A CNN-Multi-Head Attention Framework for Gearbox Incremental Fault Diagnosis Under Non-Stationary Conditions

Hao Zhang[1,2,3], Shunuan Liu[1,2,3], Bin Luo[1,2,3], Konstantinos Gryllias[4], and Chenyu Liu[1,2,3,*]

[1]*School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, China*

[2]*Key Laboratory of Aircraft High Performance Assembly, Ministry of Industry and Information Technology, Xi'an, China*

[3]*Key Laboratory of Aeronautics and Astronautics High Performance Assembly of Shaanxi, Xi'an, China*

[4]*KU Leuven, Leuven, Belgium*

* *chenyuliu@nwpu.edu.cn*

ABSTRACT

Deep learning-based gearbox fault diagnosis approaches have demonstrated exceptional performance in achieving accurate fault identification across diverse industrial applications. Nonetheless, machines frequently operate under conditions characterized by time-varying speeds or loads, known as non-stationary working conditions. When a series of different non-stationary conditions tasks are sequentially input into the model for training, an issue arises where the model tends to forget previous tasks, a phenomenon referred to as "catastrophic forgetting". To address the challenge posed by task increments within non-stationary conditions, this paper proposes an incremental learning-based multi-task fault diagnosis framework under non-stationary conditions. This methodology enhances the model's diagnostic capabilities under non-stationary conditions by amalgamating convolutional neural network (CNN) with multi-head self-attention mechanisms. It employs exemplar replay and hybrid cross-head knowledge distillation techniques to preserve the model's understanding of prior tasks, thereby facilitating the incremental learning of multiple tasks. The efficacy of this proposed framework is substantiated through its application on the MCC5-THU fault diagnosis datasets of gearbox under time-varying speed working conditions. Experimental results demonstrate that this approach significantly mitigates the "catastrophic forgetting" effect, thereby offering a robust solution for multi-tasks increment fault diagnosis of gearbox operating under non-stationary conditions.

1. INTRODUCTION

Gearbox is extensively utilized across multiple domains of modern industry, such as aviation, chemical manufacturing, and power generation, serving as a fundamental component of industrial production systems (Du, Chen, Zhang, & Yan, 2015). The sustained and stable functioning of gearbox is critical for ensuring the continuity and efficiency of industrial productivity. Even a minor malfunction in such equipment can trigger a series of chain reactions, potentially leading to significant economic losses across the production line and broader industrial systems, as well as posing serious risks to the safety of personnel (Peng, Qiao, Cheng, & Qu, 2021). Therefore, the rapid and accurate diagnosis of faults of gearbox, along with timely detection and maintenance during the early stages of failure, is of critical importance (Feng, Chen, & Zuo, 2018).

The advancement of artificial intelligence has drawn significant attention to data-driven diagnostic methods based on machine learning. Compared to traditional signal processing techniques, data-driven methods not only significantly enhance diagnostic accuracy but also enable end-to-end detection. Widely adopted algorithms for these methods encompass support vector machines (Yin & Hou, 2016), multi-layer perceptrons (Sinitsin, Ibryaeva, Sakovskaya, & Eremeeva, 2022), and convolutional neural networks (CNN) (Qin et al., 2024), among others. For instance, Lin, Li, Yang, and Wang (2018) constructed a GAN-CNN few-samples fault diagnosis model: GAN generates virtual samples in the target domain through cross-domain feature learning, and CNN fuses virtual and real samples to improve the fault diagnosis accuracy under few-samples conditions. Yu et al. (2023) introduced an intelligent diagnostic method based on an autoencoder network, utilizing a multi-layer sparse autoencoder to classify and identify frequency domain signals of various faults, thereby achieving fault diagnosis for aircraft engine

bearings. However, these methods are currently used mainly for gearbox fault diagnosis under stable speed and load conditions. In real-world industrial settings, due to variations in production tasks and operational conditions, a significant portion of mechanical equipment operates under non-stationary working conditions, and the speed or load of some equipment may even fluctuate periodically (Liu, Wang, Yang, & Qin, 2019). Speed variations cause dynamic shifts in characteristic frequencies, while load fluctuations alter the amplitude modulation characteristics of vibration signals. Consequently, non-stationary operating conditions induce significant domain drift in gearbox fault information. This phenomenon not only obscures discriminative fault features but also complicates the establishment of reliable mappings between fault signatures and corresponding categories (Dong, Jiang, Yao, Mu, & Yang, 2024).

So far, more and more researches focuses on using Deep Learning (DL)-based algorithms to deal with domain drift issue caused by non-stationary conditions. Zhao, Kang, Tang, and Pecht (2017) developed a ResNet framework integrated with wavelet packet decomposition that dynamically adjusted wavelet coefficients across frequency bands, achieving adaptive fault feature extraction and improving gearbox diagnosis accuracy under non-stationary conditions. Zhao et al. (2022) proposed a 1D-CNN method that combined the Fisher criterion with an adaptive activation function. By embedding the Fisher discriminant criterion into the network to optimize the feature projection direction and combining it with an adaptive activation function to adjust the nonlinear mapping, accurate diagnosis of gearbox faults under non-stationary conditions was achieved.

The aforementioned methods enhance the efficiency and accuracy of gearbox fault diagnosis but most research predominantly focus on one specific diagnostic task, which means the model is trained once-for-all. In actual industrial environments, it is unrealistic to collect and learn fault information of gearbox under all possible operating conditions at one time. As new working conditions continuously emerging, the model needs to be retrained in order to adapt to the streaming data . This re-training process is recognised as a “new task” (Wang, Xiong, & He, 2023). A critical challenge arises when the model is incrementally updated with new task data: the inherent discrepancy between feature distributions of historical and novel tasks triggers parameter drift. This drift biases the model toward the feature domain of the latest task, resulting in the loss of knowledge from old tasks—a phenomenon termed catastrophic forgetting (Shi et al., 2024).

