Engineering Applications of Artificial Intelligence 162 (2025) 112559

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Multimodal unified generalization and translation network for intelligent fault diagnosis under dynamic environments

Chao Zhao [a,b,c,*], Weiming Shen [c], Enrico Zio [d,e], Hui Ma [a,b,**]

[a] Key Laboratory of Vibration and Control of Aero-Propulsion System Ministry of Education, Northeastern University, Shenyang, 110819, China
[b] School of Mechanical Engineering and Automation, Northeastern University, Shenyang, 110819, China
[c] State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science & Technology, Wuhan, 430074, China
[d] MINES Paris PSL University, CRC, Sophia Antipolis, France
[e] Energy Department, Politecnico di Milano, Milan, Italy

ARTICLE INFO

Keywords:
Domain generalization
Multimodal data
Missing modality
Intelligent fault diagnosis
Rotating machinery

ABSTRACT

Multimodal data fusion can generate reliable fault representations for intelligent fault diagnosis. However, simple data fusion strategies often introduce fault-irrelevant information, thereby reducing robustness against unknown domain shifts. Moreover, traditional methods generally lack adaptive mechanisms to address missing modalities, leading to considerable performance degradation under sensor failure conditions. To address these problems, this paper proposes a multimodal unified generalization and translation network. To learn invariant unified representations for resisting unknown data distribution shifts, information-enhanced concatenation first generates intra-domain and cross-domain representations. Subsequently, mutual information maximization is applied to remove fault-unrelated information from these representations. Finally, A hybrid ensemble diagnosis strategy fully leverages the interaction of multimodal information across different levels. In addition, semantic supervision investigates the relationships among different modalities and enables intermodal translation in the event of a sensor failure within the monitoring system. Extensive experimental results based on a public bearing dataset and a self-collected motor dataset indicate that the proposed method improves accuracy by 10.53 % and 8.47 % compared to the state-of-the-art methods, respectively. The code and datasets are available at https://github.com/CHAOZHAO-1/MUGTN.

1. Introduction

With the advent of the Industrial Internet of Things era, industrial manufacturing systems have undergone significant transformations (Bingsen Wang et al., 2023; Zio, 2022). Ubiquitous sensing devices collect massive amounts of data from manufacturing facilities on the shop floor. Extracting valuable insights from this sensory data is essential for rotating machinery health management system to achieve self-information collection, self-learning, and self-diagnosis. Intelligent fault diagnosis enables manufacturing systems to self-optimize and adapt to various complex environments, aiming to promptly identify fault conditions in rotating mechanical equipment, thereby ensuring reliability and minimizing the risk of accidents (Zhao et al., 2024).

As machinery complexity increases in modern industry, the use of multiple types of sensors becomes necessary for comprehensive equipment monitoring from various perspectives (Pei et al., 2024). Firstly, each type of sensor signal exhibits specific characteristics related to distinct operational states of machinery (Kong et al., 2020). Vibration and acoustic sensors are two widely used types of sensors for machinery health monitoring. Vibration signals are highly sensitive to faults, offer strong interpretability, and are easy to collect. Acoustic signals are noninvasive, easily captured, and have a broad monitoring range (Yao et al., 2021). Vibration signals are well-suited for monitoring mechanical structural faults, including bearing defects, gear wear, and rotor imbalances. In contrast, acoustic signals excel in detecting phenomena such as cracks, air leaks, and motor whines. Therefore, integrating both

* Corresponding author. Key Laboratory of Vibration and Control of Aero-Propulsion System Ministry of Education, Northeastern University, Shenyang, 110819, China.
** Corresponding author. Key Laboratory of Vibration and Control of Aero-Propulsion System Ministry of Education, Northeastern University, Shenyang, 110819, China.
E-mail addresses: zhaoc@me.neu.edu.cn (C. Zhao), shenwm@hust.edu.cn (W. Shen), enrico.zio@polimi.it (E. Zio), huima@me.neu.edu.cn (H. Ma).

https://doi.org/10.1016/j.engappai.2025.112559
Received 13 June 2025; Received in revised form 25 August 2025; Accepted 28 September 2025
Available online 3 October 2025
0952-1976/

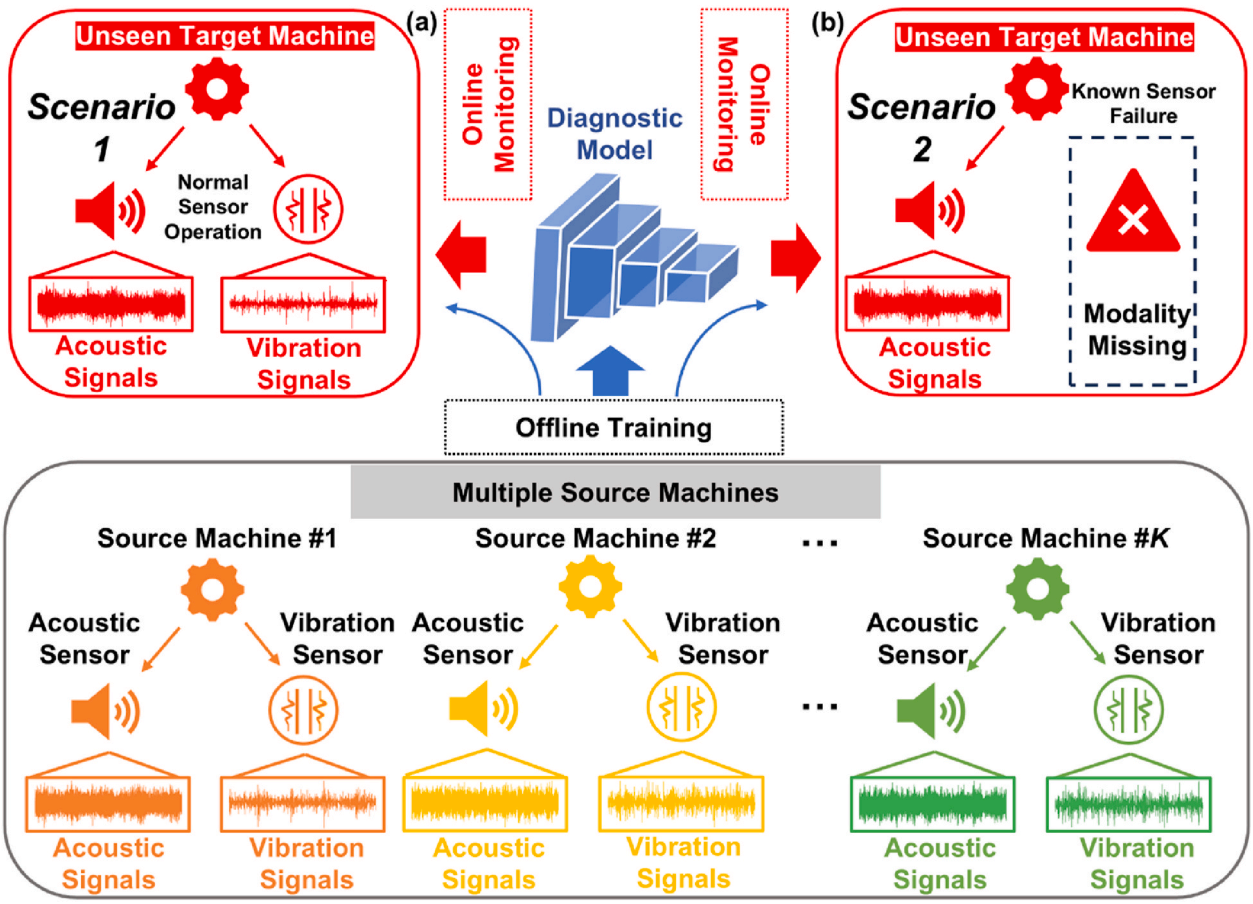

Fig. 1. Illustration of multimodal domain generalization for intelligent fault diagnosis. (a) Scenario 1: normal sensor operation. Two modalities (vibration and acoustic) are both available in machinery monitoring process. (b) Scenario 2: observed sensor failure. One of the sensors in the monitoring system has failed.

vibration and acoustic signals enables the extraction of more reliable and robust features, enhancing the accuracy and effectiveness of fault identification (Gao et al., 2025). Secondly, whereas data-driven diagnostic methods can automatically extract meaningful fault patterns from sensor measurements, the reliability of such systems built from the ground up largely depends on assuming normal sensor readings without outliers or spurious data. Any sensor failure, loss of communication, or malfunction can lead to abnormal and corrupted sensor readings of inferior quality (Huo et al., 2022). These inaccurate measurements may introduce biased fault information, thereby producing erroneous diagnostic outcomes (T. Zhang et al., 2025; Zhang et al., 2023).

Recently, multimodal learning attracts significant research interest in intelligent fault diagnosis due to its ability to improve data utilization, establish relationships between different modalities, and enhance diagnostic accuracy (Yang et al., 2024a). Various data fusion methods based on vibration and acoustic signals have been proposed to bolster the fault diagnosis capabilities of mechanical equipment (Ma et al., 2019; Sun et al., 2024; Yang et al., 2024b). Most of these methods require that the training and testing data be drawn from the same distribution. However, this assumption often does not hold in real-world industrial applications (Fan et al., 2025).

In practical scenarios, intelligent diagnostic models are typically trained using data sourced from various operational conditions or other machines of the same type (He and Shen, 2024a, 2024b). However, due to the diverse operational conditions and distinct physical characteristics of machines, significant domain shifts greatly reduce the effectiveness of data-driven diagnostic models (Ding et al., 2024; Wen et al., 2024; Zhang et al., 2024). To tackle this challenge, domain generalization techniques have emerged as a prominent research focus. The mechanism of domain generalization-based fault diagnosis (DGFD) involves training an intelligent diagnostic model and then subsequently applying it directly to unseen target diagnostic tasks (Bin Wang et al., 2023). DGFD methods utilize the diverse source domains to capture universal knowledge that is robust against domain shifts (Zhao and Shen, 2022).

Based on the above discussion, two critical issues persist in existing research. First, many data fusion strategies often introduce task-irrelevant information while neglecting cross-modal interactions, which compromises model robustness under varying working conditions or across equipment. Second, most methods lack adaptive mechanisms to accommodate modality missing caused by sensor failures, potentially leading to erroneous diagnostic outcomes. Taken together, these limitations underscore the urgent need for a robust and reliable multimodal diagnostic model capable of addressing domain shifts and modality missing in intelligent fault diagnosis.

To meet this demand, this study introduces a novel task called multimodal domain generalization-based fault diagnosis (MMDGFD), as illustrated in Fig. 1. In MMDGFD, a diagnostic model is trained on multi-source multimodal data using an offline training mechanism. The trained model is then applied for online monitoring in two scenarios: normal sensor operation (see Fig. 1(a)) and observed sensor failure (see Fig. 1(b)). In scenario 1, both types of sensors are functioning normally, whereas in scenario 2, one of the sensors is inoperative. To the best of our knowledge, no research has been reported on MMDGFD that considers both scenarios.

