Composite Neuro-Fuzzy System-Guided Cross-Modal Zero-Sample Diagnostic Framework Using Multi-Source Heterogeneous Non-Contact Sensing Data

Sheng Li, Jinchen Ji, Ke Feng, Ke Zhang, Qing Ni, Yadong Xu

***Abstract*—Zero-sample diagnostic methods have gained recognition in addressing the scarcity of gearbox fault samples, thereby being regarded as a promising technique to guarantee gearbox safety. However, historical zero-sample approaches typically neglect the use of multi-modal non-contact sensing data and rarely consider the interpretability of the diagnostic process. This oversight limits their application in industrial environments that require high reliability or operate under extreme conditions. Therefore, this paper presents a composite neuro-fuzzy system-guided cross-modal zero-sample diagnostic framework, termed FCZD-IA, which employs infrared thermography and acoustic data to monitor gearbox conditions. Specifically, FCZD-IA uses a proposed composite neural system as a decision-maker in the diagnostic task, while integrating a deep backbone network to discriminatively learn high-level fault features from multi-modal data. Moreover, a specific training strategy is designed to guide the learning process of the FCZD-IA to promote robust and interpretable zero-sample diagnostics. Comprehensive experimental results validate the effectiveness of the proposed framework and its superiority over other competitive methods.**

***Index Terms*—Gearbox, Non-Contact Data Processing, Multi-Modal Learning, Deep Neuro-Fuzzy System, Zero-Sample Diagnosis.**

I. Introduction

GEARBOXES are essential in various applications of engineering systems. Failures in the crucial components of gearboxes can lead to significant and far-reaching consequences [1] [2] [3]. Thus, incorporating real-time monitoring and fault diagnosis is imperative to guarantee the industrial system's safety, reliability, and efficient operation [4].

Recently, deep learning has witnessed remarkable advancements in pattern recognition, providing a promising tool for healthy condition monitoring of rotating machinery. Specifically, supervised deep learning methods have shown advantages in industrial fault diagnosis by leveraging large annotated datasets. For instance, Xu et al. [2] specially designed a hierarchical, densely connected deep network to supervise and recognize the health status of electromechanical systems. Additionally, various attention modules, such as self-attention [3], cross-attention [5], and frequency attention mechanisms, have been integrated with supervised diagnostic methods to optimize performance. It should be noted that the application of these supervised methods is restricted in real industry scenarios due to their reliance on extensive annotated samples and the challenge of diagnosing unknown defects without annotations. To tackle these limitations, zero-sample diagnosis has emerged as a viable tool for diagnosing unknown defects [6] [7]. Relevant approaches generally utilize a shared semantic space and prior knowledge of known statuses to infer feature representations of potentially unknown statuses, even when the models have not previously processed samples of unknown statuses [8]. For instance, Feng et al. [5] introduced a knowledge transfer technique that leverages semantic vectors to bridge the correlations between observed and unobserved defect categories, thereby enabling zero-sample diagnosis. Similarly, Hu et al. [4] developed a semantic-consistent embedding-based approach for zero-sample diagnosis tasks and demonstrated its effectiveness in a three-phase transmission system. Notably, historical zero-sample diagnostic methods have predominantly relied on contact sensing data, particularly vibration signals. However, under certain conditions, such as extreme temperatures and high humidity, the accuracy of these contact sensors can be significantly compromised [9]. In contrast, non-contact monitoring techniques, including thermal imaging, acoustic sensing, laser Doppler vibrometry, optical sensing, radar-based monitoring, and electromagnetic field monitoring, offer considerable advantages such as enhanced environmental compatibility and broader monitoring capabilities [1]. Particularly, infrared thermal (IRT) imaging and acoustic (AC) sensing data are frequently employed due to their user-friendliness and precision. Nevertheless, processing IRT and AC sensing data introduces distinct challenges, as illustrated by the gearbox

This work was supported by the Jiangsu-Hong Kong-Macao University Alliance (JHMUA) Funding and the National Natural Science Foundation of China (No. 72404034). (*Corresponding author: Ke Feng and Yadong Xu*)

Sheng Li is with the International Machinery Center, School of Mechanical Engineering, Xi'an Jiaotong University, Shaanxi 710049, China (e-mail: shengli@hhu.edu.cn).

Ke Zhang is with the Bussiness School, Hohai University, Nanjing, 211100, China (e-mail: kezhang@hhu.edu.cn).

Ke Feng is with the State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710054, China, and also with the School of Mechanical Engineering, Xi'an Jiaotong University, Shaanxi 710049, China (e-mail: ke.feng@outlook.com.au).

Yadong Xu is with the Department of Industrial and Systems Engineering and also with the Research Institute for Advanced Manufacturing, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: 230208033@seu.edu.cn).

Jinchen Ji and Qing Ni are with the School of Mechanical and Mechatronic Engineering, University of Technology Sydney, NSW 2007, Australia (e-mail: Jin.Ji@uts.edu.au; Qing.Ni@outlook.com.au).

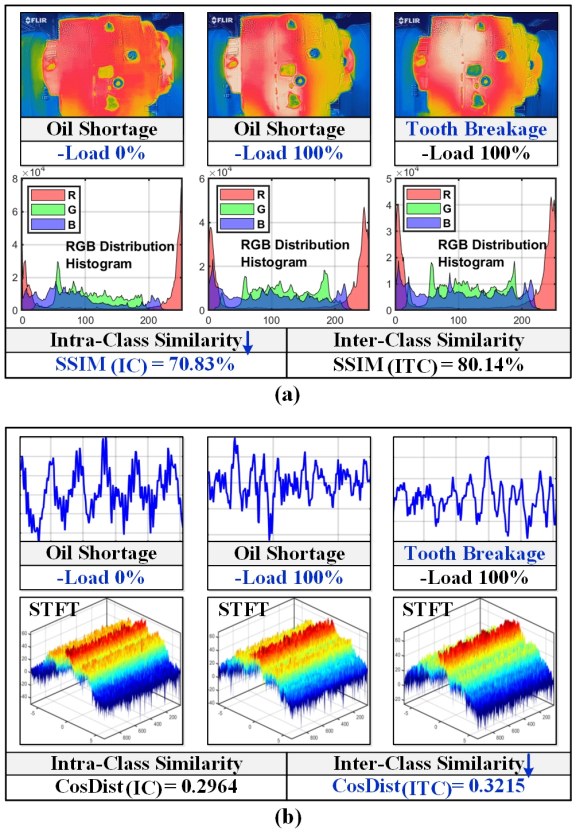

Fig. 1. Data characteristic analysis of multi-source heterogeneous non-contact sensing data, i.e., (a) IRT visual data and (b) acoustic data, under load variation. The similarity of IRT visual data is calculated using the Structural Similarity Index Measure (SSIM) algorithm. For acoustic data, the similarity is given by the cosine distance of its Fast Fourier Transform (STFT) waveform.

example.

In Fig. 1(a), regarding the status of the gearbox's lubrication oil shortage, the load variation from 0% to 100% diminishes the visual distribution similarity of intra-class samples and results in the distribution similarity of inter-class samples even surpassing that of intra-class samples ($SSIM_{(ITC)} > SSIM_{(IC)}$). Intra-class samples' excessive variation may let deep networks learn misleading feature semantics, potentially resulting in incorrect recognition of oil shortage as tooth breakage status. Meanwhile, it is noteworthy that the acoustic data is less sensitive to load variation, preserving a robust similarity of the intra-class acoustical distribution ($CosDist_{(IC)} > CosDist_{(ITC)}$), which indicates that emphasize the modality complementary in the non-contact diagnosis task is critical, which have been proved in some historical works [1]. Moreover, despite the effectiveness of deep learning methods in processing high-dimensional and multi-modal non-contact sensing data and achieving impressive results [1], their diagnostic processes still can be regarded as 'black boxes'. This opacity in interpretation limits their application in diagnostic scenarios where reliability is critical.