To mitigate this limitation, Incremental Learning (IL)-based approaches have emerged as a promising solution. These methods aim to develop adaptive models capable of continuous evolution, which learn new tasks while retaining knowledge from previously learned tasks by designing task-adaptive knowledge retention mechanisms (Wang, Liu, & Xiao, 2024). Current IL techniques are primarily categorized into replay-based methods (Ostapenko, Puscas, Klein, Jahnichen, & Nabi, 2019; Rebuffi, Kolesnikov, Sperl, & Lampert, 2017; Yin et al., 2020), constraint-based methods (Kirkpatrick et al., 2017; Li & Hoiem, 2017; Lopez-Paz & Ranzato, 2017), and structure-based methods (Mallya & Lazebnik, 2018; Veniat, Denoyer, & Ranzato, 2020). Some of the IL methods have been applied for mechanical systems' fault diagnostics. Zhang et al. (2024) developed a task-aware ResNet architecture with dynamic parameter adaptation, which can selectively preserve task-critical parameters. The method was applied to a class-incremental diagnostic case where new gearbox faults emerge among different tasks. Chen et al. (2022) introduced a dual-branch CNN architecture with dynamic-stable aggregation, where the weights of the dynamic and stable branches were adaptively adjusted to balance diagnostic performance between new and historical fault categories. This approach facilitated IL of gearbox fault categories under the operating condition of motor speed 1496 rpm. Chen et al. (2023) proposed a continuous domain adaptive diagnosis framework that retains prior task knowledge through classifier solidification and aligns the distribution of new and old tasks through the maximum mean difference, thus enabling multi-task bearing fault diagnosis under stationary operating conditions. Experiments showed that this method can incrementally learn tasks under six different operating conditions: 500 rpm-20N, 1000 rpm-20N, 1500 rpm-40N, etc.

On the one hand, these IL-based methods can maintain the fault classification accuracy with less knowledge forgotten dealing with multiple diagnostic tasks. On the other hand, most current methods are mostly proposed for IL under stationary working conditions with constant speed and load. In real industrial environment, different non-stationary conditions always appear through the diagnostic tasks. Therefore, a thorough research on multi-task IL for gearbox fault diagnosis under non-stationary conditions is of great importance.

This paper proposes a multi-head self-attention IL model for gearbox fault diagnostics under non-stationary working conditions. A CNN framework was adopted to extract fault-related features, which were further sent to a fully connected feed-forward network. The correlation weights within the fault feature space were adaptively adjusted from multiple perspectives, thereby improving the model's understanding of the fault features. The model used exemplar replay strategy to achieve data-level knowledge retention. By constructing a hybrid cross-head knowledge distillation loss between the teacher model and student model, the knowledge of old tasks was continuously transferred during the training process, achieving parameter-level knowledge retention. Finally, the proposed method was validated on a real-world gearbox fault diagnosis datasets under non-stationary working conditions. Experiments demonstrated

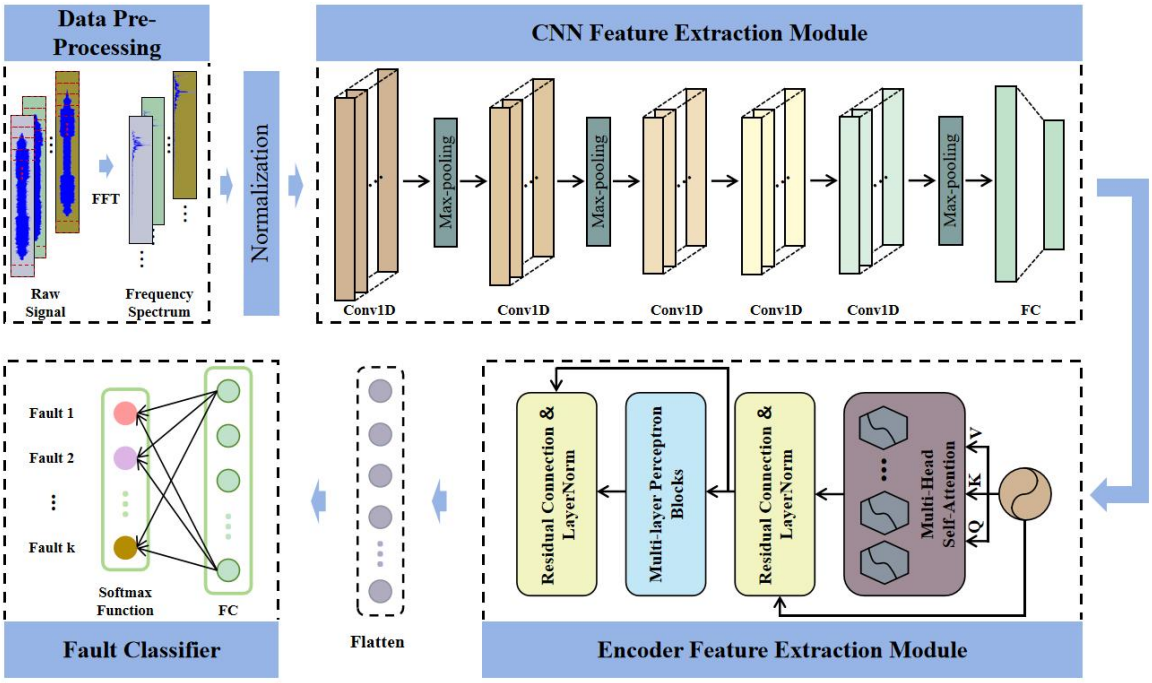

Figure 1. Diagnostic model network structure.

that the method outperformed existing approaches in multi-task IL for gearbox fault diagnosis under non-stationary scenarios.

