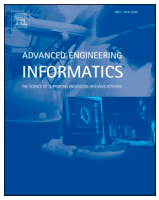
Advanced Engineering Informatics 65 (2025) 103136

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



A new lifelong learning method based on dual distillation for bearing diagnosis with incremental fault types☆

Shijun Xie [a], Changqing Shen [a,*], Dong Wang [b], Juanjuan Shi [a], Weiguo Huang [a], Zhongkui Zhu [a]

[a] School of Rail Transportation, Soochow University, Suzhou 215131 PR China
[b] The State Key Laboratory of Mechanical Systems and Vibration, Shanghai Jiao Tong University, Shanghai 200240 PR China

ARTICLE INFO

Keywords:
Fault diagnosis
Lifelong learning
Catastrophic forgetting
Dataset distillation
Feature distillation

ABSTRACT

In the rapidly evolving industrial environment, bearings may develop new fault types, posing significant challenges to deep learning-based intelligent fault diagnosis models. These models often suffer from catastrophic forgetting when encountering unknown fault types, resulting in performance degradation. Lifelong learning strategies offer a solution by enabling models to retain old knowledge while acquiring new information. However, traditional replay-based lifelong learning methods typically involve risks of privacy leakage and escalating storage costs. To address these issues, this study proposes a novel lifelong learning method called lifelong learning based on dual distillation (LLDD), which integrates a dual-distillation mechanism comprising dataset distillation and feature distillation, and introduces an equiangular basis vector (EBV) classifier. The dataset distillation technique compresses the dataset of each task into a small number of synthetic data that capture the essential information of the task, serving as replay exemplars. This approach reduces reliance on original data and storage costs. Feature distillation ensures that the model's representations do not deviate significantly from previous ones. The proposed method effectively prevents an increase in the number of model parameters during the lifelong learning process by incorporating the EBV classifier, thereby maintaining model complexity stability. The performance of LLDD is validated on two bearing diagnosis cases with incremental fault types. Results demonstrate that the proposed method surpasses other lifelong learning methods in performance and memory efficiency.

1. Introduction

Fault diagnosis of critical mechanical equipment components has become essential for ensuring continuous and efficient production in the era of fast-paced industrial advancement [1]. Technological advancements have increased the complexity of these machines, raising the reliability standards for their key parts. Particularly in industries requiring high precision and reliability, such as aerospace, aviation, and high-speed rail, bearings, as core mechanical components, play a crucial role. Their performance stability and reliability directly influence the system's safe operation and economic efficiency [2,3]. Bearings endure prolonged use under harsh conditions, and any degradation can lead to equipment failure or severe accidents. Delayed detection and resolution of bearing failures not only result in significant economic losses but also pose serious risks to personnel safety and may cause substantial societal repercussions. Consequently, developing and implementing effective bearing fault diagnosis technologies are vital for enhancing the reliability and safety of mechanical equipment [4].

Over the past few years, artificial intelligence (AI) has seen a growing integration into various facets of manufacturing and production processes [5,6]. Among its numerous applications, deep learning—a critical branch of AI—has demonstrated substantial potential and value in fault diagnosis [7]. By constructing intricate neural network models, deep learning can autonomously extract fault features from vast datasets and perform intelligent classification, thereby significantly enhancing the accuracy and efficiency of fault diagnosis processes. Ruan et al. [8] advanced bearing fault diagnosis with a physics-guided convolutional neural network, which refines input size and kernel dimensions to

☆ *Corresponding author:Prof. Changqing ShenSoochow University, Suzhou 215131, P.R. ChinaTel.: +86–18994371309, Fax: +86–512-67601052E-mail address: cqshen@suda.edu.cn.
* Corresponding author.
E-mail address: cqshen@suda.edu.cn (C. Shen).

https://doi.org/10.1016/j.aei.2025.103136
Received 20 November 2024; Received in revised form 7 January 2025; Accepted 16 January 2025
Available online 22 January 2025

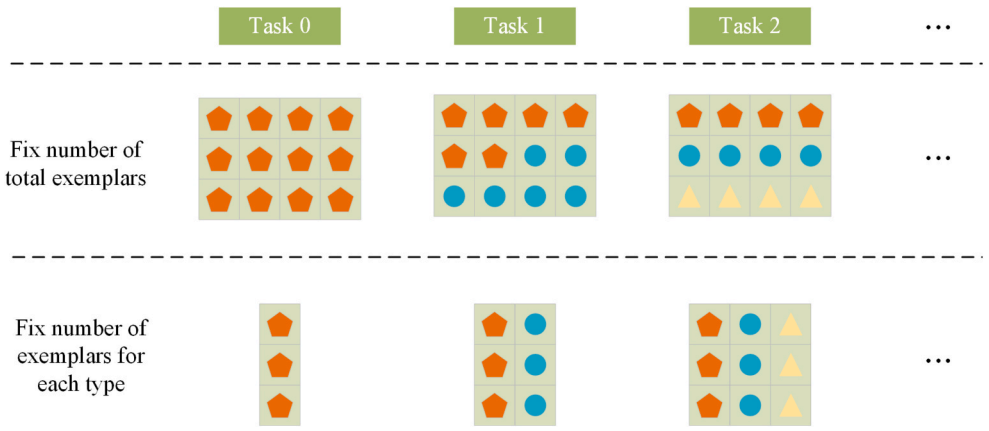


Fig. 1. Two ways of storing exemplars.

bolster accuracy and computational efficiency. Hou et al. [9] developed Diagnosisformer, leveraging an augmented transformer framework for rolling bearing fault detection. It integrates a multi-feature encoder paired with a cross-flip decoder, facilitating streamlined feature extraction and precise fault categorization. In response to constraints posed by restricted sample availability and variable operational velocities, Shao et al. [10] constructed an attention-guided generative adversarial network to extract global thermal correlation features from infrared thermal images and designed a dual-threshold training mechanism to improve the quality of generated images and training efficiency. For cross-domain diagnostics, Huo et al. [11] enhanced transfer learning via a linear superposition network tailored to rolling bearings. They fine-tuned the 1D-CNN architecture and utilized target domain pseudo-labels to refine the loss function, thereby elevating diagnostic accuracy. Zhao et al. [12] presented a class-aware adversarial multi-wavelet convolutional neural network, featuring a nuanced classification approach and intensified alignment of conditional distributions. This design ensures meticulous feature mapping across domains, pioneering solutions for cross-domain fault identification in rotating machinery.

In real-case scenarios, new types of bearing faults inevitably arise when operating under complex conditions for extended periods, which are referred to as incremental fault types [13]. On the one hand, training diagnosis models directly with new fault type data leads to degradation in performance on prior fault type data, an issue named catastrophic forgetting. On the other hand, accumulating data on all recognized fault types for the purpose of retraining networks presents challenges in terms of difficulty and expense. For establishing a reliable and stable fault diagnosis model, the catastrophic forgetting of deep learning should be addressed. Wang et al. [14] introduced a graph persistent learning network that leverages graph convolutional networks to detect novel fault types and autonomously updates the diagnostic model through class-incremental learning, effectively mitigating catastrophic forgetting. Li et al. [15] proposed a continual learning classification approach, drawing inspiration from the biological immune system. This method integrates classification and clustering techniques to hierarchically cluster unidentified samples on the basis of similarities within memory cell sets. Zhu et al. [16] successfully classified new fault types by leveraging a reserved embedding space in conjunction with virtual sample guidance. Sun et al. [17] proposed a domain incremental learning method for bearing fault diagnosis that integrates the cross-entropy loss function with Elastic Weight Consolidation (EWC) to constrain important parameters from the old domain based on filtering columns and knowledge base. Zhang et al. [18] introduced deep adaptive sparse residual networks using a task-aware dynamic masking strategy to adjust the retention and utilization of blank weights in the network, balancing memory and learning capability of the model. Zheng et al. [19] employed a memory indexing method to encode extracted features by product quantization and store them in indices. This approach avoids storing data in its original format but still faces storage pressure. Liu et al. [20] combined generated feature replay and knowledge distillation to prevent catastrophic forgetting. This method avoids privacy leakage and the storage of exemplars by generating features of past task data through a generative model for replay. However, since each stage requires training a generative model and generating features, it significantly increases the computational cost.

Existing lifelong learning-based fault diagnosis methods predominantly rely on replay, in which stored data or synthetic data are reused during the learning of new tasks. They comply with the intuition that reviewing strengthens the memory of past knowledge. However, in real-world scenarios, because of copyright and privacy concerns, directly storing data in their original format is not allowed, rendering some replay-based methods infeasible [21]. Additionally, the performance of networks, when continuously learning multiple new tasks, is limited by the storage capacity of exemplar sets. A fix number of exemplars for each type leads to an increase in storage cost as the number of learned classes grows, while a fix number of total exemplars leads to a reduction in the number of exemplars stored per class, rending information loss of exemplar sets. The two ways of storing exemplars are shown in Fig. 1.

On the basis of the preceding analysis, a lifelong learning method based on dual distillation for bearing fault diagnosis with incremental fault types is proposed in this article. Inspired by dataset distillation with distribution match, lifelong learning based on dual distillation (LLDD) distillates the dataset of prior tasks to generate a small but more representative set of synthetic data as exemplars. The storage cost is reduced by storing the distilled synthetic data rather than raw data. Another technique applied to mitigate catastrophic forgetting is feature distillation, a method of knowledge distillation, which limits the excessive deviation of features extracted by the current model from those extracted by previous models. Applying a fully connected layer associated with softmax as a classifier is common in classification tasks. However, with the increase in categories, the number of trainable parameters grows linearly, which leads to a rise in computational costs. For addressing these issues, an equiangular basis vector (EBV) classifier is introduced, which minimizes the spherical distance between input embeddings and corresponding basis vectors. Owing to the predefined angular basis vectors, the number of trainable parameters does not increase in lifelong learning.

The contributions of this work are as follows:

1) A new lifelong learning method based on dual distillation is proposed to address the fault diagnosis with incremental fault types. It enables models to continuously learn new knowledge while increasing its storage efficiency.