Recently, the integration of Neuro-Fuzzy Systems (NFS) and Deep Neural Network (DNN) has emerged as a promising hybrid, leveraging the high-dimensional data processing capabilities of DNN while integrating the interpretability of NFS. For instance, Zhang et al. [10] incorporated a fuzzy system into a deep echo state network in a machinery fault diagnosis task. Xu et al. [11] transformed the learning weights of a fuzzy neural network into guidance operators to facilitate the diagnosing process of healthy statuses. Perez-Perez et al. [12] used a Takagi-Sugeno model to represent system behavior, achieving admirable accuracy in the wind turbine diagnostic task. Despite historical studies demonstrating the potential of integrating NFS and DNN, current applications remain predominantly limited to supervised diagnostic tasks and primarily employ contact sensing techniques [13]. To the best of our knowledge, NFS has not yet been applied in zero-sample diagnostic tasks or for simultaneous processing of multi-source heterogeneous non-contact data. One reason for this limitation is that zero-sample diagnostic tasks typically require more high-level feature extraction capability in comparison with supervised diagnostic tasks, including modeling dependency relationships between observed (Ob) and unobserved (Ub) faults and processing high-dimensional data to extract transferable features. However, NFS-based methods, represented by ANFIS, benefit from their interpretability but face limitations in extracting deep high-level features due to the rules exponentially explosion, challenging to achieve competitive performance in such complex diagnostic tasks compared to DNN-based methods. Techniques for dimensionality reduction, e.g., PCA and FCA, can mitigate these deficiencies to some extent but inevitably reduce the interpretability of NFS.

To tackle the aforementioned challenges and further improve the interpretability of the zero-sample diagnosis process, we propose a novel neural-fuzzy hybrid. To highlight our contributions to the neural-fuzzy hybrid, we first outline the key differences between our proposed method in applying the fuzzy system and the previous works. Part of the previous works primarily utilized fuzzy rules solely for data preprocessing or post-processing [14], where the fuzzy systems are not substantially integrated with the neural learning process [15]. Moreover, numerous works merely utilized the membership function as an unlearnable activation unit to model nonlinear relationships [16] [17] [9]. Moreover, in the regression and classification tasks [18] [19] [16] [20], historical works typically regarded NFS-based method, e.g., ANFIS, as a feature extractor or terminal classifier [21] [22]. The results, although encouraging, appear coarse. This is because non-fuzzy methods, e.g., convolution neural networks or transformer-like networks, have been proven more competitive in such tasks [23] [24] [25]. In comparison, our proposed method does not concentrate on improving the feature-extracting capabilities of the Neuro-Fuzzy System (NFS). Instead, we introduce a new concept within our framework where there is no need for the NFS to directly process high-dimensional data, thus avoiding the complications associated with rule explosions. Specifically, we employ a sophisticated network structure and corresponding specialized training strategy that allows the NFS to guide the deep neural network to facilitate discriminative learning and complementary fusion. This approach effectively combines the powerful feature-extracting strengths of deep neural networks with the interpretability of NFS.

Inspired by the historical works, we have embedded the neuro-fuzzy system (NFS) as a collaborative guide component within a sophisticated cross-modal zero-sample diagnostic framework, aiming to leverage the human-like reasoning capabilities of the NFS to guide the deep neural network-based zero-sample diagnosis process, thereby enhancing the transparency of the diagnostic procedures. To the best of our knowledge, this is the first attempt to integrate an NFS with a cross-modal zero-sample diagnosis algorithm. Table 1 gives a comparative analysis showcasing the advancements achieved by our work. To achieve these objectives, we first propose a series of cooperative modules along with a specialized loss function: 1) **Attention-based Cross-modal Fusion (ACF) Module**, which employs a sophisticated cross-connected self-attention mechanism to facilitate the interaction and fusion of complementary information across different modalities. 2) **Composite Neuro-Fuzzy System (CNFS)**, which is constructed by dual neuro-fuzzy systems to map infrared thermography visual and acoustic features into fuzzy soft labels, which provide detailed and discriminative information among defect categories, which are essential for modeling the correlations between observed (Ob) and unobserved (Ub) defects in a zero-sample framework. 3) **Fuzzy Soft Triplet (FST) Loss Function**, which uses fuzzy soft labels to guide the network in multi-modal feature fusion and discriminative learning. Additionally, by integrating a specific component, the FST loss can optimize intra-class feature distribution shifts induced by load variations. Based on these developed modules, we further outline the core contributions of our work as follows:

- **A composite neuro-fuzzy system-guided cross-modal zero-sample diagnostic (FCZD-IA) network is developed for gearbox health monitoring.** Specifically, the FCZD-IA uses the deep sub-network to extract high-level, multi-modal, gearbox defect-related features from infrared thermography and acoustic modalities. Simultaneously, it leverages the CNFS to generate fuzzy soft labels that interpretably delineate the discriminative relationships among defect statuses, crucial for inferring potentially unknown statuses in zero-sample diagnostic tasks. Notably, by integrating the interpretability of the neural fuzzy system with the data processing capabilities of deep learning methods, our work circumvents the rule explosion typically encountered in Neuro-Fuzzy Systems when directly processing multi-modal fault samples, thereby enabling the capability to achieve interpretable cross-modal zero-sample diagnostics.
- **A Cross-Modal Fuzzy Learning (CFL) strategy is introduced to update the parameters of the FCZD-IA network during the training phase.** The CFL strategy innovatively leverages generated fuzzy soft labels to guide the complementary fusion and discriminative learning of the extracted multi-modal features, and it adaptively optimizes load variation-induced intra-class variance. Furthermore, by supervising semantic consistency within multi-modal feature representations, the CFL strategy mitigates the representation ambiguities resulting from modality differences. Summarily, these integrated functionalities enable the CFL strategy to enhance zero-sample diagnostic performance by effectively leveraging information from multiple modalities. Three zero-sample diagnosis scenarios demonstrate the effectiveness of this strategy.

TABLE I
COMPARISON OF TECHNICAL DETAILS WITH HISTORICAL MODELS. ✓DENOTES THE PRESENCE OF A TECHNIQUE, WHILE A BLANK SPACE INDICATES ITS ABSENCE. KEY TERMS USED INCLUDE MM FOR THE USE OF MULTIPLE MODALITIES, ZSD FOR ZERO-SAMPLE DIAGNOSIS, NFS FOR NEURO-FUZZY SYSTEM, ATT FOR THE ATTENTION MECHANISM, NCS FOR NON-CONTACT SENSING, AND NWCS FOR NON-STATIONARY WORKING CONDITIONS.

Methods	MM	ZSD	NFS	Att	NCS	NWCs
AMDC-CNN [26]	✓			✓		
CMFCNN [9]	✓			✓		
SCE [4]	✓	✓		✓	✓	
DARN [27]		✓		✓		✓
LDS-IFD [28]	✓					
CAE-CNN [29]		✓		✓		
CADAE [30]		✓				✓
ZLCFDM [31]		✓		✓		
AIFN-IA [1]	✓			✓	✓	
DFESN [10]			✓			
FSDN [11]			✓			
TS-FD [32]	✓			✓		
ANFIS-TS [12]			✓			
SVM-FDNN [33]	✓		✓			
ANFIS-DA [34]			✓			✓
FCZD-IA(Ours)	✓	✓	✓	✓	✓	✓

The remainder of this paper is organized as follows: Section II offers a detailed description of our proposed method, elaborating on its intricacies. Section III presents an extensive evaluation of the FCZD-IA's performance using variable-load gearbox datasets. Section IV is dedicated to the verification and comprehensive discussion of the approach developed in this study. Finally, Section V concludes the paper, summarizing the main findings and outlining potential avenues for future research.

II. THE PROPOSED METHOD

A. Attention-based Cross-Modal Fusion Module

As described in Fig. 1, we have demonstrated the importance of modality complementarity in the zero-sample diagnosis task when using infrared thermography (IRT) and acoustic (AC) data. Motivated by this finding, we developed an Attention-based Cross-modal Fusion (ACF) module designed to facilitate information interaction and fusion between IRT and AC modalities. Specifically, the ACF module consists of two cross-connected multi-head self-attention (MHSA) layers, each followed by its respective feed-forward layers as depicted in Fig. 2. Input IRT and acoustic feature vectors are denoted as $O_v = (o_1^v, \ldots, o_N^v)$ and $O_a = (o_1^a, \ldots, o_N^a)$, respectively, with dimensions $O_v, O_a \in \mathbb{R}^{1\times N}$. The inputs $O_i, i \in \{v, a\}$