The rest of this paper is structured as follows: Section 2 presents the proposed framework and delineates its training process. Section 3 and Section 4 demonstrates the efficacy of the proposed framework through validation on a gearbox dataset under non-stationary conditions. The paper concludes with a summary in Section 5.

2. METHODOLOGY

2.1. Network architecture

In this study, a five-layer CNN was implemented to build the feature extraction module, which constitutes the initial phase of the feature processing pipeline. In the second phase, the feature extraction module was constructed using a multi-head self-attention encoder layer to improve contextual understanding of the features obtained from the first stage. The detailed architecture of the network is illustrated in Figure 1. The acquired time-domain signals first underwent truncation, followed by spectral transformation via FFT to derive frequency-domain representations of fault characteristics. The spectrum was standardized via min-max amplitude normalization and subsequently fed into the deep neural network. After processing through the two feature extraction modules, the feature information was flattened and passed to the final classifier to achieve fault classification.

2.2. Training process

The IL process can be categorized into two phases: the initial phase and the later IL phase. In order to emulate the data condition in real-time industrial gearbox fault diagnostic scenarios, this paper makes the following assumptions: (1) Owing to constraints related to data confidentiality and limited storage capacity during machine operating, once the model completes a diagnostic task, only a small subset of samples (constituting less than 3% of the total sample size) is retained, while the remaining data becomes inaccessible. (2) After the completion of learning for the current task, the parameters of the diagnostic model are preserved temporarily; however, these parameters are promptly deleted after the model finishes learning the subsequent new task to alleviate storage demands. As the model undergoes IL across multiple tasks, its classification capability is progressively enhanced, enabling it to perform fault diagnosis under multiple operational conditions simultaneously.

2.2.1. Initial phase (Phase 0)

The initial stage aligns with conventional DL training methodologies. Following data preprocessing, the samples of all kinds of faults are partitioned into training and testing sets, after which classification training is conducted. After completion of the training process, the model parameters are retained, and all training samples are replayed and forward propagated to capture the feature map generated by the intermediate feature extraction layer. The fault feature space of each sample is derived from the average of the input and

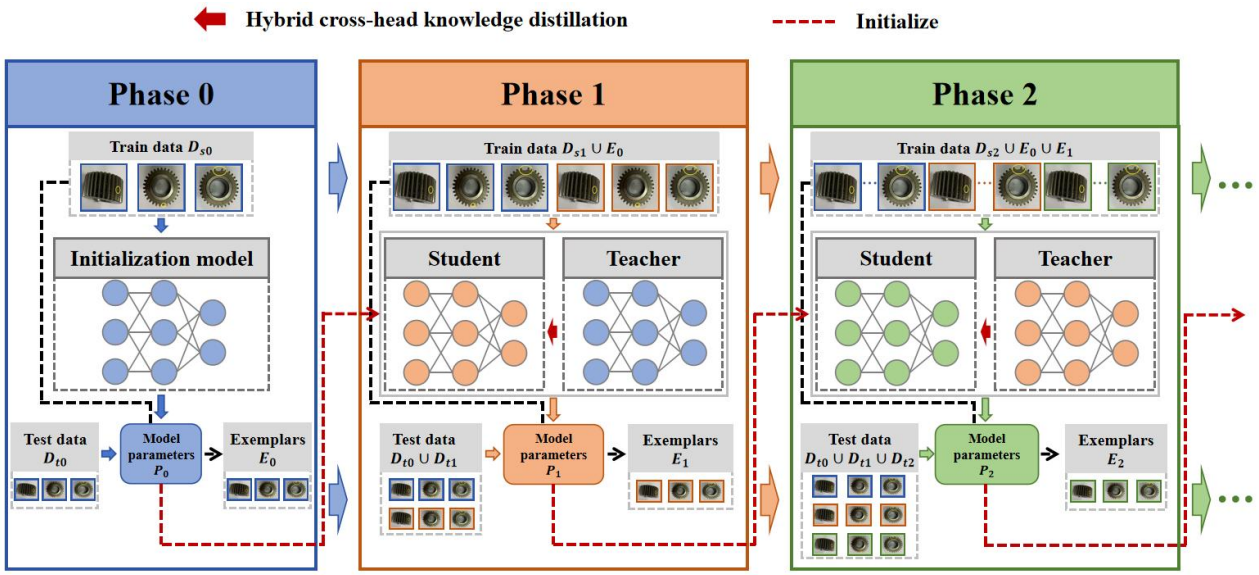

Figure 2. The pipeline of proposed framework.

output sequences in the multi-head self-attention layer, as these sequences may contain richer and more effective information. Subsequently, the sample closest to the centroid of the feature space for each fault is iteratively selected. The mathematical representation of the average feature space for each sample is provided in Eq. (1), the centroid of the feature space for each fault category is detailed in Eq. (2), and the iterative exemplar selection process is formalized in Eq. (3).

$$F_m(x_{i,j}^k, \theta_i) = mean(F_1(x_{i,j}^k, \theta_i), F_2(x_{i,j}^k, \theta_i)) \quad (1)$$

$$u_i^k = \frac{1}{N_i^k} \sum_{j=1}^{N_i^k} F_m(x_{i,j}^k, \theta_i) \quad (2)$$

$$E_d \leftarrow argmin_{x \epsilon X_k} \left\| u_i^k - \frac{1}{d} [F_m(x) + \sum_{j=1}^{d-1} F_m(E_j)] \right\| \quad (3)$$

where $d \in [1, e]$, e is the number of exemplars, $x_{i,j}^k$ represents the j-th sample of fault k in task i, and θ_i represents the model parameters after learning task i. The exemplars are subsequently selected and stored utilizing Eq. (1) - (3).

