtain two feature vectors. All other samples within a training batch are treated as negative samples. Following the data augmentation process, each original sample yields two enhanced images, transforming the initial $N + M$ samples into a $2(N + M)$ dataset. It is important to note that only the augmented images are utilized in this stage, not the original ones.

Within this $2(N + M)$ dataset, each data point forms a positive pair with another data point derived from the same source, while simultaneously serving as a negative pair with the remaining $2(N + M) - 2$ data points. The loss function for this stage is formulated as follows:

$$\mathcal{L}_0 = \frac{1}{2(N + M)} \sum_{k=1}^{N+M} (I_{2k,2k-1} + I_{2k-1,2k}) \quad (6)$$

$$I_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2(N+M)} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \quad (7)$$

$$\text{sim}(z_i, z_j) = \frac{z_i^T z_j}{\tau \|z_i\| \|z_j\|} \quad (8)$$

where z_i and z_j represent the feature vectors obtained from two augmented images generated from the same sample after feature extraction by encoder E . The parameter τ denotes the temperature coefficient, which controls the concentration level of the distribution. The function $\mathbb{1}_{[k \neq i]}$ is an indicator function whose range is $[0, 1]$, taking the value 1 if and only if $k \neq i$.

3.2. The learning process of training stage 1: Semi-supervised learning

The initial stage, referred to as Stage 1 and based on Level 1, initiates the training of both G_{coarse} and G_{fine} through the implementation of a contrastive learning scheme. This stage employs a multifaceted loss function that encompasses several key components, each serving a distinct purpose in the learning process. Specifically, the total loss function at this stage is comprised of three primary elements: the cross-entropy loss for labeled data, the similarity loss for positive pairs of samples, and a distribution regularization term. It is worth noting that the distribution regularization term is further subdivided into two components: a large class distribution regularization term and a small class distribution regularization term.

To begin with, the Cross-Entropy (CE) loss function [34] is applied directly to the labeled data. However, it is important to recognize that the provided true labels correspond to large class groups rather than small class groups. Consequently, the cross-entropy loss can only be calculated using the p_{coarse} obtained after the features are processed through G_{coarse} , in conjunction with the true labels. As a result, this particular loss function is exclusively utilized for the training of G_{coarse} . The cross-entropy loss function \mathcal{L}_{CE} is mathematically expressed as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{k=1}^N \sum_{i \in D_i} y_i \log(p_{coarse_i}) \quad (9)$$

where y_i represents the correct label of sample i , p_{coarse_i} denotes the standard assignment probability of sample i , and N represents the total number of labeled samples.

Subsequently, a loss term corresponding to contrastive learning at Level 1 is designed, which is specifically trained on G_{fine} . In contrast to traditional approaches that use two augmented maps generated from a single sample as positive pairs, this research employs the nearest neighbor samples of a given sample as positive pairs. This research is justified by the fact that after the Stage 1 pretraining, the model’s encoder has developed a certain level of representation ability, ensuring a high similarity between adjacent samples. Therefore, the selection of nearest neighbor samples as positive pairs does not impede the model’s learning process due to excessive error information. Furthermore, the inherent differences between nearest neighbor samples are greater than

those between two augmented graphs of the same sample, which facilitates the model's ability to learn useful features. The nearest neighbor similarity loss function \mathcal{L}_{sim} is defined as:

$$\mathcal{L}_{sim}^{stg1} = -\frac{1}{N+M} \sum_{i \in D} \log(\text{sim}(p_{fine_i}, p'_{fine_i})) \quad (10)$$

where sim represents the cosine distance between two vectors, while p_{fine_i} and p'_{fine_i} denote the probability assigned to the minority class of sample i and its corresponding sample, respectively.

Although the aforementioned loss functions provide guidance for sample learning, they do not impose restrictions on the number of classes into which unlabeled samples are divided. Instead, they merely cluster similar samples into adjacent spaces, potentially leading to a scenario where all unlabeled samples are clustered into a single class (model collapse). This phenomenon manifests when multiple prototypes are left idle, with few samples assigned to them. To mitigate this issue, the optimal approach would be to use the Kullback–Leibler (KL) [34] divergence between the sample distribution derived from the current model and the prior distribution of the data as a loss function. The distribution regularization term loss function \mathcal{L}_{reg} is expressed as:

$$\mathcal{L}_{reg} = KL(p_{prior} \parallel \text{mean}(p_{coarse})) \quad (11)$$

where p_{prior} represents the prior distribution of samples.

However, obtaining the prior distribution of samples is often challenging in practice. To address this, this research employs a uniform distribution as a substitute for the sample's prior distribution. While this assumption may not be entirely accurate, empirical evidence from our experiments suggests that it does not significantly impact the model's final performance, unless the dataset contains a large number of examples for each type. For instance, the dataset utilized in this study does not conform to a uniform distribution, yet it still yields favorable results using this method. The large class grouping regularization term \mathcal{L}_{reg0} and the small class grouping regularization term \mathcal{L}_{reg1} are defined as follows:

$$\begin{aligned} \mathcal{L}_{reg0} &= KL(Q \parallel \text{mean}(p_{coarse})) \\ \mathcal{L}_{reg1} &= KL(Q \parallel \text{mean}(p_{fine})) \end{aligned} \quad (12)$$

where Q represents the uniform distribution. These regularization terms serve to guide the model in evenly distributing all samples across each prototype, thereby preventing model collapse.

Finally, the comprehensive loss function for this stage is formulated as:

$$\mathcal{L}_1 = \mathcal{L}_{CE} + \mathcal{L}_{sim}^{stg1} + \alpha \mathcal{L}_{reg0} + \beta \mathcal{L}_{reg1} \quad (13)$$

Eq. (13) encapsulates the multifaceted approach to learning in Stage 1, incorporating cross-entropy loss, similarity loss, and both large and small class distribution regularization terms, with α and β serving as weighting factors for the respective regularization terms.

3.3. The learning process of training stage 2: Semi-supervised learning in small classes

The second stage of pre-training, denoted as Stage 2 and based on Level 2, is characterized by its focus on achieving two primary objectives: enhancing the aggregation within small classes and promoting separation between small classes that belong to the same large class. To accomplish these goals, the model's representation capabilities are further refined at the Level 2 stage by innovatively treating samples within the same small class, along with other samples in that class, as positive pairs. This approach facilitates the separation of small classes within the same larger class, effectively removing samples that do not belong to the class in question.

It is important to note that this stage retains the use of the cross-entropy loss \mathcal{L}_{CE} and the distribution regularization terms \mathcal{L}_{reg0} and \mathcal{L}_{reg1} , as previously employed in Eq. (13). However, to fully realize the objectives of this stage, two novel loss functions are introduced, each serving a distinct purpose in the learning process.

The first of these new loss functions, termed the intra-small class similarity loss, utilizes the assignment probability derived from the initial enhanced sample as the basis for the current model's classification of the sample. Specifically, p_{fine} is employed to calculate the probability of assigning the sample to each minor class, with the minor class exhibiting the highest probability being selected as the classification result. This process effectively partitions all samples into nk subsets, denoted as $g_1, g_2, g_3, \dots, g_{(nk)}$. Samples within the same subset g_i are considered to belong to the same small class. The corresponding loss function, \mathcal{L}_{sim} , is mathematically expressed as follows:

$$\mathcal{L}_{sim}^{stg2} = -\frac{1}{n * k} \sum_{i=1}^{n*k} \sum_{j \in g_i} \log(\text{sim}(p_{fine_j}, p_{fine_{j'}})) \quad (14)$$

where j' represents a randomly selected sample belonging to the same small class as j .

The second novel loss function designed in this stage leverages the concept of Kullback–Leibler (KL) divergence, which measures the information loss when one probability distribution is used to approximate another. However, due to the asymmetric nature of KL divergence, this research employs a symmetric variant, known as symmetric KL divergence (SKL). The SKL considers bidirectional information flow between two probability distributions and is particularly well-suited for measuring the similarity or difference between two probability distributions. It is important to emphasize that the goal here is not to calculate the similarity between the predicted distribution and the true distribution, but rather to quantify the difference between the probabilities assigned to two samples belonging to the same cluster. The SKL is mathematically defined as:

$$SKL(p \parallel q) = \frac{1}{2} (KL(p \parallel q) + KL(q \parallel p)) \quad (15)$$

Building upon above, the research employs the average negative SKL calculated between each small class within the same large class as a loss function. This approach guides the model to differentiate and separate each small class within the same large class. The process begins by dividing the samples into n classes $G_1, G_2, G_3, \dots, G_n$ according to p_{coarse} . Subsequently, each large class G_i is further subdivided into several small classes $g_{i1}, g_{i2}, g_{i3}, \dots$. The average distribution probability of the small class within each large class is calculated as follows:

$$p_{ij} = \frac{1}{l} \sum_{j \in g_{ij}} p_{fine_j} \quad (16)$$

where g_{ij} denotes the j th small class of the i th large class, and l represents the number of samples belonging to the j th small class of the i th large class.

The average loss for the i th large class is then computed using the following formula:

$$\mathcal{L}_{SKLi} = -\frac{1}{C_{class}^2} \sum SKL(p_{ix} \parallel p_{iy}) \quad (17)$$

where p_{ix} and p_{iy} represent the average assignment probabilities of any two small classes within the same large class, $class$ denotes the number of sub-classes in the large class, and C_{class}^2 signifies the number of possible combinations.