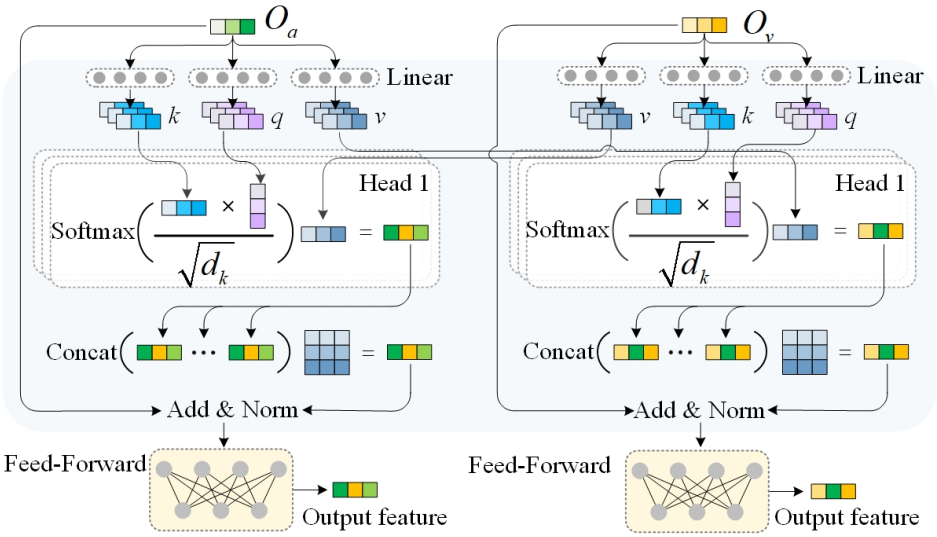

Fig. 2. Structure of the Attention-based Cross-modal Fusion (ACF) module. $Concat(\cdot)$ denotes the feature concatenate operation.

are initially processed by the cross-connected MHSA layers, formulated as:

$$\widetilde{O_i} = f_{LN}(f_{SA}(O_i) + O_i), i \in \{v, a\} \tag{1}$$

where $f_{LN}(\cdot)$ is a layer normalization operation. $f_{SA}(\cdot)$ represents the cross-attention operation, formulated as follows:

$$f_{SA}(O_i) = f_{Softmax}(\frac{q(O_i)k(O_i)^T}{\sqrt{d_k}})v(O_{\bar{i}}) \tag{2}$$

where $f_{Softmax}(\cdot)$ denotes the softmax operation. If $i \in \{v, a\}$, then $\bar{i} \in \{a, v\}$. $q(\cdot)$, $k(\cdot)$, and $v(\cdot)$ represent query, key, and value matrices. d_k denotes the number of attention heads for normalization. Subsequently, a residual-connected feed-forward layer is utilized to further enhance the module's capability for non-linear modeling [1].

TABLE II
DETAILS OF THE MODALITY ANALYSIS.

Modality	Indicators
RGB distribution indicators	Red channel mean
	Green channel mean
	Blue channel mean
	Homogeneity energy metric
	Textural contrast index
Frequency domain indicators	Peak frequency
	Variance of central frequency
	Frequency center
	Root mean square frequency
	Total spectral power

B. Composite Neuro-Fuzzy System

In this subsection, we introduce a novel composite neural-fuzzy system, termed CNFS (Fig. 3), designed to simulate the process of human understanding IRT and auditory data to achieve interpretability. Initially, we analyze how humans intuitively understand these two modalities. Notably, our work focuses on gearbox damages, categorized as structural damages, i.e., lubricating oil shortage (OS) and worsened oil viscosity (VIS), and non-structural damages, i.e., tooth breakage (TB). Historical studies have showcased that an OS defect may lead to increased convective thermal resistance and, consequently, a temperature rise, while a VIS defect may cause an increase in casing temperature. Both conditions significantly alter the temperature of the gearbox, which is evident in the IRT images or videos [1]. Humans can intuitively identify potential damages by observing imaging variations, which are essentially changes in RGB distribution. Thereby, within the CNFS, we use the RGB distribution indicators in IRT videos to simulate human visual perception. Moreover, structural damage, such as gear tooth breakage (TB), typically manifests as distinct spectral peaks within specific frequency bands, which may be perceived by humans as aberrant noises. Skilled technicians can identify potential structural defects by detecting these aberrant acoustic variations. Within the proposed CNFS, Fast Fourier Transform (FFT) analysis [6] is initially applied to the acoustic signals to summarize critical frequency domain indicators. This fashion simulates and systematizes the human capability to perceive gearbox damages through aberrant sounds.

As shown in Table 2, we merely select a limited set of key metrics as inputs to the NFS to prevent dimensionality explosion while still featuring effectiveness, as further validated in the experimental section. However, these features alone are inadequate for the zero-sample task. Therefore, we utilize a Deep Neural Network (DNN) to extract advanced features from the IRT and AC data, thereby complementing the NFS's capabilities.

Based on the analysis above, as depicted in Fig. 3, we have constructed an advanced Composite Neuro-Fuzzy System (CNFS) to simulate the decision-making process of the

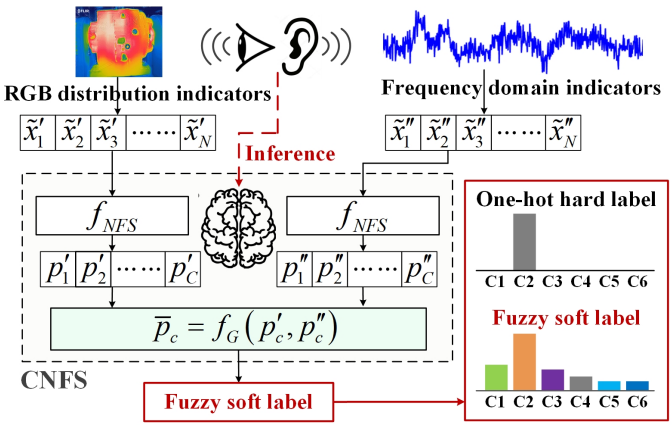

Fig. 3. Implementation Details of the Composite Neuro-Fuzzy System (CNFS). f_{NFS} denotes the feature mapping operation via NFS. f_G is soft voting operation.

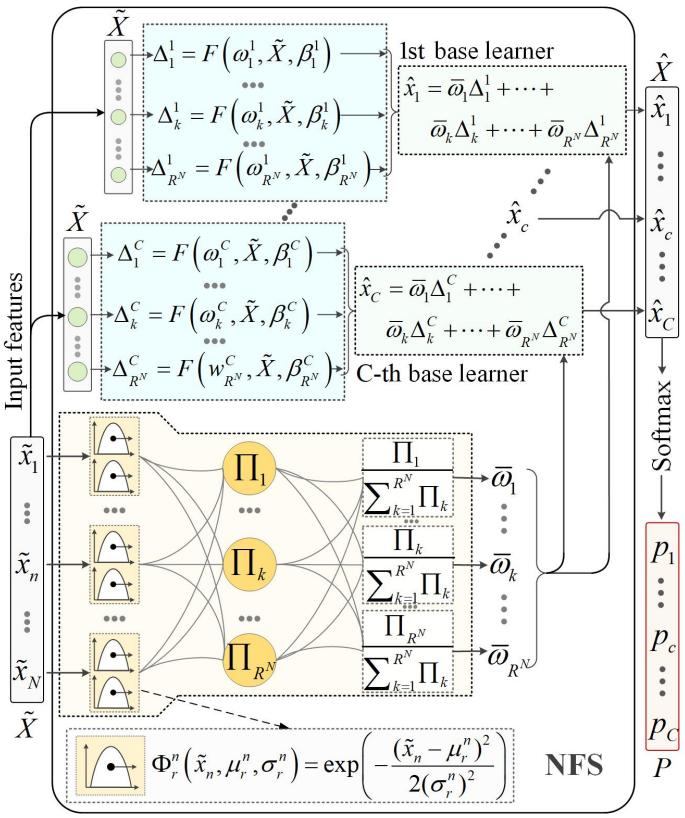

Fig. 4. Structure of the Neuro-Fuzzy System (NFS).

"brain," which leverages existing experience (rules) to infer the probability that the input features (IRT RGB distribution and acoustic frequency domain indicators) belong to certain fault categories. Within the CNFS, a soft voting strategy is employed to aggregate independent decisions from the IRT and AC modalities, thereby enabling a comprehensive final decision. We convert the format of the final decision into a probability vector, termed fuzzy soft labels in our work, to aid the guided deep-learning branch in understanding the decision results. The CNFS comprises dual NFSs to process the input IRT and acoustic features respectively. The NFS can be described as follows:

Neuro-Fuzzy System (NFS): Given the input feature vector (IRT visual or acoustic frequency domain feature vector) $\widetilde{X} = (\widetilde{x}_1, \ldots, \widetilde{x}_N) \in \mathbb{R}^{1\times N}$. NFS comprises multiple base learners [35]. For each base learner, $\widetilde{X}$ is initially element-wised transformed into membership values via Gaussian membership function $\Phi_r^n(\cdot)$, which are defined as follows:

$$\Phi_r^n\left(\widetilde{x}_n, \mu_r^n, \sigma_r^n\right) = exp\left(-\frac{(\widetilde{x}_n - \mu_r^n)^2}{2(\sigma_r^n)^2}\right) \tag{3}$$

Where $r = 1, 2, \ldots, R$ is the index of the fuzzy rules for each element in $\widetilde{X}$, and n is the element index of $\widetilde{X}$. Consequently, there are a total of R^N combinations of fuzzy rules [36]. Given an input IRT or AC feature vector of size $\mathbb{R}^{1\times\overline{N}}$, processing the feature vector directly with the NFS leads to an exponential growth in fuzzy rules ($R^N \rightarrow R^{\overline{N}}$) [12] [37]. However, within our method, we restrict the NFS to processing only N modality indicators ($N \ll \overline{N}$), thereby significantly reducing the number of calculated rules and helping mitigate the rule explosion issue to a degree.