2.2.2. Incremental Learning phases

During the IL phase, the methods of exemplar replay and hybrid cross-head knowledge distillation were employed to preserve the model's diagnostic capabilities for previously learned tasks. The diagnostic procedure is illustrated in Figure 2. When a new task t_n is introduced, the process transitions to incremental stage n, during which both the student and teacher models are initialized using the model parameters P_{n-1}. The training data are $D_{sn} \cup E_{n-1} \cup \ldots \cup E_0$, where D_{sn} represents the training set for the new task t_n, and E_i denotes the exemplars retained from task i, $i \in [1, n]$. To mitigate the bias arising from the imbalance in sample sizes between new and old tasks, each batch is constructed to include data from the new task alongside exemplars from the older tasks.

After feeding the frequency spectrum into both the student model and the teacher model, the respective outputs p^s and p^t are generated. Concurrently, the output features from the encoder layer of the student model are passed through the classifier of the teacher model to produce the cross-prediction distribution $\hat{p}^s$. Recent research has demonstrated that minimizing the Kullback-Leibler (KL) divergence between the cross prediction and the teacher prediction, referred to as the cross distillation loss, enables the student model to focus more effectively on learning the feature extraction capabilities of the teacher model during backpropagation (Wang et al., 2024). As illustrated in Figure 3, this paper defines the knowledge distillation loss as the sum of the soft loss and the cross distillation loss.

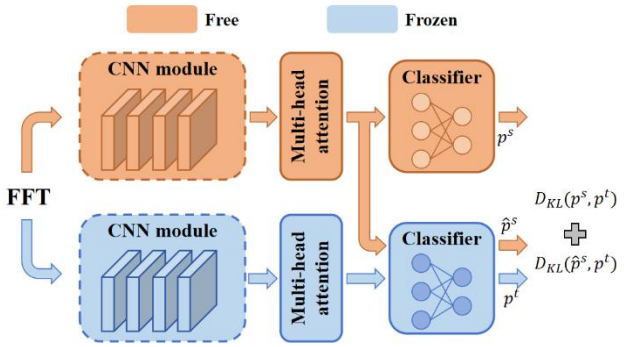

Figure 3. Hybrid cross-head knowledge distillation.

The mathematical formulation of the KL divergence is provided in Eq. (4), while the hybrid cross-head knowledge distillation loss is detailed in Eq. (5).

$$D_{KL}(p^s, p^t) = \frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} \left(\sum_j^c p^s(j) \times log\left(\frac{p^s(j)}{p^t(j)}\right) \right) \quad (4)$$

$$L_{dis}(p^s, p^t) = D_{KL}(p^s, p^t) + D_{KL}(\hat{p}^s, p^t) \quad (5)$$

where N_{batch} is the size of batch, c is the size of fault prediction sequence, p^s and p^t are the output sequences of the student model and the teacher model respectively, and $\hat{p}^s$ is the cross prediction. The cross-entropy loss function is utilized to quantify the discrepancy between p^s and the truth labels of the samples, thereby constructing the hard loss function $L_{hard}(p^s, y)$. The overall loss function, which comprises two components, $L_{hard}(\hat{y}, y)$ and $L_{dis}(p^s, p^t)$, is formulated as presented in Eq. (6).

$$Loss = \alpha \cdot L_{hard}(p^s, y) + (1 - \alpha) \cdot L_{dis}(p^s, p^t) \quad (6)$$

where α is a weight factor used to control the weights of the two losses and balance the model's diagnostic capabilities for new and old tasks. y represents the true label of the sample.

3. EXPERIMENTS

3.1. Dataset and experiment setup

3.1.1. MCC5-THU dataset

The proposed framework was validated using the MCC5-THU variable working condition gearbox fault dataset (Chen, Liu, He, Zou, & Zhou, 2024). The test rig, as illustrated in Figure 4, was primarily composed of a motor, a torque sensor, a two-stage parallel gearbox, a magnetic powder brake, and two triaxial vibration sensors. The vibration sensors were positioned at the motor's output end and the exterior of the gearbox's intermediate shaft to capture vibration data under non-stationary conditions, with a sampling frequency of 12.8 kHz. The gear module was 1.5 mm with a face width of 10 mm. Both the faulty gears and bearings were located on the intermediate shaft. The data from five distinct single gear fault types and a healthy condition were used to validate the accuracy of the model. Each fault was obtained by laser etching with an accuracy of 0.01mm. The detailed fault information and the corresponding label settings are shown in Table 1.

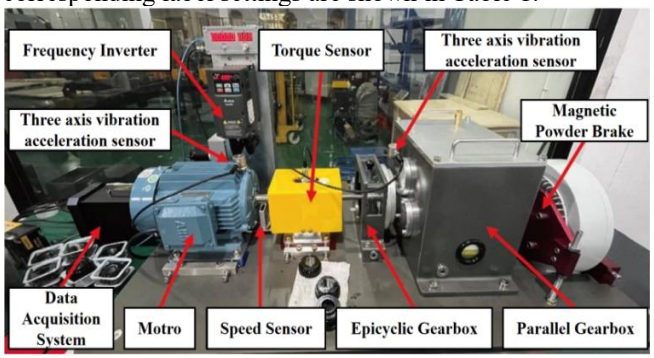

Figure 4. The gearbox fault test rig (Chen, Liu, He, Zou, & Zhou, 2024).

The experiments in MCC5-THU dataset use a constant torque of 10 Nm, and three speed levels, i.e., 0-500-1000 RPM, 0-1500-2000 RPM, and 0-2500-3000 RPM. The speed spectrum is illustrated in Figure 5, where the speed is modulated in a stepwise fashion. The overlaid annotations in y-axis indicate the three speed levels following the same fashion.

Table 1. Fault types and label settings.