Finally, the average symmetric KL divergence loss for all large classes is calculated as:

$$\mathcal{L}_{SKL}^{stg2} = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{SKLi} \quad (18)$$

Incorporating all these components, the comprehensive loss function for this stage is formulated as:

$$\mathcal{L}_2 = \mathcal{L}_{CE} + \mathcal{L}_{sim}^{stg2} + \alpha \mathcal{L}_{reg0} + \beta \mathcal{L}_{reg1} + \gamma \mathcal{L}_{SKL}^{stg2} \quad (19)$$

It is crucial to note that the parameters α , β , and γ serve as weighting factors for their respective loss components, allowing for fine-tuning of the model's learning process.

3.4. The learning process of training stage 3: Semi-supervised learning in large classes

The third and final stage of the step-wise contrastive learning process, denoted as Stage 3 and based on Level 3, represents the culmination of the pre-training phases. This stage is primarily focused on two critical objectives: enhancing the aggregation within large classes and promoting separation between these large classes. It is noteworthy that Stage 3 shares significant similarities with Stage 2, and as such, it continues to employ the cross-entropy loss \mathcal{L}_{CE} and the distribution regularization terms \mathcal{L}_{reg0} and \mathcal{L}_{reg1} , as previously utilized in Eq. (13). However, a key distinction in this stage lies in the redefinition of the similarity loss \mathcal{L}_{sim} and the symmetric KL divergence loss \mathcal{L}_{SKL} , which are tailored to address the specific goals of this final pre-training phase.

To begin with, the model's classification of samples in this stage is based on the assignment probability derived from the first augmented sample. Specifically, p_{coarse} is employed to calculate the probability of assigning each sample to each class, with the class exhibiting the highest probability being selected as the classification result. This process effectively partitions all samples into n distinct subsets, denoted as $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \dots, \mathcal{G}_n$. Samples within the same subset \mathcal{G}_i are considered to belong to the same broad class. The corresponding loss function, termed the large intra-class similarity loss \mathcal{L}_{sim}^{stg3} , is mathematically expressed as follows:

$$\mathcal{L}_{sim}^{stg3} = -\frac{1}{n} \sum_{i=1}^n \sum_{j \in \mathcal{G}_i} \log(\text{sim}(p_{coarse_j}, p_{coarse_{j'}})) \quad (20)$$

where j' represents a randomly selected sample belonging to the same large class as j .

Furthermore, a large within-class dissimilarity loss is introduced, which is defined by calculating the symmetric KL divergence between all pairs of large classes. This approach aims to maximize the dissimilarity between different large classes, thereby enhancing the model's ability to distinguish between broad categories. To implement this, the average distribution probability of each large class is first calculated as follows:

$$p_i = \frac{1}{l} \sum_{j \in \mathcal{G}_i} p_{coarse_j} \quad (21)$$

where \mathcal{G}_i represents all samples of the i th large class, and l denotes the number of samples in this large class.

The corresponding loss function, termed the intra large class dissimilarity loss \mathcal{L}_{SKL}^{stg3} , is then formulated as:

$$\mathcal{L}_{SKL}^{stg3} = -\frac{1}{C_n^2} \sum SKL(p_i \| p_j) \quad (22)$$

where p_i and p_j represent any two large classes, respectively, and C_n^2 denotes the number of possible pairwise combinations of large classes.

The comprehensive loss function for this final stage is thus defined as:

$$\mathcal{L}_3 = \mathcal{L}_{CE} + \mathcal{L}_{sim}^{stg3} + \alpha \mathcal{L}_{reg0} + \beta \mathcal{L}_{reg1} + \gamma \mathcal{L}_{SKL}^{stg3} \quad (23)$$

where the parameters α , β , and γ serve as weighting factors for their respective loss components, allowing for fine-tuning of the model's learning process.

Therefore, based on the discussions presented above, the corresponding pseudo algorithm designed to implement the proposed step-wise contrastive representation learning methodology is detailed in Algorithm 1.

Algorithm 1 Step-wise contrastive representation learning framework

Stage1: Unsupervised pre-training

Input: Gram Matrices dataset: $D = \{X_i\}_{i=1}^N$

- 1: Create augments for P_t : $D_p = \{X_i, X'_i\}_{i=1}^N \xleftarrow{\text{transformation}} D$
- 2: Initialize the parameter $\theta_p \xrightarrow{\text{initialize}} f_p(\theta)$
- 3: **while** not converge **do** Train on P_t
- 4: Calculate the output: $l_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2(N+M)} \mathbb{1}_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)}$
- 5: Calculate the loss $\mathcal{L}_0 = \frac{1}{2(N+M)} \sum_{k=1}^{N+M} (l_{2k, 2k-1} + l_{2k-1, 2k})$
- 6: Update θ_p by back propagation
- 7: **end while**
- 8: **Output:** optimized parameter θ_p for P_t

Stage2: Step-wise Contrastive Learning

9: **for** $t = 1$ to 3 **do**

10: **Input:** Gram Matrices dataset: $D_t = \{X_i, Y_i\}_{i=1}^M$

11: Initialize the network: $f_t(\theta) : \theta_t \xleftarrow{\text{weights-sharing}} \theta_p$

12: Calculate coarse classification probability: $p_{coarse} = \frac{\exp(\frac{1}{\tau} \text{feature} \cdot \mathcal{G}_{coarse}^\top)}{\sum \exp(\frac{1}{\tau} \text{feature} \cdot \mathcal{G}_{coarse}^\top)}$

13: Calculate fine classification probability: $p_{fine} = \frac{\exp(\frac{1}{\tau} \text{feature} \cdot \mathcal{G}_{fine}^\top)}{\sum \exp(\frac{1}{\tau} \text{feature} \cdot \mathcal{G}_{fine}^\top)}$

14: **if** $t == 1$: Calculate the loss: $\mathcal{L}_t = \mathcal{L}_{CE} + \mathcal{L}_{sim}^{stgt} + \alpha \mathcal{L}_{reg0} + \beta \mathcal{L}_{reg1}$

15: **else:** Calculate the loss: $\mathcal{L}_t = \mathcal{L}_{CE} + \mathcal{L}_{sim}^{stgt} + \alpha \mathcal{L}_{reg0} + \beta \mathcal{L}_{reg1} + \gamma \mathcal{L}_{SKL}^{stgt}$

16: **end for**

Output: Make predictions for D_t

4. Experiments

To validate the performance of the proposed method, SCRL, this section presents two comprehensive case studies that specifically focus on the diagnosis of gearbox transmission systems. In this research, comparative analyses are undertaken against several established diagnostic models, including DTC, IIC, ORCA, and OpenNCD. Key evaluation metrics—such as accuracy, Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), t-distributed Stochastic Neighbor Embedding (t-SNE), and confusion matrices (CM)—are utilized throughout these case studies to facilitate a thorough and systematic comparative assessment of the diagnostic methodologies. Furthermore, an ablation study is conducted to investigate the optimal hyperparameter selections that may enhance the robust performance of the proposed SCRL model. By employing these comprehensive metrics, the research is able to evaluate the effectiveness of the proposed method in relation to the selected benchmark models.

4.1. Case study I

4.1.1. Experimental apparatus and data acquisition

The research employs the Drivetrain Prognostics Simulator (DPS), a sophisticated apparatus developed by SpectraQuest Inc., as illustrated in Fig. 2. This intricate system comprises multiple interrelated components, each playing a crucial role in facilitating its advanced operational capabilities. Among these components are a variable speed drive motor, a planetary gearbox system, a two-stage parallel gearbox system, resistance-load gearboxes that are connected to a resistance-load inducing electric load motor, and an electric control unit that governs the entire assembly. A comprehensive overview of the physical parameters pertaining to the planetary gearbox system is detailed in Table 1.

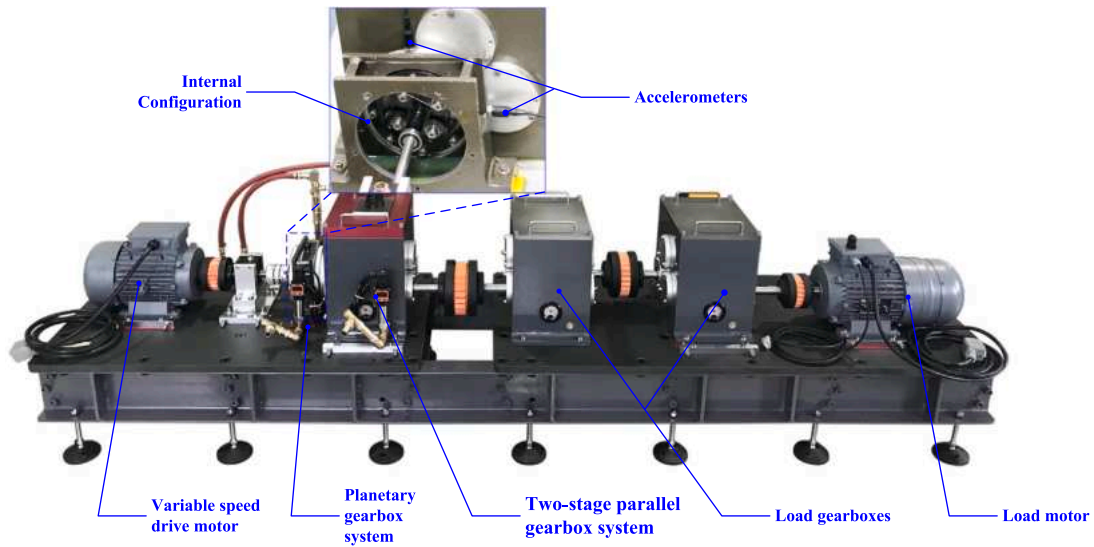


Fig. 2. Illustration of the Drivetrain Prognostics Simulation (DPS).

Table 1
Physical parameters of the planetary gear set in DPS.