2) A dual-distillation mechanism, which combines the advantages of

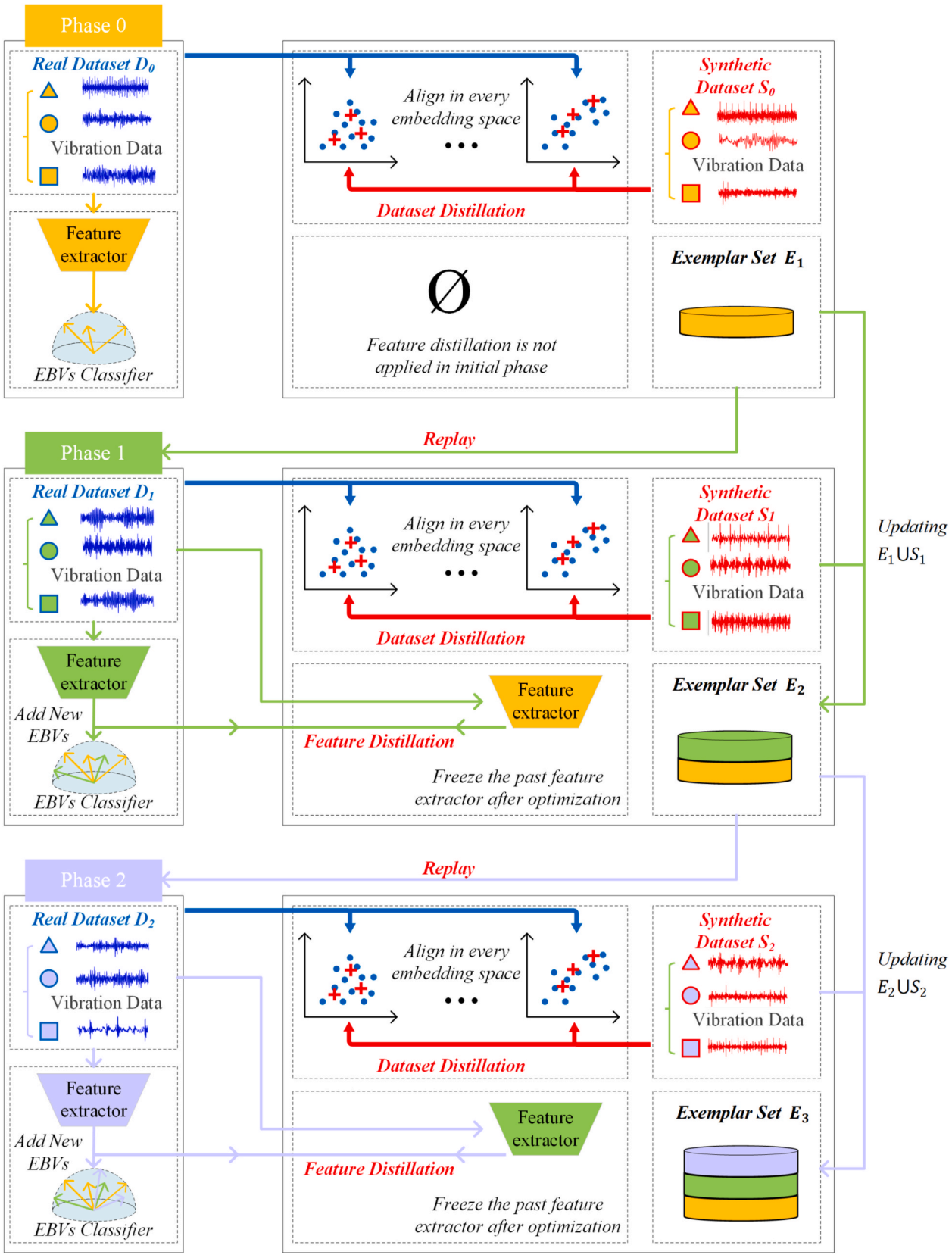


Fig. 2. Framework of the proposed LLDD.

dataset distillation and feature distillation, is proposed. It transfers the knowledge from past tasks to the current model while facilitating efficiency and data privacy.

3) An EBV classifier is introduced into lifelong learning to replace the fully connected layer associated with softmax, keeping the number of trainable parameters constant with the growth of fault types.

4) The feasibility and efficiency of the proposed method are validated on two distinct bearing datasets. Experimental results indicate that LLDD outperforms other methods in terms of diagnostic accuracy and memory efficiency.

The rest of this article is organized as follows: Section 2 presents the fundamental theory of lifelong learning and the definition of fault diagnosis under incremental fault types. Section 3 provides a detailed description of LLDD. Section 4 indicates and analyzes the experimental results. Conclusions are drawn in Section 5.

2. Theoretical background

2.1. Lifelong learning

In static environments, deep learning models excel at extracting features from comprehensive data. However, their performance often deteriorates in dynamic contexts, particularly when confronted with domain shifts or the introduction of new categories. This decline partly accounts for the phenomenon called catastrophic forgetting, in which retraining on previously unseen data can result in the forgetting of previously learned knowledge. The reason for catastrophic forgetting is that when the model learns a new task through backpropagation, the model parameters are adjusted according to the fault data of the new task, leading to the model forgetting the knowledge learned from the old tasks. To address this challenge, scholars have proposed the paradigm of lifelong learning. This paradigm aims to emulate the human capacity for continuous assimilation of new information without forgetting prior knowledge. The essence of lifelong learning lies in enabling models to handle a sequence of tasks without forgetting past tasks, thereby preventing catastrophic forgetting and facilitating rapid adaptation to new tasks [22]. Recent approaches of lifelong learning are broadly categorized into the following three types.

1) Regularization-based methods: Regularization-based methods introduce regularization terms into the loss function to protect old knowledge from being overwritten by new information. The advantage of these approaches lies in the fact that they do not require access to old data, which ensures data security. Learn without forgetting [23] integrates the concept of knowledge distillation into lifelong learning, utilizing the predictions of the prior network to guide the training of the current network. Aljundi et al. proposed memory-aware synapses [24], which prevents changes to important parameters related to past tasks by calculating the importance of neurons.

2) Parameter isolation-based methods: Parameter isolation-based methods adapt the model to new tasks by expanding the neural network without changing the parameters of the previous tasks. PackNet [25] frees up redundant parameters of the deep network via a weight-based pruning technique for new tasks; when parameters are updated, the parameters of prior tasks are frozen. Abati et al. [26] added a gating module to each convolutional layer, selecting appropriate filters for each task. Parameter isolation-based methods are only employed in simple lifelong learning tasks owing to the introduction of a large number of parameters and computations.

3) Replay-based methods: By replaying the exemplars that contain information about previous tasks, replay-based lifelong learning approaches alleviate the forgetting of past knowledge. On the basis of the construction rules of exemplar sets, replay-based approaches are subdivided into two strategies. The first strategy involves selecting representative exemplars from the previous datasets using specific methods. The second strategy employs generative models to generate synthetic data, circumventing the need to directly store data in original format. Rebuffi et al. [27] proposed iCaRL, which combines exemplar replay and knowledge distillation for the first time, using the nearest-mean-of-exemplars rule to avoid changes in the classifier structure in lifelong learning. Xiang et al. [28] proposed a lifelong learning method based on conditional adversarial networks, which generates synthetic data for joint training while learning new tasks.

2.2. Definition of fault diagnosis under incremental fault types

For simulating the scenario in which new types of faults gradually emerge in bearings under real working conditions, the fault dataset is divided into multiple diagnostic tasks, each with varying numbers of categories and learning difficulties.

In a scenario comprising $N + 1$ phases, the first phase is an initial phase, followed by subsequent N incremental phases, with each phase representing a distinct fault diagnosis task. The diagnosis model is initially trained in phase 0, and its parameters are updated sequentially in the subsequent N incremental phases. The range of faults identifiable by the model expands in correlation with the number of diagnostic tasks. T_n represents task n, and its dataset is denoted as $D_n = \{\boldsymbol{x}_i^n, y_i^n\}_{i=1}^{M^n}$, where M^n represents the number of samples for task n. P_n indicates the cumulative number of distinct fault types that can be recognized by the model after phase n, and Q_n denotes the quantity of fault types that need to be learned during phase n. Hence, $P_n = P_{n-1} + Q_n$. During phase n, as the model learns diagnostic task n, it retains relevant knowledge from previous tasks, enabling recognition of all fault types encountered thus far.

3. Proposed lifelong learning method based on dual distillation

Replay-based lifelong learning methods prevent catastrophic forgetting by replaying exemplars. However, these methods are not suitable for real-world scenarios that involve data security and limited storage space. To address these issues, this study introduces a novel lifelong learning method based on dual distillation. The flowchart of the proposed method is shown in Fig. 2, and it is detailed as follows.

3.1. Overview of the proposed approach

Each phase of LLDD includes two parts: training a feature extractor and updating an exemplar set. The fast Fourier transform is initially applied to the vibration signal to obtain its frequency-domain representation. Subsequently, the 1D frequency-domain signal is transformed into a 2D format. While data processing remains consistent, training in incremental phases is different from that in initial phase, which utilizes dataset distillation and feature distillation. The classifier employed is an EBV classifier, which is characterized by pregenerated EBVs, $\boldsymbol{W}_{EBVs} = [\boldsymbol{w}_1; \boldsymbol{w}_2; \cdots; \boldsymbol{w}_N]$. Consequently, this design ensures that no additional trainable parameters are introduced during the training phase.

(1) Phase 0: The training process in the initial phase is similar to traditional deep learning strategies that only concern classification loss. The 2D frequency-domain dataset $D_0 = \{\boldsymbol{x}_i^0, y_i^0\}_{i=1}^{M^0}$, where $y_i^0 \in (0,1,2,\ldots,Q_0-1)$, is used to train feature extractor $F_0(\cdot)$. Classifier $C_0(\cdot)$ selects Q_0 basis vectors $[\boldsymbol{w}_1; \boldsymbol{w}_2; \ldots; \boldsymbol{w}_{Q_0}]$ for classification in the initial phase. After optimization, distilled dataset $S_0 = \{\boldsymbol{s}_i^0, y_i^0\}_{i=1}^{M_s^0}$ is generated through distilling the current dataset D_0. M_s^0 is the number of samples in S_0, and $M_s^0 = Q_0 \cdot K$, where K is the number of exemplars stored for each fault type. Synthetic data within S_0 encapsulate pivotal information from the initial dataset D_0 and serve as exemplars in subsequent incremental training phases. The exemplar set for phase 1 $E_1 = \{S_0\}$.