To bridge this gap, this study proposes a Multimodal Unified Generalization and Translation Network (MUGTN) for rotating machinery fault diagnosis. Information-enhanced concatenation and

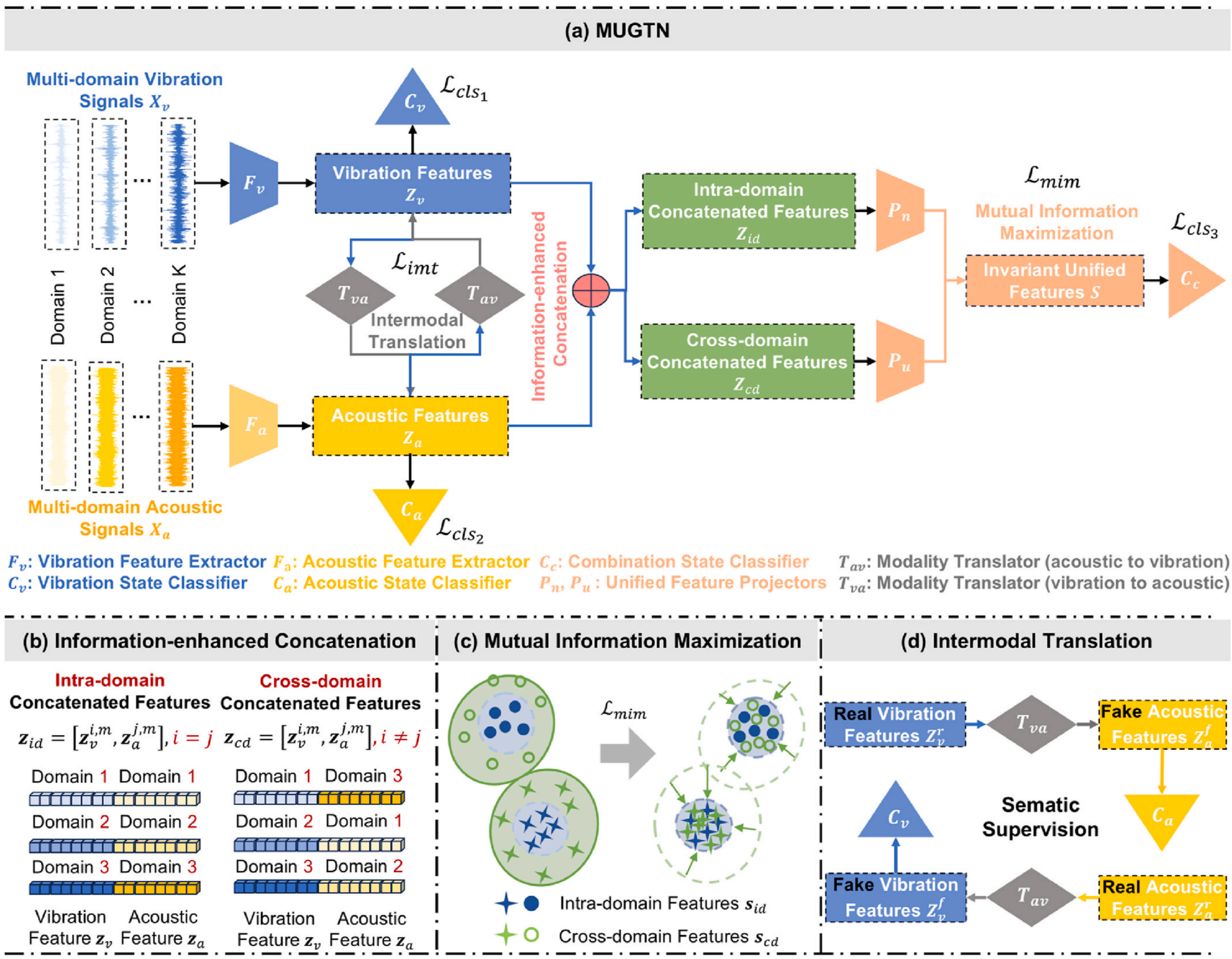

Fig. 2. (a) Overview structure of the proposed MUGTN. (b) Illustration of information-enhanced concatenation. (c) Illustration of mutual information maximization. (d) Illustration of intermodal translation.

mutual information maximization jointly enable the diagnostic model to learn invariant unified representations from acoustic and vibration data. Hybrid ensemble diagnosis combines the outputs of three different classifiers to obtain the final diagnosis results, maximizing the use of multimodal data across different levels of information. To address scenarios with missing modality, modal translators are constructed to generate synthetic data using a designed semantic supervision strategy.

The main insights and contributions of this study are summarized as follows.

(1) A challenge and valuable task with two scenarios called MMDGFD is explored. It is the first work to address multimodal domain generalization in machinery fault diagnosis.
(2) A novel method named MUGTN designed based on the analysis of the knowledge contained in the signal, demonstrating superior performance in comprehensive experiments.
(3) A new multimodal motor failure dataset containing acoustic and vibration data is introduced, which can serve as a valuable resource for future research on MMDGFD.

The remainder of this paper is structured as follows: Section 2 provides a review of related work. Section 3 details the proposed method. Section 4 presents two case studies to assess the diagnostic performance of the proposed method. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related work

2.1. Multimodal data-driven fault diagnosis

Multimodal fusion methods have garnered significant research attention in the field of intelligent fault diagnosis, as multimodal data can provide comprehensive equipment information. Effective and straightforward fusion methods typically include data-level fusion, feature-level fusion, and decision-level fusion (Kibrete et al., 2024). Data-level fusion methods mitigate information loss, while feature-level fusion methods streamline data dimensions. Decision-level fusion methods enhance robustness (Xie et al., 2022). In addition to these typical methods, researchers have developed various fusion techniques to facilitate the exchange of modal information. For instance, Ma et al. (2019) employed constraints on a coupled layer to extract the joint information from multimodal data. Sun et al. (2024) utilized attention mechanisms to capture fault features based on physical properties and generate universal representations. Yang et al. (2024b) proposed a contrastive graph generative network to achieve multimodal information fusion for harmonic drives fault diagnosis. Zhang et al. (T. Zhang

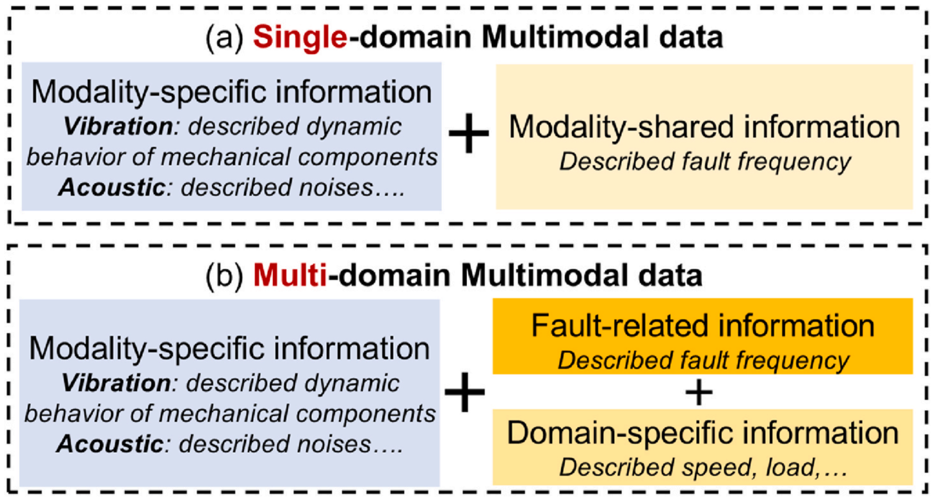

Fig. 3. The differences between (a) single-domain multimodal data and (b) multi-domain multimodal data.

et al., 2025) developed a well-structured multi-scale deep feature memory and recovery network for axial flow pump fault diagnosis. However, these methods typically necessitate training and testing data that originate from the same distribution, a requirement that is often challenging or even impossible to satisfy in real-world industrial applications.

2.2. Domain generalization-based fault diagnosis

Recently, advanced domain generalization techniques have been incorporated into deep networks to tackle domain shift in real-time fault diagnose (Zhao and Shen, 2023a, 2024). These techniques can be broadly categorized into three groups: data augmentation, domain-invariant representation learning, and learning strategy. Data-driven models trained on sufficiently diverse samples during the training phase can effectively handle domain shift. To increase data diversity, Zhao and Shen (Zhao and Shen, 2023b) utilized mutual information to generate fake samples, whereas Han et al. (T. Han et al., 2021) increased diversity of training data by adding Gaussian noise and shifting amplitudes. Another effective approach involves learning features that remain stable across different domains. Li et al. (2021) developed a structural causal model to guide deep models in capturing fault-related features. Zhang et al. (G. Zhang et al., 2025) proposed domain-label-free dual disentanglement method using contrastive and adversarial disentanglement. Guo et al. (2025) developed a method based on a structural causal model for cross-machine bearing fault diagnosis. Additionally, researchers have devised various learning strategies to enhance the generalization capability of diagnostic models. For example, Gao et al. (2024) proposed a meta-learning strategy to enhance the domain-specific representation. Zheng et al. (2025) developed a specific task-guided collaborative domain generalization network for fault diagnosis. However, both of these approaches focus on single-modality signals, which are unsuitable for complex industrial mechanical systems.

2.3. Multimodal domain generalization

To the best of our knowledge, no studies have specifically focused on multimodal domain generalization within the field of intelligent fault diagnosis. Some research has been conducted in computer vision. For instance, to address the first-person action recognition problem, Planamente et al. (2022) proposed an RNA-Net with a relative norm alignment loss. This approach aligns the norm representations of features from two modalities, ensuring that their contributions are balanced during training across various domains. Dong et al. (2023) employed supervised contrastive learning to capture components shared across modalities and applied distance constraints to manage modality-specific features.

3. Proposed method

Fig. 2(a) provides an overview of the proposed MUGTN, which consists of three main components: invariant unified feature learning, hybrid ensemble diagnosis, and intermodal translation.

3.1. Problem definition of MMDGFD

Followed by a previous study (Dong et al., 2023), the definition of MMDGFD is given. Let $\boldsymbol{X}$ denote an input space and $\mathscr{Y}$ denote a label space. In the training stage, given K source training domains $\mathscr{D}_{train} = \{\mathscr{D}^k | k = 1, \ldots, K\}$, where $\mathscr{D}^k = \left\{\boldsymbol{x}_j^k, y_j^k\right\}_{j=1}^{n_k}$ denotes the k-th domain with n_k data samples. Each data sample $\boldsymbol{x}_j^k = \left\{\left(\boldsymbol{x}_j^k\right)_n \middle| n = 1, \ldots, N\right\} \in \boldsymbol{X}$ is comprised of N different modalities, and $y_j^k \in \mathscr{Y}$ denotes the corresponding machine health state. The distributions among multiple domains are different: $P(\boldsymbol{X}^1) \neq P(\boldsymbol{X}^2) \neq \ldots \neq P(\boldsymbol{X}^K)$. The goal of MMDGFD is to construct a generalized multimodal feature extractor $\boldsymbol{z} = F(\boldsymbol{x})$ and a robust state classifier $y = C(\boldsymbol{z})$. The constructed model can reduce classification risk $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathscr{D}_{test}}[C(F(\boldsymbol{x})) \neq y]$ on an unseen target testing domain $\mathscr{D}_{test}$, where data distributions differ from the training domain $\{P(\boldsymbol{X}^{test}) \neq P(\boldsymbol{X}^{train})\}$. Each data samples $\boldsymbol{x}_j^{test}$ in the unseen testing target domain comprises M ($1 \leq M \leq N$) modalities.

3.2. Invariant unified representation learning

As shown in Fig. 3, single-domain multimodal data contain modality-specific and modal-shared information (Hazarika and Zimmermann, 2020). Modality-specific information is private to each modality, whereas modality-shared information is invariant among different modalities. Although multimodal monitoring signals come from different sensors, they share common information about machine health states. Simultaneously, each modality has distinctive characteristics that are unrelated to other modalities. Filtering out modality-specific information and retaining modality-shared information is an effective approach for monitoring machine states in a stable environment. However, unlike single-domain multimodal data, multi-domain multimodal data in a dynamic environment include three parts: fault-related, domain-specific, and modality-specific information. Domain-specific information is brought about by the changing environment and is unrelated to predicting machine states. Fault-related information is shared across various modalities and domains. Therefore, it is challenging in MMDGFD tasks to learn comprehensive fault representations that are robust to domain shift because domain-specific factors need to be delicately removed.

To tackle this difficult problem, the proposed MUGTN learns underlying commonalities from multi-domain multimodal data from a disentangled view. An invariant unified representation space that captures robust representations from data of different modalities across domains is explored. Specifically, information-enhanced concatenation and mutual information maximization are designed to cooperatively learn invariant unified representations.

As shown in Fig. 2 (b), information-enhanced concatenation generates both intra-domain concatenated features $\boldsymbol{z}_{id}$ and cross-domain concatenated features $\boldsymbol{z}_{cd}$. Concatenating high-dimensional features from different modalities is a simple yet effective method to address the limitations of single-modal information, thereby enhancing the performance of diagnostic models (Xie et al., 2022). To increase feature diversity and enhance the utilization of useful information, in addition to generating intra-domain concatenated features where different modal features come from the same domain, cross-domain concatenated features where different modal features come from different domains are also generated. The $\boldsymbol{z}_{id}$ and $\boldsymbol{z}_{cd}$ are defined as follows,

$$\begin{aligned} \boldsymbol{z}_{id} &= \left[\boldsymbol{z}_v^{i,m}, \boldsymbol{z}_a^{j,m}\right], i = j \\ \boldsymbol{z}_{cd} &= \left[\boldsymbol{z}_v^{i,m}, \boldsymbol{z}_a^{j,m}\right], i \neq j \end{aligned} \tag{1}$$

where $\boldsymbol{z}_v^{i,m}$ denotes vibration features belonging to the i-th domain m-th class. Similarly, $\boldsymbol{z}_a^{j,m}$ denotes acoustic features belonging to the j-th domain m-th class. In practical implementation, information-enhanced concatenation is achieved by shuffling vibrational and acoustic samples from different source domains within the same class and then recombining them. As a result, the computational complexity remains low. The number of generated cross-domain features is consistent with that of intra-domain features. From a causal modeling perspective, shuffling and recombining modalities across different domains can be interpreted as generating perturbed samples (Liu et al., 2024; Sui et al., 2024). These perturbations expose the model to a wider spectrum of modality combinations and yield additional diverse instances, thereby enhancing information coverage. These perturbations act as counterfactual-style samples that disrupt spurious correlations between domain-specific factors and fault labels. Consequently, the diagnostic model is encouraged to focus on fault-related information that remains invariant across domains.