Parameters	Sun	Planet (4)	Ring	Carrier
Number of teeth	28	36	100	–
Module [mm]	1	1	1	–
Pressure angle [°]	20	20	20	–
Face width [mm]	10	10	10	–
Young's modulus [Pa]	2.1×10^{11}	2.1×10^{11}	2.1×10^{11}	–
Poisson's ratio	0.3	0.3	0.3	–
Mass [kg]	–	–	9.86×10^{-2}	–
Moment of inertia [kg·m ²]	2.41×10^{-6}	1.60×10^{-5}	9.20×10^{-3}	4.99×10^4
Base circle [mm]	13.2	16.9	47.0	–
Torsional stiffness [N/m]	0	–	1×10^9	0
Torsional damping [N·s/m]	0	–	1×10^3	0

The primary focus of this study is on the planetary gearbox system, which incorporates spur gears, in conjunction with the two-stage parallel gearbox system, characterized by the use of spur gears. Furthermore, it is essential to note that the dataset utilized in Case Study I comprises fault samples classified into four distinct categories: Chipped, Crack, Intact, Missing, and Worn. Within this experimental framework, the vibration signals extracted from the planetary gearbox transmission system are carefully selected for subsequent analysis. These vibration signals are acquired while the drive motor is operated at a rotational frequency of 30 Hz. Throughout the course of this study, the gearbox is maintained at a consistent operational rotational speed, whereas the sampling frequency is established at 30,720 Hz. The data collection procedure extends over a total sampling duration of 2.13 s, resulting in the acquisition of 65,536 samples. This meticulous data collection approach enables a thorough examination of the gearbox's performance under predefined operating conditions.

4.1.2. Data pre-processing

Data pre-processing initiates with the transformation of the raw vibration dataset into a Gram Matrix, which is configured with a pixel size of 32 by 32. This transformation is essential for performing various feature enhancement techniques, including image manipulation methods such as cutting and flipping, that are applied to the data. The process begins with the utilization of the Fourier Transform (FT), which converts each time-domain signal, captured through a sliding window approach, into its corresponding frequency-domain vector. Subsequently, the frequency-domain vector from each segment sequence is subjected to a dot product operation with the frequency-domain vectors of other sequences, thereby yielding the final Gram

Matrix. In this study, five distinct types of fault data correspond to the generated Gram Matrices: Chipped, Crack, Intact, Missing, and Worn.

Once the Gram Matrix images containing effective features are acquired through the aforementioned processing techniques, the resultant dataset is introduced into a step-wise contrastive learning methodology. Among the five types of data samples, three types—Chipped, Crack, and Intact—are classified as known classes and denoted as C_l . Conversely, the remaining two types of data, Missing and Worn, are categorized as unknown classes and represented as C_u . In this allocation, 10% of the samples from the known class C_l are designated as the labeled dataset, referred to as D_l . The remaining 90% of the data from C_l , in addition to all samples from the unknown classes C_u , collectively form the unlabeled dataset, termed D_u . Consequently, all samples are defined within the overarching dataset $D = D_l + D_u$. Finally, the size of the labeled sample set is denoted as $\mathcal{N} = \text{len}(D_l)$, and the size of the unlabeled sample set is expressed as $\mathcal{M} = \text{len}(D_u)$.

4.1.3. Experimental validation and comparative analysis for diagnosing planetary gearboxes

To rigorously assess the effectiveness and advantages of the proposed Step-wise Contrastive Representation Learning (SCRL) model, a comprehensive validation is conducted utilizing the aforementioned dataset of planetary gearboxes. For a thorough comparative analysis, several related state-of-the-art semi-supervised learning techniques, including DTC [29], ORCA [30,31], IIC [32], and OpenNCD [33], are employed as benchmarks. These models are thoughtfully selected due to their established performance and relevance within the domain, providing a solid foundation for meaningful comparison.

Baseline: The models such as DTC, ORCA, IIC, and OpenNCD are utilized as baseline models. Notably, both ORCA and OpenNCD employ

Table 2
Comparison of accuracy across baseline models.

Methods	All	Label	Unlabel	NMI	ARI
DTC	67.26%	69.39%	81.89%	0.4728	0.3951
IIC	<u>78.70%</u>	62.88%	90.54%	0.5288	<u>0.5813</u>
ORCA	70.93%	88.62%	62.91%	0.6628	0.5073
OpenNCD	72.89%	76.87%	65.24%	0.4946	0.4761
SCRL(Ours)	92.13%	94.01%	<u>86.17%</u>	0.6703	0.7299

the same dataset partition structure as described in this study. However, the dataset partitions for DTC and IIC treat known classes and new classes as completely disjoint sets, focusing solely on identifying the new classes. To adapt these two models for recognizing known classes, samples belonging to the known classes are treated as part of the new class, and performance metrics such as accuracy are calculated and reported in the same manner as for the normal new classes. All models operate using 10% labeled data from the known classes (see Table 2).

Evaluation Metrics: Various evaluation metrics are employed to assess the accuracy of the model regarding known classes, new classes, full class accuracy, Normalized Mutual Information (NMI) [35], and the Adjusted Rand Index (ARI) [36]. (1) **Labeled Classes Accuracy:** This metric evaluates the classification accuracy of samples that belong to known classes (the top four classes). These samples are extracted, and their classification accuracy is subsequently calculated. (2) **Unlabeled Classes Accuracy:** This metric pertains to samples associated with new categories (the last three categories) that do not have real labels. Since these new categories do not provide actual labels, only clustering results are analyzed. For this purpose, the Hungarian algorithm is utilized to compute the maximum matching scheme, and accuracy is then calculated based on this matching scheme. (3) **All Classes Accuracy:** This metric evaluates the overall accuracy by using the Hungarian algorithm to determine the maximum matching scheme according to the clustering results of all categories of samples, and accuracy is calculated in accordance with this matching scheme. (4) **Normalized Mutual Information (NMI):** NMI measures the correlation between two random variables, indicating the degree of information shared between the clustering results and the true labels. The NMI value ranges from [0,1], where a value closer to 1 denotes a higher degree of similarity between the clustering outcome and the true class labels, whereas a value approaching 0 indicates lower similarity. (5) **Adjusted Rand Index (ARI):** ARI assesses the agreement between the clustering results and the true class labels. The ARI value ranges from [-1,1], with a value nearer to 1 indicating a higher level of agreement, a value around 0 suggesting a random match, and a value closer to -1 reflecting inconsistency between the clustering results and the true class labels.

Implementation Details: In the proposed method, ResNet-18 serves as the backbone feature extractor. After completing training in stage 0, the parameters of the first three blocks of ResNet-18 are fixed, allowing only the last block and the mapping head to be fine-tuned during subsequent stages (stage 1, stage 2, and stage 3). The output dimension from ResNet-18 is 32. The over-clustering factor is set to $k = 5$, meaning that the model will partition all samples into $5 \times 7 = 35$ clusters. The optimizer selected for this process is Adam, with the learning rate established at 0.05. The batch size utilized during training is set to 512. The temperature coefficient is defined as $\tau = 5$; the weights for \mathcal{L}_{reg0} and \mathcal{L}_{reg1} are specified as $\{\alpha, \beta\} = \{10, 10\}$; and the weight parameter for \mathcal{L}_{SKL} is designated as $\gamma = 0.1$. The total training duration consists of 400 epochs, wherein stage 0 runs for 300 epochs, stage 1 for 20 epochs, stage 2 for 30 epochs, and stage 3 for 50 epochs.

Ablation Study: This analysis aims to investigate the contributions of various components of the loss function employed in the proposed methodology through a series of carefully designed ablation experiments. Specifically, this study examines the effects of the similarity loss \mathcal{L}_{sim} and the Kullback–Leibler divergence loss \mathcal{L}_{SKL} in both stages 2

Table 3
Ablation study.

Methods	All	Label	Unlabel
w/o \mathcal{L}_{sim}^{stg2}	78.80%	72.52%	63.97%
w/o \mathcal{L}_{SKL}^{stg3}	71.72%	82.84%	77.12%
w/o \mathcal{L}_{sim}^{stg3}	79.51%	<u>86.72%</u>	69.85%
w/o \mathcal{L}_{SKL}^{stg2}	<u>82.28%</u>	86.07%	<u>77.23%</u>
SCRL (Ours)	92.13%	94.01%	86.17%

and 3 of the model. For the sake of clarity and precision, the terms \mathcal{L}_{sim}^{stg2} and \mathcal{L}_{SKL}^{stg2} are defined to represent the similarity loss and inter-class dissimilarity loss, respectively, during stage 2. In a similar manner, the losses in stage 3 are designated as \mathcal{L}_{sim}^{stg3} and \mathcal{L}_{SKL}^{stg3} . To comprehensively assess the significance of these individual components, a series of ablation studies are conducted, systematically removing each term from the loss function while closely monitoring the resulting impact on model performance. In the associated Table 3, the notation w/o is used to indicate the exclusion of a specific term from the overall model configuration, facilitating a clearer understanding of its role in performance outcomes.