(2) Incremental learning phase: After phase 0, subsequent incremental learning phases are implemented. Phase n serves as an illustrative example. Feature extractor $F_n(\cdot)$ is initialized by the parameters of prior feature extractor $F_{n-1}(\cdot)$. The EBV classifier selects $[\boldsymbol{w}_1; \boldsymbol{w}_2; \ldots; \boldsymbol{w}_{P_n}]$

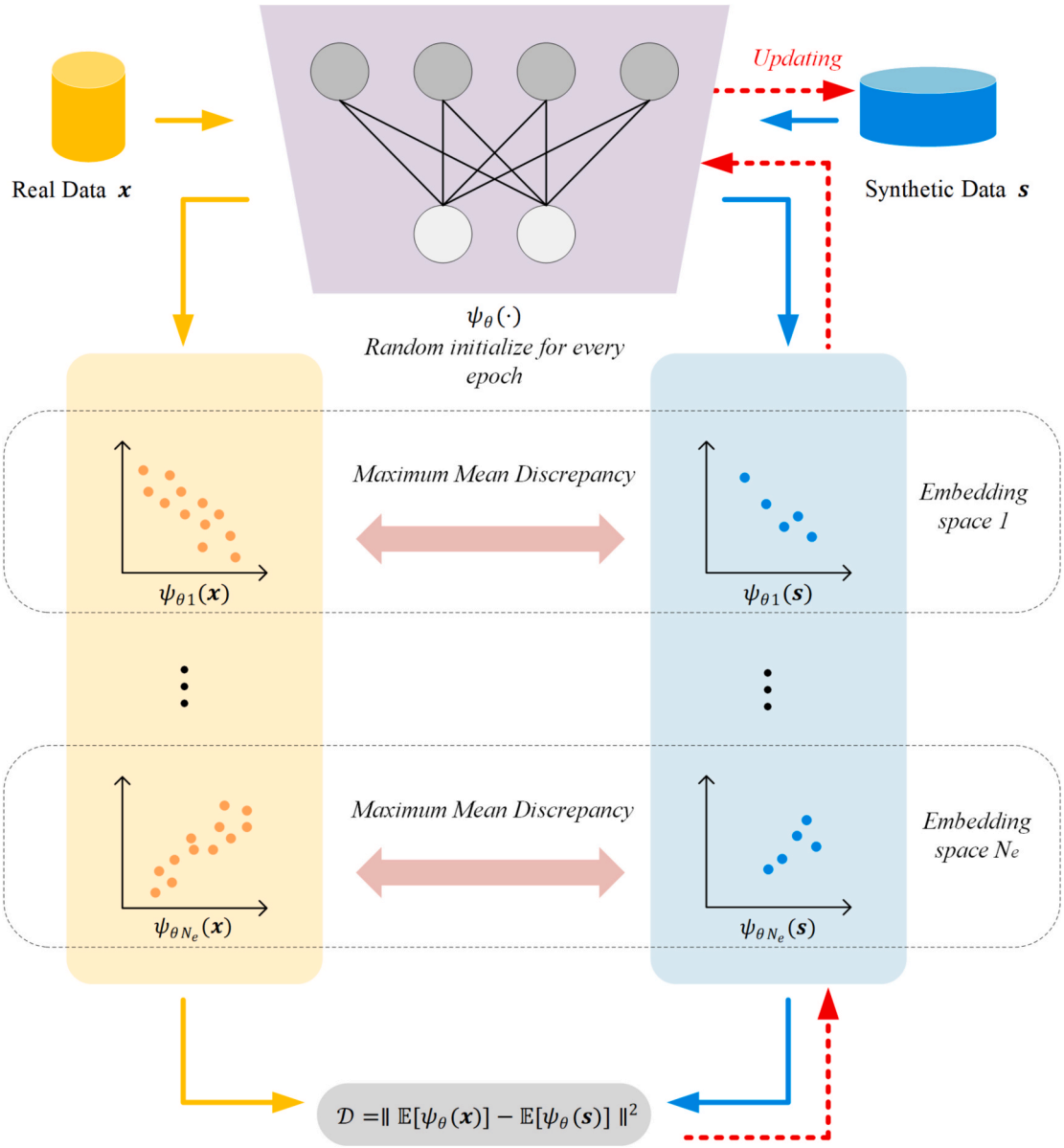


Fig. 3. Framework of dataset distillation.

from $\boldsymbol{W}_{EBVs} = [\boldsymbol{w}_1; \boldsymbol{w}_2; \cdots; \boldsymbol{w}_M]$ as basis vectors for classification. In phase n, dataset $D_n = \{\boldsymbol{x}_i^n, y_i^n\}_{i=1}^{M^n}$, where $y_i^n \in (P_{n-1}, P_{n-1}+1, \ldots, P_{n-1}+Q_n-1)$, and exemplar set E_n are jointly utilized to train $F_n(\cdot)$. Exemplar set $E_n = \{S_0, S_1, \cdots, S_{n-1}\}$ contains all distilled datasets of prior tasks. Similar to the initial phase, dataset D_n is distilled into S_n, after which exemplar set E_{n+1} is updated by $E_{n+1} = E_n \bigcup S_n$.

The proposed method is summarized in Algorithm 1.

Algorithm 1: Lifelong learning method based on dual distillation

Data: $D_0, D_1, \ldots D_N$, where $D_n = \{\boldsymbol{x}_i^n, y_i^n\}_{i=1}^{M^n}$
For $n = 0$, do:
Input: $D_0 = \{\boldsymbol{x}_i^0, y_i^0\}_{i=1}^{M^0}$
Select Q_0 EBVs $[\boldsymbol{w}_1; \boldsymbol{w}_2; \ldots; \boldsymbol{w}_{Q_0}]$ as classifier C_0
$\hat{y} = C_0(F(\boldsymbol{x}_i^0))$ (by Eq. (4))
Compute current classification loss $\mathscr{L}_0^C$ (by Eq. (5))
Update parameters of feature extractor $F_0(\cdot)$
Distill D_0 into $S_0 = \{\boldsymbol{s}_i^0, y_i^0\}_{i=1}^{M_s^0}$
Construct exemplar set $E_1 = S_0$
End for

(*continued on next column*)

(*continued*)

Algorithm 1: Lifelong learning method based on dual distillation

For $n = 1, 2, \ldots, N$, do:
Input: $D_n = \{\boldsymbol{x}_i^n, y_i^n\}_{i=1}^{M^n}$, $E_n = \{S_0, S_1, \cdots, S_{n-1}\}$
Add Q_n EBVs into C_{n-1} as C_n
Initialize F_n with F_{n-1}
Compute feature distillation loss $\mathscr{L}_n^D$ (by Eq. (3))
$\hat{y} = C_n(F(\boldsymbol{x}_i^n))$
Compute current classification loss $\mathscr{L}_n^C$ (by Eq. (5))
$\hat{y} = C_n(F(\boldsymbol{s}_i^n))$
Compute replay classification loss $\mathscr{L}_{0:n-1}^C$ (by Eq. (6))
Overall loss $\mathscr{L}_n = \mathscr{L}_n^D + (1-\eta)\mathscr{L}_n^C + \eta\mathscr{L}_{0:n-1}^C$
Update parameters of feature extractor $F_n(\cdot)$
Distill D_n into $S_n = \{\boldsymbol{s}_i^n, y_i^n\}_{i=1}^{M_s^n}$
Update exemplar set $E_{n+1} = E_n \bigcup S_n$
End for

3.2. Details of the lifelong learning method based on dual distillation

LLDD can construct a sustainable and efficient intelligent fault

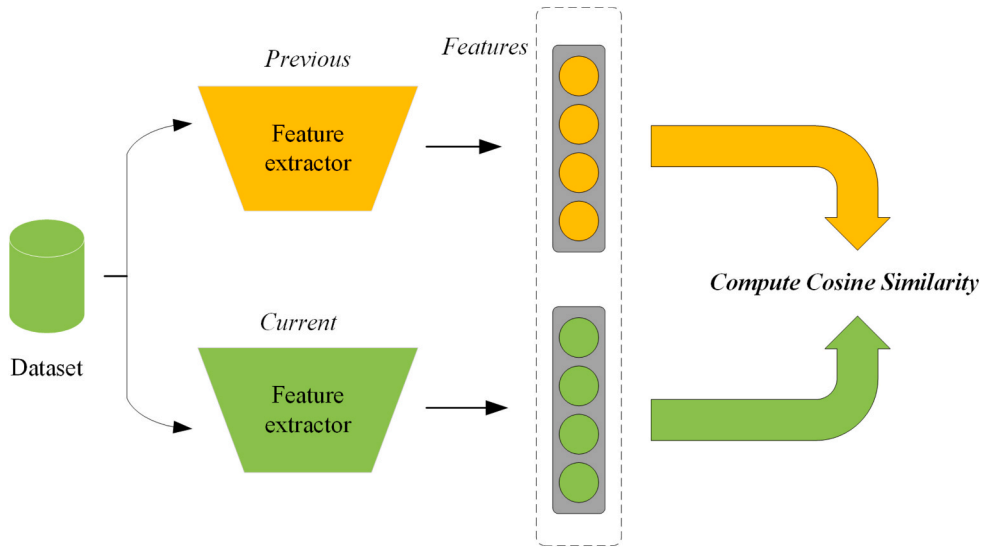


Fig. 4. Framework of feature distillation.

diagnosis model. The lifelong learning strategy adopted by this method utilizes a dual-distillation mechanism, which not only effectively prevents catastrophic forgetting of previous knowledge but also optimizes memory efficiency. To control the complexity of the model and avoid unnecessary growth in the number of parameters during continuous learning, the proposed method introduces an EBV classifier. This type of classifier enables the model to effectively absorb and integrate new knowledge while maintaining its parameter size. The dual-distillation mechanism and EBV classifier are described in detail as follows.

3.2.1. Dual-distillation mechanism

The LLDD model leverages a dual-distillation mechanism that astutely integrates the benefits of dataset distillation and feature distillation. Dataset distillation condenses the original dataset into a set of representative synthetic samples for replay, maintaining the diagnostic capability of the model for past tasks. The synthetic data retains more key information from past tasks compared to traditional lifelong learning methods. Therefore, it substantially reduces storage requirements. Meanwhile, feature distillation ensures that the model does not deviate significantly from previously learned representations when tackling new tasks through regularization strategies, effectively mitigating the risk of knowledge forgetting.

(1) Dataset distillation: Dataset distillation is applied to create a small and efficient synthetic dataset that preserves the essential information of the original large dataset needed for model training [29]. The goal of dataset distillation is to reduce time cost and computational resource requirements. The model trained on the distilled dataset can achieve performance comparable to that of the model trained on the original dataset.

The dual-distillation mechanism is inspired by the dataset condensation with distribution matching proposed by Zhao et al. [30]. A representative synthetic dataset is generated to serve as exemplars for replay by distilling the dataset, which alleviates catastrophic forgetting. Distilling the dataset may result in some loss of fault information, but compared to traditional methods of sampling exemplars in replay methods, it retains more information. To obtain synthetic dataset S_n whose distribution approximates the real dataset D_n, the dataset distillation module optimizes the distance between the two distributions of S_n and D_n using maximum mean discrepancy. The framework of dataset distillation and feature distillation used in the presented method are shown in Fig. 3 and Fig. 4.