Notably, μ_r^n and σ_r^n are the parameters for the r-th fuzzy rule of the n-th element. Following this, we define the firing degree $\prod_k$ specific rule for the given input $\widetilde{X} = (\widetilde{x}_1, \ldots, \widetilde{x}_N)$ as below:

$$\prod_k\left(\widetilde{X}\right) = \Phi_k^1(\widetilde{x}_1) \times \Phi_k^2(\widetilde{x}_2) \times \ldots \times \Phi_k^N(\widetilde{x}_N) \tag{4}$$

Within NFS, each base learner maps the input feature vector $\widetilde{X} = (\widetilde{x}_1, \ldots, \widetilde{x}_N) \in \mathbb{R}^{1\times N}$ to its fuzzy representation through a linear regression model [35] [38], parameterized as follows:

$$\Delta_k^c = F\left(w_k^c, \widetilde{X}, \beta_k^c\right) = \sum_{i=1}^{N} \omega_{i,k}^c \widetilde{x}_i + \beta_k^c \tag{5}$$

where $k = 1, 2, \ldots, R^N$ indexes the rule combinations, and $c = 1, 2, \ldots, C$ indexes the defect categories. The final output vector $\widehat{X} = (\widehat{x}_1, \ldots, \widehat{x}_C)$, is calculated by weighting these predictions according to the fuzzy logic combination of the firing strengths of the rules [37]:

$$\widehat{x}_c = \sum_{k=1}^{R^N}\left(\frac{\prod_k\left(\widetilde{X}\right)}{\sum_{j=1}^{R^N}\prod_j\left(\widetilde{X}\right)}\right)\Delta_k^c \tag{6}$$

Where $\widehat{X} = (\widehat{x}_1, \ldots, \widehat{x}_C) \in \mathbb{R}^{1\times C}$ represents the fuzzy representation corresponding to input feature vector $\widetilde{X} = (\widetilde{x}_1, \ldots, \widetilde{x}_N) \in \mathbb{R}^{1\times N}$. We subsequently use the softmax operation to transform the $\widehat{X} = (\widehat{x}_1, \ldots, \widehat{x}_{\widetilde{N}}) \in \mathbb{R}^{1\times\widetilde{N}}$ into a probability vector $P = (p_1, \ldots, p_C) \in \mathbb{R}^{1\times C}$, formulated as follows:

$$p_c = \frac{exp(\widehat{x}_c)}{\sum_{c=1}^{C} exp(\widehat{x}_c)} \tag{7}$$

Where, the exponentiation of each $\widehat{x}_c$ ensures that each element contributes positively, and the transformed probability vector sums to 1 [38]. We further compare our method with the first-order rule of the Takagi-Sugeno-Kang (TSK) type to demonstrate our method's distinctiveness [39]. The first-order

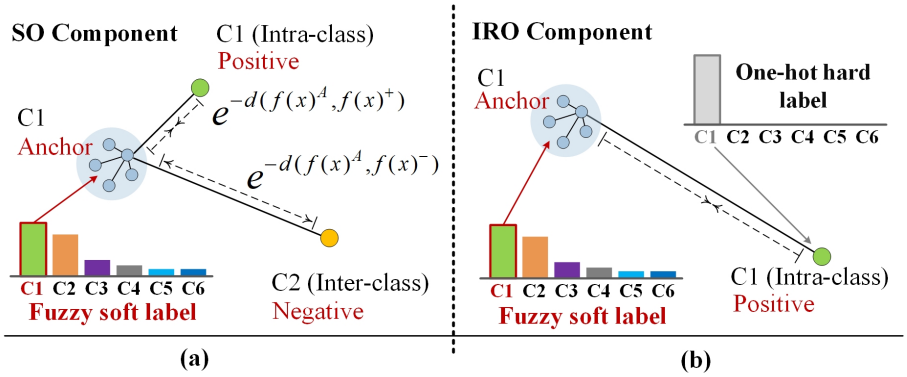

Fig. 5. Implementation Details of the Fuzzy Soft Triplet (FST) Loss. 'SO' and 'IRO' represent the standard optimization and intra-class re-optimization, respectively.

TSK can be formulated as:

$$\begin{aligned} &\text{Rule: If } x_1 \text{ is } R_1, ..., x_i \text{ is } R_i, ..., x_n \text{ is } R_n. \\ &\text{Then: } y = c_0 + c_1 x_1 + ... + c_i x_i + ... + c_n x_n \end{aligned} \tag{8}$$

Where x_i is the i-th element of an input feature. y denotes the corresponding crisp output. R_i is the i-th fuzzy sets, and c_i is the consequent coefficient of the TSK rule. In comparison, in our method, each base learner outputs the probability of each category for a given input through the softmax function. In this case, for a C classification task, a rule of our method can be formulated as:

$$\begin{aligned} &\text{Rule: If } x_1 \text{ is } R_1, ..., x_i \text{ is } R_i, ..., x_n \text{ is } R_n. \\ &\text{Then: } y \text{ is class 1 with } p_1, ..., \text{ is class } c \text{ with } p_c, \\ &\qquad ..., \text{ is class } C \text{ with } p_C, \end{aligned} \tag{9}$$

Where p_c denotes the probability of belonging to category c. Obviously, our method is specifically designed for classification tasks, particularly suitable for diagnostic tasks. As illustrated in Fig. 4, we use an individual NFS to respectively process IRT visual and AC frequency domain features, producing two types of decisions (probability vectors). Consequently, we employ a soft voting mechanism, denoted as $f_G(\cdot)$, to aggregate the decision results from different modalities, which is formulated as follows:

$$\overline{p}_c = \alpha^{(I)} \times p_c^{(IRT)} + \alpha^{(A)} \times p_c^{(AC)} \tag{10}$$

Where $\overline{p} = (\overline{p}_1, ..., \overline{p}_C) \in \mathbb{R}^{1\times C}$. $\alpha^{(I)}$ and $\alpha^{(A)}$ are weight factors that determine the relative contribution of the corresponding modality to the final decision (default weight factor is set to 0.5).

C. Fuzzy Soft Triplet Loss

In this subsection, we introduce a novel Fuzzy Soft Triplet (FST) Loss (Fig. 5), which consists of two components: a Standard Optimization (SO) component and an Intra-class Re-optimization (IRO) component. The SO component of the FST loss leverages the category discriminative information provided by the fuzzy soft labels to simultaneously encourage intra-class compactness and inter-class separation. Additionally, the IRO component is specifically designed to optimize the intra-class variance induced by non-stable working conditions (NWCs). Given a triplet consisting of anchor, positive (intra-class), and negative (inter-class) samples [40], denoted as $\{f(x)^A, f(x)^+, f(x)^-\}$, the fuzzy soft labels for the anchor $f(x)^A$ are represented as $[s_1, s_2, ..., s_C] \in \mathbb{R}^{1\times C}$. $y(x)^+$ and $y(x)^-$ denote the ground-truth categories of the $f(x)^+$ and $f(x)^-$, respectively. We then construct a normalized probability vector P^{SO} using the fuzzy soft labels of $f(x)^A$, which serves as the ground-truth vector for the SO component, formulated as follows:

$$P^{SO} = \left[\frac{p_{y(x)^+}}{p_{y(x)^+} + p_{y(x)^-}}, \frac{p_{y(x)^-}}{p_{y(x)^+} + p_{y(x)^-}}\right] \tag{11}$$