The results presented in Table 3 demonstrate that the various modules incorporated in this study contribute positively to the overall functionality and performance of the model. Notably, the removal of \mathcal{L}_{sim}^{stg3} leads to a substantial degradation in performance metrics. This finding is particularly significant, considering that stage 3 constitutes the final phase of contrastive learning within the model, and it plays a crucial role in facilitating the aggregation of large categories. Furthermore, stage 3 establishes the foundational basis for the ultimate classification of samples. Consequently, the inclusion of \mathcal{L}_{sim}^{stg3} is essential for ensuring effective model functionality and achieving optimal performance outcomes.

t-SNE Visualization: t-SNE (t-Distributed Stochastic Neighbor Embedding) represents a sophisticated dimensionality reduction technique that is predominantly employed for visualizing high-dimensional data, whereby complex mathematical transformations preserve the essential topological relationships between data points. Although several dimensionality reduction methods exist, t-SNE has gained particular prominence because its primary objective is to faithfully map high-dimensional data into more manageable two- or three-dimensional representations, thereby facilitating intuitive understanding of complex data distribution patterns. When a model demonstrates strong clustering capabilities in the original high-dimensional space, the t-SNE projection effectively reveals inherent data clusters, structural similarities, and potential categorical relationships, which consequently enhances the clarity and interpretability of the data's underlying structure. Furthermore, unlike linear dimensionality reduction techniques such as PCA, t-SNE's ability to preserve local structure while maintaining global patterns makes it especially valuable for analyzing complex, nonlinear relationships in modern machine learning applications and exploratory data analysis.

The t-SNE visualization results, as illustrated in Fig. 3, demonstrate that the proposed SCRL method effectively characterizes features into distinctly separated clusters that correspond to each fault category. This categorization encompasses not only the known classes, such as chipped, cracked, and intact but also extends to include unknown classes like missing and worn. These visualization outcomes indicate that the SCRL model excels in capturing and accurately representing the underlying patterns embedded within the data pertaining to the known classes. Moreover, it exhibits a remarkable adaptability for diagnosing faults even when faced with unknown classes. The compactness of data points within each known fault class demonstrates a strong intra-class similarity, which further reinforces the effectiveness of the SCRL model. Simultaneously, the clear separation observed between different clusters underscores the model's proficiency in differentiating among various fault types, thereby enhancing diagnostic accuracy. In contrast,

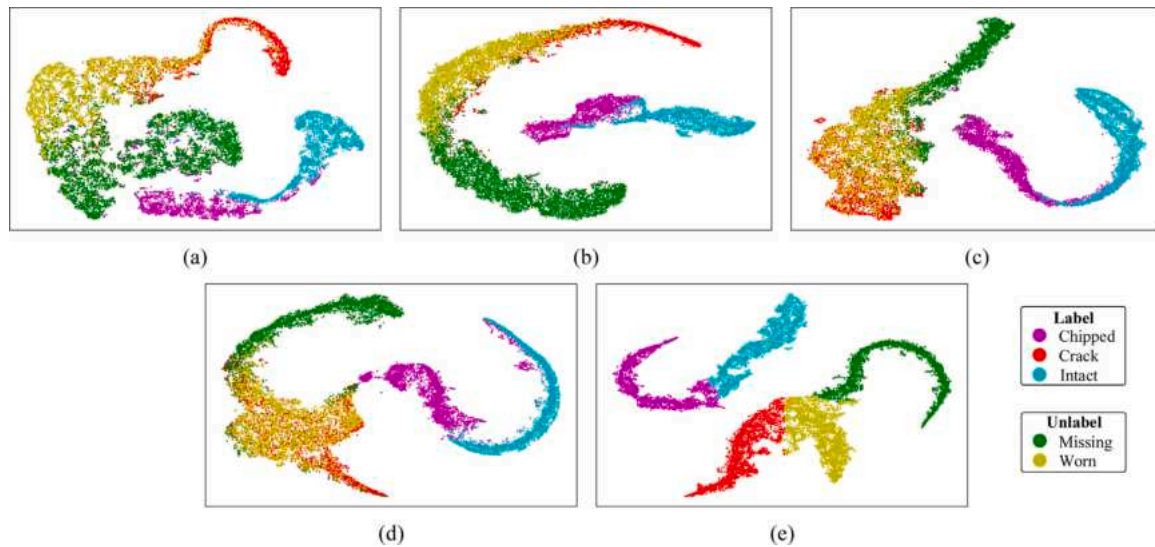


Fig. 3. t-SNE visualization of learned feature representation for Case I with 60% known classes (10% labeled) and 40% unknown classes on (a) DTC, (b) IIC, (c) ORCA, (d) OpenNCD and (e) SCRL (Ours). Colors represent classes.

t-SNE visualizations derived from semi-supervised learning models, including DTC and IIC, tend to exhibit overlapping clusters, particularly in the presence of unknown fault classes. This overlapping phenomenon indicates a substantial decline in the effectiveness of feature extraction and classification performance when compared to the SCRL method. Furthermore, in the context of the ORCA model, while certain known classes such as chipped and intact may be identifiable through well-distributed clusters, significant challenges arise with identifying classes like cracked, which often overlap with unknown classes such as missing and worn. This overlap further complicates the diagnostic process. Importantly, failures from known classes, such as cracked, are frequently misidentified as belonging to unknown classes, including missing and worn, which highlights the model's inherent limitations in effectively generalizing across a diverse range of data instances. In contrast, the t-SNE visualization associated with the SCRL method consistently demonstrates clear separability for both known and unknown faults. This outcome not only reinforces the method's superior classification capability but also its robust diagnostic capabilities, particularly in addressing the intricate complexities involved in fault identification and characterization.

Confusion Matrix: The confusion matrix represents a fundamental and sophisticated visual tool that is extensively employed in machine learning to systematically evaluate the performance of classification models, thereby providing crucial insights into the model's predictive accuracy by meticulously illustrating the correspondence between predicted and actual labels. While conceptually straightforward, yet powerful in its application, the confusion matrix is conventionally represented as a square matrix wherein predicted categories are displayed along one dimension and actual categories along the other, with binary classification tasks typically utilizing a standard two-dimensional format that facilitates immediate interpretation. Furthermore, the confusion matrix serves as a foundational framework from which multiple essential performance metrics—such as accuracy, precision, recall, and F1 score—can be systematically derived, thus enabling researchers and practitioners to comprehensively assess the model's effectiveness across various operational contexts and use cases. Moreover, this analytical tool proves particularly valuable when dealing with imbalanced datasets, as it explicitly reveals both the successes and failures of the model across different classes, thereby offering insights that might otherwise be obscured by simpler evaluation metrics.

To perform a more rigorous quantitative analysis of the diagnostic outcomes, a Confusion Matrix (CM) is utilized. This case study presents results that are depicted in Fig. 4. The data reveal that, when the

proposed SCRL method is trained with 3000 samples per category, it achieves high accuracy in identifying all fault types. Specifically, the method attains an accuracy of 90.5% for identifying chipped components, while the accuracy for identifying intact components is notably higher at 97.6%. Additionally, the SCRL method detects cracks with an accuracy of 91.8% and identifies missing components with an accuracy of 89.2%. Conversely, the semi-supervised learning model known as DTC demonstrates limited effectiveness in fault identification. It consistently misclassifies chipped and crack faults, accurately identifying only the missing and worn category. Furthermore, the IIC model shows only marginal accuracy in distinguishing known class faults associated with gear failures. Its accuracy rates for identifying chipped, crack, and intact faults are alarmingly low, at 67.3%, 45.3%, and 76.3%, respectively. Notably, it achieves a 85.4% accuracy for missing components and a 99.2% accuracy for worn components. Moreover, while other semi-supervised models such as ORCA and OpenNCD show relatively high accuracy in recognizing known class faults, they exhibit significant shortcomings in accurately identifying unknown class faults, indicating a notable deficiency in their generalization capabilities. For instance, the ORCA model misidentifies missing and worn components, achieving limited accuracy rates of 32.0% and 0.0%, respectively. However, it accurately identifies cracks and intact components with over 94.8% accuracy. Similarly, the OpenNCD model performs well in identifying crack and worn faults, with accuracy rates of 92.8% and 97.3%, respectively. In contrast, its accuracy for other fault type, missing components, remains alarmingly low at 37.4%. Overall, the proposed SCRL model exhibits superior performance across all fault categories and accurately identifies all five types of faults, resulting in an impressive average accuracy rate. This performance emphasizes the effectiveness of the SCRL approach in fault diagnosis, highlighting its potential applicability in real-world scenarios where precise fault identification is crucial for ensuring operational reliability and safety.

Comparison of Key Hyper-Parameter Tuning Performance : The performance of the proposed model is systematically evaluated, SCRL, by examining the effects of various hyper-parameters, specifically the over-clustering factor k and the weight γ associated with the symmetric Kullback–Leibler (KL) divergence. The subsequent analysis details how adjustments to these hyper-parameters influence model accuracy, with results summarized in Table 4.