The synthetic dataset $S_n = \{s_i^n, y_i^n\}_{i=1}^{M_s^n}$ is initialized by the real data in D_n. Given that directly estimating the high-dimensional data is expensive and inaccurate, real data $\boldsymbol{x} \in D_n$ and synthetic data $\boldsymbol{s} \in S_n$ are embedded into a lower-dimensional space by a series of randomly initialized convnet $\psi_\theta(\cdot)$. The discrepancy of distribution for a specific $\psi_\theta(\cdot)$ is calculated as follows:

$$D = \|\mathbb{E}[\psi_{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}[\psi_{\boldsymbol{\theta}}(\boldsymbol{s})]\|^2. \tag{1}$$

Each $\psi_\theta(\cdot)$ provides different interpretations for inputs; hence, the combination of them reflects the comprehensive distance. The final object function is expressed as follows:

$$\min_{S_n} L = \sum_{c=P_{n-1}}^{P_n-1} \mathbb{E}_{\theta\sim\Theta} \| \mathbb{E}_{(\boldsymbol{x},y)\in D_n^c}\left[\psi_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right] - \mathbb{E}_{(\boldsymbol{s},y)\in S_n^c}\left[\psi_{\boldsymbol{\theta}}\left(\boldsymbol{s}\right)\right] \|^2. \tag{2}$$

where Θ is the distribution for random ConvNet initialization; D_n^c and S_n^c represent the data in D_n and S_n that belong to type c, respectively.

(2) Feature distillation: Hinton et al. [31] introduced the concept of knowledge distillation for the first time to transfer knowledge from a large, complex teacher model to a smaller, more efficient student model while retaining as much of the teacher model's performance as possible. Knowledge distillation can be divided into two categories: logits distillation and feature distillation. The method proposed by Hinton et al. is logits distillation. The student model learns to replicate the soft target, the probability distributions over all classes generated by the teacher model. Instead of only transferring knowledge through the final output probability, feature distillation focuses more on intermediate representations or features [32].

Feature distillation is employed to mitigate catastrophic forgetting, which compels the features output by the current feature extractor $F_n(\cdot)$ to align with the features output by the past one. Feature distillation transfers key knowledge from $F_{n-1}(\cdot)$ to $F_n(\cdot)$. Given that the angular component captures the core semantic content in neural network representations more effectively than the magnitude does, cosine similarity, instead of the Euclidean distance, is employed to assess the discrepancy [33]. The feature distillation loss function of phase n is expressed as follows:

$$\mathscr{L}_n^D = -\mathbb{E}_{\boldsymbol{x}\in D_n}\left[\frac{F_n(\boldsymbol{x})\cdot F_{n-1}(\boldsymbol{x})}{\|F_n(\boldsymbol{x})\| \times \|F_{n-1}(\boldsymbol{x})\|}\right]. \tag{3}$$

3.2.2. EBV classifier

Many neural network models designed for classification tasks incorporate a fully connected layer associated with softmax as their classifier. However, such a classifier is unsuitable for lifelong learning.

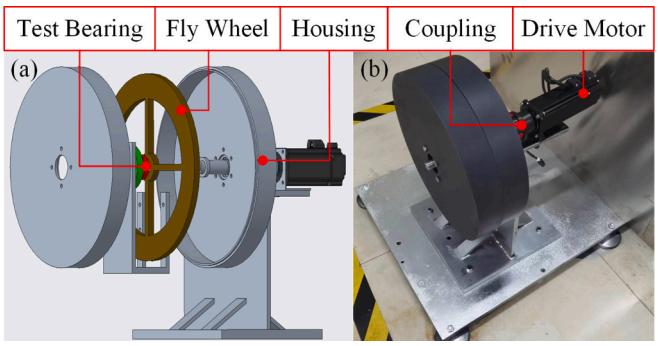


Fig. 5. Flywheel bearing test rig. (a) Design drawing. (b) Physical drawing.

Table 1
Data Description of the Dataset.

Fault type	Notation	Label	Task ID	Phase
Normal	NO	0	T0	0
0.2 mm Inner Fault	0.2IF	1		
0.2 mm Ball Fault	0.2BF	2		
0.2 mm Outer Fault	0.2OF	3		
0.3 mm Inner Fault	0.3IF	4	T1	1
0.3 mm Ball Fault	0.3BF	5		
0.3 mm Outer Fault	0.3OF	6		
0.4 mm Inner Fault	0.4IF	7	T2	2
0.4 mm Ball Fault	0.4BF	8		
0.4 mm Outer Fault	0.4OF	9		
0.5 mm Inner Fault	0.5IF	10	T3	3
0.5 mm Ball Fault	0.5BF	11		
0.5 mm Outer Fault	0.5OF	12		
0.6 mm Inner Fault	0.6IF	13	T4	4
0.6 mm Ball Fault	0.6BF	14		
0.6 mm Outer Fault	0.6OF	15		

Table 2
ResNet18 Framework.

Name of the layer	Size of the output	Channels × Kernel size
Input	$3 \times 32 \times 32$	/
Convnet	$64\times 32 \times 32$	$64\times 3 \times 3$, stride $= 1$
ResBlock1	$64\times 32 \times 32$	$\begin{bmatrix} 64 \times 3 \times 3 \\ 64 \times 3 \times 3 \end{bmatrix} \times 2$, stride $= 1$
ResBlock2	$128 \times 16 \times 16$	$\begin{bmatrix} 128 \times 3 \times 3 \\ 128 \times 3 \times 3 \end{bmatrix} \times 2$, stride $= 2$
ResBlock3	$256 \times 8 \times 8$	$\begin{bmatrix} 256 \times 3 \times 3 \\ 256 \times 3 \times 3 \end{bmatrix} \times 2$, stride $= 2$
ResBlock4	$512 \times 4 \times 4$	$\begin{bmatrix} 512 \times 3 \times 3 \\ 512 \times 3 \times 3 \end{bmatrix} \times 2$, stride $= 2$
Average pool	$512 \times 1 \times 1$	$512 \times 4 \times 4$, stride $= 1$
Classifier	Number of fault types	/

Table 3
Compared Methods.

Method	Description
R1	Fine-tuning of all network parameters
R2	Joint training
M1	EWC
M2	MAS
M3	iCaRL
M4	LUCIR [35]
A1	Remove feature distillation
A2	Replace distilled exemplars with herding exemplars
A3	Remove feature distillation and replay with herding exemplars
LLDD	The proposed method

Table 4
Hyperparameter Setting.

Hyperparameters	Value	Hyperparameters	Value
Learning rate α_1	0.001	Epochs	60
Learning rate α_2	1	Distillation epochs	15,000
K	3	Batch size	128

Table 5
Average Accuracy of LLDD and Relevant Methods (%).

Method	Phase0	Phase1	Phase2	Phase3	Phase4
R1	99.85 ± 0.12	56.88 ± 6.85	36.08 ± 5.19	29.45 ± 3.06	23.31 ± 5.46
R2	**99.85 ± 0.30**	**99.58 ± 0.27**	**99.59 ± 0.19**	**99.41 ± 0.17**	**99.74 ± 0.08**
M1	99.85 ± 0.12	64.83 ± 7.42	41.74 ± 7.97	32.10 ± 4.87	29.86 ± 5.17
M2	99.85 ± 0.12	67.78 ± 3.09	40.38 ± 9.86	31.51 ± 5.30	27.90 ± 5.60
M3	99.45 ± 0.60	94.11 ± 1.50	89.37 ± 4.26	85.61 ± 3.59	87.07 ± 1.51
M4	99.20 ± 0.80	93.98 ± 3.51	92.50 ± 3.04	89.36 ± 1.41	89.39 ± 2.29
A1	100 ± 0.00	93.60 ± 2.07	92.24 ± 1.42	93.97 ± 1.35	94.15 ± 0.81
A2	100 ± 0.00	97.89 ± 1.28	73.10 ± 10.02	60.11 ± 8.53	56.79 ± 5.64
A3	99.50 ± 0.10	42.74 ± 0.06	29.76 ± 0.12	23.08 ± 0.00	18.75 ± 0.00
LLDD	**100 ± 0.00**	**98.71 ± 0.38**	**97.64 ± 0.57**	**98.43 ± 0.38**	**98.78 ± 0.18**

In lifelong learning, the dimension of features extracted by feature extractor d is fixed while the number of classes P is gradually increasing, which means that the structure of the fully connected layer $\boldsymbol{W} \in \mathbb{R}^{d\times P}$ changes and the number of parameters increases linearly with the number of classes.

An EBV classifier [34] is introduced to address this issue. The EBV classifier replaces the traditional one-hot vectors with EBVs, allowing more class basis vectors to be accommodated within a limited dimensional space, thereby reducing the parameter scale. EBVs $\boldsymbol{W}_{EBVs} \in \mathbb{R}^{d\times M}$, in which $\boldsymbol{W}_{EBVs} = [\boldsymbol{w}_1; \boldsymbol{w}_2; \cdots; \boldsymbol{w}_M]$ are generated in advance, do not increase the number of trainable parameters in lifelong learning before the number of classes exceeds M. Therefore, it avoids model expansion and increase of computational costs. The probability that signal $(\boldsymbol{x}_i, y_i)$ is recognized as class y_i by the model is

$$P_{EBVS}(y = y_i|\boldsymbol{x}_i) = \exp((\boldsymbol{w}_i F(\boldsymbol{x}_i))/T) \Big/ \left(\sum_{j=1}^{P} \exp(\boldsymbol{w}_j F(\boldsymbol{x}_i)/T) \right). \tag{4}$$

where T is a hyperparameter related to temperature, and $P \leq M$.

3.3. Constructing the overall loss function

According to the description above, the parameters of the classifier, specifically the EBVs, are predefined. Thus, only the parameters of the feature extractor need to be continuously updated for diagnostic tasks. The loss function comprises three components: feature distillation loss, current classification loss, and replay classification loss. The feature distillation loss is calculated using Equation (3). Details of the current classification loss and replay classification loss are provided below.

Current classification loss enables the model to learn knowledge relevant to the current task, which is calculated as follows:

$$L_n^C = -\sum_{i=1}^{M^n} \log P_{\text{EBVs}} \left(y = y_i^n \middle| \boldsymbol{x}_i^n \right). \tag{5}$$

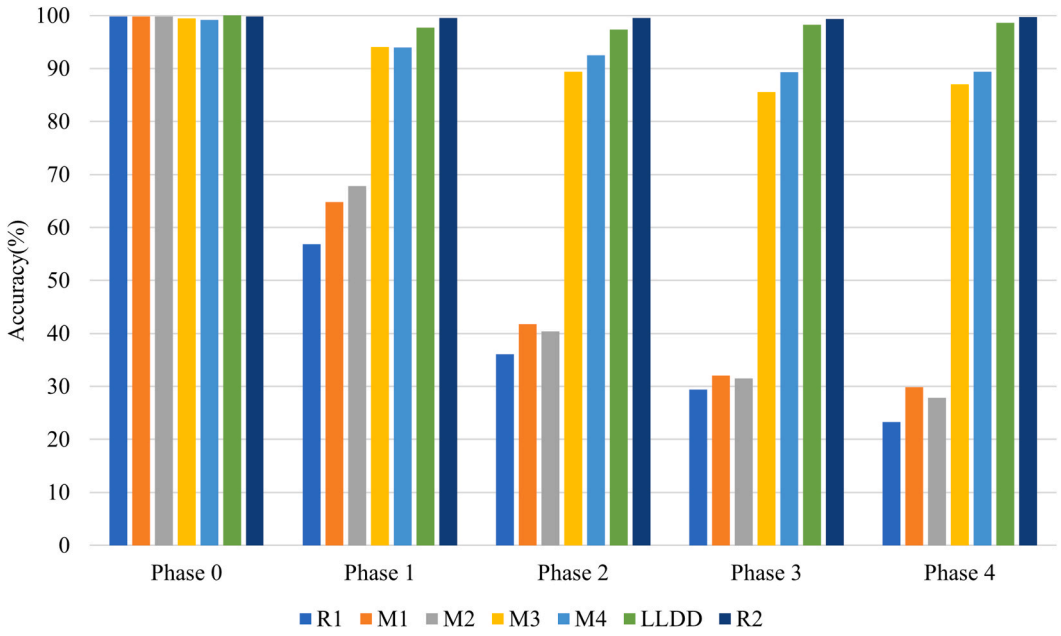


Fig. 6. Diagnostic accuracy of compared methods and LLDD.

where $(\boldsymbol{x}_i^n, y_i^n) \in D_n$. $\boldsymbol{x}_i^n$ represents the i-th sample in dataset D_n and y_i^n represents the corresponding label.