Where, the probability $p_{y(x)^+}$ implies the probability that the anchor $f(x)^A$ belongs to the same category as $y(x)^+$, indicating intra-class similarity. Conversely, $p_{y(x)^-}$ represents the probability that $f(x)^A$ belongs to a different category $y(x)^-$, highlighting inter-class variability. Utilizing a metric learning mechanism, we construct another normalized probability vector $\widetilde{P}^{SO}$ based on the feature distance metric [41]. Let $d(f(x)^A, f(x)^+) = \| f(x)^A - f(x)^+ \|_2^2$ and $d(f(x)^A, f(x)^-) = \| f(x)^A - f(x)^- \|_2^2$, respectively. Thereby, $\widetilde{P}^{SO}$ can be given by:

$$\widetilde{P}^{SO} = \left[\frac{e^{-d(f(x)^A,f(x)^+)}}{e^{-d(f(x)^A,f(x)^+)} + e^{-d(f(x)^A,f(x)^-)}}, \frac{e^{-d(f(x)^A,f(x)^-)}}{e^{-d(f(x)^A,f(x)^+)} + e^{-d(f(x)^A,f(x)^-)}}\right] \tag{12}$$

Where $\widetilde{P}^{SO}$ represents the feature distances among anchor, positive, and negative instances. If $e^{-d(f(x)^A,f(x)^+)} > e^{-d(f(x)^A,f(x)^-)}$, this suggests that both intra-class similarity and inter-class discrepancy are effectively optimized. Since, $y(x)^+$ and $y(x)^-$ are negatively correlated with $d(f(x)^A, f(x)^+)$ and $d(f(x)^A, f(x)^-)$, respectively, it follows that the closer the feature distances between two instances, the higher the probability that they belong to the same category. Consequently, we employ cross-entropy to construct

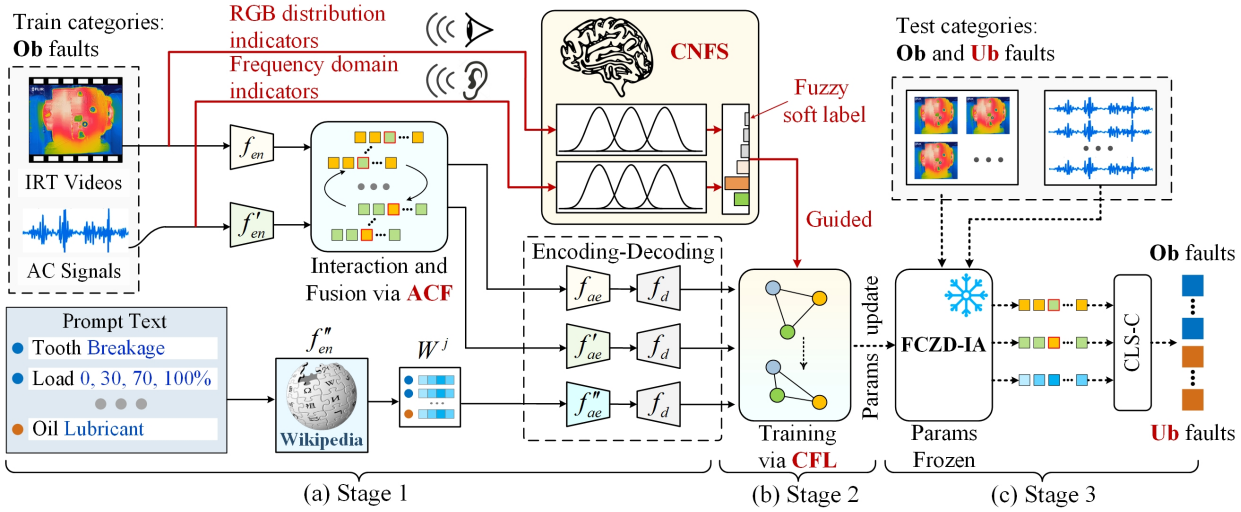

Fig. 6. Implementation Details of Our Proposed Method. The proposed method contains three stages: (a), (b), and (c). The data flow passing through the CNFS is represented by red arrow lines, while black arrow lines indicate the data flow in the deep backbone network. Dot arrow lines denote the data flow during the diagnostic process in the evaluation phase. CLS-C refers to a specially designed classifier with a calibration factor, as detailed in Eq (19).

the SO component, defined as follows:

$$\mathcal{L}^{(RO)}_{((f(x)^A,f(x)^+,f(x)^-)} = \frac{1}{2N_t}\sum_{i=1}^{N_t}\sum_{i=1}^{2} -P^{SO}[i]\cdot log(\widetilde{P}^{SO}[i]) \tag{13}$$

Where i denotes the index of a triplet, and N_t represents the number of triplets in a mini-batch. The additional IRO component further optimizes intra-class compactness to eliminate intra-class variance induced by Non-stationary Working Conditions (NWCs), and can be formulated as follows:

$$\mathcal{L}^{(IRO)}_{((f(x)^A,f(x)^+)} = \frac{1}{N_t}\sum_{i=1}^{N_t}\sum_{c=1}^{C} -y(x)^+_c\cdot log(p_c(f(x)^A)) \tag{14}$$

Summarily, the overall FST loss includes both the SO and IRO components, and can be given by:

$$\mathcal{L}_{FST} = \beta_t\cdot\mathcal{L}^{(RO)}_{((f(x)^A,f(x)^+,f(x)^-)} + (1-\beta_t)\cdot\mathcal{L}^{(IRO)}_{((f(x)^A,f(x)^+)} \tag{15}$$

Where t represents the iteration index. $\beta_t \in \{0,1\}$ is a learnable control term derived from a linear layer and a Sigmoid function, utilized to control the strength of the intra-class re-optimization in each iteration.

D. Overall Composite Neuro-Fuzzy System-Guided Cross-Modal Zero-Sample Diagnosis Framework

To harness the interpretability capabilities of the Neuro-Fuzzy System alongside the deep and high-level feature extraction capabilities of deep neural networks, we designed the FCZD-IA framework. FCZD-IA primarily features two components: the composite neuro-fuzzy system (CNFS) and the deep backbone network (indicated by black arrow lines in Fig. 6). Inspired by the knowledge distillation mechanism, FCZD-IA utilizes the CNFS as a decision-maker to guide the deep network's feature learning through fine-grained category knowledge. Specifically, our proposed method contains three stages, which are detailed below (the data flow of each stage is represented by differently colored arrow lines in Fig. 6).

Stage 1: Construction of the FCZD-IA network structure. As depicted in Fig. 6, raw IRT and acoustic data are initially summarized into multiple key RGB distribution indicators ($\mathbb{R}^{1\times5}$) and frequency domain indicators ($\mathbb{R}^{1\times5}$), respectively, before being fed into the CNFS to infer the fuzzy soft labels (as detailed in Section II). In the deep backbone network, pre-trained networks and subsequent linear encoders $f_{en}(\cdot)$ and $f^{'}_{en}(\cdot)$ are utilized to map the input IRT visual and acoustic instances into 1-D feature vectors with the same dimension. These vectors are then fed into attention-based modality fusion (ACF) modules to facilitate cross-modal information interaction and fusion. To prevent distortion of the original prompt information, the prompt text is not directly fed into the ACF. Instead, a Word2Vec model (pre-trained with Wikipedia data), extracts the prompt text into a prompt matrix $W^j \in \mathbb{R}^{N\times C}, j = 1,\ldots,C$. The features obtained are subsequently processed through encoding-decoding processes (denoted as $f_{ae}(\cdot)$, $f^{'}_{ae}(\cdot)$, $f^{''}_{ae}(\cdot)$ and $f_d(\cdot)$ in Fig. 6), resulting in process-encoded features ($\widetilde{O}_i \in \mathbb{R}^{1\times N}$ and $\widetilde{W}^j \in \mathbb{R}^{N\times C}$) and their corresponding decoded features ($\widehat{O}_i$ and $\widehat{W}^j \in \mathbb{R}^{N\times C}$).