Unified feature projectors project concatenated multimodal features to harmonious invariant unified representation by maximizing mutual information between intra-domain and cross-domain concatenated features. The invariant unified representation space sufficiently maintains the fault-related attributes that are robust to domain shift and removes the redundant information that can disturb model decisions. The projected representations $\boldsymbol{s}_{id}$ and $\boldsymbol{s}_{cd}$ are calculated as follows,

$$\begin{aligned} \boldsymbol{s}_{id} &= P_n(\boldsymbol{z}_{id}) \\ \boldsymbol{s}_{cd} &= P_u(\boldsymbol{z}_{cd}) \end{aligned} \tag{2}$$

where P_n and P_u denote unified feature projectors.

Mutual information (MI), an information theory concept, quantifies the dependencies between pairs of multi-dimensional variables. Maximizing MI has proven effective in eliminating redundant information that is irrelevant to downstream tasks, while also capturing invariant patterns across time or different domains (W. Han et al., 2021). The MI between variable X and Y is defined as (Tishby and Zaslavsky, 2015),

$$I(X;Y) = \mathbb{E}_{P(x,y)}\left[\log \frac{P(x,y)}{P(x)P(y)}\right] \tag{3}$$

where $P(x,y)$ denotes the joint probability distribution of x and y. $P(x)$ is the marginal distribution of x. Similarly, $P(y)$ is the marginal distribution of y.

MUGTN disentangles fault-related, domain-specific, and modality-specific information by maximizing MI among intra-domain and cross-domain concatenated features in the unified representation space. Since MI is intractable to calculate directly, maximizing MI is generally replaced by increasing its lower bound(Wang et al., 2021). InfoNCE (Oord et al., 2018) is adopted to estimate the lower bound of MI in this study,

$$I(\boldsymbol{s}_{id};\boldsymbol{s}_{cd}) \geq \mathbb{E}\left[\frac{1}{W}\sum_{i=1}^{W}\log\frac{e^{f\left(s_{id}^{i},s_{cd}^{j}\right)}}{\sum\limits_{j=1}^{N} e^{f\left(s_{id}^{i},s_{cd}^{j}\right)}}\right] \tag{4}$$

where $f(\cdot,\cdot)$ is a critic function and W is the sample number. To maximize the lower bound of MI, supervised contrastive loss is employed to increase the MI among samples. During the training phase, the projected representations $\boldsymbol{s}_{id}$ and $\boldsymbol{s}_{cd}$ are combined into a set $\boldsymbol{s} = \boldsymbol{s}_{id} \cup \boldsymbol{s}_{cd}$. Given a query $\boldsymbol{s}^i$, the supervised contrastive loss is defined as,

$$\mathscr{L}_{mim} = -\sum_{i\in I}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{e^{\left(s^i\cdot s^p/\tau\right)}}{\sum\limits_{a\in A(i)} e^{\left(s^i\cdot s^a/\tau\right)}} \tag{5}$$

where $i \in I = \{1, 2, \ldots, 2W\}$ and $A(i) \equiv I/i$. τ is the scalar temperature parameter. P represents the positive sample set.

As shown in Fig. 2(c), projected intra-domain features $\boldsymbol{s}_{id}$ maintain consistent latent domain-specific information, making them similar to each other and resulting in their close proximity within the unified representation space. Conversely, projected inter-domain features $\boldsymbol{s}_{cd}$ possess differing domain-specific information, causing their positions in the space to be relatively dispersed. By maximizing mutual information between intra-domain and inter-domain features, domain-specific information is effectively removed, reducing intra-class discrepancy. Ultimately, through the synergy of information-enhanced concatenation and mutual information maximization, invariant unified representations are obtained.

3.3. Hybrid ensemble diagnosis

Fusing multimodal data at different levels offers distinct advantages (Rahate et al., 2022). Feature-level fusion integrates low-level features from various modalities, whereas decision-level fusion combines high-level decisions from different modalities. To improve the flexibility and adaptability of the diagnostic model for various tasks and environments, this study develops a hybrid ensemble diagnosis strategy.

MUGTN equips three state classifiers: the vibration state classifier C_v, the acoustic classifier C_a, and the combination state classifier C_c. First, invariant unified representations output by the projectors are fed into the combination state classifier C_c to capture the relationship between multimodal monitoring data and machine health states,

$$\mathscr{L}_{cls_3} = L_{CE}(C_c(\boldsymbol{S}), \mathscr{Y}) \tag{6}$$

where L_{CE} denotes the cross-entropy function.

Feature-level fusion combines the multimodal data at the feature level, allowing the diagnostic model to utilize the rich information from each modality. On the other hand, decision-level fusion can further integrate the independent decisions of different classifiers to maximize the advantages of each modality. Therefore, the vibration classifier C_v and the acoustic classifier C_a also mine the underlying knowledge in

Table 1
Network architecture of MUGTN.

Module	Layer Type	Kernel Number	Kernel Size	Stride	Activation	Normalization
Feature Extractor (F_v, F_a)	Convolution	16	64 × 1	1	ReLU	Batch Normalization
	Max-pooling	/	2 × 1	2	/	/
	Convolution	32	16 × 1	1	ReLU	Batch Normalization
	Max-pooling	/	2 × 1	2	/	/
	Convolution	64	5 × 1	1	ReLU	Batch Normalization
	Max-pooling	/	2 × 1	2	/	/
	Convolution	64	5 × 1	1	ReLU	Batch Normalization
	Max-pooling	/	2 × 1	2	/	/
Modal Translator (T_{av}, T_{va})	Fully-connected	/	128	/	ReLU	Batch Normalization
	Fully-connected	/	128	/	ReLU	Batch Normalization
	Fully-connected	/	128	/	ReLU	Batch Normalization
Unified Feature Projector (P_n, P_u)	Fully-connected	/	256	/	ReLU	Batch Normalization
	Fully-connected	/	256	/	ReLU	Batch Normalization
	Fully-connected	/	256	/	ReLU	Batch Normalization
State Classifier (C_v; C_a, C_c)	Fully-connected	/	Class number	/	Softmax	/

their respective modal data,

$$\begin{aligned}\mathscr{L}_{cls_1} &= L_{CE}(C_v(F_v(\boldsymbol{X}_v)), \mathscr{Y})\\ \mathscr{L}_{cls_2} &= L_{CE}(C_a(F_a(\boldsymbol{X}_a)), \mathscr{Y})\end{aligned} \tag{7}$$

The total classification loss is defined as,

$$\mathscr{L}_{cls} = \mathscr{L}_{cls_1} + \mathscr{L}_{cls_2} + \mathscr{L}_{cls_3} \tag{8}$$

The results of the hybrid ensemble diagnosis are obtained by applying majority voting on three different state classifiers, which is defined as follows:

$$H(x) = \underset{m}{argmax}\left(C_v^m F_v(\boldsymbol{x}_v) + C_a^m F_a(\boldsymbol{x}_a) + C_c^m(P_n([F_v(\boldsymbol{x}_v), F_a(\boldsymbol{x}_a)]))\right) \tag{9}$$

where m denote m-th predictive confidence.

3.4. Intermodal translation

The authenticity of the data is critical to the data-driven diagnostic model. However, in actual monitoring, sensor failures and loss of communication can occur, negatively affecting diagnostic model performance. To improve the robustness of the diagnostic model facing various external limitations, modality translators T_{av} and T_{va} are equipped in the MUGTN to achieve intermodal translation when certain modality is missing.

Modal translators generate fake data for the missing modality from the available monitoring data. The real vibration features Z_v^r, fake vibration features Z_v^f, real acoustic features Z_a^r, and fake acoustic features Z_a^f can be calculated as follows:

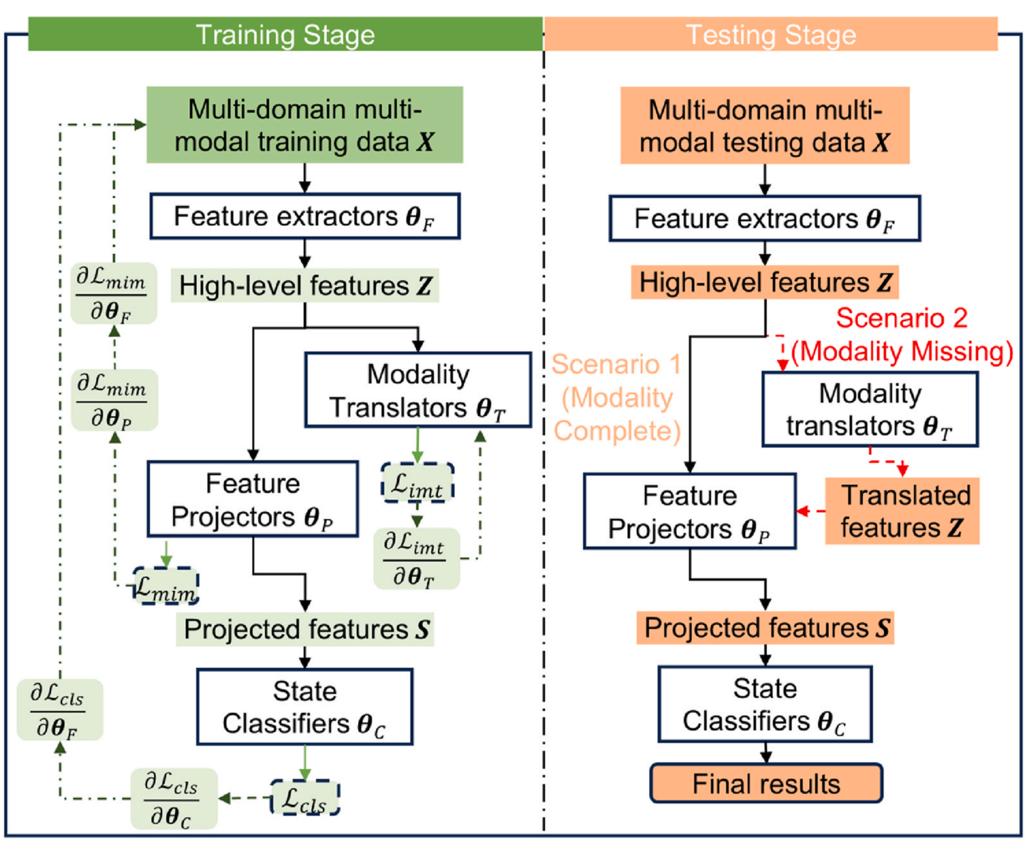

Fig. 4. Flowchart of the proposed MUGTN.

Table 2
Details of bearing operating conditions.

Domain Index	Working Conditions		Domain Index	Working Conditions	
	Speed (r/min)	Load (Full Load)		Speed (r/min)	Load (Full Load)
A1	500	0 %	A9	1000	20 %
A2	1000	0 %	A10	1000	40 %
A3	1500	0 %	A11	1000	60 %
A4	2000	0 %	A12	1000	80 %
A5	2500	0 %	A13	2000	20 %
A6	3000	0 %	A14	2000	40 %
A7	3500	0 %	A15	2000	60 %
A8	4000	0 %	A16	2000	80 %

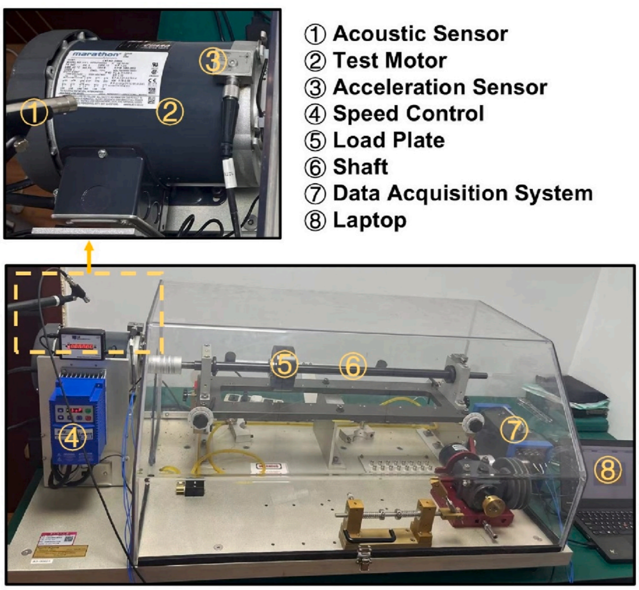

Fig. 5. Modular test rig for the self-collected multimodal motor dataset.

$$
\begin{aligned}
\mathbf{Z}_v^r &= \overline{F_v}(\mathbf{X}_v) \\
\mathbf{Z}_v^f &= T_{av}(\overline{F_a}(\mathbf{X}_a)) \\
\mathbf{Z}_a^r &= \overline{F_a}(\mathbf{X}_a) \\
\mathbf{Z}_a^f &= T_{va}(\overline{F_v}(\mathbf{X}_v))
\end{aligned}
\tag{10}
$$

where $\overline{F_v}$ and $\overline{F_a}$ denote that the parameters of vibration and acoustic feature extractors are not updated during the training process.