Fault type	Fault severity	Label
Health	\	0
Gear pitting	Fault diameter 1.5 mm	1
Gear wear	Full teeth surface area	2
Miss teeth	\	3
Teeth break	3/4 of the teeth width	4
Teeth crack	3/4 of the teeth height	5

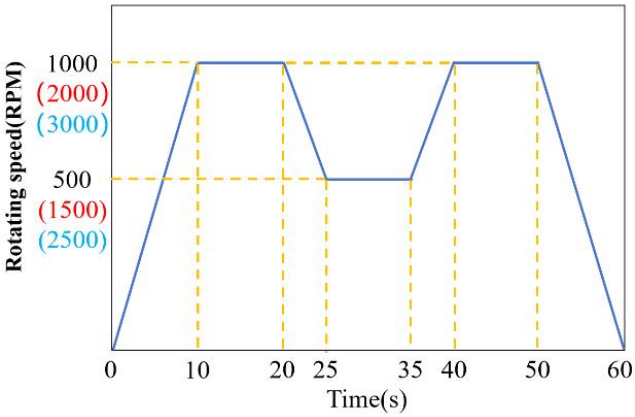

Figure 5. The time-varying rotational speed curve (Chen, Liu, He, Zou, & Zhou, 2024).

Table 2. The speed levels of the three tasks.

Phase	Task	Speed range [rpm]	Number of samples
Phase 0	Task 0	0-500-1000	4932
Phase 1	Task 0	0-500-1000	144
	Task 1	0-1500-2000	4932
Phase 2	Task 0	0-500-1000	144
	Task 1	0-1500-2000	144
	Task 2	0-2500-3000	4932

In order to emulate the incremental operating conditions in the real industrial environment, three phases with multiple tasks were defined for model training. Phase 0 was the initial phase of the IL process, and the fault data under non-

Table 3. Parameters of the proposed model.

Component	Layers	Filter	Filter number	Output size	Activation function
CNN feature extraction module	Input	-	-	36×1×2560	-
	Conv 1D_1	11×1/4×1	64	36×64×639	ReLU
	Maxpool	3×1/2×1	64	36×64×319	-
	Conv 1D_2	5×1/1×1	192	36×192×319	ReLU
	Maxpool	3×1/2×1	192	36×192×159	-
	Conv 1D_3	3×1/1×1	384	36×384×159	ReLU
	Conv 1D_4	3×1/1×1	256	36×256×159	ReLU
	Conv 1D_5	3×1/1×1	256	36×256×159	ReLU
	Maxpool	3×1/2×1	256	36×256×79	-
	Flatten	-	-	36×20224	-
	Dropout	-	0.5	36×20224	-
	FC	20224×4096	-	36×4096	ReLU
	Dropout	-	0.5	36×4096	-
Encoder feature extraction module	Multi-head self-attention mechanism	Head_num:8 Depth:512 Q_dim:64 K_dim:64 Series_dim:8	-	36×8×512	-
	Add＆Norm	Layer norm	-	36×8×512	-
	Fully connected	-	-	36×8×2048	ReLU
		-	-	36×8×512	-
	Add＆Norm	Layer norm	-	36×8×512	ReLU
Flatten	-	-	-	36×4096	-
Classifier	FC	4096×6	-	36×6	-

stationary conditions with a speed of 0-500-1000 RPM was used as Task 0 to train the initial model. In Phase 1, newly acquired fault data from distinct non-stationary working conditions were designated as Task 1 to retrain the model, while replaying exemplars from Task 0. The same framework applies to Phase 2 for IL. There were 4932 samples for each task, and 3600 of them were used for training. The rest 1332 samples were used for testing in each phase. Detailed information of the tasks is provided in Table 2.

Considering the substantial vibration signal attenuation induced at near-zero rotation speeds during equipment startup/shutdown phases (0-1.5s and 56.5-60s), this study exclusively retained vibration data within the range of 1.5s to 56.5s. A sliding window is used to sample the original vibration data, with a window length of 5120 points and a sliding step size of 2560 points.

3.1.2. Experiment settings

The detailed parameters of the proposed network framework are presented in Table 3. The CNN feature extraction module comprised 5 convolutional layers. Dropout layers with a retention rate of 0.5 were incorporated to mitigate overfitting. The encoder feature extraction module was constructed with an 8-head self-attention layer followed by a feedforward neural network layer. The output sequence from the CNN feature extraction module was segmented into eight subsequences, each of length 512, to serve as contextual information. The dimensions of the query (Q) and key (K) matrices were set to 64. Following the CNN and encoder feature extraction modules, the processed information was ultimately mapped to six fault categories through a classifier consisting of a linear layer.

During the IL process, different tasks were sequentially introduced into the model. In the training phase, the batch size was set to 36, and the model was trained for 400 epochs. The optimization process employed Stochastic Gradient Descent (SGD) with a learning rate of 1×10-4 and a decay rate of 0.99. The loss function was composed of the hard loss and the hybrid cross-head knowledge distillation loss, with a weighting factor α set to 0.9 to ensure the model's predictive accuracy for new tasks. To smooth the output distribution, and facilitate the learning of more generalized features, the temperature coefficient T for the distillation term was set to 2 according to (Li & Hoiem, 2017). Within each batch, the new task data constituted 24 samples, while

Table 4. Diagnostic accuracies and BWTs of the IL methods in different phases.

Method	Phase 0	Phase 1			Phase 2			
	Task 0	Task 0	Task 1	BWT	Task 0	Task 1	Task 2	BWT
EWC	96.24%	58.18%	95.57%	-0.3806	61.41%	83.25%	97.29%	-0.2357
IL-VOC	97.74%	94.52%	97.52%	-0.0322	91.14%	96.31%	97.52%	-0.0391
Proposed framework	**99.47%**	**97.29%**	**99.47%**	**-0.0218**	**96.47%**	**98.87%**	**99.69%**	**-0.0180**

the older data comprised a total of 12 samples. In Phase 1, Task 0 accounted for 12 samples per batch. In Phase 2, the old data from Task 0 and Task 1 each contributed 6 samples, with this pattern continuing for subsequent tasks. The random seed was fixed to ensure reproducibility.