Stage 2: Training the FCZD-IA network via Cross-Modal Fuzzy Learning (CFL) strategy. Although the encoded and decoded features retain essential status information from multiple modalities in the first stage, their representations remain abstract and challenging to interpret, complicating the task of reliable zero-sample diagnosis task. Furthermore, semantic ambiguities caused by differences in modality and load variance limit the network's ability to learn discriminative information reflective of the defect category, thereby impacting diagnostic accuracy. To address these challenges, we have specifically designed a CFL strategy that uses the fuzzy soft labels generated by the CNFS as prior knowledge to guide the feature processing procedure of the deep backbone network

to facilitate interpreted fusion and discriminative learning. Specifically, we first utilize the proposed Fuzzy Soft Triplet (FST) loss to construct the discriminative learning term $\mathcal{L}_{DL}$, formulated as follows:

$$\begin{aligned}\mathcal{L}_{DL} = \frac{1}{N_t}\sum_{j=1}^{N_t}(&\mathcal{L}_{FST}[\widetilde{O}_v^{j+}, \widetilde{W}^{j+}, \widetilde{O}_v^{j-}] \\ &+\mathcal{L}_{FST}[\widetilde{O}_a^{j+}, \widetilde{W}^{j+}, \widetilde{O}_a^{j-}] \\ &+\mathcal{L}_{FST}[\widetilde{W}^{j+}, \widetilde{O}_v^{j+}, \widetilde{W}^{j-}] \\ &+\mathcal{L}_{FST}[\widetilde{W}^{j+}, \widetilde{O}_a^{j+}, \widetilde{W}^{j-}])\end{aligned} \tag{16}$$

where $\widetilde{O}_i^{j+}, \widetilde{O}_i^{j-}, i \in \{v, a\}$ correspond to the positive and negative samples for the j-th mini-batch. Subsequently, drawing inspiration from [23], we employ the mean squared error loss $\mathcal{L}_{MSE}(\cdot)$ to construct the semantic alignment term, denoted as $\mathcal{L}_{Ali}$. Promoting semantic consistency among multimodal features $\mathcal{L}_{Ali}$ aims to help the network explore the commonality of the different modalities, and thereby boost the robustness of the diagnostic performance, can be formulated as:

$$\begin{aligned}\mathcal{L}_{Ali} = \frac{1}{N_t}\sum_{j=1}^{N_t}(&\mathcal{L}_{MSE}[\widehat{O}_v^j, \widehat{W}^j] \\ &+\mathcal{L}_{MSE}[\widehat{O}_a^j, \widehat{W}^j] + \mathcal{L}_{MSE}[\widehat{W}^j, W^j])\end{aligned} \tag{17}$$

where $\mathcal{L}_{MSE}[\widehat{W}^j, W^j]$ ensures the semantic consistency for the prompt features before and after decoding. Summarily, the overall CFL training strategy can be formulated as:

$$\mathcal{L}_{CFL} = \mathcal{L}_{DL} + \alpha_{Ali} \cdot \mathcal{L}_{Ali} \tag{18}$$

where α_{Ali} is a hyper-parameter that governs the intensity of semantic alignment during training and is manually calibrated in our work with a default weight of 0.1. Overemphasizing semantic consistency may hinder the network's ability to explore discrepancies among modalities.

Stage 3: Zero-sample diagnosis during the evaluation phase. Considering that our training phase solely employs the observed fault samples, the classification results may exhibit significant bias towards the observed classes. To tackle this issue, we incorporate the calibration strategy into the fault classifier to mitigate this bias. Specifically, this classifier (denoted as CLS-C in Fig. 6) can be defined as follows:

$$\begin{aligned}\widetilde{y} = \underset{\widetilde{W}^j \in W^{(U)} \bigcup W^{(O)}}{argmax} &(\sigma \cdot CD((\widetilde{Y}_i|\widetilde{\Theta}_f)^T, \widetilde{W}^j) \\ &-\widehat{\gamma} \cdot \mathbb{I} \cdot [\widetilde{W}^j \in W^{(O)}])\end{aligned} \tag{19}$$

Where $\widehat{\gamma}$ is a calibration factor, and $CD(\cdot)$ denotes cosine distance. $\widetilde{y}$ represents the predicted class for the test samples $Y_i, i \in \{v, a\}$. $\widetilde{Y}_i$ refers to the encoded feature vectors obtained via the FCZD-IA network, whose parameters, denoted as $\widetilde{\Theta}_f$, are frozen during the evaluation phase. $\widetilde{W}^j$ indicates the learned category prompt matrix, which includes prompts or both observed and unobserved categories $W^{(O)}$ and $W^{(U)}$, respectively. If $\widetilde{y}$ belongs to an observed class, the indicator $\mathbb{I}$ is set to 1, and 0 otherwise.

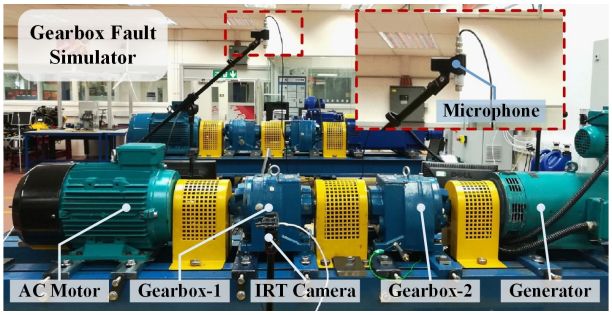

Fig. 7. Gearbox Fault Simulator in this Experiment.

III. Experimental Validation

A. Experimental settings

The FCZD-IA code is implemented using the Python 3.8.16, and Pytorch 2.0.1 environment, and all experiments are conducted on a workstation with the Ubuntu 20.04 operating system with a GTX3060Ti GPU. FCZD-IA is compared with eight state-of-the-art zero-sample models, namely: 1) **GEMZSL** [25]. 2) **CJMEZSL** [42]. 3) **AVCAZSL** [43]. 4) **VAEGAN** [44]. 5) **TF-VAEGAN** [44]. 6) **Bi-VAEGAN** [24]. In the evaluation phase, we assess zero-sample diagnosis performance utilizing a generalized zero-sample diagnosis score, i.e., HM score [45]. The experimental gearbox dataset was collected from the gearbox fault simulator (Fig. 7). Infrared thermal (IRT) videos were captured by using an IRT camera, with each video having a duration of 1 second and a frame rate of 5 fps. Acoustic data were recorded synchronously with the IRT videos utilizing microphone equipment with a sampling frequency of 12.8 kHz.

In the experiment, we employed six gearbox states. *Baseline* represents an undamaged, healthy condition. *TB50* and *TB100* denote 50% and 100% tooth breakage on the driving gear. *OS1500* and *OS2000* indicate lubricating oil shortages of 1100 mL and 600 mL, respectively, in terms of the baseline value of 2600 mL. *VIS100* denotes the oil viscosity of EP100, in contrast to the baseline oil viscosity of EP320. We set four load conditions (0%, 30%, 70%, and 100%) for each state. Consequently, there are a total of 24 unique working states. Each working state is involved with the collection of 240 IRT video samples and the corresponding 240 acoustic samples. As a result, the experimental data to be used have a total of $(240 + 240) \times 24 = 11520$ samples. As shown in Table 3, we establish three zero-sample diagnosis scenarios to validate our proposed method. The motivation for setting these scenarios is to employ baseline samples and less severe fault samples (i.e., TB50 and OS2000) for diagnosing challenging-to-collect, more critical faults (i.e., TB100, OS1500, and VIS100).

B. Comparison with the other state-of-the-art methods

This section demonstrates a performance comparison between the FCZD-IA framework and other leading approaches. To ensure the reliability of the experimental results, each model is implemented five times, with the outcomes presented in Table 4. The proposed method demonstrates substantial improvements in zero-sample diagnosis performance,

TABLE III
DETAILS OF THE EXPERIMENTAL SCENARIOS.

Scenarios	Observed Faults (Ob)	Unobserved Faults (Ub)
Scenario-I	TB50, OS2000, VIS100.	TB100, OS1500.
Scenario-II	TB50, OS2000, OS1500.	TB100, VIS100.
Scenario-III	TB50, OS2000.	TB100, OS1500, VIS100.

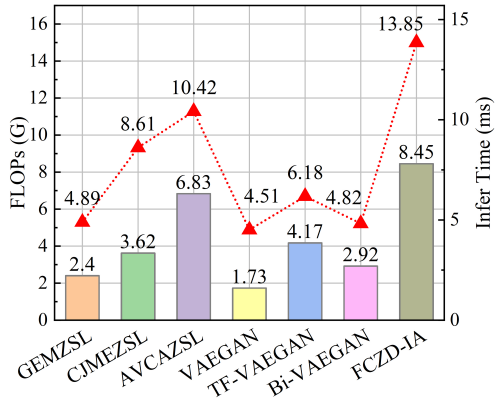

Fig. 8. Time Efficiency Analysis. The bar plot (left y-axis) displays the Floating Point Operations Per Second (FLOPs) for various comparison methods, while the dot plot (right y-axis) shows the inference time for a batch size of data across these methods.