It is natural that the generated synthetic data should be correctly recognized by their respective state classifiers, meaning semantic similarity between real and generated features should be strictly controlled. Therefore, semantic supervision is applied to construct modality translators,

$$
\mathscr{L}_{imt} = L_{CE}(\overline{C_v}(T_{av}(\overline{F_a}(\mathbf{X}_a))), \mathscr{Y}) + L_{CE}(\overline{C_a}(T_{va}(\overline{F_v}(\mathbf{X}_v))), \mathscr{Y}) \tag{11}
$$

where $\overline{C_v}$ and $\overline{C_a}$ denote that the parameters of vibration and acoustic state classifiers are not updated. In this way, only the modality translators are penalized during the training process to establish the connection between different modalities.

3.5. Object function optimization

The final optimization objective of the proposed MUGTN can be written as,

$$
\mathscr{L} = \mathscr{L}_{cls} + \alpha \mathscr{L}_{mim} + \mathscr{L}_{imt} \tag{12}
$$

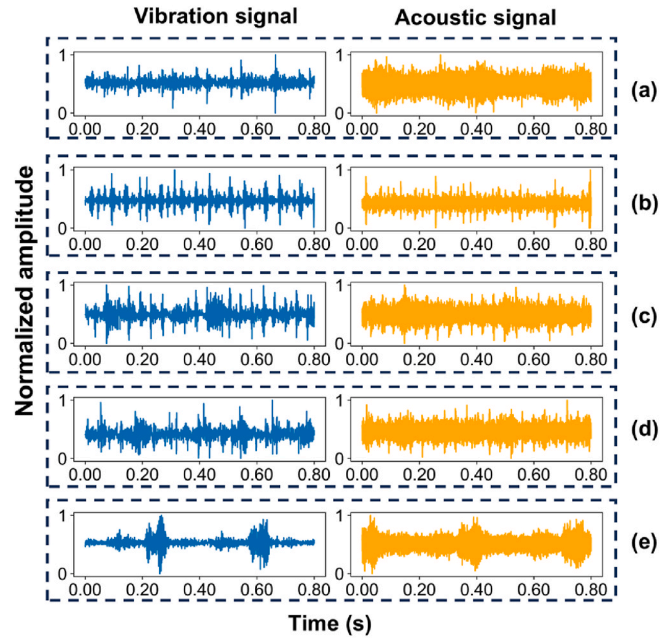

Fig. 6. Visualization of multimodal monitoring signals for bearings. (a) healthy, (b) outer race fault, (c) inner race fault, (d) ball fault, and (e) cage fault.

where α is a balanced coefficient. The Adam optimizer is employed to update the network parameters at each training epoch as,

$$
\begin{aligned}
\boldsymbol{\theta}_P^{q+1} &\leftarrow \boldsymbol{\theta}_P^q - \mu\left[\alpha\left(\frac{\partial \mathscr{L}_{mim}}{\partial \boldsymbol{\theta}_P^q}\right)\right], \\
\boldsymbol{\theta}_F^{q+1} &\leftarrow \boldsymbol{\theta}_F^q - \mu\left[\left(\frac{\partial \mathscr{L}_{cls}}{\partial \boldsymbol{\theta}_F^q}\right) + \alpha\left(\frac{\partial \mathscr{L}_{mim}}{\partial \boldsymbol{\theta}_F^q}\right)\right], \\
\boldsymbol{\theta}_C^{q+1} &\leftarrow \boldsymbol{\theta}_C^q - \mu\left[\left(\frac{\partial \mathscr{L}_{cls}}{\partial \boldsymbol{\theta}_C^q}\right)\right], \\
\boldsymbol{\theta}_T^{q+1} &\leftarrow \boldsymbol{\theta}_T^q - \mu\left[\left(\frac{\partial \mathscr{L}_{imt}}{\partial \boldsymbol{\theta}_T^q}\right)\right]
\end{aligned}
\tag{13}
$$

where μ denotes learning rate, and q represents q-th iteration update. The detail network architecture of the MUGTN is listed as Table 1.

Fig. 4 plots the flowchart of the proposed MUGTN. During the offline training stage, multi-domain multimodal monitoring data are used as input. The data first pass through the feature extractors, and are subsequently fed into the feature projectors and modality translator.

Table 3
Details of bearing diagnosis setting.

Task Name	Known Source Domains	Unknown Target Domains
Task 1	A1, A3, A5, A7, A10, A12, A14, A16	A2, A4, A6, A8, A9, A11, A13, A15
Task 2	A2, A4, A6, A8, A9, A11, A13, A15	A1, A3, A5, A7, A10, A12, A14, A16
Task 3	A1, A3, A5, A7, A9, A11, A13, A15	A2, A4, A6, A8, A10, A12, A14, A16
Task 4	A2, A4, A6, A8, A10, A12, A14, A16	A1, A3, A5, A7, A9, A11, A13, A15
Task 5	A1, A2, A3, A4, A9, A10, A11, A12	A5, A6, A7, A8, A13, A14, A15, A16
Task 6	A5, A6, A7, A8, A13, A14, A15, A16	A1, A2, A3, A4, A9, A10, A11, A12
Task 7	A1, A2, A3, A4, A13, A14, A15, A16	A5, A6, A7, A8, A9, A10, A11, A12
Task 8	A5, A6, A7, A8, A9, A10, A11, A12	A1, A2, A3, A4, A13, A14, A15, A16

Table 4
Details of motor diagnosis setting.

Task Name	Known Source Domains	Unknown Target Domain
Task 9	B2, B3, B4	B1
Task 10	B1, B3, B4	B2
Task 11	B1, B2, B4	B3
Task 12	B1, B2, B3	B4

Finally, the processed representations are passed into the state classifiers. Forward propagation is performed to compute Eq. (12), followed by backward propagation to update the network parameters according to Eq. (13). After multiple training epochs, the optimal MUGTN model is obtained. In the online monitoring stage, two scenarios are considered. In Scenario 1, the input consists of multimodal data from an unknown target domain, and the pre-trained MUGTN is employed to generate the final diagnosis decisions. In Scenario 2, the target domain data exhibit missing modalities. In this case, the MUGTN performs intermodal translation according to Eq. (10), and then produces the final diagnosis decisions.

4. Case study

In this section, comprehensive experiment tasks were performed on two mechanical failure datasets to assess the efficacy and practical applicability of the proposed MUGTN.

4.1. Dataset description and experimental setting

(1) Public multimodal bearing failure dataset. This dataset was provided by iFLYTEK AI DEVELOPER COMPETITION (iFLYTEK, n. d.). It includes data from a rotor comprehensive fault simulation test bench that monitors five types of bearings with different health statuses: healthy, outer race fault, inner race fault, ball fault, and cage fault. Bearing designation is NSK 6200. Vibration data and acoustic data are synchronously collected. Vibration data is acquired using a wired vibration acceleration sensor mounted horizontally on the bearing seat, with a sampling rate of 5.12 kHz. Acoustic data are collected using a pickup sensor at a sampling rate of 48 kHz. The lengths of the vibration and acoustic samples are 4096 and 38400, respectively. The dataset includes signals collected under sixteen working conditions (see Table 2), each characterized by different speeds and loads. A bearing monitoring multimodal sample includes a vibration sample and an acoustic sample. For each machine health status, 200 samples were prepared, with 100 randomly chosen for training and the rest for testing.

(2) Self-collected multimodal motor failure dataset. This dataset was collected from a Spectra-Quest Mechanical Fault Simulator, as depicted in Fig. 5. The test rig monitored motors with six health conditions: healthy, bearing fault, bowed rotor, broken rotor bars, rotor misalignment, and voltage unbalance. Signals were collected under four different rotating speeds (B1 = 5 Hz, B2 = 10 Hz, B3 = 20 Hz and B4 = 30 Hz). Vibration and acoustic data were synchronously recorded, with sampling rates of 25.6 kHz for both types of signals. The lengths of the vibration and acoustic samples are both 1024. A motor monitoring multimodal sample includes a vibration sample and an acoustic sample. For each machine health status, 1000 samples were prepared, with 800 randomly chosen for training and the rest for testing.

Fig. 6 displays the vibration and acoustic signals for bearings with different health status. As shown in Tables 3 and 4, eight bearing multimodal fault diagnosis tasks and four motor multimodal fault diagnosis tasks are designed.

4.2. Comparative methods and implementation details

Given the limited existing research on MMDGFD, a number of representative multimodal fusion and domain generalization methods are selected for evaluation on the same dataset to assess the feasibility and superiority of the proposed method. Detailed descriptions of the comparative methods are provided in Table 5.

The experimental tests were carried out in the PyTorch software environment, using Nvidia-3080Ti GPUs for computational power. The

Table 5
Details of comparative methods.

Method Name	Domain Generalization Method	Multimodal Fusion Method	Data
M1	ERM (Vapnik, 1999)	/	Vibration
M2	ERM (Vapnik, 1999)	/	Acoustic
M3	ERM (Vapnik, 1999)	Data-level fusion via direct data concatenation	Vibration and acoustic
M4	ERM (Vapnik, 1999)	Feature-level fusion via direct feature concatenation	Vibration and acoustic
M5	ERM (Vapnik, 1999)	Decision-level fusion via majority voting	Vibration and acoustic
M6	CDDG (Jia et al., 2024)	Data-level fusion via direct data concatenation	Vibration and acoustic
M7	CDDG (Jia et al., 2024)	Decision-level fusion via majority voting	Vibration and acoustic
M8	CCDG (Ragab et al., 2022)	Data-level fusion via direct data concatenation	Vibration and acoustic
M9	CCDG (Ragab et al., 2022)	Feature-level fusion via direct feature concatenation	Vibration and acoustic
M10	A state-of-the-art multimodal domain generalization method (SimMMDG (Dong et al., 2023))		Vibration and acoustic

Table 6
Results of bearing fault diagnostic tasks (Scenario 1).

Method	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Average
M1	66.44 ± 0.23	56.02 ± 0.57	66.60 ± 0.41	55.96 ± 0.46	48.98 ± 0.25	31.61 ± 0.13	50.45 ± 0.57	45.24 ± 0.25	52.66
M2	87.88 ± 0.21	84.94 ± 0.25	90.92 ± 0.13	83.01 ± 0.53	66.24 ± 0.51	49.63 ± 0.36	62.92 ± 0.85	73.61 ± 0.72	74.89
M3	86.90 ± 0.18	82.83 ± 0.17	89.44 ± 0.06	85.30 ± 0.39	68.65 ± 0.58	45.00 ± 0.46	61.31 ± 1.01	69.93 ± 0.74	73.67
M4	89.56 ± 0.09	85.18 ± 0.24	92.59 ± 0.24	87.36 ± 0.18	63.92 ± 0.58	42.23 ± 0.45	66.59 ± 0.82	70.03 ± 0.25	74.68
M5	88.21 ± 0.34	81.34 ± 0.26	92.66 ± 0.47	81.39 ± 0.46	66.89 ± 0.49	38.24 ± 0.27	64.06 ± 0.34	68.42 ± 0.29	72.65
M6	89.21 ± 0.17	85.38 ± 0.32	91.26 ± 0.06	86.16 ± 0.09	65.93 ± 0.20	45.14 ± 0.28	**75.63 ± 0.19**	72.33 ± 0.18	76.38
M7	89.24 ± 0.11	86.34 ± 0.41	90.09 ± 0.51	86.08 ± 0.39	71.48 ± 0.24	37.97 ± 0.11	72.76 ± 0.48	71.14 ± 0.37	75.64
M8	87.91 ± 0.19	84.11 ± 0.59	89.38 ± 0.29	86.37 ± 0.13	56.24 ± 0.13	48.94 ± 1.37	73.57 ± 0.14	66.63 ± 0.26	74.14
M9	87.07 ± 0.10	85.72 ± 0.48	93.75 ± 0.46	84.94 ± 0.27	69.89 ± 0.09	47.57 ± 0.76	72.37 ± 0.30	72.40 ± 0.14	76.71
M10	90.47 ± 0.39	81.09 ± 0.49	84.19 ± 0.45	79.26 ± 0.42	55.00 ± 0.36	58.67 ± 0.53	58.42 ± 0.60	63.14 ± 1.00	71.28
Proposed	**90.96 ± 0.48**	**89.02 ± 0.65**	**95.44 ± 0.23**	**87.82 ± 0.54**	**78.28 ± 1.09**	**61.66 ± 1.10**	71.44 ± 0.51	**79.85 ± 1.08**	**81.81**

Table 7
Results of motor fault diagnostic tasks (Scenario 1).