3.2. Comparative experiments

3.2.1. Comparison methods

The diagnostic performance of the proposed model was evaluated through a comparative analysis with two previously published IL methods. To ensure fairness, the normalized frequency spectrum was also used as input data for the other two methods.

(1) Elastic Weight Consolidation

Elastic Weight Consolidation (EWC) is a weight parameter constraint approach in IL, designed to quantify the significance of model parameters for previously learned tasks. By incorporating an importance regularization term into the loss function, the method restricts the update gradient of model parameters critical to old tasks when learning new tasks, thereby preserving prior knowledge (Kirkpatrick et al., 2017).

The EWC model comprised three fully connected layers, which served as input layer, intermediate hidden layer and classifier respectively, and the dimension of the intermediate hidden layer was 800×800. ReLU activation function and dropout layer with retention rate of 0.5 were added after the input layer and hidden layer. After completing the training of t_n, all training samples from t_n were replayed to compute and store the Fisher information matrix. During the Phase of n+1, the regularization loss was defined based on the Fisher information matrix to penalize updates to parameters sensitive to old tasks. In this method, the hyperparameter λ, which governs the degree of protection for old knowledge, was set to 1×108. The training process involved 400 epochs, utilizing SGD as the optimizer, with a learning rate of 1×10-3 and a decay rate of 0.99.

(2) IL-based varying operating conditions diagnosis method

IL-based varying operating conditions diagnosis method (IL-VOC) is developed for comparative analysis based on the methodology outlined in reference (Wang, Xiong, & He, 2023), representing a multi-task diagnostic approach for rotating machinery that incorporates IL principles. The method leverages CNN as its foundational framework, integrating multiple IL mechanisms to effectively mitigate forgetting. It demonstrates robust multi-task learning capabilities, exhibiting consistent performance stability under steady-state operational conditions. The optimization process employs the Adam optimizer with a learning rate of 1×10-2. During the IL phase, the weight matrix in the total loss function was set to [1.0, 0.001, 0.1, 0.1]. The temperature coefficient for the distillation loss was configured as T=2, and the batch size was maintained at 36.

3.2.2. Evaluation metrics

After the completion of training in each phase, the model's classification accuracy across all previously learned tasks was assessed. To comprehensively evaluate the model's performance, this paper employs two key metrics to analyse the IL capabilities. The first metric (Accs) is the average classification accuracy across all learned tasks, which serves as an indicator of the model's overall diagnostic performance. The second metric, Backward Transfer (BWT), quantifies the model's tendency to forget previously acquired knowledge. The mathematical formulation of BWT is provided in Eq. (7).

$$\mathrm{BWT} = \frac{1}{n}\sum_{i=0}^{n-1} R_{n,i} - R_{i,i} \tag{7}$$

where n represents the phase number, and the value of BWT ranges between [-1, 1]. $\mathrm{BWT} = 0$ indicates no forgetting of previous tasks. When $\mathrm{BWT} > 0$, it suggests that the learning of subsequent tasks has positively influenced the accuracy of prior tasks. Conversely, $\mathrm{BWT} < 0$ signifies the presence of forgetting, with lower BWT values indicating a higher degree of forgetting.

4. RESULTS

4.1. Comparative analysis of different IL methods

The test results of the EWC, the IL-VOC, and the proposed framework are summarized in Table 4. As evident from the

Table 5. The classification accuracies and the BWTs of the five methods in different phases.

Method	Phase 0	Phase 1			Phase 2			
	Task 0	Task 0	Task 1	BWT	Task 0	Task 1	Task 2	BWT
JT(best)	99.47%	99.24%	99.69%	-0.0023	99.39%	99.84%	99.69%	0.0004
BM	99.47%	87.38%	99.69%	-0.1209	80.40%	98.34%	99.62%	-0.1021
ES+BM	99.47%	96.54%	99.77%	-0.0293	96.17%	99.09%	99.62%	-0.0199
KD+BM	99.47%	96.39%	99.62%	-0.0308	89.18%	98.42%	99.69%	-0.0575
Proposed framework	**99.47**%	**97.29**%	**99.47**%	**-0.0218**	**96.47**%	**98.87**%	**99.69**%	**-0.0180**

table, the proposed framework demonstrated superior performance in terms of accuracy and BWT compared to both EWC and IL-VOC in all the phases. In general, the proposed methodology not only enhanced the diagnostic accuracy, but also effectively mitigated the issue of forgetting.

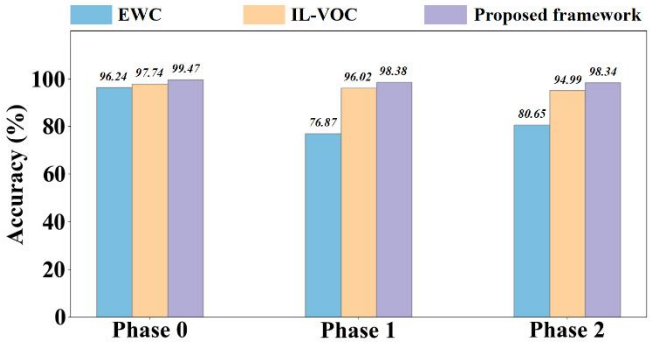

Figure 6. The accuracies of IL methods.

The averaged classification accuracy of each model across different phases are visualized in Figure 6. The figure indicates that the accuracy of EWC experienced a significant decline in Phase 1, accompanied by a higher degree of forgetting. However, the forgetting showed a recovery in Phase 2. According to the data presented in Table 4, the accuracy of Task 0 in Phase 2 is higher than that in Phase 1, suggesting that the EWC achieves a certain level of knowledge retention through parameter constraints. Despite this, its overall accuracies remains lower than that of the other two methods. In contrast, the proposed framework demonstrates consistently higher accuracies compared to IL-VOC, with no significant drop observed across the three stages. This indicates that this solution is more effective in mitigating the model's tendency to forget previous knowledge.