By replaying a small amount of distilled data, the feature extractor retains the essential information of previous tasks in incremental phases. The replay classification loss function is expressed as follows:

$$L_{0:n-1}^{C} = -\sum_{i=1}^{M_s^n} \log P_{\mathrm{EBVs}}\left(y = y_i^n \middle| \boldsymbol{s}_i^n\right). \tag{6}$$

where $(\boldsymbol{s}_i^n, y_i^n) \in E_n$, and M_s^n denotes the number of exemplars in exemplar set E_n.

Given the imbalance between the number of stored exemplars and the number of samples in the current task, directly summing their losses for the total classification loss is inappropriate. Consequently, an old task factor η is incorporated to balance the current classification loss and the replay classification loss. Classification loss is calculated as follows:

$$\mathscr{L}^C = (1-\eta)\mathscr{L}_n^C + \eta\mathscr{L}_{0:n-1}^C. \tag{7}$$

As the model learns more tasks, old task factor η gradually increases, indicating that the model needs to retain more knowledge. η is calculated as follows:

$$\eta = \frac{1}{1+\mathrm{e}^{-0.15\cdot P_{n-1}}}. \tag{8}$$

In phase 0, the overall loss function only contains current classification loss $\mathscr{L}_0^C$, which is computed as

$$\mathscr{L}_0 = \mathscr{L}_0^C. \tag{9}$$

By contrast, the overall loss function in phase n is expressed as

$$\mathscr{L}_n = \mathscr{L}_n^D + \mathscr{L}^C = \mathscr{L}_n^D + (1-\eta)\mathscr{L}_n^C + \eta\mathscr{L}_{0:n-1}^C. \tag{10}$$

4. Experimental results and analysis

4.1. Case 1: Flywheel bearing dataset

4.1.1. Dataset description

As shown in Fig. 5, the dataset is obtained from a flywheel bearing test rig comprising essential elements such as drive motor, flywheel, test bearing (6203DDU NSK), coupling, and acceleration sensor. Data are sampled at a high rate of 10 kHz to ensure accuracy. To simulate partial bearing failures, this study employs wire-cutting technology on the inner ring, roller, and outer ring of the bearing. In this experiment, 16 fault types are selected at a rotational speed of 800 rpm. The dataset includes normal bearing data and 15 distinct fault types of inner, outer, and ball ranging from 0.2 mm to 0.6 mm. For each fault type, 100 samples are designated for training and another 100 for testing, with each sample comprising 1024 individual data points. The detailed specifications of the faults for each task are outlined in Table 1.

4.1.2. Comparison methods

LLDD is compared with fine-tuning, joint training, and four well-established lifelong learning methods to verify its effectiveness. For a fair comparison, ResNet18 is employed as the baseline architecture across all methodologies, as illustrated in Table 2. The methods involved in the experiment are outlined in Table 3.

R1 and R2 are related non-lifelong learning methods. R1 fine-tunes the parameters of the entire deep neural network to illustrate the effect of catastrophic forgetting on model performance. R2 is joint training, which combines the data of the current task and past tasks to jointly optimize the neural network. The training cost of R2 is highest, but the effect is most ideal, serving as the upper bound for the performance of all replay-based lifelong learning methods.

M1 and M2 are regularization-based lifelong learning methods. M1 employs a method that maintains elastic weights to mitigate the catastrophic forgetting problem in neural networks. M2 introduces an assessment of the importance of neurons within the neural network. M3 and M4 are replay-based lifelong learning methods. M3 combines replay with knowledge distillation and replaces the linear classifier with a nearest neighbor classifier. M4 utilizes a cosine normalization classifier to alleviate catastrophic forgetting.

Ablation studies are conducted to determine the significance and necessity of each strategy within the LLDD framework. A1 is designed to verify the effectiveness of distilled exemplars, and A2 is used to validate feature distillation. A3 removes feature distillation and replace distilled exemplars with herding exemplars.

4.1.3. Experimental implementation details

The hyperparameters of the proposed method are listed in Table 4.

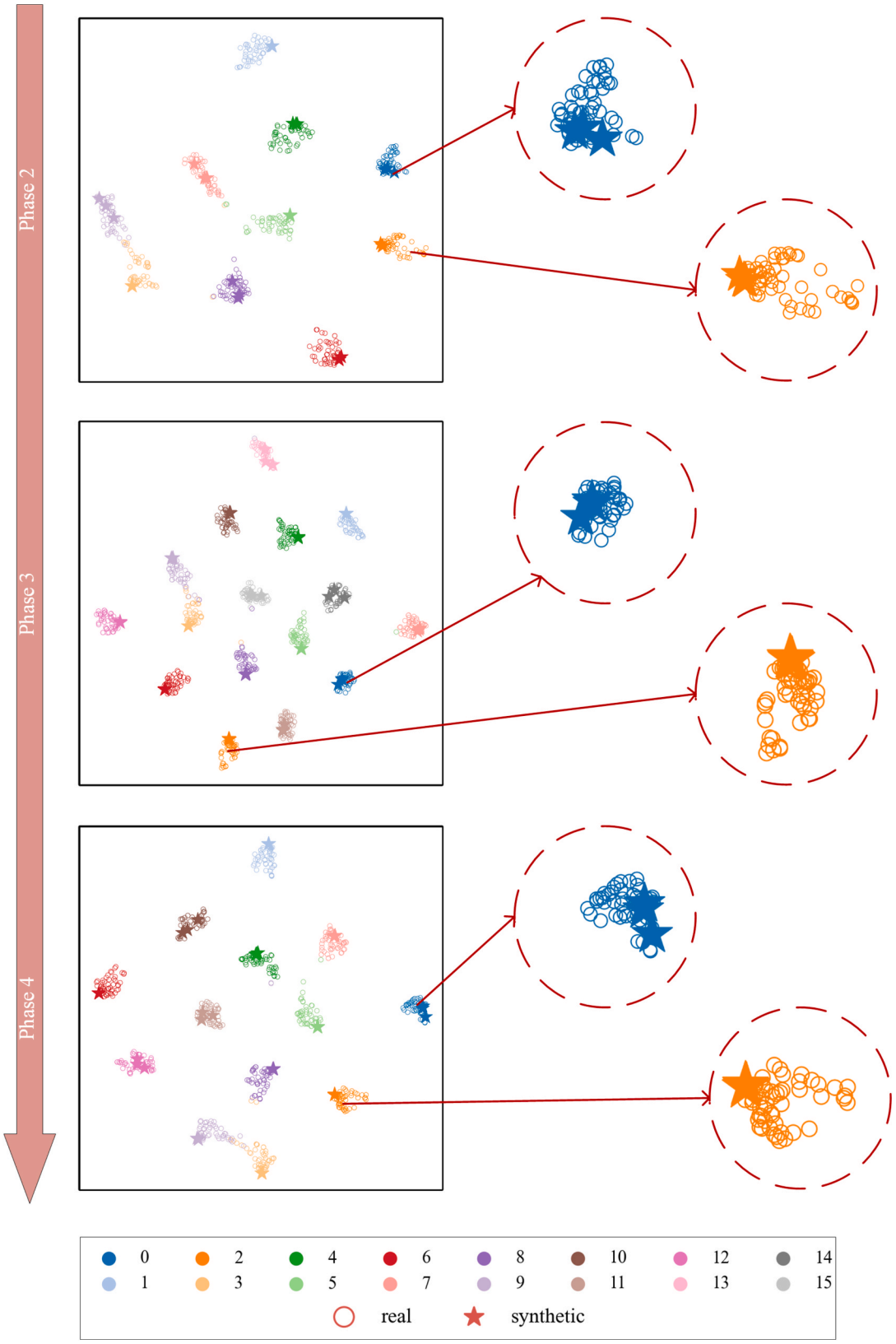


Fig. 7. Visualization of features extracted from real and synthetic data.

The learning rates α_1 and α_2 for feature extractor training and dataset distillation are 0.001 and 1, respectively, with corresponding iterations of 60 and 15000. For replay-based methods, the quantity of exemplars allocated to each fault type is set to 3, which implies that the cumulative count of exemplars increases with the diagnostic task. Each experiment is conducted 5 times, and the average accuracy and standard deviation are given.