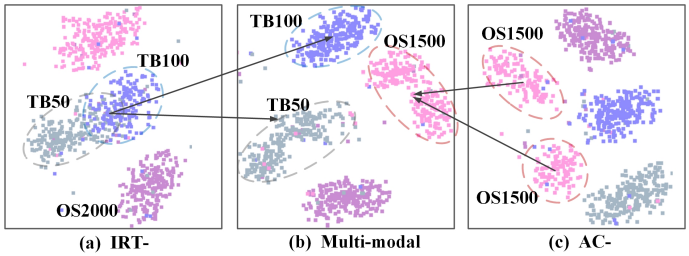

Fig. 9. Comparison of the learned features from uni-modal and multi-modal fashions. The arrow indicates the improvement direction of feature aggregation.

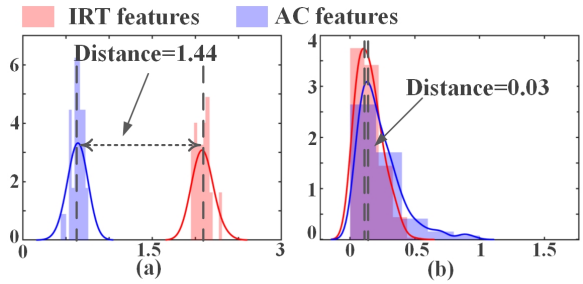

Fig. 10. Semantic Alignment of IRT and AC Modalities (a) Before and (b) After Training.

as measured by Harmonic Mean (HM) scores. Specifically, it enhances diagnostic accuracy in Scenario 1 from 6.79% to 21.74%. In Scenario 2, our method boosts diagnostic accuracy from 12.31% to 27.61%. In Scenario 3, improvements range from 4.83% to 21.93% compared to competing methods. Experimental results underscore the robust zero-sample diagnostic capabilities of the proposed method. Specifically, FCZD-IA surpasses VAEGAN, TF-VAEGAN, and Bi-VAEGAN, indicating that the proposed method can construct a more comprehensive fault representation than unimodal approaches. Notably, the proposed method does not require manual annotation of collected visual samples like TF-VAEGAN, Bi-VAEGAN, and GEMZSL rely on predefined detailed fault attributes [6] [7] [8]. Consequently, their performance may significantly degrade when the attribute information is imprecise or unavailable for unknown faults.

We further analyze the time efficiency of various methods. Fig. 8 indicates that the time cost of our method significantly exceeds that of unimodal methods (GEMZSL, VAEGAN, TF-VAEGAN, and Bi-VAEGAN) with inference times being 8.96ms, 9.34ms, 7.67ms, and 9.03ms longer, respectively. This is primarily due to our method's requirement to process multiple modalities simultaneously, which leads to greater computational calculations as indicated by the corresponding FLOPs. Furthermore, our method is also more time-consuming compared to multimodal methods (CJMEZSL and AVCAZSL), with inference times longer by 5.24ms and 3.43ms. Still, we consider our method effective as it provides necessary interpretability and surpasses other competitive methods in diagnostic accuracy.

IV. VALIDATION OF THE IMPROVEMENTS

A. Validation of the cross-modal feature fusion mechanism

We further validate the impact of training the proposed network using only a single modality (IRT or acoustic). As indicated in Table 5, the multi-modal approach yields advantages in each scenario. Additionally, we employ a T-SNE scatter plot to intuitively demonstrate the discriminative degree of the learned features. According to the cluster hypothesis, instances within the same cluster exhibit a higher degree of common characteristics. Notably, as shown in Fig. 9, the OS1500 samples are mistakenly divided into distinct clusters (defect categories). This division primarily occurs because the oil shortage lacks the apparent acoustic characteristics, making it difficult to extract its intra-class commonalities using solely the acoustic modality. Similarly, instances of different tooth breakage levels (TB50 and TB100) overlap within the same cluster. This is primarily due to the lack of discriminative IRT visual characteristics, making it challenging to achieve inter-class separation using only the IRT modality.

In comparison, as depicted in Fig. 10, guided by the fuzzy soft labels, the semantic distances between the IRT and AC modalities are more closely aligned, demonstrating that their fault-related commonalities are refined by our method. Moreover, the NFS contributes to constructing more discriminative feature representations by guiding the learning process in the CFZD strategy. As illustrated in Figs. 9c and 9b, what were previously incorrectly separated intra-class clusters are now aggregated into more compact groups. Furthermore, the overlapping inter-class clusters, as shown in Fig. 9a, are effectively segregated and categorized into their respective categories, as demonstrated in Fig. 9b.

TABLE IV
DIAGNOSTIC RESULTS OF THE COMPARISON APPROACHES [%]. *U* SIGNIFIES THE ACCURACY OF DIAGNOSING UNOBSERVED FAULTS, *O* REPRESENTS THE ACCURACY OF DIAGNOSING OBSERVED FAULTS, *HM* DENOTES THE HM SCORES ($HM = 2(O \times U)/(O + U)$).BOLD FONT INDICATES THE OPTIMAL RESULT.

Methods	Scenario-I			Scenario-II			Scenario-III		
	O	U	HM	O	U	HM	O	U	HM
GEMZSL	81.91 ± 3.63	63.53 ± 11.39	69.67 ± 6.89	74.71 ± 5.64	58.15 ± 3.66	65.39 ± 4.48	49.91 ± 4.86	**65.27 ± 4.14**	56.65 ± 2.02
CJMEZSL	74.65 ± 4.46	55.32 ± 1.76	63.53 ± 2.78	63.07 ± 5.21	64.97 ± 9.30	62.21 ± 3.04	51.60 ± 2.67	45.41 ± 5.69	47.34 ± 2.66
AVCAZSL	82.16 ± 2.68	70.75 ± 1.69	76.84 ± 1.31	68.54 ± 9.37	72.79 ± 5.98	69.60 ± 2.32	71.19 ± 3.93	48.67 ± 5.12	57.78 ± 4.90
VAEGAN	69.71 ± 9.82	61.39 ± 1.75	64.81 ± 3.34	70.70 ± 3.63	44.32 ± 7.50	54.30 ± 6.76	33.75 ± 12.02	62.44 ± 7.27	40.68 ± 9.35
TF-VAEGAN	93.45 ± 5.16	48.77 ± 10.21	63.08 ± 7.64	**94.86 ± 2.01**	39.72 ± 4.07	55.63 ± 3.87	76.04 ± 5.35	28.82 ± 2.65	41.47 ± 3.28
Bi-VAEGAN	**94.12 ± 3.63**	66.09 ± 4.02	78.03 ± 2.11	90.59 ± 5.42	56.49 ± 6.82	69.04 ± 3.58	**77.71 ± 2.23**	41.31 ± 4.53	53.83 ± 3.45
FCZD-IA	82.51 ± 0.79	**88.04 ± 2.73**	**84.82 ± 1.17**	92.09 ± 1.81	**74.37 ± 2.29**	**81.91 ± 1.76**	74.40 ± 5.03	53.32 ± 1.78	**62.61 ± 2.45**

TABLE V
EFFECT OF TRAINING FCZD-IA WITH VARIED MODALITIES [%]. *IRT-* INDICATES THAT THE PROPOSED MODEL WAS TRAINED EXCLUSIVELY USING IRT VISUAL DATA. *AC-* SIGNIFIES THAT THE PROPOSED MODEL WAS TRAINED SOLELY USING ACOUSTIC SIGNALS. ↓ DENOTES THAT THE VARIANT'S ACCURACY IS LOWER THAN THAT OF OUR METHOD.

Methods	Scenario-I	Scenario-II	Scenario-III
IRT-	77.42 (7.39 ↓)	67.29 (4.36 ↓)	52.83 (9.78 ↓)
AC-	57.81 (27.01 ↓)	36.16 (47.75 ↓)	37.72 (24.91 ↓)

TABLE VI
EFFECT OF THE COMPOSITE NEURO-FUZZY SYSTEM (CNFS) [%]. *CNFS(SL)* DENOTES THE APPLICATION OF THE CNFS TO SUPERVISE THE DIAGNOSIS OF THE DEFECTS IN EACH SCENARIO. FOR EXAMPLE, **IRT:93.82/AC:83.30** INDICATES A SUPERVISED DIAGNOSIS ACCURACY OF 93.82% FOR IRT SAMPLES AND 83.30% FOR AC SAMPLES. FCZD-NF REPRESENTS A VARIANT WITHOUT CNFS.