Initially, the influence of the over-clustering factor k is assessed by setting it to several distinct values: $k = 2$, $k = 3$, $k = 4$, $k = 5$, and $k = 6$. The performance outcomes associated with these various k settings are depicted in Fig. 5(a). The findings indicate that the

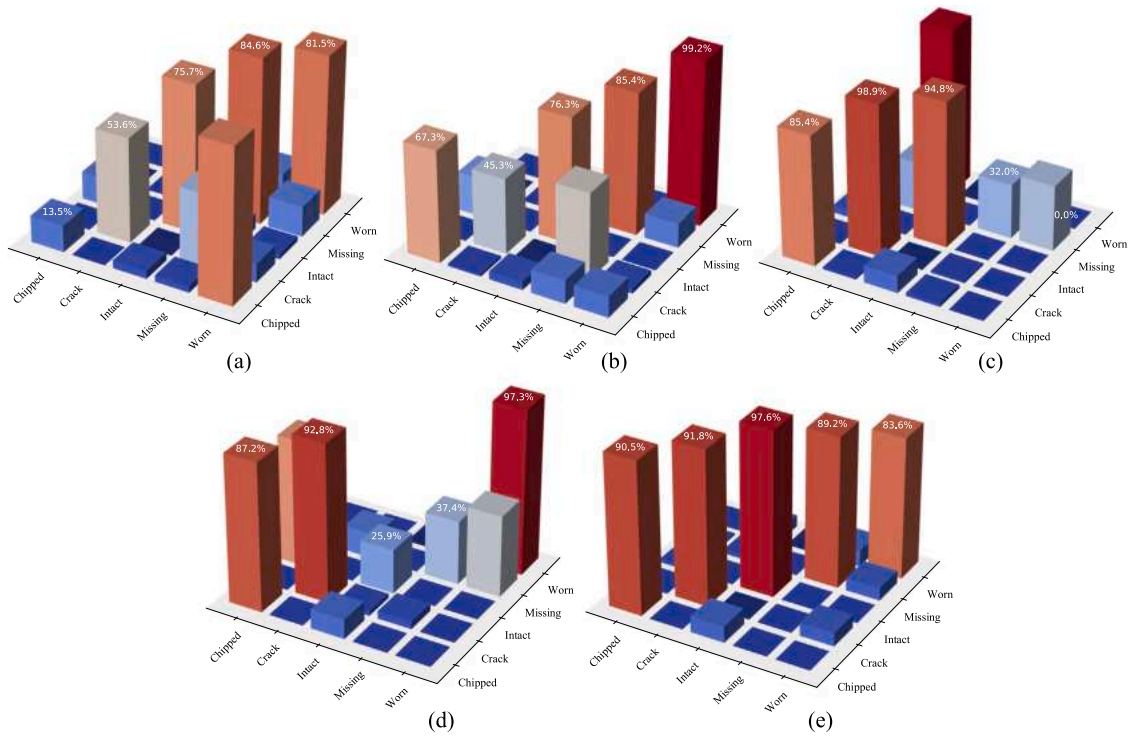


Fig. 4. The classifying performance through Confusion Matrices of learned feature representation for Case I with 60% known classes (10% labeled) and 40% unknown classes on (a) DTC, (b) IIC, (c) ORCA, (d) OpenNCD and (e) SCRL (Ours).

Table 4
Comparison of key hyperparameter tuning performance: Over-clustering factor k and symmetric KL divergence weight γ .

k	Over-clustering factor			γ	Symmetric KL divergence		
	All	Label	Unlabel		All	Label	Unlabel
2	84.41%	<u>90.04%</u>	76.85%	0.00	81.40%	87.21%	73.62%
3	86.15%	87.49%	<u>84.35%</u>	0.02	74.66%	<u>89.50%</u>	54.75%
4	92.13%	94.01%	86.17%	0.05	84.87%	88.49%	<u>80.02%</u>
5	85.49%	90.00%	79.45%	0.10	92.13%	94.01%	86.17%
6	<u>86.55%</u>	89.20%	83.62%	0.15	83.73%	88.44%	77.42%

model’s accuracy remains relatively stable within the range of $k = 2$ to $k = 4$ and reaches an optimal peak when $k = 4$. However, as k exceeds 5, there is a marked decline in model performance. This deterioration can be attributed to the role of k in regulating the degree of over-clustering. Specifically, an excessively high value of k results in the division of samples into an excessive number of smaller clusters. Ideally, this should not negatively impact model performance; however, practical constraints inherent to the hardware limit the dataset’s learning process to batch operations. Notably, the internal similarity calculations for these smaller clusters rely on the same batch of data, while the batch size employed in this study is 512. Consequently, an overly large number of small clusters yields too few samples within each cluster, which ultimately restricts the information available to the model and leads to a reduction in performance. Therefore, setting the over-clustering factor k to approximately 4 is strongly recommended.

Furthermore, the model’s performance is investigated by varying the parameter γ , specifically, by evaluating the settings $\gamma = 0$, $\gamma = 0.02$, $\gamma = 0.05$, $\gamma = 0.1$, and $\gamma = 0.15$. As illustrated in Fig. 5(b), it becomes evident that the accuracy of fault classification for known classes is relatively insensitive to variations in γ . However, γ significantly impacts the classification accuracy of unknown class faults. The symmetric KL divergence is designed to assist in distinguishing between different classes. When γ is set too low, the training process during Stage 2 fails to effectively segregate heterogeneous classes present within the same larger class, which is an essential function of this stage. Neglecting

this task effectively causes the model to leap prematurely from Stage 1 to Stage 3, resulting in decreased performance. Conversely, when γ is excessively high, it generates a substantial repulsive effect among the smaller classes within the larger class during the training of Stage 2. This can lead to frequent fluctuations in the model’s classification structure for samples and subsequently increase the risk of model non-convergence. Hence, selecting an appropriate value for γ is crucial for optimizing the model’s performance. Based on the results presented above, a value of approximately 0.1 is deemed most effective.

4.2. Case study II

To further validate the performance of the proposed method, a second case study is conducted in this section. Similar to the first case study, this experimental validation focuses on the diagnosis of gearboxes. In this context, comparative analyses are performed against several established diagnostic models, including DTC, IIC, ORCA, and OpenNCD. Key performance indicators such as accuracy, t-distributed stochastic neighbor embedding (t-SNE) analyses, and confusion matrices (CM), which are illustrated and discussed in case study I, are utilized in this study to facilitate a thorough comparative assessment. By employing these metrics, the effectiveness of the proposed method can be rigorously evaluated against the selected benchmarking models, thereby enhancing our understanding of its diagnostic capabilities in the context of gearbox condition monitoring.

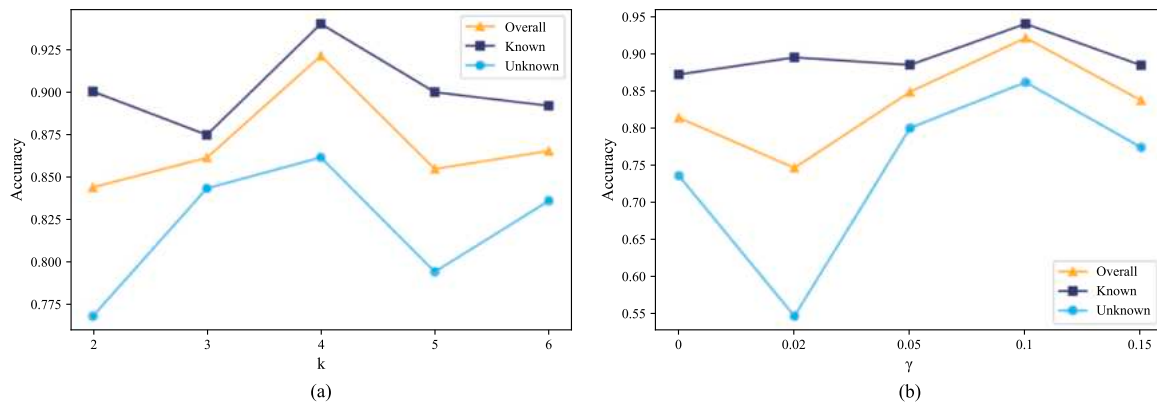


Fig. 5. Effect of the over-clustering factor k (a) and symmetric KL divergence weight γ (b) on model accuracy.

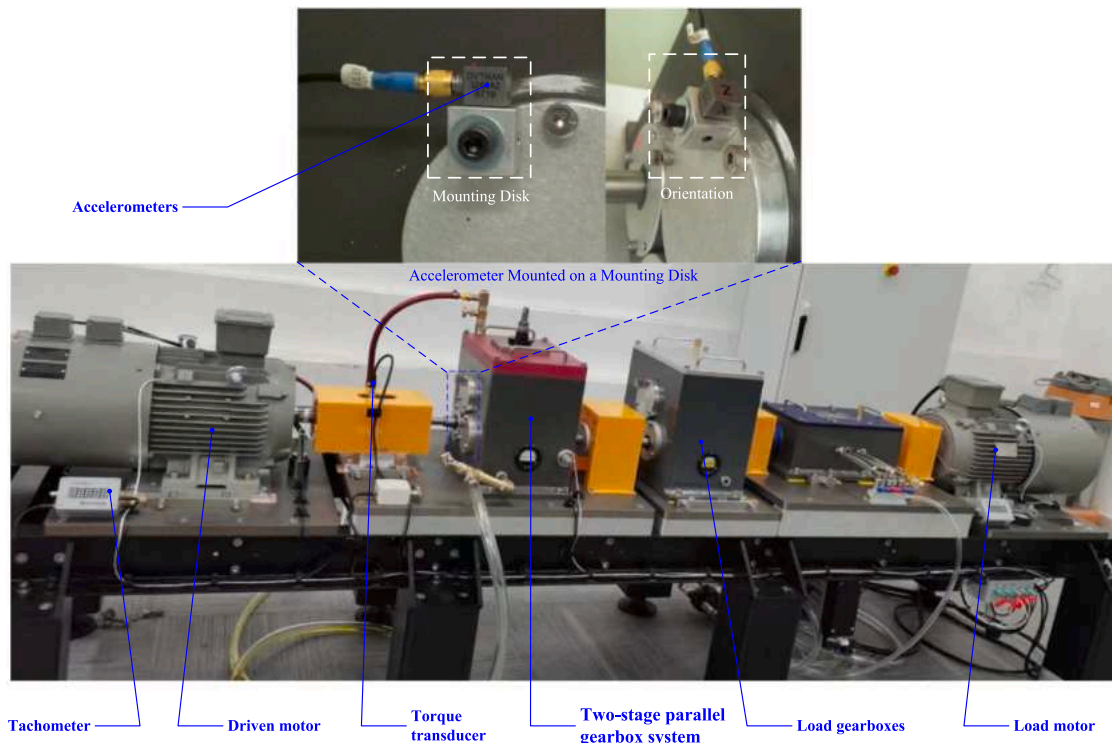


Fig. 6. Experimental test-rig of gear transmission system.