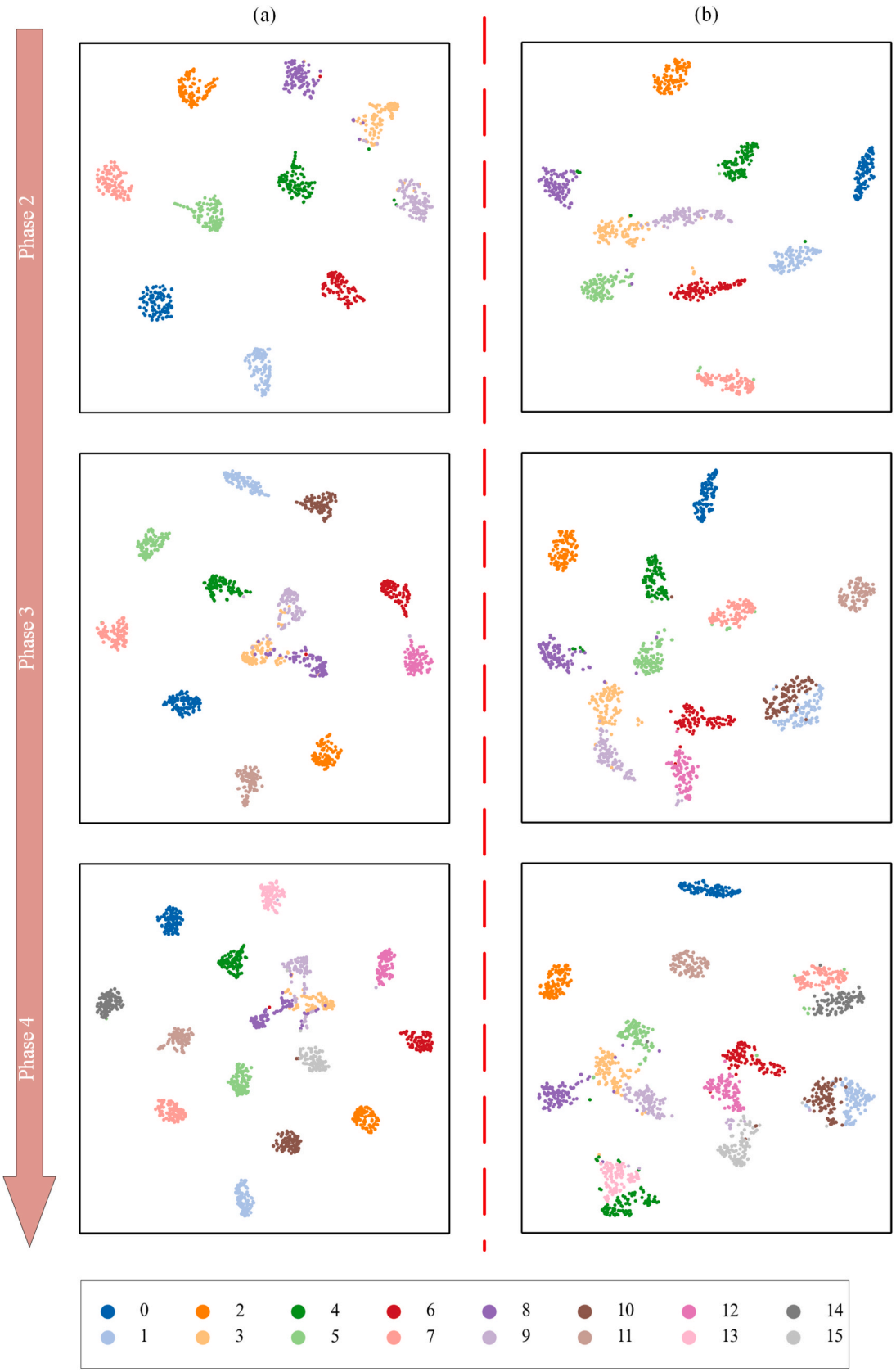


Fig. 8. Visualization of features extracted in last three phases. (a) A2. (b) A3.

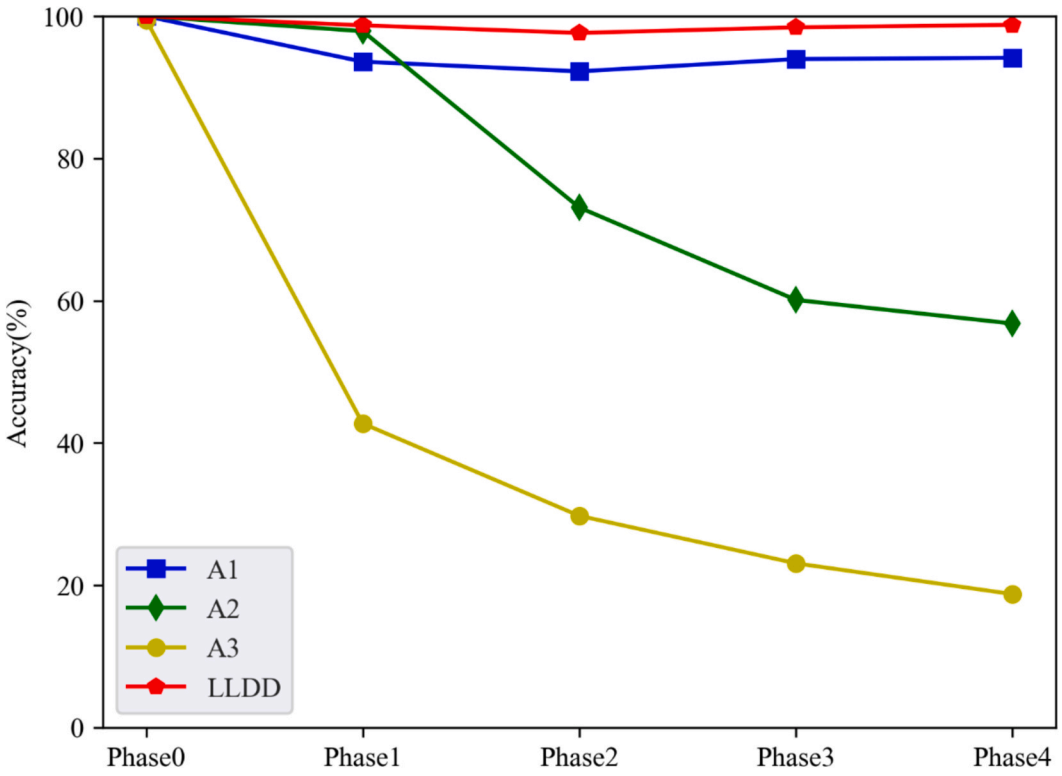


Fig. 9. Accuracy of ablation experiments.

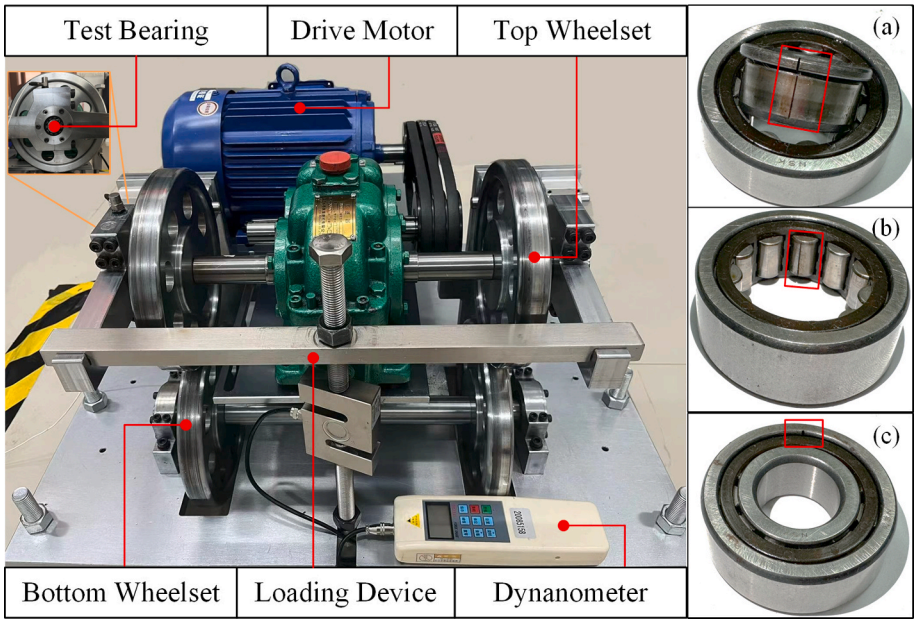


Fig. 10. Integrated wheelset drive test rig. (a) Inner fault. (b) Outer fault. (c) Ball fault.

Table 6
Average Accuracy of Different Methods under Various Exemplar Set Sizes (%).

Exemplar set size	16	32	48	64	80	96
Herding	72.33	81.82	88.61	91.48	91.87	93.77
iCaRL	80.31	84.83	89.84	90.85	92.83	92.25
LUCIR	81.46	86.21	89.63	92.64	92.88	94.55
LLDD	**94.06**	**96.98**	**98.19**	**98.48**	**98.46**	**98.31**

4.1.4. Experimental result and analysis

The experimental results are shown in Table 5. R1 and R2 are non-lifelong learning methods. They delineate the lower and upper thresholds of diagnostic precision. M1–M4 are popular lifelong learning methods. Comparative analysis with these techniques demonstrates the robustness and efficiency of our proposed methodology. Furthermore, A1–A3, which are used for ablation experiments, quantify the significance of individual components within the proposed approach in mitigating catastrophic forgetting.

1) Incremental average accuracy: Fig. 6 illustrates the trend of

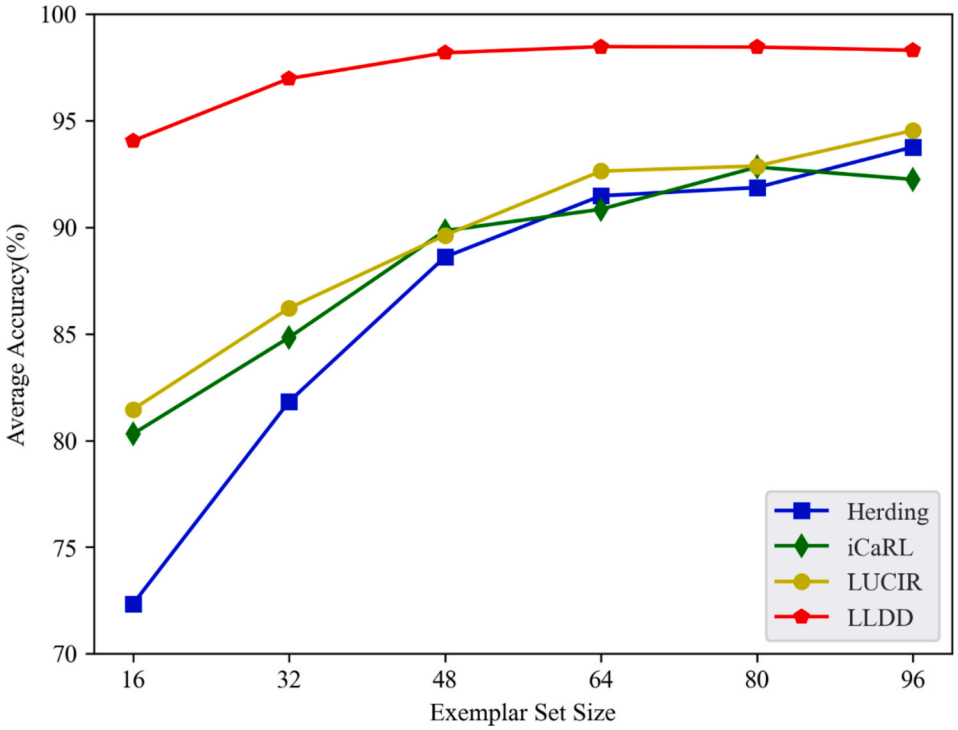


Fig. 11. Average accuracy of different methods under various exemplar set sizes.

average diagnostic accuracy for different methods under the condition of continuous learning of incremental tasks. R1 is a method of fine-tuning the network at the parameter level, in which the diagnostic accuracy drops sharply as the phases increase, reflecting the catastrophic forgetting issue faced by neural networks under incremental task conditions. The results of R2 are ideal, representing the upper bound of lifelong learning performance. M1 and M2 are based on regularization methods; their diagnostic accuracy gradually decreases as the phases progress but remains higher than that of R1. M3, M4, and LLDD utilize exemplars, achieving significantly higher diagnostic accuracy than M1 and M2 without exemplars, indicating that replay-based methods outperform regularization-based ones. Compared with the well-performing M3 and M4, the proposed method consistently achieves the highest diagnostic accuracy at every phase, making it the closest in accuracy to the R2 method among all approaches. M3 and M4 select representative exemplars from the original dataset, inevitably leading to information loss during selection. By contrast, the proposed method distills the dataset, preserving high-quality synthetic data and reducing the information loss associated with retaining exemplars.