Method	Task 9	Task 10	Task 11	Task 12	Average
M1	63.87 ± 2.64	73.31 ± 0.38	79.01 ± 0.74	56.42 ± 1.05	68.15
M2	49.23 ± 4.71	61.89 ± 2.57	59.14 ± 1.05	28.13 ± 1.77	49.60
M3	70.72 ± 5.44	71.70 ± 1.19	68.60 ± 0.85	47.60 ± 3.04	64.66
M4	71.72 ± 5.32	77.26 ± 3.60	71.32 ± 0.79	47.42 ± 4.65	66.93
M5	75.53 ± 1.63	75.57 ± 1.44	76.15 ± 1.06	47.88 ± 2.27	68.78
M6	71.30 ± 2.28	71.86 ± 1.36	67.66 ± 0.70	52.83 ± 3.64	65.91
M7	73.29 ± 2.60	68.60 ± 3.00	70.47 ± 2.23	51.85 ± 5.10	66.05
M8	62.93 ± 2.88	82.45 ± 2.69	74.92 ± 3.54	65.09 ± 4.01	71.35
M9	59.88 ± 3.20	81.96 ± 3.60	79.14 ± 4.08	46.50 ± 4.19	66.87
M10	77.97 ± 3.22	79.10 ± 0.57	74.28 ± 0.92	46.51 ± 3.86	69.47
Proposed	**82.16** ± **0.70**	**82.70** ± **1.29**	**80.68** ± **0.98**	**66.23** ± **1.89**	**77.94**

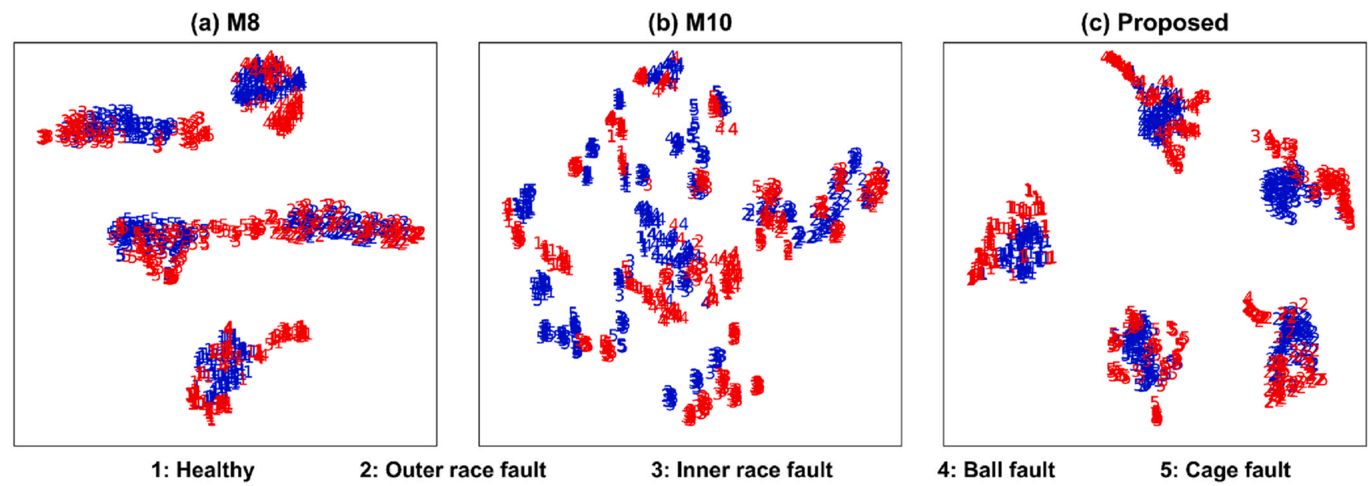

Fig. 7. Visualization results of learned features by different methods on Task 3 via t-SNE. (a) M8, (b) M10, and (c) Proposed. The blue and red values represent source and target samples, respectively. The value denotes the class label for each sample. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Adam optimizer was used to optimize the model parameters. To mitigate the effects of random variations, the diagnostic results were averaged across ten independent trials. Hyperparameters for all methods were selected based on relevant literature to ensure optimal performance and meet experimental requirements. For a fair comparison, all models were implemented with an identical network architecture. Training was carried out for 50 epochs with a batch size of 256. The learning rate was set to $lr = 0.0005/(1 + 10 \times \varphi)^{0.75}$, where φ represents the ratio of the current epoch to the total number of epochs. The hyperparameter α for the proposed method was set to 1.

4.3. Experimental results and performance analyses for scenario 1

(1) Diagnosis results of Scenario 1

The diagnosis results of bearing in Scenario 1 are listed in Table 6. When comparing the performance of M1 and M2, a significant performance gap between the two modalities is evident. The average accuracies of the diagnostic model based on acoustic signals are 20 % higher than that of the diagnostic model based on vibration signals. These results indicate that different modalities exhibit varying sensitivity to domain shifts. Additionally, a comparison of the performance of M3, M4, and M5 with that of M1 and M2 reveals that the three representative multimodal fusion methods fail to outperform the superior single modality. This indicates that simply fusing multimodal data does not significantly improve the generalization performance of diagnostic models in dynamic environments. Comparing the performance of M6-M9 with that of M3, M4, and M5, it is evident that advanced domain generalization techniques are more effective in learning robust fault representations for MMDGFD tasks. Finally, the proposed method achieves the highest accuracy on seven bearing diagnosis tasks and attains the highest average accuracy of 81.81 % among the 11 comparative methods, demonstrating its superior effectiveness in addressing MMDGFD problems.

The diagnostic results for motors in Scenario 1 are presented in Table 7. Based on the performance of M1 and M2, it is evident that the intelligent diagnostic model utilizing vibration signals outperform that based on acoustic signals, which stands in contrast to the observations from the bearing experiments. These results suggest that the robustness of different modalities to domain shifts varies across different devices and environments. Additionally, Table 7 shows that the performance of the three classical fusion methods (M3, M4, and M5) is comparable to that of the best-performing single modality, which is consistent with the findings from the bearing experiments. By combining the results from Tables 6 and 7, it becomes evident that multimodal fusion offers greater robustness across different devices and environments, thereby reducing the need for pre-experimental analysis to identify the optimal modality. Moreover, although the widely adopted multimodal domain generalization method SimMMDG (denoted as M10) slightly outperforms the three classical fusion approaches (M3, M4, and M5), its overall performance remains suboptimal and is still inferior to that of M8. Finally, the proposed method achieves the highest accuracies across all four motor diagnostic tasks, further confirming its effectiveness and superiority.

To intuitively compare the multimodal fusion features learned by different methods, t-SNE was employed to visualize the features learned by M8, M10, and the proposed method on Task 3. As shown in Fig. 7(a), the features exhibit a certain degree of clustering, with partial overlap between the source and target domains. However, some class

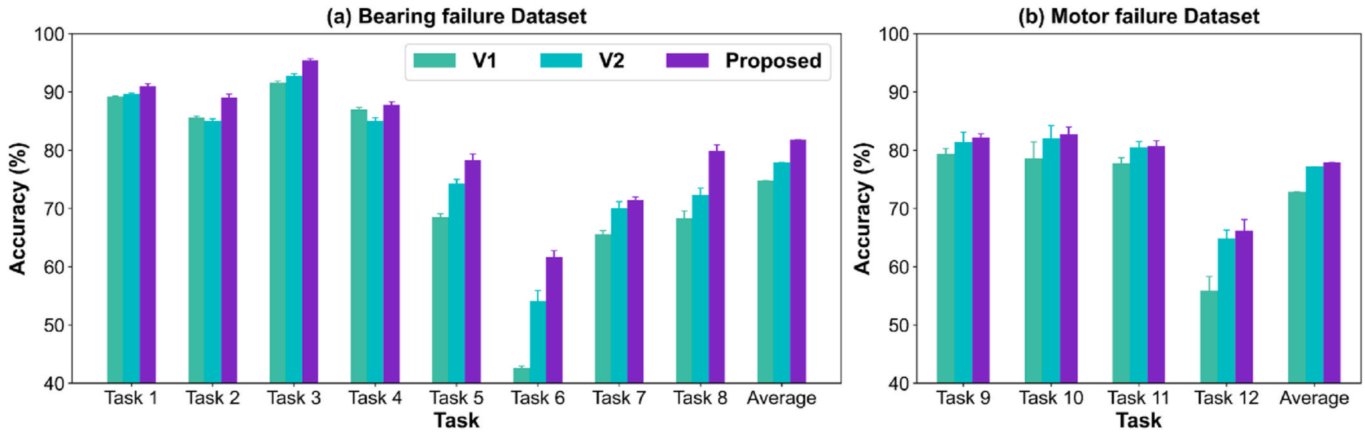

Fig. 8. Experimental results of the proposed method and its variants on bearing and motor fault diagnosis tasks.

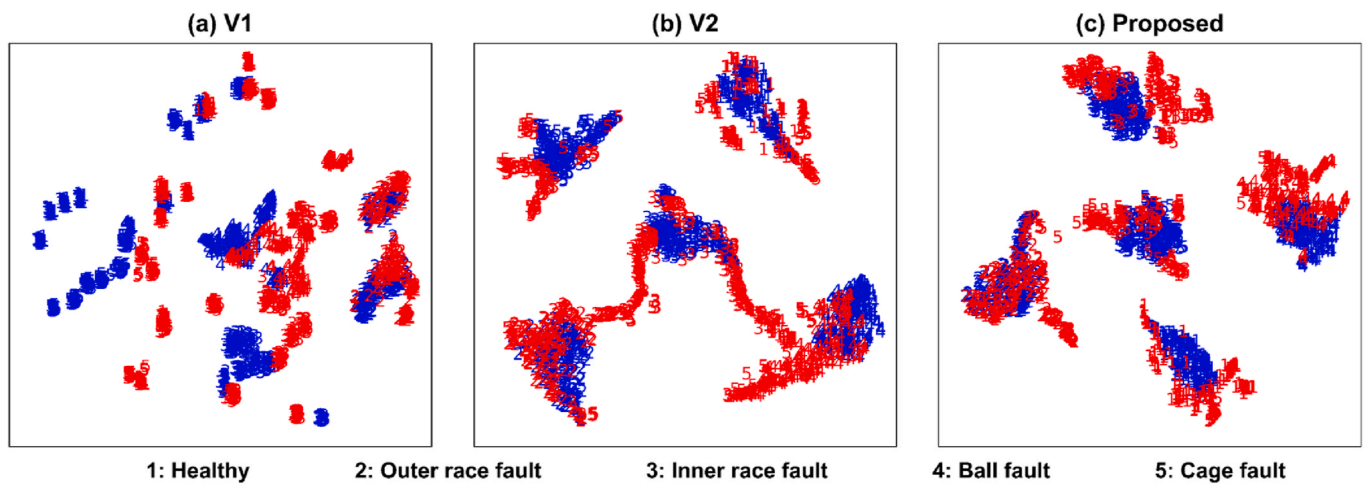

Fig. 9. Visualization results of learned features by different methods on Task 2 via t-SNE. (a) V1, (b) V2, and (c) the proposed method. The blue and red values represent source and target samples, respectively. The value denotes the class label for each sample. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

boundaries remain unclear. In Fig. 7(b), the separation between classes is relatively poor. In contrast, Fig. 7(c) shows that the classes are well separated, with a larger inter-class distance and good overlap between source and target samples, which facilitates accurate classification by the classifier.