4.2. Ablation experiments

4.2.1. Comparison methods

Given that the proposed framework incorporates multiple strategies to mitigate the forgetting of previous tasks, an ablation study is conducted to validate the effectiveness of this integrated approach and to evaluate the contribution of each component within the proposed framework.

(1) Benchmark Model

The Benchmark Model (BM) serves as the baseline model in this study. Unlike other models, BM did not incorporate any IL techniques to constrain its training process for new tasks, except for initializing the model parameters of Phase n with those from Phase n-1. This experiment was designed to investigate the minimum level of knowledge retention achievable by the diagnostic model with incremental tasks.

(2) Joint Training

Joint Training (JT) is regarded as the theoretical upper bound for the classification accuracy of the diagnostic model. This approach retains all historical task samples and integrates data from task t_0 to t_n in phase n, ensuring comprehensive knowledge preservation throughout the continuous learning. As the number of tasks increases, the volume of training data progressively expands, leading to a corresponding rise in computational and training costs.

(3) ES+BM

Building on the baseline model, only the exemplar replay method was implemented, utilizing the same parameter configurations as those defined in the proposed framework.

(4) KD+BM

KD+BM was an IL model that only added hybrid cross-head knowledge distillation to the baseline model. The parameter setting of the distillation loss term was consistent with the proposed framework. This experiment was

proposed to explore the effect of hybrid cross-head knowledge distillation on alleviating forgetting.

4.2.2. Analysis of results

The test results for the four methods and the proposed framework are presented in Table 5. In Phase 0, all methods exhibit accuracy levels consistent with the BM. In Phase 1, the proposed framework demonstrated better knowledge retention capabilities, achieving higher Task 0 accuracy compared to BM, ES+BM, and KD+BM (though marginally lower than JT's). The accuracy of the proposed framework on Task 1 was lower than that of the other four methods, which might due to the fact that the knowledge distillation and example replay methods limited the performance of the model on Task 1 to some extent. In the Phase 2, the proposed framework sustains its advantage in Task 0, outperforming BM, ES+BM, and KD+BM while performing second to JT. For Task 1, it ranks third behind JT and ES+BM, and achieves parity with JT in Task 3. Notably, the proposed framework achieved superior BWT values across all three stages compared to BM, ES+BM, and KD+BM, approaching JT's performance level. These results collectively demonstrated the proposed framework's effectiveness in mitigating catastrophic forgetting while preserving knowledge from previously learned tasks.

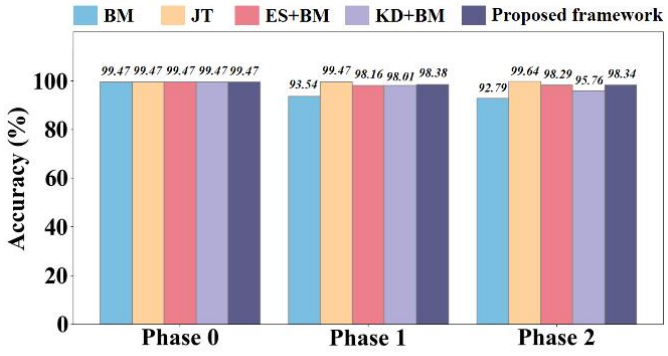

Figure 7. Classification accuracy of the IL methods.

For a more detailed comparison, Figure 7 illustrates the averaged accuracies of each method across different phases. The significant decline in the accuracies of BM suggested that, in the absence of IL techniques, the diagnostic model exhibited substantial forgetting of previous tasks. During the subsequent incremental stages, the accuracies of the KD+BM was significantly higher than that of BM, demonstrating that the hybrid cross-head knowledge distillation effectively mitigates the issue of forgetting. The accuracies of the ES+BM were superior to those of KD+BM but remained lower than the proposed framework. This suggested that the exemplar replay method was more effective in retaining knowledge compared to the hybrid cross-head knowledge distillation method. Furthermore, the contribution of the exemplar replay method to the proposed framework was more significant than that of the hybrid cross-head knowledge distillation approach. In summary, the experimental results demonstrated that the proposed framework effectively retained knowledge of previous tasks while maintaining high accuracy for new tasks, even when sequentially learning multiple tasks of non-stationary operating conditions of the gearbox.

5. CONCLUSIONS

This study proposes a novel approach to address the challenges of multi-task IL under non-stationary conditions for a gearbox. The framework integrated a CNN with the encoder feature extraction module of a transformer architecture to achieve fault diagnosis using frequency-domain spectra. To mitigate the issue of forgetting in IL, the approach incorporated exemplar replay and hybrid cross-head knowledge distillation, effectively preserving knowledge of previously learned tasks and reducing the knowledge forgetting. The efficacy of this framework was validated using the MCC5-THU dataset. Across three incremental learning phases with various working condition tasks, the framework sequentially learned tasks with different time-varying speed ranges of non-stationary conditions, achieving average diagnostic accuracy of 99.47%, 98.38%, and 98.34%, respectively, outperforming other IL models. Furthermore, ablation experiments confirmed that exemplar replay contributes most significantly to knowledge retention, while the fusion of multiple components enhances the method's overall ability to retain knowledge. This research offers a new solution for incremental learning and fault diagnosis under non-stationary conditions. Future work will focus on extending this method to multi-task IL under more complex and random time-varying speed or load conditions, with the aim of improving the model's generalization capability and applicability.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support from Fundamental Research Funds for the Central Universities, Northwestern Polytechnical University (G2023KY05102).