2) Representative ability of synthetic data: T-SNE is a commonly used method for data dimensionality reduction. For visualizing the classification effect of LLDD, the features extracted from fault data by each phase model are represented in 2D coordinates. Fig. 7 visualizes the distribution of real and synthetic data obtained through T-SNE in last three phases, with each phase model exhibiting good feature extraction performance. The distributions of real and synthetic data are highly similar, suggesting that the distilled data accurately condense the information from previous fault data and can serve as exemplars to mitigate catastrophic forgetting in models. Synthetic data exhibit a representative ability comparable to that of real data.

3) Ablation experimental analysis: The effectiveness of distilled exemplars and feature distillation is validated by conducting ablation experiments. The results are shown in Fig. 9. LLDD achieves the best performance, whereas A3, which removes both methods simultaneously, has the lowest accuracy. A1, which removes feature distillation, shows a slight decline in performance at the first incremental phase and remains unchanged in subsequent phases, proving that dataset distillation and replay can effectively suppress catastrophic forgetting. A2, which replaces distilled exemplars with herding exemplars, keeps high accuracy after the first incremental learning task, but its accuracy gradually decreases in subsequent phases. This result indicates that feature distillation maintains important parameters related to past tasks in the neural network to some extent. The ability to extract features of A2 and A3 as tasks increase is compared in Fig. 8. In comparison with the result of the proposed method depicted in Fig. 7, the feature extraction capability of A1 slightly decreases with the increase in the number of fault types to be recognized, while A2 exhibits confusion. The results show that both strategies in the dual-distillation mechanism have their respective roles, with dataset distillation being more effective than feature distillation in overcoming catastrophic forgetting.

4.2. Case 2: Wheelset bearing dataset

4.2.1. Dataset description

Experiments are conducted on a wheelset bearing dataset to further validate the memory effectiveness of LLDD. Fig. 10 depicts the setup of the test rig, which comprises drive motor, wheel set, test bearing (NSK-NJ204ET-009), loading device, and dynamometer. The rotational speed of the flywheel, ranging from 400 r/min to 1200 r/min, is controlled by a drive motor. In this experiment, an acceleration signal measured at a flywheel speed of 400 r/min is used to construct a dataset of incremental fault types. The sampling frequency is 12800 Hz. Consistent with experiments on the flywheel dataset, 16 types of faults are used in this experiment. Each diagnostic task includes the types of faults shown in Table 1.

4.2.2. Comparison methods

On the wheelset dataset, LLDD is compared with several replay-based lifelong learning methods to verify that the proposed method can retain the most effective knowledge when storing exemplars. The same fault types and task settings as in the Table. 1 are used. The exemplar sizes of 16–96 correspond to storing 1–6 exemplars for each type, using the average of five random experiments as the result. The average diagnostic accuracy of each approach in the last phase is used as

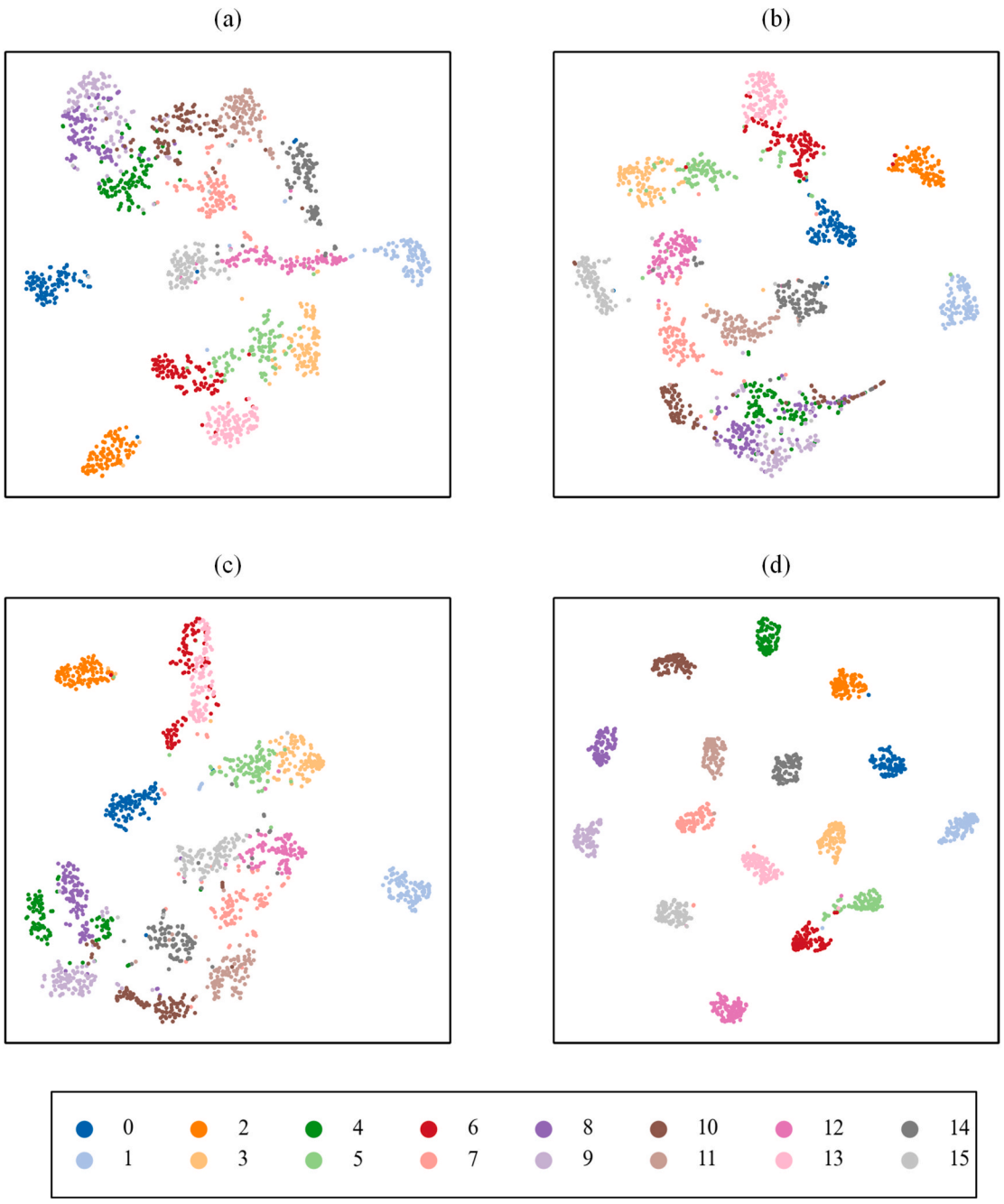


Fig. 12. Visualization of features extracted with an exemplar set size of 48. (a) Herding. (b) iCaRL. (c) LUCIR. (d) LLDD.

the evaluation criterion. Herding selects samples closest to the average feature to build the exemplar set. iCaRL adds knowledge distillation on top of herding and uses the nearest mean sample classification rule instead of a fully connected layer classifier to avoid changes in network structure when types increase. LUCIR also uses a sampling strategy the same as herding but employs a cosine normalization classifier to avoid imbalance between old and new classes.

4.2.3. *Experimental result and analysis*

It is necessary to set a proper size of exemplar set for different lifelong learning methods. The larger the size of the exemplar set, the stronger the ability to resist catastrophic forgetting. However, increasing the size of the exemplar set leads to increase storage costs. The general rule is that the smallest size that maintains high and stable diagnostic model accuracy is the optimal size for the exemplar set. Table 6 shows the average accuracy of various methods at the end of the last training phase under different exemplar set sizes.

1) Average accuracy across various exemplar set sizes: Fig. 11 reflects the change trend of accuracy for various methods as the exemplar set size increases. LLDD outperforms herding, iCaRL, and LUCIR across all exemplar set sizes. When the exemplar set size is less than 64, the accuracy of the three comparison methods gradually increases with the increase in exemplar set size. This result indicates that using more exemplars during replay retains more information from past tasks, enhancing the ability to overcome catastrophic forgetting. However, when the exemplar set size exceeds 64, because of inherent algorithmic limitations, the growth in diagnostic accuracy slows down and approaches the performance ceiling of each method. The presented method stores one exemplar per class, achieving desirable performance with an exemplar set size of 16, and quickly reaches its performance