(2) Ablation study

Ablation experiments were conducted to assess the contributions of different components by gradually removing them from the proposed method and observing the impact on model performance. Two variants of the proposed method, labeled as V1 and V2, were implemented for comparison. In the V2, the information-enhanced concatenation is removed, meaning it only utilizes the intra-domain concatenated features. The V1 further removes the mutual information maximization from the V2.

The comparison results are presented in Fig. 8. As shown in Fig. 8, the accuracies of the V2 are consistently lower than those of the

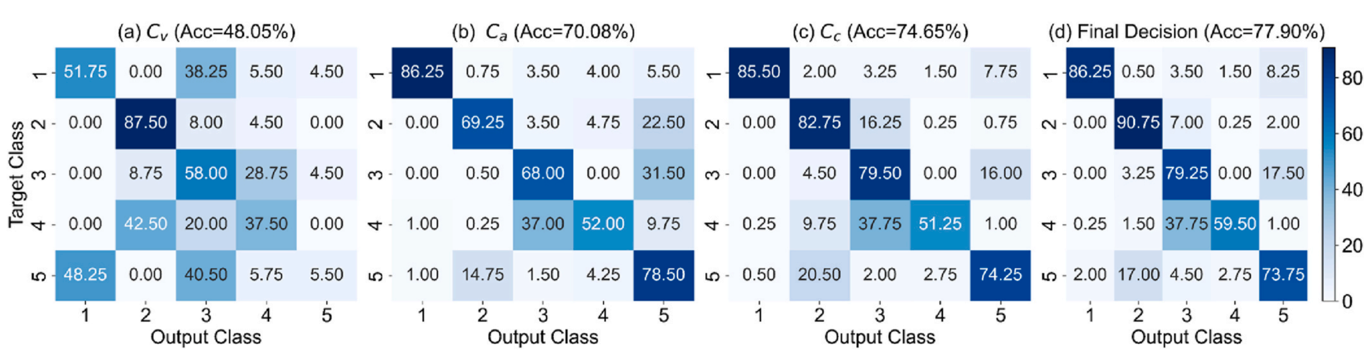

Fig. 10. Confusion matrices for the three classifiers and the final ensemble results on Task 5. (a) Vibration state classifier C_v, (b) Acoustic state classifier C_a, (c) Combination state classifier C_c, and (d) Final decision.

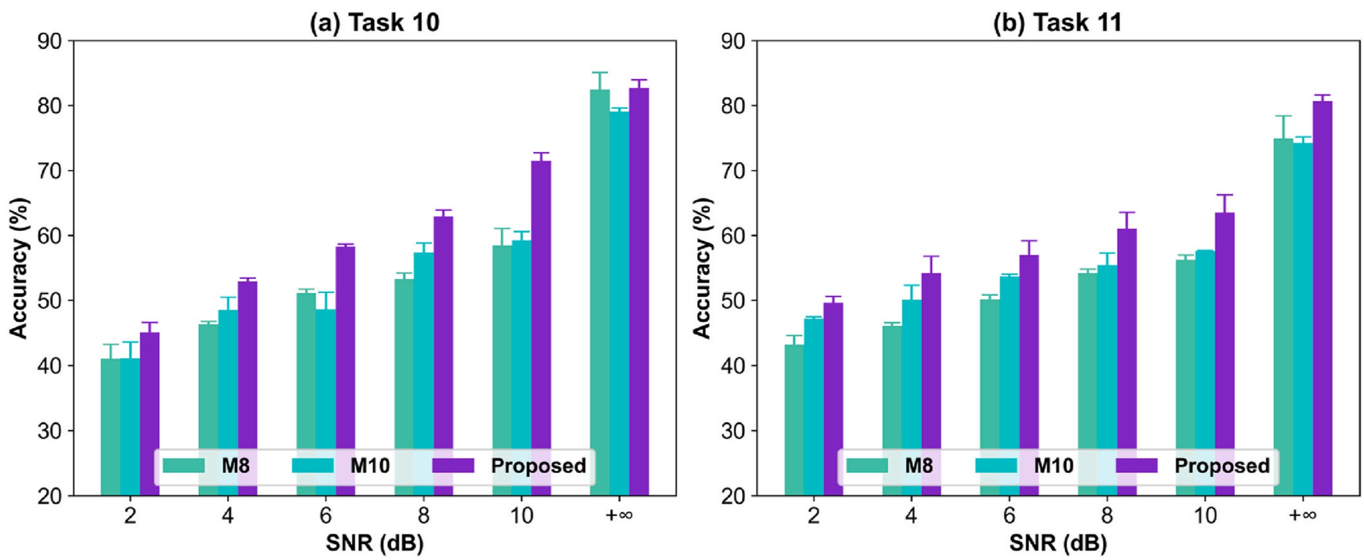

Fig. 11. Diagnosis results of different methods under varying levels of additional noise.

proposed method across all tasks. This suggests that the information-enhanced concatenation is crucial as it enriches feature diversity and helps to eliminate domain-specific information, thereby increasing resistance to distribution shifts. Comparing V2 with V1, it is evident that maximizing mutual information between high-level features from different domains enhances the quality of the unified representations, leading to better overall performance.

To provide an intuitive understanding of the proposed components, t-SNE technology is adopted to visualize the learned features of V1, V2, and the proposed method on Task 2. Fig. 9 shows the visualization results. From Fig. 9(a), it can be seen that source samples are not well clustered together and exhibit large intraclass variation. Additionally, target samples are not well clustered and are not aligned with source samples, indicating a large domain discrepancy between source and target domains. These results imply that simply concatenating multimodal features into unified representations cannot effectively resist domain shifts. Fig. 9(b) shows that maximizing mutual information of features from different domains can reduce intraclass variation and increase interclass separability, which helps improve the quality of unified representations. From Fig. 9(c), it can be seen that with the cooperation of information-enhanced concatenation and mutual information maximization, the interclass separability and intraclass clustering of the learned features are satisfactory. Additionally, unknown target domain samples are located in the same region as source domain samples. Clear decision boundaries and well-aligned domains contribute to the excellent generalization performance of the proposed method.

To display the effectiveness of hybrid ensemble diagnosis, confusion matrices of the three classifiers and the final decision are exhibited in Fig. 10. With feature-level fusion, the classifier C_c obtains higher accuracy than single-modal classifiers (C_v and C_a), indicating that multimodal unified representation containing more information can help to attenuate the domain shift. Fig. 10(d) shows the final ensemble predictions, which is better than those of the three classifiers. These results demonstrate that leveraging information from multiple levels enhances the robustness of intelligent fault diagnostic models.

(3) Performance analysis under noisy environment

In practical industrial settings, environmental noise is inevitable. Therefore, in this subsection, Gaussian white noise is added to the original vibration and acoustic signals. The signal-to-noise ratio (SNR) is defined as follows:

Table 8
Computation time (s) of different methods.

Method	Bearing fault diagnosis task		Motor fault diagnosis task	
	Training	Testing	Training	Testing
M3	63.02	0.37	40.43	0.05
M4	69.58	0.42	42.95	0.06
M5	68.80	0.62	42.99	0.08
M6	5567.64	3.77	1225.51	0.55
M7	10316.28	2.19	1033.41	0.40
M8	72.85	0.38	50.46	0.06
M9	81.43	0.37	54.75	0.06
M10	271.91	1.20	106.29	0.08
Proposed	95.47	0.44	63.97	0.06

$$SNR_{dB} = 10\log_{10}\frac{P_{signal}}{P_{noise}} \tag{14}$$

where P_{signal} and P_{noise} denote the powers of the original signal and the added Gaussian white noise, respectively.

Five Gaussian white noise SNR levels (2, 4, 6, 8, and 10 dB) are employed in Tasks 10 and 11. Fig. 11 presents the diagnostic results of three different methods. Accuracy means and standard deviations are evaluated over ten trials. As shown in Fig. 11, the diagnostic performance of all three methods decreases with increasing noise intensity. This is because the information in the original signals is disrupted and submerged by noise, leading to a higher misdiagnosis rate. Nevertheless, the proposed method consistently achieves the best performance across different noise levels, demonstrating its robustness in noisy environments.

(4) Analysis of computational efficiency

Table 8 presents the computational costs of various methods for bearing and motor diagnosis tasks. The testing time denotes the inference time required for all test samples. As shown in Table 8, the proposed method achieves shorter training time and faster inference speed compared with the advanced method M10. Although the proposed method requires slightly longer training time than M3, M4, M5, M8, and M9, the proposed method delivers superior diagnostic accuracy. Furthermore, since M6 and M7 are based on the CCDG framework, the inherent complexity of their model architectures results in relatively longer training and testing times.

Table 9
Results of bearing fault diagnostic tasks (Scenario 2).

Task Name	Vibration	Acoustic	SimMMDG (Zero padding)	Proposed (Zero padding)	SimMMDG (Modal translation)	Proposed (Modal translation)
Task 1	Missing	Real	90.42 ± 0.29	88.01 ± 1.17	**90.52 ± 0.33**	88.89 ± 0.67
	Real	Missing	21.02 ± 2.31	21.30 ± 1.84	56.70 ± 0.58	**66.67 ± 1.62**
Task 2	Missing	Real	83.04 ± 0.50	81.06 ± 2.23	82.13 ± 0.51	**86.94 ± 1.30**
	Real	Missing	20.00 ± 0.00	27.37 ± 10.00	41.12 ± 0.72	**52.17 ± 2.36**
Task 3	Missing	Real	82.79 ± 0.32	87.35 ± 1.12	83.93 ± 0.34	**90.64 ± 0.93**
	Real	Missing	20.22 ± 0.65	29.31 ± 8.33	52.74 ± 0.66	**66.00 ± 1.40**
Task 4	Missing	Real	83.18 ± 0.42	73.84 ± 6.43	80.77 ± 0.37	**85.59 ± 0.53**
	Real	Missing	20.00 ± 0.00	22.94 ± 5.13	49.30 ± 0.49	**52.82 ± 1.53**
Task 5	Missing	Real	53.67 ± 0.44	59.04 ± 3.00	54.40 ± 0.40	**64.89 ± 1.96**
	Real	Missing	20.00 ± 0.00	27.34 ± 9.40	41.67 ± 0.84	**46.71 ± 1.53**
Task 6	Missing	Real	60.13 ± 0.62	51.27 ± 1.60	**61.55 ± 0.58**	59.61 ± 1.27
	Real	Missing	20.00 ± 0.00	20.22 ± 0.67	**41.21 ± 0.26**	32.68 ± 0.34
Task 7	Missing	Real	57.62 ± 0.56	51.09 ± 5.30	57.80 ± 0.69	**64.88 ± 0.67**
	Real	Missing	20.06 ± 0.19	21.76 ± 3.52	44.41 ± 0.79	**50.09 ± 1.56**
Task 8	Missing	Real	61.95 ± 0.82	66.74 ± 1.92	62.73 ± 0.91	**69.70 ± 1.42**
	Real	Missing	20.00 ± 0.00	26.02 ± 9.21	34.16 ± 0.49	**41.75 ± 2.54**
Average			45.88	47.17	58.45	**63.75**

Table 10
Results of motor fault diagnostic tasks (Scenario 2).

Task Name	Vibration	Acoustic	SimMMDG (Zero padding)	Proposed (Zero padding)	SimMMDG (Modal translation)	Proposed (Modal translation)
Task 9	Missing	Real	29.55 ± 3.01	32.43 ± 2.16	43.57 ± 1.82	**49.15 ± 2.78**
	Real	Missing	54.68 ± 3.07	**79.83 ± 3.11**	76.98 ± 3.00	76.09 ± 2.59
Task 10	Missing	Real	35.70 ± 4.42	43.35 ± 5.73	47.85 ± 1.53	**63.22 ± 4.41**
	Real	Missing	62.97 ± 1.30	73.78 ± 3.90	76.49 ± 1.46	**78.32 ± 0.79**
Task 11	Missing	Real	24.18 ± 5.25	34.23 ± 2.34	48.63 ± 1.96	**61.96 ± 0.97**
	Real	Missing	59.85 ± 3.43	71.87 ± 4.41	75.50 ± 1.28	**79.05 ± 0.94**
Task 12	Missing	Real	29.29 ± 2.29	34.77 ± 2.33	20.17 ± 2.07	**35.44 ± 1.49**
	Real	Missing	39.28 ± 2.78	59.13 ± 2.37	46.48 ± 3.35	**62.35 ± 0.52**
Average			41.94	53.68	54.46	**63.20**

4.4. Experimental results and performance analyses for scenario 2

The effectiveness of data-driven diagnostic models relies on the proper functioning of sensors. However, the availability of condition monitoring sensors cannot always be guaranteed. This subsection evaluates the performance of the proposed method in scenarios where modalities are missing. In this section, the primary focus is on analyzing the role of the semantic supervision-based modal translators. There is no need to conduct ablation experiments to validate the contributions of the proposed information-enhanced concatenation and mutual information maximization.