4.2.1. Specifications for data description and test-rig

The gearbox dataset, which has been meticulously gathered from a gear transmission system, offers a comprehensive representation of various operational conditions and fault types. This system, as illustrated in Fig. 6, includes several principal components: a tachometer, a driven motor, a torque transducer, a two-stage parallel gearbox system, load gearboxes, and a load motor. Notably, the placement of the accelerometer is of particular significance, as it is mounted on a separate disk. For a more detailed examination, this configuration is depicted in an enlarged view in the zoomed section of Fig. 6. To accurately capture the dynamics of the system, the dataset is sampled at a frequency of 12.8 kHz. Furthermore, it encompasses a range of operational conditions, with rotational speeds systematically varied from 1600 to 2400 revolutions per minute (r/min). Besides normal operating conditions, the dataset incorporates five common gear fault types, as depicted in Fig. 7. These include: “miss”, which denotes a missing tooth; “chipped”, indicating chipped tooth; “root”, referring to a crack at the tooth root; and “eccentric”, which involves misaligned

geometric and rotational centers. The gear meshing configuration is portrayed in Fig. 8(a), while Fig. 8(b) provides a view of the internal configuration of the parallel gearbox system, where the faulty gear is clearly marked with a dotted box for easy identification. To facilitate in-depth gear diagnostic analysis, vibration data is collected along the x -axis of the accelerometer while the gear rotates at a constant speed of 1600 r/min. Each category, including the healthy condition, comprises 768,000 data points collected over a 60-second duration. This extensive dataset serves as a robust foundation for the development and validation of fault diagnostic algorithms, thereby enabling researchers to investigate a wide range of operational conditions and fault types within controlled experimental settings.

4.2.2. Results and comparative analysis of case study II

As discussed similarly in case study I, several models, namely DTC, ORCA, IIC, and OpenNCD, are employed as baseline models for comparative analysis. Notably, ORCA and OpenNCD utilize the same dataset partitioning structure as delineated in this study, thereby ensuring consistency throughout the evaluation process. In contrast, the partitioning



Fig. 7. (a) Miss (missing tooth), (b) Chipped (cracked teeth), (d) Root (crack at tooth root), (e) Eccentric (misaligned geometric and rotational centers).

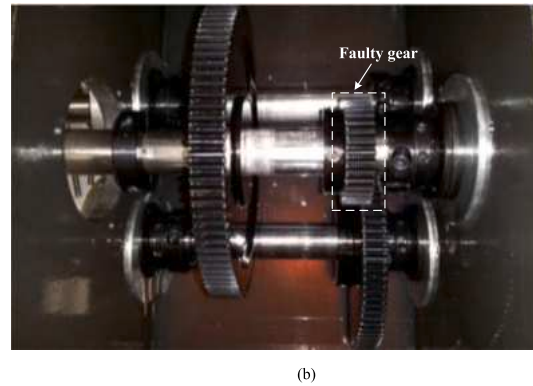
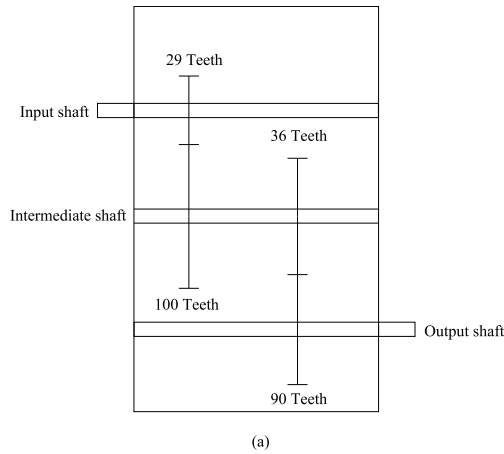


Fig. 8. (a) Gear meshing, (b) Internal configuration of parallel gearbox system.

Table 5
The results of comparative analysis in Case study II.

Methods	All	Label	Unlabel	NMI	ARI
DTC	61.93%	77.29%	59.57%	<u>0.5878</u>	0.4751
IIC	73.41%	76.19%	<u>72.35%</u>	0.3966	0.2010
ORCA	65.82%	73.84%	66.19%	0.4901	0.4504
OpenNCD	<u>80.89%</u>	<u>86.87%</u>	65.24%	0.4946	<u>0.4761</u>
SCRL(Ours)	91.90%	93.60%	84.20%	0.6156	0.6164

approach for DTC and IIC considers known classes and new classes as entirely separate, disjoint sets, which allows these models to focus exclusively on the identification of new classes. To enable DTC and IIC to also recognize known classes, samples from these known classes are strategically reclassified as part of the new class category. This adaptation facilitates a more comprehensive evaluation of the models' capabilities across different scenarios. As a result, performance metrics, such as accuracy, are uniformly calculated and reported for both known and new classes, ensuring that comparisons across all models are valid and meaningful. Furthermore, it is important to note that all models operate using only 10% of labeled data from the known classes. This highlights the challenge of relying on a limited dataset while striving to achieve robust performance in diagnostic tasks, thereby demonstrating the models' effectiveness in leveraging sparse labeled information (see Table 5).

The t-SNE visualization results, as depicted in Fig. 9, indicate that the features are distinctly and well-separated into clusters corresponding to each fault category when the proposed SCRL method is employed. This visualization demonstrates that the SCRL model excels in capturing and representing the underlying patterns embedded within the data, especially when compared with established models such as DTC, ORCA, IIC, and OpenNCD. The compactness of the data points within each known fault class suggests a strong intra-class similarity, whereas the clear separation between different clusters underscores the model's ability to differentiate among various fault types. In contrast, t-SNE visualizations for semi-supervised learning models like DTC and IIC often display overlapping clusters, particularly among unknown

fault classes. This overlap indicates a reduction in the effectiveness of feature extraction and classification performance compared to the SCRL method. Additionally, while models like OpenNCD may show well-distributed clusters for known fault classes, the presence of dispersed or overlapping points associated with unknown faults highlights their limited capability to generalize effectively across diverse data instances. Regarding ORCA, the features captured from known classes, such as chipped and eccentric, tend to overlap, except for the health class, which can be easily identified. However, unknown classes like miss and root cannot be classified effectively. In contrast, the t-SNE visualization associated with the SCRL method is expected to demonstrate clear separability not only for known faults but also for unknown faults. This result reinforces the method's superior classification and diagnostic capabilities, particularly in addressing the complexities involved in fault identification and characterization.

To conduct a more rigorous quantitative analysis of diagnostic outcomes, a Confusion Matrix (CM) is employed. In this context, the case study presents results that are visually represented in Fig. 10. The data indicate that when the proposed SCRL method is trained utilizing 3000 samples per category, it achieves remarkable accuracy in identifying various fault types. Specifically, the method achieves an accuracy rate of 89.9% in identifying missing components. The accuracy for identifying healthy components is even higher, reaching 99.2%. Furthermore, the SCRL method detects chipped components with an accuracy of 82.7% and identifies eccentric components with an accuracy of 80.8%. In contrast, the semi-supervised learning model known as DTC demonstrates limited effectiveness in fault identification, consistently misclassifying eccentric, healthy, missing, and root faults. Moreover, the IIC model exhibits high accuracy in distinguishing known class faults associated with healthy components, although its accuracy rates for identifying chipped, eccentric, and root components are lower, at 74.3%, 66.1%, and 63.8% respectively. Notably, it achieves an accuracy of 90.6% for healthy components. Additionally, while other semi-supervised models such as ORCA and OpenNCD show relatively high accuracy in recognizing known class faults, they display significant shortcomings in accurately identifying unknown class faults. This indicates a notable limitation in their generalization capabilities.

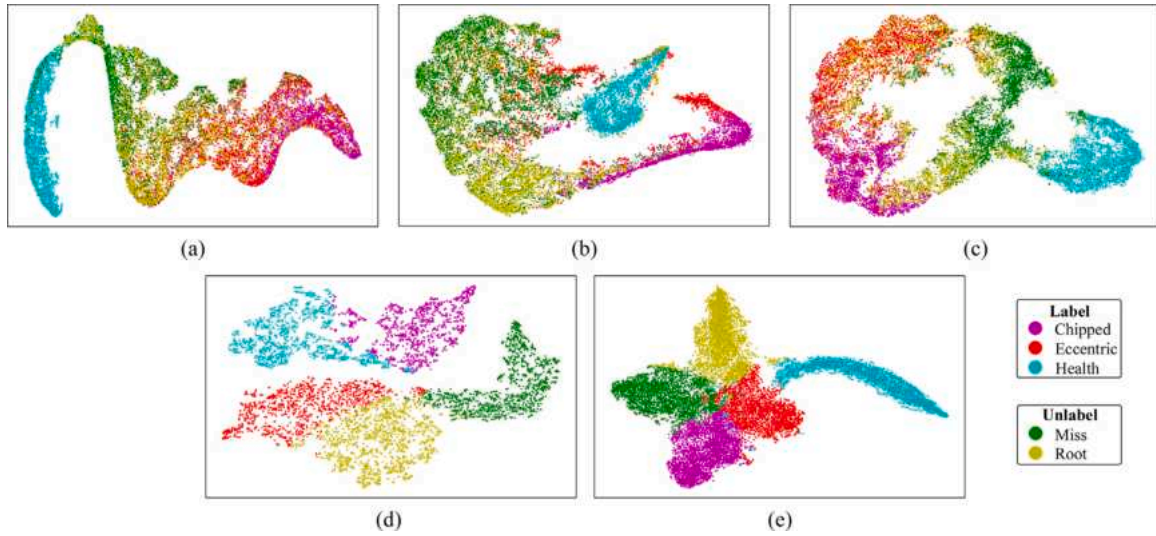


Fig. 9. t-SNE visualization of learned feature representation for Case II with 60% known classes (10% labeled) and 40% unknown classes on (a) DTC, (b) IIC, (c) ORCA, (d) OpenNCD and (e) SCRL (Ours). Colors represent classes.