REFERENCES

Du, Z., Chen, X., Zhang, H., & Yan, R. (2015). Sparse feature identification based on union of redundant dictionary for wind turbine gearbox fault diagnosis. IEEE Transactions on Industrial Electronics, 62(10), 6594-6605.

Peng, Y., Qiao, W., Cheng, F., & Qu, L. (2021). Wind turbine drivetrain gearbox fault diagnosis using information fusion on vibration and current signals. IEEE Transactions on Instrumentation and Measurement, 70, 1-11.

Feng, Z., Chen, X., & Zuo, M. J. (2018). Induction motor stator current AM-FM model and demodulation analysis for planetary gearbox fault diagnosis. IEEE Transactions on industrial informatics, 15(4), 2386-2394.

Yin, Z., & Hou, J. (2016). Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes. Neurocomputing, 174, 643-650.

Sinitsin, V., Ibryaeva, O., Sakovskaya, V., & Eremeeva, V. (2022). Intelligent bearing fault diagnosis method combining mixed input and hybrid CNN-MLP model. Mechanical Systems and Signal Processing, 180, 109454.

Qin, G., Zhang, K., Lai, X., Zheng, Q., Ding, G., Zhao, M., & Zhang, Y. (2024). An adaptive symmetric loss in dynamic wide-kernel ResNet for rotating machinery fault diagnosis under noisy labels. IEEE Transactions on Instrumentation and Measurement, 73, 1-12.

Lin, X., Li, B., Yang, X., & Wang, J. (2018, December). Fault diagnosis of aero-engine bearing using a stacked auto-encoder network. In 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC) (pp. 545-548). IEEE.

Yu, G., Wu, P., Lv, Z., Hou, J., Ma, B., & Han, Y. (2023). Few-shot fault diagnosis method of rotating machinery using novel MCGM based CNN. IEEE Transactions on Industrial Informatics, 19(11), 10944-10955.

Liu, R., Wang, F., Yang, B., & Qin, S. J. (2019). Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions. IEEE Transactions on Industrial Informatics, 16(6), 3797-3806.

Dong, Y., Jiang, H., Yao, R., Mu, M., & Yang, Q. (2024). Rolling bearing intelligent fault diagnosis towards variable speed and imbalanced samples using multiscale dynamic supervised contrast learning. Reliability Engineering & System Safety, 243, 109805.

Zhao, M., Kang, M., Tang, B., & Pecht, M. (2017). Deep residual networks with dynamically weighted wavelet coefficients for fault diagnosis of planetary gearboxes. IEEE Transactions on Industrial Electronics, 65(5), 4290-4300.

Zhao, X., Yao, J., Deng, W., Ding, P., Ding, Y., Jia, M., & Liu, Z. (2022). Intelligent fault diagnosis of gearbox under variable working conditions with adaptive intraclass and interclass convolutional neural network. IEEE Transactions on Neural Networks and Learning Systems, 34(9), 6339-6353.

Wang, P., Xiong, H., & He, H. (2023). Bearing fault diagnosis under various conditions using an incremental learning-based multi-task shared classifier. Knowledge-based systems, 266, 110395.

Shi, M., Ding, C., Chang, S., Shen, C., Huang, W., & Zhu, Z. (2024). Cross-domain class incremental broad network for continuous diagnosis of rotating machinery faults under variable operating conditions. IEEE Transactions on Industrial Informatics, 20(4), 6356-6368.

Wang, L., Liu, S., & Xiao, H. (2024). Vaccine enhanced continual learning with TFE to overcome catastrophic forgetting for variable speed-bearing fault diagnosis. IEEE Transactions on Industrial Informatics, 20(5), 7112-7123.

Ostapenko, O., Puscas, M., Klein, T., Jahnichen, P., & Nabi, M. (2019). Learning to remember: A synaptic plasticity driven framework for continual learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11321-11329).

Rebuffi, S. A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). icarl: Incremental classifier and representation learning. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 2001-2010).

Yin, H., Molchanov, P., Alvarez, J. M., Li, Z., Mallya, A., Hoiem, D., ... & Kautz, J. (2020). Dreaming to distill: Data-free knowledge transfer via deepinversion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8715-8724).

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13), 3521-3526.

Li, Z., & Hoiem, D. (2017). Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12), 2935-2947.

Lopez-Paz, D., & Ranzato, M. A. (2017). Gradient episodic memory for continual learning. Advances in neural information processing systems, 30.

Mallya, A., & Lazebnik, S. (2018). Packnet: Adding multiple tasks to a single network by iterative pruning. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 7765-7773).

Veniat, T., Denoyer, L., & Ranzato, M. A. (2020). Efficient continual learning with modular networks and task-driven priors. arXiv preprint arXiv:2012.12631.

Zhang, Y., Shen, C., Shi, J., Li, C., Lin, X., Zhu, Z., & Wang, D. (2024). Deep adaptive sparse residual networks: A lifelong learning framework for rotating machinery fault diagnosis with domain increments. Knowledge-Based Systems, 293, 111679.

Chen, B., Shen, C., Wang, D., Kong, L., Chen, L., & Zhu, Z. (2022). A lifelong learning method for gearbox diagnosis with incremental fault types. IEEE transactions on instrumentation and measurement, 71, 1-10.

Chen, B., Shen, C., Li, L., Shi, J., Huang, W., & Zhu, Z. (2023, October). Continual unsupervised domain adaptation for bearing fault diagnosis under variable working conditions. In International Conference on Electrical and Information Technologies for Rail

Transportation (pp. 395-403). Singapore: Springer Nature Singapore.

Wang, J., Chen, Y., Zheng, Z., Li, X., Cheng, M. M., & Hou, Q. (2024). CrossKD: Cross-head knowledge distillation for object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16520-16530).

Chen, S., Liu, Z., He, X., Zou, D., & Zhou, D. (2024). Multi-mode fault diagnosis datasets of gearbox under variable working conditions. Data in brief, 54, 110453.