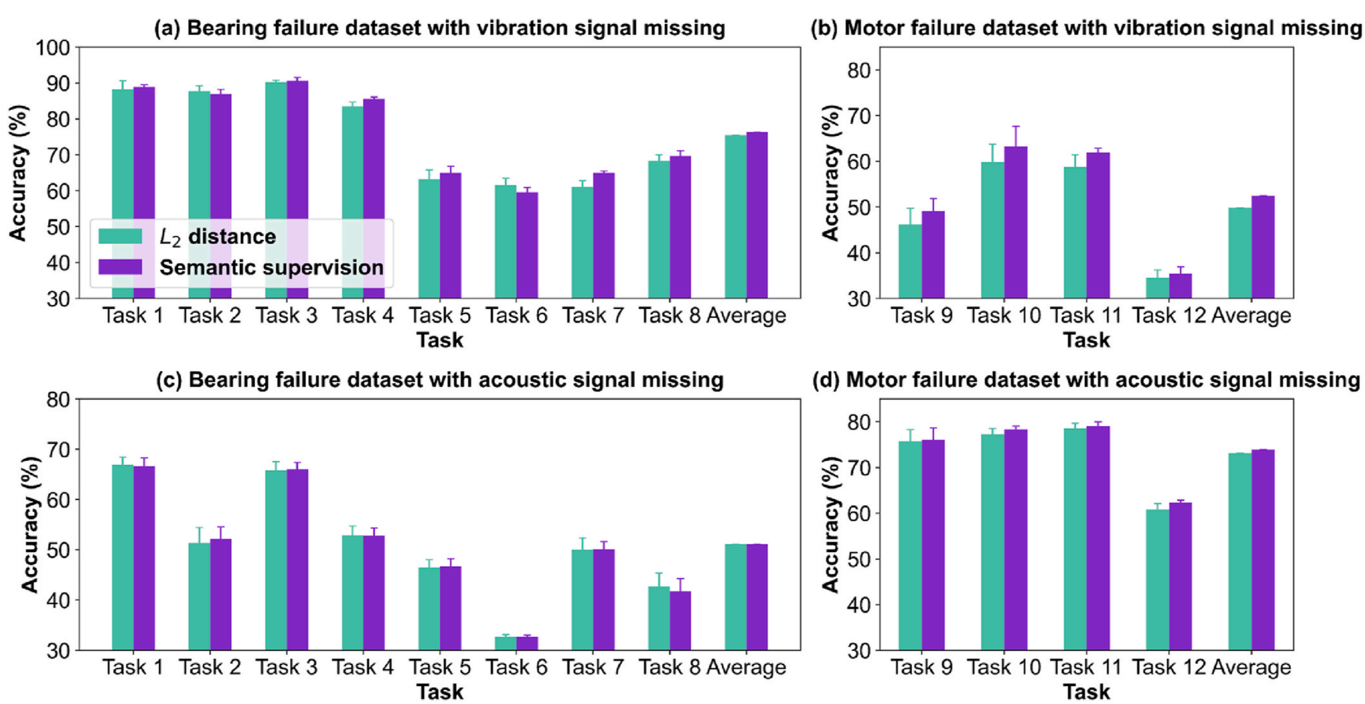

Fig. 12. Diagnostic results of the proposed method using different intermodal translation approaches.

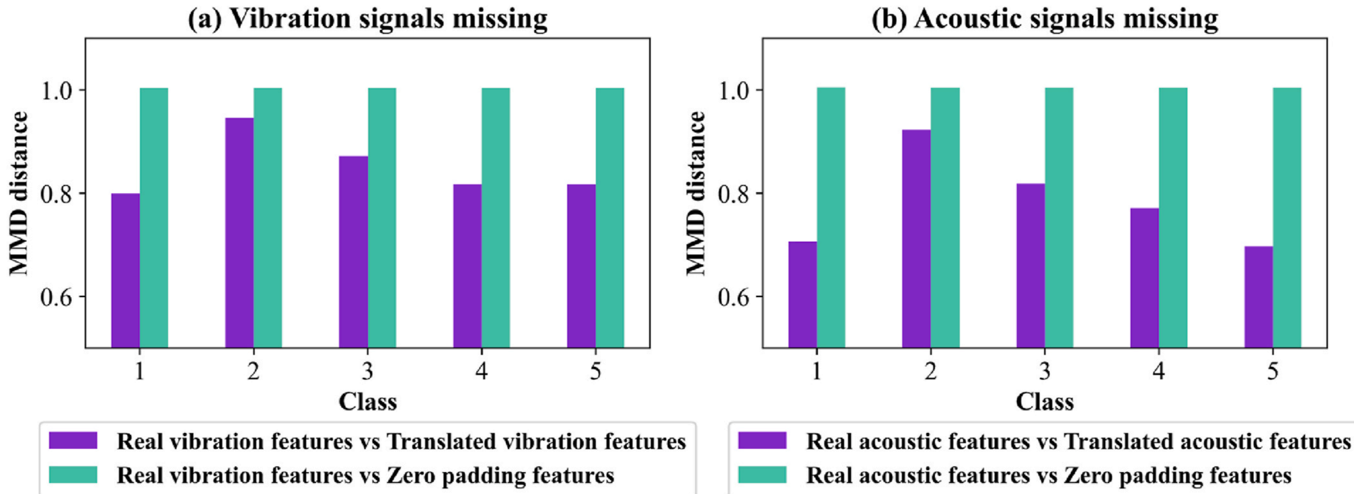

Fig. 13. Comparison of the MMD distances of different features learned by the proposed method in Task 7.

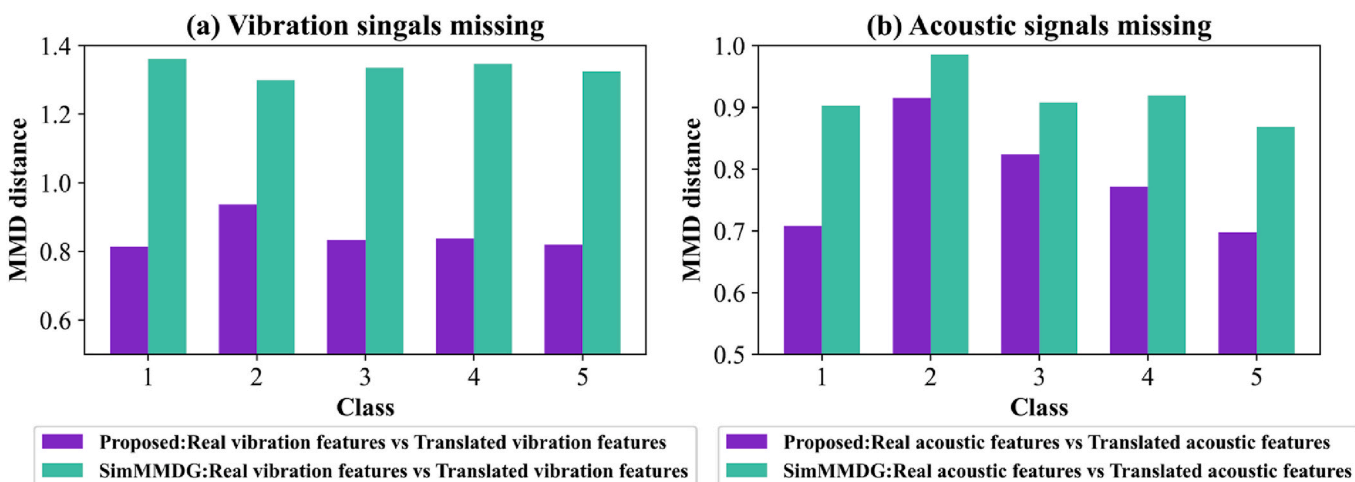

Fig. 14. Comparison of the MMD distances of features learned by two methods in Task 7.

(1) Diagnosis results of Scenario 2

When one of the sensors in a machinery monitoring system is broken, a simple approach is to set the collected data of the missing modality to zero. Tables 9 and 10 display the comparison results of SimMMDG and the proposed method on the missing modality scenario. Comparing the performance of SimMMDG with zero padding and modal translation, it is evident that modal translation improves the diagnostic performance of SimMMDG. These results indicate that learning modality relationships can help intelligent diagnostic models address the missing modality problem and enhance their adaptability to external limitations. Similarly, it can also be found that the proposed method with modal translation outperforms that with zero padding. Additionally, the accuracy of the proposed method with modal translation is higher than that of SimMMDG with modal translation in almost all tasks, demonstrating the superiority of the proposed method in MMDGFD with missing modality scenarios.

From Table 9, under the acoustic-missing scenario, the accuracy of SimMMDG (Zero padding) consistently approaches 20 %, highlighting the critical importance of acoustic signals for bearing diagnosis tasks. The slightly better performance of Proposed (Zero padding) can be attributed to its multi-sensor ensemble learning strategy, which prevents the model from complete failure when highly informative signals are missing. The larger variance observed in Proposed (Zero padding) compared to SimMMDG (Zero padding) arises because zero padding introduces non-informative signals, disrupting the consistency of intra-domain and cross-domain features generated by the information-enhanced concatenation module. Consequently, the mutual information maximization objective is impeded, leading to reduced model robustness.

(2) Performance of semantic supervision for intermodal translation

Previous study SimMMDG (Dong et al., 2023) utilizes L_2 distance to guide the modal translators in achieving intermodal translations. However, deep networks exhibit complex nonlinear mappings, where even slight variations in input can lead to significant differences in output. Therefore, maintaining semantic consistency between real data and fake data is important. Fig. 12 shows the comparison results of the proposed method with L_2 distance and semantic supervision. It can be found that the proposed semantic supervision is a better approach than L_2 distance.

To further validate the reliability of the features translated by the proposed method, Fig. 13 presents the maximum mean discrepancy (MMD) distances between the translated features and the real features. It also shows the MMD distances between the features generated from zero-padded inputs and the real features. As observed in Fig. 13, for each class, the MMD distance between the translated features and the real features is consistently smaller than that between the zero-padded input features and the real features. This provides further evidence for the effectiveness of the proposed method in intermodal translation.

Moreover, Fig. 14 shows the MMD distances for the proposed method and SimMMDG. It can be seen that the translated features generated by the proposed method are of higher quality, exhibiting lower MMD distances than those produced by SimMMDG. These results provide quantitative evidence of the superiority of the proposed method in intermodal feature translation.

5. Conclusion

In this paper, a valuable and realistic industrial task called multimodal domain generalization-based fault diagnosis is introduced. To address this challenge, a novel multimodal learning method called MUGTN is dedicated designed. Information-enhanced concatenation and mutual information maximization cooperatively help the deep diagnostic model to learn invariant unified representations. Hybrid ensemble diagnosis combines the advantages of different-level fusions, enhancing the robustness of the proposed method for different devices and environments. Additionally, semantic supervision is developed to achieve reliable intermodal translation, improving the flexibility of the proposed method in missing modality scenarios. Comprehensive experiments based on two mechanical test rigs verify the effectiveness of the proposed MUGTN in two scenarios.

This paper focuses on investigating MMDGFD under different working conditions. In the future, our work will investigate MMDGFD under cross-machine settings. Although the proposed method can be extended to other time-series signals, it is not capable of processing text-based data. In the future, we will explore methods that incorporate video, text, and other modalities.

CRediT authorship contribution statement

Chao Zhao: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Weiming Shen:** Writing – review & editing, Supervision, Funding acquisition. **Enrico Zio:** Writing – review & editing, Supervision, Funding acquisition. **Hui Ma:** Supervision.

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.

Acknowledgements

This research was supported in part by the Ministry of Science and Technology of the People's Republic of China under Grant 2022YFE0114200, in part by the Fundamental Research Funds for the Central Universities of China under Grant HUST:2021GCRC058, and in part by the China Scholarship Council (CSC) under Grant 202206160065.