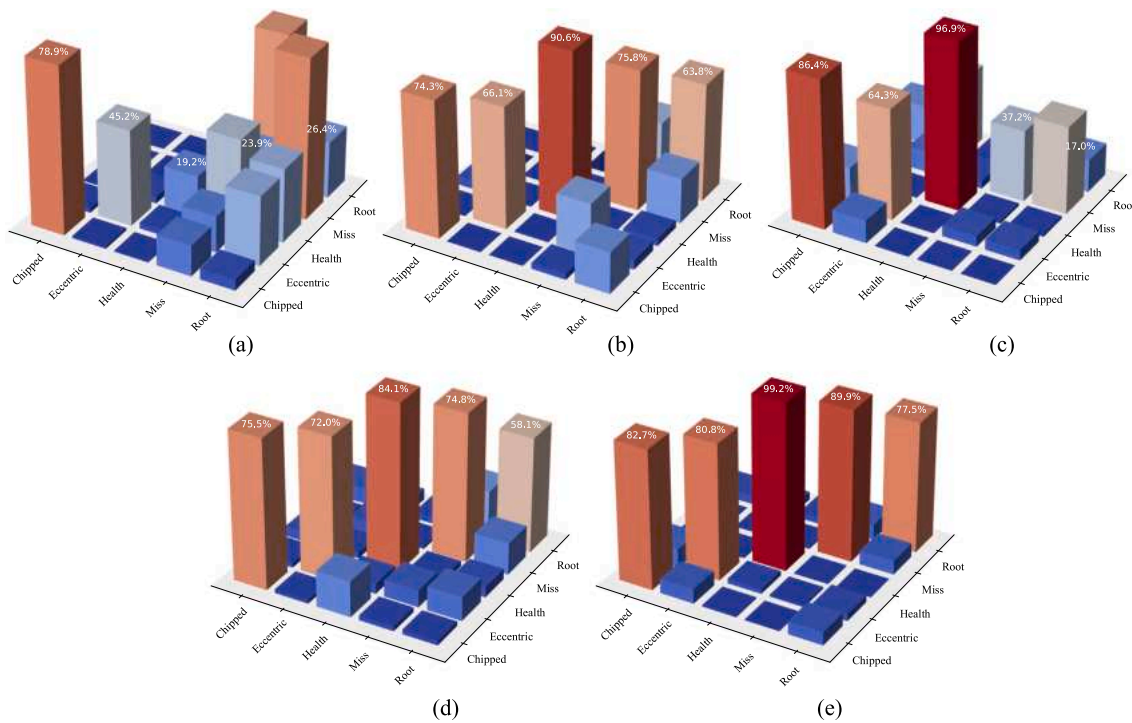


Fig. 10. The classifying performance through Confusion Matrices of learned feature representation for Case II with 60% known classes (10% labeled) and 40% unknown classes on (a) DTC, (b) IIC, (c) ORCA, (d) OpenNCD and (e) SCRL (Ours).

For instance, the ORCA model misidentifies miss and root components, yielding limited accuracy rates of 37.2% and 17.0%, respectively. However, it accurately identifies chipped and healthy components with accuracy rates of 86.4% and 96.9%. Similarly, the OpenNCD model performs well in identifying chipped and miss faults, achieving accuracy rates of 75.5% and 74.8%, respectively, whereas its accuracy for other fault types, such as eccentric and root components, remains low at 72.0% and 58.1%. Overall, the proposed SCRL model demonstrates superior performance across all fault categories, accurately identifying all five types of faults and resulting in an impressive average accuracy rate.

4.3. Statistical assessment of classifiers

In this section, a comprehensive statistical analysis framework is presented that employs both the Friedman test and the subsequent Nemenyi post hoc test [37,38] to systematically evaluate the statistical significance of observed performance variations across multiple algorithms with respect to accuracy, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) metrics (Table 6). The analytical procedure begins with the implementation of the Friedman test, which methodically evaluates and ranks each algorithm’s performance across individual datasets, thereby establishing a hierarchical performance

Table 6
Ordinal rankings of superior outcomes achieved by five methods in comparative evaluation.

Methods	Metric	Dataset		Mean rank
		Case I	Case II	
DTC	ACC	5	5	5.0
	NMI	5	2	3.5
	ARI	5	3	4.0
IIC	ACC	2	3	2.5
	NMI	3	5	4.0
	ARI	2	5	3.5
ORCA	ACC	4	4	4.0
	NMI	2	4	3.0
	ARI	3	4	3.5
OpenNCD	ACC	3	2	2.5
	NMI	4	3	3.5
	ARI	4	2	3.0
SCRL(Ours)	ACC	1	1	1.0
	NMI	1	1	1.0
	ARI	1	1	1.0

structure. Subsequently, the Nemenyi test is deployed to conduct pairwise comparisons of the algorithms' mean ranks, utilizing calculations predicated on the χ^2_F distribution with $k-1$ degrees of freedom, where k denotes the total number of algorithms under examination. To establish precise statistical distinctions among the algorithmic approaches, a rigorous post hoc analysis becomes imperative for identifying specific algorithm pairs that contribute to these significant variations. In this analytical context, the Nemenyi test serves as a robust statistical tool for revealing substantive performance differentials, particularly when the mean rank disparities between algorithm pairs exceed predetermined critical thresholds. This research systematically calculates mean ranks for each algorithm within the comparative framework, encompassing $k = 5$ distinct algorithms evaluated across $n = 2$ datasets. The Friedman test yields χ^2_F values of 7.6, 4.4, and 4.4 for accuracy, NMI, and ARI metrics, respectively. Given the experimental parameters of $4(k - 1)$ degrees of freedom, an established significance level α level of 0.05, and a critical threshold value of 6.388 for the Friedman test (specifically $k = 5$ and $n = 2$), it can be concluded that statistically significant differences exist in accuracy measurements ($7.6 > 6.388$), while NMI and ARI metrics do not exhibit statistically significant variations ($4.4 < 6.388$). This rejection of the null hypothesis for accuracy measurements provides a strong statistical foundation for conducting subsequent post hoc analyses.

The Nemenyi test is implemented to conduct comprehensive pairwise comparisons among all classifier combinations. Within the parameters of our experimental design, where $k = 5$ and $\alpha = 0.1$, the critical difference threshold CD_α is calculated as 3.8880, with a corresponding q_α value of 2.459. Through this detailed statistical analysis, it is demonstrated that the accuracy performance of the proposed methodological approach exhibits statistically significant differences when compared to existing approaches including DTC, IIC, ORCA, and OpenNCD. The complete statistical comparison framework and its results are visually represented in Fig. 11, which illustrates the relative performance rankings and significant differences identified through the Nemenyi test analysis.

5. Conclusion and future work

This research offers a novel solution to the limitations faced by deep learning methods in diagnosing planetary gearboxes within industrial settings. By tackling issues such as dependency on initial labeling, limitations in loss function strategies, and the challenge of detecting unknown faults, the introduction of Step-wise Contrastive Representation Learning marks a step forward. The development of a comprehensive framework analyzing intra-cluster and inter-cluster dynamics, combined with a multi-step loss function strategy, enhances

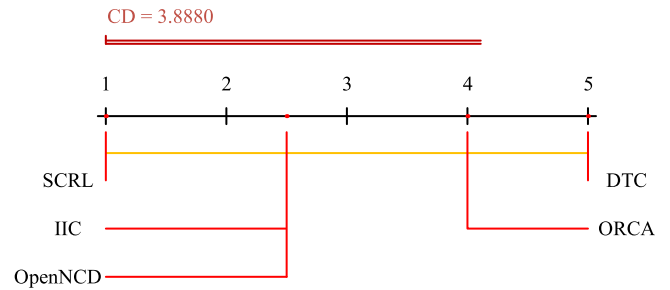


Fig. 11. Accuracy rankings and statistical comparisons: Mean ranks from Friedman test and Nemenyi Post-hoc analysis results.

model adaptability and performance. These innovations not only improve the accuracy and reliability of gear fault diagnosis but also provide substantial gains in generalization capabilities. This study thereby demonstrates potential for significant advancements in the deployment of resilient and effective diagnostic systems in complex industrial environments.

While the Step-wise Contrastive Representation Learning approach has demonstrated considerable promise in addressing fundamental limitations inherent to deep learning methodologies, it is important to note that the experimental validation conducted in this research predominantly focuses on steady-state operational conditions. Although these initial results are encouraging, the complex and dynamic nature of industrial environments necessitates a more comprehensive evaluation framework. Therefore, future work should systematically examine the framework's robustness and generalizability under more challenging scenarios, including but not limited to: variable load conditions, continuously fluctuating speed profiles that better represent actual operational patterns, and diverse environmental factors such as temperature gradients and mechanical vibrations that are omnipresent in industrial settings. By expanding the scope of validation to encompass these multifaceted operational conditions, researchers can not only better assess the model's practical applicability but also establish its reliability across a broader field of industrial applications and environmental contexts, thus bridging the gap between theoretical advancement and practical implementation.

CRedit authorship contribution statement

Peng Chen: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Ruijin Zhang:** Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Shuai Fan:** Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Junyu Guo:** Validation, Software, Resources, Methodology, Investigation, Formal analysis. **Xingkai Yang:** Validation, Software, Methodology, Investigation, Formal analysis.