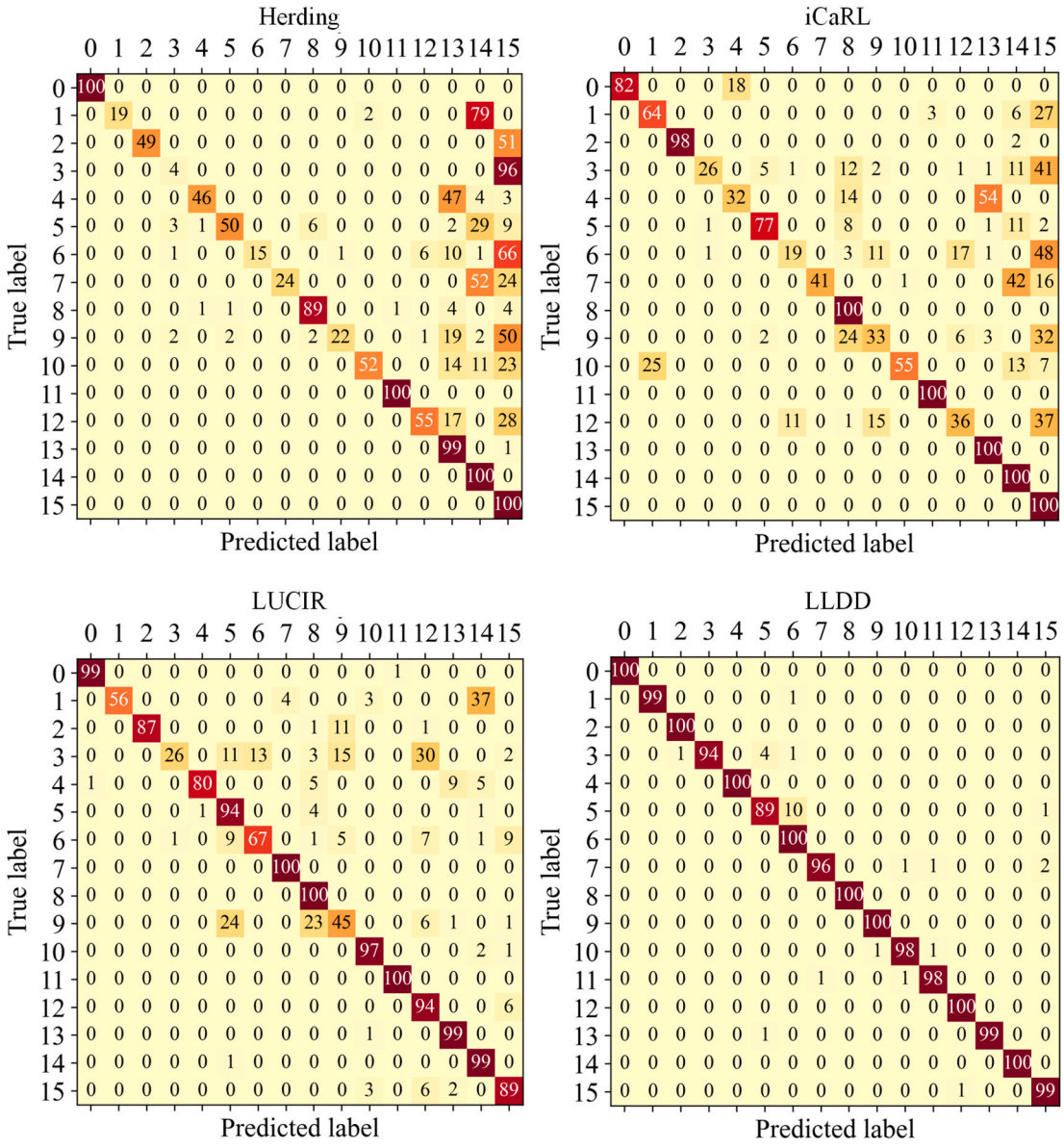


Fig. 13. Confusion matrix depicting the experimental results of four methods under an exemplar set size of 16.

limit. This finding suggests that distilled exemplars can retain more information from past tasks compared with those derived from herding. The feature extraction ability of models applying these four approaches with an exemplar set of 48 is demonstrated in Fig. 12.

2) Analysis under extreme limitation: Fig. 13 illustrates the confusion matrices for the proposed method and comparative methods during the final phase when the exemplar set size is limited to 16, meaning only one exemplar per class is stored. Under the extreme limitation of the size of the exemplar set, the three comparative methods exhibit a tendency to misclassify previously encountered fault types as new ones, suggesting a loss of knowledge from prior diagnostic tasks. By contrast, the proposed method sustains high diagnostic accuracy even with this extreme constraint on exemplar set size, presenting its excellent memory efficiency.

5. Conclusion

This study introduces a new lifelong learning method based on dual distillation for class-incremental scenarios. By combining dataset distillation and feature distillation through a dual-distillation mechanism, the proposed method can effectively alleviate catastrophic forgetting in lifelong learning. This approach generates a small amount of distilled data by distilling past task datasets, preserving key knowledge from previous tasks. Through replaying these distilled exemplar data, old knowledge is transferred to the current model. Feature distillation avoids drastic changes in important parameters by aligning representations learned by past and current models. The significance of dataset distillation and feature distillation is verified through ablation experiments. Experiments on a flywheel dataset demonstrate the effectiveness of the method, while experiments on a wheelset dataset show that LLDD has higher memory efficiency compared with other replay-based lifelong learning methods. Future work will explore the performance of the dual-distillation mechanism in cross-domain and cross-device scenarios.

CRediT authorship contribution statement

Shijun Xie: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Changqing Shen:** Writing – review & editing, Supervision, Project

administration, Funding acquisition, Conceptualization. **Dong Wang:** Visualization, Methodology, Investigation, Data curation. **Juanjuan Shi:** Validation, Software, Methodology, Formal analysis. **Weiguo Huang:** Supervision, Resources, Investigation, Formal analysis. **Zhongkui Zhu:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization.

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

This work is financially supported by the National Natural Science Foundation of China (No. 52272440) and Suzhou Science Foundation (Nos. SYG202323 and ZXL2022027).

Data availability

Data will be made available on request.

References

[1] Y. Lai, H. Shao, X. Zheng, et al., Automated fault diagnosis of rotating machinery using sub domain greedy Network Architecture search, Adv. Eng. Inf. 62 (2024) 102753.
[2] H. Zhu, C. Shen, J. Wang, et al., Few-Shot Class-Incremental Learning with Adjustable Pseudo-Incremental Sessions for Bearing Fault Diagnosis, IEEE Sens. J. 24 (12) (2024) 19543–19552.
[3] P. Zhou, S. Chen, Q. He, et al., Rotating machinery fault-induced vibration signal modulation effects: A review with mechanisms, extraction methods and applications for diagnosis, Mech. Syst. Sig. Process. 200 (2023) 110489.
[4] C. Wang, J. Yang, H. Jie, et al., An energy-efficient mechanical fault diagnosis method based on neural dynamics-inspired metric SpikingFormer for insufficient samples in industrial Internet of Things, IEEE Internet Things J. (2024), https://doi.org/10.1109/JIOT.2024.3476034.
[5] Y. Xue, C. Wen, Z. Wang, et al., A novel framework for motor bearing fault diagnosis based on multi-transformation domain and multi-source data, Knowl.-Based Syst. 283 (2024) 111205.
[6] Q. Chen, X. Dong, Z. Peng, Interpreting What Typical Fault Signals Look Like via Prototype-matching, Adv. Eng. Inf. 62 (2024) 102849.
[7] L. Song, Y. Jin, T. Lin, et al., Remaining Useful Life Prediction Method Based on the Spatiotemporal Graph and GCN Nested Parallel Route Model, IEEE Trans. Instrum. Meas. 73 (2024) 1–12.
[8] D. Ruan, J. Wang, J. Yan, et al., CNN parameter design based on fault signal analysis and its application in bearing fault diagnosis, Adv. Eng. Inf. 55 (2023) 101877.
[9] Y. Hou, J. Wang, Z. Chen, et al., Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved Transformer, Eng. Appl. Artif. Intel. 124 (2023) 106507.
[10] H. Shao, W. Li, B. Cai, et al., Dual-threshold attention-guided GAN and limited infrared thermal images for rotating machinery fault diagnosis under speed fluctuation, IEEE Trans. Ind. Inf. 19 (9) (2023) 9933–9942.
[11] C. Huo, Q. Jiang, Y. Shen, et al., Enhanced transfer learning method for rolling bearing fault diagnosis based on linear superposition network, Eng. Appl. Artif. Intel. 121 (2023) 105970.
[12] K. Zhao, Z. Liu, B. Zhao, et al., Class-aware adversarial multiwavelet convolutional neural network for cross-domain fault diagnosis, IEEE Trans. Ind. Inf. 20 (3) (2023) 4492–4503.
[13] K. Hu, Q. He, C. Cheng, et al., Adaptive incremental diagnosis model for intelligent fault diagnosis with dynamic weight correction, Reliab. Eng. Syst. Saf. 241 (2024) 109705.
[14] S. Wang, Y. Lei, N. Lu, et al., Graph Continual Learning Network: An Incremental Intelligent Diagnosis Method of Machines for New Fault Detection, IEEE Trans. Autom. Sci. Eng. (2024), https://doi.org/10.1109/TASE.2024.3417208.
[15] D. Li, S. Liu, F. Gao, et al., Continual learning classification method and its application to equipment fault diagnosis, Appl. Intell. 52 (1) (2022) 858–874.
[16] H. Zhu, C. Shen, L. Li, et al., Reserving embedding space for new fault types: A new continual learning method for bearing fault diagnosis, Reliab. Eng. Syst. Saf. 252 (2024) 110433.
[17] M. Sun, X. Xiao, T. Chen, et al., A Novel Domain Incremental Learning Method for Bearing Fault Diagnosis Based on F&K, IEEE Trans. Ind. Inf. (2024), https://doi.org/10.1109/TII.2024.3470908.
[18] Y. Zhang, C. Shen, J. Shi, et al., Deep adaptive sparse residual networks: A lifelong learning framework for rotating machinery fault diagnosis with domain increments, Knowl.-Based Syst. 293 (2024) 111679.
[19] J. Zheng, H. Xiong, Y. Zhang, et al., Bearing fault diagnosis via incremental learning based on the repeated replay using memory indexing (R-REMIND) method, Machines. 10 (5) (2022) 338.
[20] Y. Liu, B. Chen, D. Wang, et al., A lifelong learning method based on generative feature replay for bearing diagnosis with incremental fault types, IEEE Trans. Instrum. Meas. 72 (2023) 1–11.
[21] Smith J S, Tian J, Halbe S, et al. A closer look at rehearsal-free continual learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 2410-2420.
[22] M. De Lange, R. Aljundi, M. Masana, et al., A continual learning survey: Defying forgetting in classification tasks, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2021) 3366–3385.
[23] Z. Li, D. Hoiem, Learning without forgetting, IEEE Trans. Pattern Anal. Mach. Intell. 40 (12) (2017) 2935–2947.
[24] Aljundi R, Babiloni F, Elhoseiny M, et al. Memory aware synapses: Learning what (not) to forget. Proceedings of the European conference on computer vision (ECCV). 2018: 139-154.
[25] Mallya A, Lazebnik S. Packnet: Adding multiple tasks to a single network by iterative pruning. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 7765-7773.
[26] Abati D, Tomczak J, Blankevoort T, et al. Conditional channel gated networks for task-aware continual learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3931-3940.
[27] S.A. Rebuffi, A. Kolesnikov, G. Sperl, et al., icarl: Incremental classifier and representation learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2001–2010..
[28] Xiang Y, Fu Y, Ji P, et al. Incremental learning using conditional adversarial networks. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6619-6628.
[29] R. Yu, S. Liu, X. Wang, Dataset distillation: A comprehensive review, IEEE Trans. Pattern Anal. Mach. Intell. 46 (1) (2023) 150–170.
[30] Zhao B, Bilen H. Dataset condensation with distribution matching. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 6514-6523.
[31] Hinton G. Distilling the Knowledge in a Neural Network. arxiv preprint arxiv: 1503.02531. 2015.
[32] J. Gou, B. Yu, S.J. Maybank, et al., Knowledge distillation: A survey, Int. J. Comput. Vis. 129 (6) (2021) 1789–1819.
[33] W. Liu, R. Lin, Z. Liu, et al., Learning with hyperspherical uniformity, in: International Conference on Artificial Intelligence and Statistics, 2021, pp. 1180–1188.
[34] Y. Shen, X. Sun, X.S. Wei, Equiangular basis vectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11755–11765.
[35] Hou S, Pan X, Loy C C, et al. Learning a unified classifier incrementally via rebalancing. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 831-839.