Data availability

Dataset is available at https://github.com/CHAOZHAO-1/HUSTmotor-multi-modal-dataset.

References

Ding, A., Yi, X., Qin, Y., Wang, B., 2024. Self-driven continual learning for class-added motor fault diagnosis based on unseen fault detector and propensity distillation. Eng. Appl. Artif. Intell. 127, 107382. https://doi.org/10.1016/j.engappai.2023.107382.
Dong, H., Nejjar, I., Sun, H., Chatzi, E., Fink, O., 2023. SimMMDG: a simple and effective framework for multi-modal domain generalization. NeurIPS 1, 1–22.
Fan, C., Wang, P., Zhang, Y., Ma, H., Li, X., Wang, Q., 2025. Digital twin assisted degradation assessment of bearing cage performance. In: IEEE Trans. Ind. Informatics, pp. 1–11. https://doi.org/10.1109/TII.2025.3552655.
Gao, S., Wang, Y., Zhang, Z., Wang, B., He, B., Zio, E., 2025. Hybrid fault diagnosis method for underwater thrusters based on the common features of multi-source signals. Mech. Syst. Signal Process. 222, 1–17. https://doi.org/10.1016/j.ymssp.2024.111740.
Gao, Y., Qi, J., Sun, Ying, Hu, X., Dong, Z., Sun, Yanfei, 2024. Industrial process fault diagnosis based on feature enhanced meta-learning toward domain generalization scenarios. Knowl. Base Syst. 289, 111506. https://doi.org/10.1016/j.knosys.2024.111506.
Guo, C., Sun, Y., Yu, R., Ren, X., 2025. Deep causal disentanglement network with domain generalization for cross-machine bearing fault diagnosis. In: IEEE Trans. Instrum. Meas., pp. 1–16. https://doi.org/10.1109/TIM.2025.3545703, 74.
Han, T., Li, Y.-F., Qian, M., 2021. A hybrid generalization network for intelligent fault diagnosis of rotating machinery under unseen working conditions. In: IEEE Trans. Instrum. Meas., pp. 1–11. https://doi.org/10.1109/TIM.2021.3088489, 70.
Han, W., Chen, H., Poria, S., 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc., pp. 9180–9192. https://doi.org/10.18653/v1/2021.emnlp-main.723.
Hazarika, D., Zimmermann, R., 2020. Misa : Modality-Invariant and -Specific representations for multimodal sentiment analysis. Proc. 28th ACM Int. Conf. Multimed.
He, Y., Shen, W., 2024a. CFSPT: a lightweight cross-machine model for compound fault diagnosis of machine-level motors. Inf. Fusion 111, 102490. https://doi.org/10.1016/j.inffus.2024.102490.
He, Y., Shen, W., 2024b. A federated cross-machine diagnostic framework for machine-level motors with extreme label shortage. Adv. Eng. Inform. 61. https://doi.org/10.1016/j.aei.2024.102511.
Huo, Z., Martinez-Garcia, M., Zhang, Y., Shu, L., 2022. A multisensor information fusion method for high-reliability fault diagnosis of rotating machinery. IEEE Trans. Instrum. Meas. 71. https://doi.org/10.1109/TIM.2021.3132051.
iFLYTEK n.d. iFLYTEK AI DEVELOPER COMPETETION [WWW Document]. URL https://challenge.xfyun.cn/topic/info?type=rotating-machinery&ch=vWxQGFU.
Jia, L., Chow, T.W.S., Yuan, Y., 2024. Causal Disentanglement Domain Generalization for time-series signal fault diagnosis. Neural Netw. 172, 106099. https://doi.org/10.1016/j.neunet.2024.106099.
Kibrete, F., Engida Woldemichael, D., Shimels Gebremedhen, H., 2024. Multi-Sensor data fusion in intelligent fault diagnosis of rotating machines: a comprehensive review. Meas. J. Int. Meas. Confed. 232, 114658. https://doi.org/10.1016/j.measurement.2024.114658.
Kong, L., Peng, X., Chen, Y., Wang, P., Xu, M., 2020. Multi-sensor measurement and data fusion technology for manufacturing process monitoring: a literature review. Int. J. Extrem. Manuf. 2. https://doi.org/10.1088/2631-7990/ab7ae6.
Li, J., Wang, Y., Zi, Y., Zhang, H., Wan, Z., 2021. Causal disentanglement: a generalized bearing fault diagnostic framework in continuous degradation mode. IEEE Transact. Neural Networks Learn. Syst. 1–13. https://doi.org/10.1109/TNNLS.2021.3135036.
Liu, R., Zhang, Q., Lin, D., Zhang, W., Ding, S.X., 2024. Causal intervention graph neural network for fault diagnosis of complex industrial processes. Reliab. Eng. Syst. Saf. 251, 110328. https://doi.org/10.1016/j.ress.2024.110328.
Ma, M., Sun, C., Chen, X., Zhang, X., Yan, R., 2019. A deep coupled network for health State assessment of cutting tools based on fusion of multisensory signals. In: IEEE Trans. Ind. Informatics, pp. 6415–6424. https://doi.org/10.1109/TII.2019.2912428, 15.
Oord, A. van den, Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv.
Pei, X., Gao, L., Li, X., 2024. Remaining useful life prediction of machinery based on performance evaluation and online cross-domain health indicator under unknown working conditions. J. Manuf. Syst. 75, 213–227. https://doi.org/10.1016/j.jmsy.2024.06.005.
Planamente, M., Plizzari, C., Alberti, E., Caputo, B., 2022. Domain generalization through audio-visual relative norm alignment in first person action recognition. In: Proc. - 2022 IEEE/CVF Winter Conf. Appl. Comput. Vision. WACV, pp. 163–174. https://doi.org/10.1109/WACV51458.2022.00024, 2022.
Ragab, M., Chen, Z., Zhang, W., Eldele, E., Wu, M., Kwoh, C.K., Li, X., 2022. Conditional contrastive domain generalization for fault diagnosis. In: IEEE Trans. Instrum. Meas.. https://doi.org/10.1109/TIM.2022.3154000, 71.
Rahate, A., Walambe, R., Ramanna, S., Kotecha, K., 2022. Multimodal Co-learning: challenges, applications with datasets, recent advances and future directions. Inf. Fusion 81, 203–239. https://doi.org/10.1016/j.inffus.2021.12.003.
Sui, Y., Mao, W., Wang, S., Wang, X., Wu, J., He, X., Chua, T.S., 2024. Enhancing out-of-distribution generalization on graphs via causal attention learning. ACM Trans. Knowl. Discov. Data 18. https://doi.org/10.1145/3644392.
Sun, D., Li, Y., Liu, Z., Jia, S., Noman, K., 2024. Physics-inspired multimodal machine learning for adaptive correlation fusion based rotating machinery fault diagnosis. Inf. Fusion 108, 102394. https://doi.org/10.1016/j.inffus.2024.102394.
Tishby, N., Zaslavsky, N., 2015. Deep learning and the information bottleneck principle. In: 2015 IEEE Inf. Theory Work. ITW, pp. 1–5. https://doi.org/10.1109/ITW.2015.7133169, 2015.
Vapnik, V.N., 1999. An overview of statistical learning theory. In: IEEE Trans. Neural Networks, pp. 988–999. https://doi.org/10.1109/72.788640, 10.
Wang, Bingsen, Baraldi, P., Zio, E., 2023. Deep multiadversarial conditional domain adaptation networks for fault diagnostics of industrial equipment. In: IEEE Trans. Ind. Informatics, pp. 8841–8851. https://doi.org/10.1109/TII.2022.3222400, 19.

Wang, Bin, Wen, L., Li, X., Gao, L., 2023. Adaptive class Center generalization network: a sparse domain-regressive framework for bearing fault diagnosis under unknown working conditions. In: IEEE Trans. Instrum. Meas., pp. 1–11. https://doi.org/10.1109/TIM.2023.3273659, 72.

Wang, Z., Luo, Y., Qiu, R., Huang, Z., Baktashmotlagh, M., 2021. Learning to diversify for single domain generalization. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 814–823. https://doi.org/10.1109/ICCV48922.2021.00087.

Wen, L., Yang, G., Hu, L., Yang, C., Feng, K., 2024. A new unsupervised health index estimation method for bearings early fault detection based on Gaussian mixture model. Eng. Appl. Artif. Intell. 128, 107562. https://doi.org/10.1016/j.engappai.2023.107562.

Xie, T., Huang, X., Choi, S.K., 2022. Intelligent mechanical fault diagnosis using multisensor fusion and convolution neural network. In: IEEE Trans. Ind. Informatics, pp. 3213–3223. https://doi.org/10.1109/TII.2021.3102017, 18.

Yang, G., Tao, H., Wu, K., Du, R., Zhong, Y., 2024a. Fault diagnosis of harmonic drives using multimodal collaborative Meta network with severely missing modality. In: IEEE Trans. Ind. Informatics, pp. 1–9. https://doi.org/10.1109/TII.2024.3396339.

Yang, G., Tao, H., Yu, T., Du, R., Zhong, Y., 2024b. Online fault diagnosis of harmonic drives using semisupervised contrastive graph generative network via multimodal data. In: IEEE Trans. Ind. Electron., pp. 3055–3063. https://doi.org/10.1109/TIE.2023.3265056, 71.

Yao, J., Liu, C., Song, K., Feng, C., Jiang, D., 2021. Fault diagnosis of planetary gearbox based on acoustic signals. Appl. Acoust. 181, 108151. https://doi.org/10.1016/j.apacoust.2021.108151.

Zhang, G., Kong, X., Ma, H., Wang, Q., Du, J., Wang, J., 2025. Dual disentanglement domain generalization method for rotating Machinery fault diagnosis. Mech. Syst. Signal Process. 228, 112460. https://doi.org/10.1016/j.ymssp.2025.112460.

Zhang, T., Jiang, L., Liu, J., Zhang, X., Zhang, Q., 2025. A multi-scale deep feature memory and recovery network for multi-sensor fault diagnosis in the channel missing scenario. Eng. Appl. Artif. Intell. 145, 110228. https://doi.org/10.1016/j.engappai.2025.110228.

Zhang, X., Huang, W., Wang, R., Liao, Y., Ding, C., Wang, J., Shi, J., 2023. Engineering Applications of Artificial Intelligence Multi-stage distribution correction : a promising data augmentation method for few-shot fault diagnosis. Eng. Appl. Artif. Intell. 123, 106477. https://doi.org/10.1016/j.engappai.2023.106477.

Zhang, Y., Ding, J., Li, Y., Ren, Z., Feng, K., 2024. Multi-modal data cross-domain fusion network for gearbox fault diagnosis under variable operating conditions. Eng. Appl. Artif. Intell. 133, 108236. https://doi.org/10.1016/j.engappai.2024.108236.

Zhao, C., Shen, W., 2024. Imbalanced domain generalization via Semantic-Discriminative augmentation for intelligent fault diagnosis. Adv. Eng. Inform. 59, 1–20. https://doi.org/10.1016/j.aei.2023.102262.

Zhao, C., Shen, W., 2023a. Mutual-assistance semisupervised domain generalization network for intelligent fault diagnosis under unseen working conditions. Mech. Syst. Signal Process. 189, 110074. https://doi.org/10.1016/j.ymssp.2022.110074.

Zhao, C., Shen, W., 2023b. Adversarial mutual information-guided single domain generalization network for intelligent fault diagnosis. In: IEEE Trans. Ind. Informatics, pp. 2909–2918. https://doi.org/10.1109/TII.2022.3175018, 19.

Zhao, C., Shen, W., 2022. A domain generalization network combing invariance and specificity towards real-time intelligent fault diagnosis. Mech. Syst. Signal Process. 173, 108990. https://doi.org/10.1016/j.ymssp.2022.108990.

Zhao, Z., Xu, Y., Zhang, J., Zhao, R., Chen, Z., Jiao, Y., 2024. A semi-supervised Gaussian Mixture Variational Autoencoder method for few-shot fine-grained fault diagnosis. Neural Netw. 178, 106482. https://doi.org/10.1016/j.neunet.2024.106482.

Zheng, X., Nie, J., He, Z., Gao, M., 2025. Specific task-guided collaborative domain generalization network for intelligent fault diagnosis under unseen conditions. IEEE Internet Things J. 12, 30835–30849. https://doi.org/10.1109/JIOT.2025.3571418.

Zio, E., 2022. Prognostics and Health Management (PHM): where are we and where do we (need to) go in theory and practice. Reliab. Eng. Syst. Saf. 218, 108119. https://doi.org/10.1016/j.ress.2021.108119.