Declaration of competing interest

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

Acknowledgments

Partial funding for this research has been provided by several sources, including the National Natural Science Foundation of China through Grant 52105111, the Basic and Applied Basic Research Foundation of Guangdong Province through Grant 2022A1515010859, the

Guangdong Provincial Science and Technology Special Fund Project through Grant STKJ2021171, the Shantou University (STU) Scientific Research Initiation Grant through Grant NTF21029, and the Sichuan Science and Technology Program (2023NSFSC0866), Joint Funds of the National Natural Science Foundation of China U23A20619, Key Research and Development Projects of Chengdu 204-YF05-2387-SN. In this work, we would like to express our sincere gratitude to Mr. Chaojun Xu and Mr. Chun Zhang for their valuable suggestions regarding the design and construction of the loss functions used in the proposed step-wise contractive model. Their insightful feedback significantly contributed to improving the robustness and efficacy of our model. We appreciate their time and effort in providing guidance throughout the development process.

Data availability

Data will be made available on request.

References

- [1] I. Misbah, C.K. Lee, K.L. Keung, Fault diagnosis in rotating machines based on transfer learning: literature review, *Knowl.-Based Syst.* 283 (2024) 111158.
- [2] P. Kundu, Review of rotating machinery elements condition monitoring using acoustic emission signal, *Expert Syst. Appl.* 252 (2024) 124169.
- [3] M.M. Kermani, A. Jalali, R. Azarderakhsh, Lightweight error detection architectures through swapping the shares for a subset of S-boxes, in: 2018 IEEE 61st International Midwest Symposium on Circuits and Systems, MWSCAS, IEEE, 2018, pp. 578–581.
- [4] P. Chen, C. Xu, Z. Ma, Y. Jin, A mixed samples-driven methodology based on denoising diffusion probabilistic model for identifying damage in carbon fiber composite structures, *IEEE Trans. Instrum. Meas.* 72 (3513411) (2023) 1–11.
- [5] Z. Chang, K. Jia, T. Han, Y.-M. Wei, Towards more reliable photovoltaic energy conversion systems: A weakly-supervised learning perspective on anomaly detection, *Energy Convers. Manage.* 316 (2024) 118845.
- [6] P. Chen, Z. Ma, C. Xu, Y. Jin, C. Zhou, Self-supervised transfer learning for remote wear evaluation in machine tool elements with imaging transmission attenuation, *IEEE Internet Things J.* 11 (2024) 23045–23054.
- [7] W. Xie, T. Han, Z. Pei, M. Xie, A unified out-of-distribution detection framework for trustworthy prognostics and health management in renewable energy systems, *Eng. Appl. Artif. Intell.* 125 (2023) 106707.
- [8] A. Cintas-Canto, M. Mozaffari-Kermani, R. Azarderakhsh, Reliable code-based post-quantum cryptographic algorithms through fault detection on FPGA, in: 2023 IEEE Nordic Circuits and Systems Conference, NorCAS, IEEE, 2023, pp. 1–5.
- [9] P. Chen, Y. Li, K. Wang, M.J. Zuo, An automatic speed adaption neural network model for planetary gearbox fault diagnosis, *Measurement* 171 (2021) 108784.
- [10] P. Chen, Y. Li, K. Wang, M.J. Zuo, A novel knowledge transfer network with fluctuating operational condition adaptation for bearing fault pattern recognition, *Measurement* 158 (2020) 107739.
- [11] P. Chen, Y. Li, K. Wang, M.J. Zuo, P.S. Heyns, S. Baggeröhr, A threshold self-setting condition monitoring scheme for wind turbine generator bearings based on deep convolutional generative adversarial networks, *Measurement* 167 (2021) 108234.
- [12] J. Zhuang, J. Yan, C.-G. Huang, M. Jia, Residual attention temporal recurrent network for fault diagnosis of gearboxes under limited labeled data, *Eng. Appl. Artif. Intell.* 129 (2024) 107539.
- [13] T. Han, T. Zhou, Y. Xiang, D. Jiang, Cross-machine intelligent fault diagnosis of gearbox based on deep learning and parameter transfer, *Struct. Control Health Monit.* 29 (3) (2022) e2898.
- [14] D. Li, Y. Zhao, Y. Zhao, A dynamic-model-based fault diagnosis method for a wind turbine planetary gearbox using a deep learning network, *Prot. Control Mod. Power Syst.* 7 (2) (2022) 1–14.
- [15] L. Zhang, Q. Fan, J. Lin, Z. Zhang, X. Yan, C. Li, A nearly end-to-end deep learning approach to fault diagnosis of wind turbine gearboxes under nonstationary conditions, *Eng. Appl. Artif. Intell.* 119 (2023) 105735.
- [16] M.S. Raghav, S. Patel, Fault diagnosis of spur gearbox by image classification using deep CNN, in: 2024 1st International Conference on Robotics, Engineering, Science, and Technology, RESTCON, IEEE, 2024, pp. 201–206.
- [17] M.H. Amiri, M. Pourgholi, N.M. Hashjin, M.K. Ardakani, Monitoring UAV status and detecting insulator faults in transmission lines with a new classifier based on aggregation votes between neural networks by interval type-2 TSK fuzzy system, *Soft Comput.* (2024) 1–34.
- [18] P. Chen, Y. Li, K. Wang, M.J. Zuo, P.S. Heyns, S. Baggeröhr, A threshold self-setting condition monitoring scheme for wind turbine generator bearings based on deep convolutional generative adversarial networks, *Measurement* 167 (2021) 108234.
- [19] K. Zhou, E. Diehl, J. Tang, Deep convolutional generative adversarial network with semi-supervised learning enabled physics elucidation for extended gear fault diagnosis under data limitations, *Mech. Syst. Signal Process.* 185 (2023) 109772.
- [20] L. Zhang, B. Wang, P. Liang, X. Yuan, N. Li, Semi-supervised fault diagnosis of gearbox based on feature pre-extraction mechanism and improved generative adversarial networks under limited labeled samples and noise environment, *Adv. Eng. Inform.* 58 (2023) 102211.
- [21] B. Zhao, C. Cheng, S. Zhao, Z. Peng, Hybrid semi-supervised learning for rotating machinery fault diagnosis based on grouped pseudo labeling and consistency regularization, *IEEE Trans. Instrum. Meas.* 72 (2023) 1–12.
- [22] Q. Luo, J. Chen, Y. Zi, J. Xie, A synchronization-induced cross-modal contrastive learning strategy for fault diagnosis of electromechanical systems under semi-supervised learning with current signal, *Expert Syst. Appl.* 249 (2024) 123801.
- [23] X. Fu, J. Tao, K. Jiao, C. Liu, A novel semi-supervised prototype network with two-stream wavelet scattering convolutional encoder for TBM main bearing few-shot fault diagnosis, *Knowl.-Based Syst.* 286 (2024) 111408.
- [24] H. Wang, Z. Liu, Y. Ge, D. Peng, Self-supervised signal representation learning for machinery fault diagnosis under limited annotation data, *Knowl.-Based Syst.* 239 (2022) 107978.
- [25] Y. Zhu, B. Xie, A. Wang, Z. Qian, Fault diagnosis of wind turbine gearbox under limited labeled data through temporal predictive and similarity contrast learning embedded with self-attention mechanism, *Expert Syst. Appl.* 245 (2024) 123080.
- [26] C. Cheng, D. Shan, Y. Teng, B. Zhao, Z. Peng, Q. He, Semisupervised fault diagnosis for gearboxes: a novel method based on a hybrid classification network and weighted pseudo-labeling, *IEEE Sens. J.* 23 (14) (2023) 16373–16383.
- [27] G. Liang, F. Li, X. Pang, B. Zhang, P. Yang, A gear fault diagnosis method based on reactive power and semi-supervised learning, *Meas. Sci. Technol.* 35 (12) (2024) 126107.
- [28] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
- [29] K. Han, A. Vedaldi, A. Zisserman, Learning to discover novel visual categories via deep transfer clustering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8401–8409.
- [30] K. Cao, Enhancing Machine Learning With Data-Efficient Methods (Ph.D. thesis), Stanford University, 2024.
- [31] R. Xiao, L. Feng, K. Tang, J. Zhao, Y. Li, G. Chen, H. Wang, Targeted representation alignment for open-world semi-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23072–23082.
- [32] W. Li, Z. Fan, J. Huo, Y. Gao, Modeling inter-class and intra-class constraints in novel class discovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3449–3458.
- [33] J. Liu, Y. Wang, T. Zhang, Y. Fan, Q. Yang, J. Shao, Open-world Semi-supervised Novel Class Discovery, 2023, arXiv e-prints arXiv:2305.13095.
- [34] S. Kullback, R.A. Leibler, On information and sufficiency, *Ann. Math. Stat.* 22 (1) (1951) 79–86.
- [35] P.A. Estévez, M. Tesmer, C.A. Perez, J.M. Zurada, Normalized mutual information feature selection, *IEEE Trans. Neural Netw.* 20 (2) (2009) 189–201.
- [36] J.M. Santos, M. Embrechts, On the use of the adjusted rand index as a metric for evaluating supervised classification, in: International Conference on Artificial Neural Networks, Springer, 2009, pp. 175–184.
- [37] N. Mehrabi Hashjin, M.H. Amiri, A. Mohammadzadeh, S. Mirjalili, N. Khodadadi, Novel hybrid classifier based on fuzzy type-III decision maker and ensemble deep learning model and improved chaos game optimization, *Cluster Comput.* (2024) 1–38.
- [38] M.H. Amiri, N. Mehrabi Hashjin, M. Montazeri, S. Mirjalili, N. Khodadadi, Hippopotamus optimization algorithm: a novel nature-inspired optimization algorithm, *Sci. Rep.* 14 (1) (2024) 5032.