Methods	Scenario-I	Scenario-II	Scenario-III
CNFS(SL)	IRT:93.82/AC:83.30	IRT:95.55/AC:81.14	IRT:89.07/AC:74.13
FCZD-NF	80.99 (3.83 ↓)	74.82 (7.09 ↓)	61.54 (1.07 ↓)

B. Interpretation of the proposed neuro-fuzzy system

In this section, we validate the interpretability of the FCZD-IA, demonstrating how the proposed CNFS guides the process of cross-modal zero-sample diagnosis. As indicated in Table 6, we begin by evaluating the impact of the CNFS. Our method surpasses the variant without CNFS (FCZD-NF) by 1.07% to 7.09% in various scenarios. Notably, by solely using the CNFS to supervise diagnose defect categories, we achieve an accuracy ranging from 89.07% to 95.55% for IRT modality, and from 74.13% to 83.30% for acoustic modality, highlighting the effectiveness of the selected frequency domain and visual indicators. This also confirms the CNFS's capability to provide interpretive guidance for the zero-sample diagnostic task. The training process of the proposed composite neuro-fuzzy system is depicted in Fig. 12, which also demonstrates its convergence.

As depicted in Table 5, the diagnostic accuracy of training the network solely with the IRT modality outperforms that of using the acoustic modality alone, indicating that the IRT modality contributes more significantly in each diagnostic scenario. Fig. 11 demonstrates that within the CNFS, the average rule strength for the IRT modality is higher than that for the AC modality. This suggests that the CNFS primarily uses knowledge from IRT visual features to guide the network's learning, aligning with the IRT modality's higher contribution and thereby highlighting the interpretability of our method. Moreover, in each scenario, the IRT modality achieves convergence more rapidly than the acoustic modality (Fig. 12), which suggests that the IRT features are more readily interpreted by the CNFS, thereby prompting more robust decision-making.

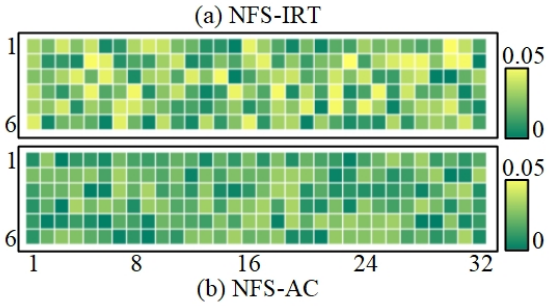

Fig. 11. Activation Strength of the Rules in CNFS. (a) Firing Strength of the NFS Processing the IRT Modality in CNFS. (b) Firing Strength of the NFS Processing the AC Modality. The x-axis represents the indices of the rules, and the y-axis represents the indices of the base learner in each NFS. The brightness of each block showcases the average normalized firing strengths of each rule in NFS. A brighter block indicates higher firing strength.

C. Validation of the Fuzzy Soft Triplet Loss Function

As depicted in Table 7, Triplet(BH)-NSO refers to training the network using basic triplet loss with a batch-hard sampling strategy, replacing the SO component in FST loss. Triplet(BA)-NSO involves training with basic triplet loss with a batch-all sampling strategy. Experimental results indicate that the proposed Fuzzy Soft Triplet (FST) loss outperforms the basic triplet loss with Batch-All (BA) or Batch-Hard (BH) strategies by 3.34% to 11.18% in various scenarios. Moreover, compared to the FST variant (FST-NIRO) that does not consider Intra-class Re-optimization (IRO), our methods show an improve-

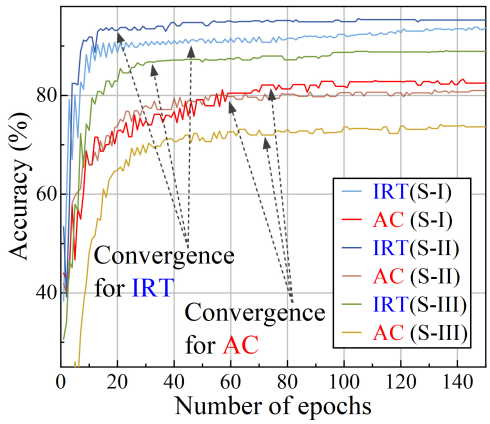

Fig. 12. Validation of CNFS Convergence across Scenarios. S-I, S-II, and S-III represent Scenario-I, Scenario-II, and Scenario-III, respectively.

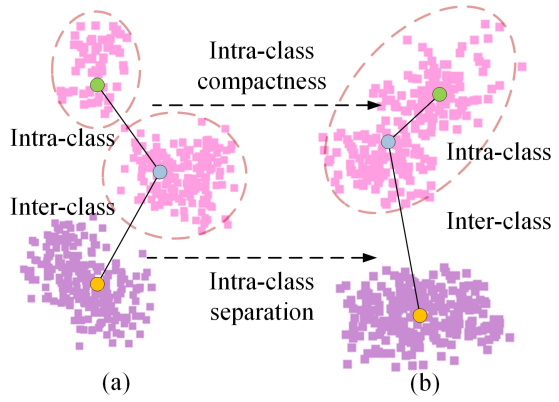

Fig. 13. Visualization of Feature Discriminability via the Fuzzy Soft Triplet Loss. (a) Features learned via the Triplet(BH)-NSO loss (a), and the proposed FST loss (b).

TABLE VII
EFFECT OF TRAINING FCZD-IA WITH VARIED TRIPLET LOSS [%].

Methods	Scenario-I	Scenario-II	Scenario-III
Triplet(BH)-NSO	81.47 (3.34 ↓)	69.18 (9.94 ↓)	59.04 (3.57 ↓)
Triplet(BA)-NSO	73.64 (11.18 ↓)	73.39 (8.52 ↓)	52.63 (9.98 ↓)
FST-NIRO	83.48 (1.33 ↓)	79.13 (2.78 ↓)	60.88 (1.73 ↓)

ment in accuracy from 1.33% to 2.78%, with the maximum improvement observed in Scenario 3. This underscores the importance of optimizing intra-class variance in gearbox diagnostic tasks. By utilizing the generated fuzzy soft labels to guide the feature learning process, the FST loss enables the network to effectively realize discriminative learning, which is crucial for achieving robust diagnostic performance under Non-stationary working conditions, as discussed in Section 1. This is evident in Fig. 13, where features learned by the FST exhibit better intra-class compactness and inter-class separation compared to those learned by the basic triplet loss. Notably, loosely associated intra-class clusters in Fig. 13a are accurately aggregated in Fig. 13b, visually demonstrating the effectiveness of optimizing intra-class variance.

V. CONCLUSION

In this study, we developed an end-to-end diagnostic framework, termed FCZD-IA, carefully designed to achieve accurate diagnostic results in zero-sample scenarios. To validate the effectiveness of this proposed method, experiments were conducted using a gearbox dataset derived from a gearbox fault simulator. The experimental results demonstrate that the FCZD-IA achieved diagnostic accuracy of 84.82%, 81.92%, and 62.61% across various zero-sample diagnostic scenarios. Experimental findings indicate that FCZD-IA surpasses existing state-of-the-art methods, highlighting the efficacy of the developed human-like composite neuro-fuzzy system and the overall cross-modal zero-sample diagnosis framework.

The developed diagnostic framework has showcased significant potential for zero-sample diagnosis tasks, primarily due to its innovative integration strategy that combines the neuro-fuzzy system with the deep neural network. Within our proposed framework, the composite neuro-fuzzy system is designed to avoid the direct processing of high-dimensional IRT videos and acoustic signals. Instead, it uses carefully selected key indicators to guide the deep neural network in processing these data, emphasizing the enhancement of discriminative learning and modality fusion during the training phase. This innovative approach facilitates effective and interpretable zero-sample diagnoses by harnessing the synergistic benefits of both neural-fuzzy systems and deep network architectures. Despite these advancements, the framework presents areas for further refinement. The modelâs reliance on numerous hyper-parameters indicates a necessity for the development of specialized optimization algorithms that can adaptively refine the network structure. Furthermore, there is potential for improvement in selecting the RGB distribution indicators for IRT modalities and the frequency domain indicators for acoustic modalities. Future research should consider developing an indicator set and employing a specialized algorithm to adaptively select the optimal key indicators. By effectively addressing these limitations, significant enhancements can be achieved in terms of both the robustness and accuracy of neuro-fuzzy hybrid models.

REFERENCES

[1] D. Sun, Y. Li, S. Jia, K. Feng, and Z. Liu, "Non-contact diagnosis for gearbox based on the fusion of multi-sensor heterogeneous data," *Information Fusion*, vol. 94, pp. 112–125, 2023.

[2] Y. Xu, X. Yan, B. Sun, and Z. Liu, "Hierarchical multiscale dense networks for intelligent fault diagnosis of electromechanical systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[3] Y. Hou, J. Wang, Z. Chen, J. Ma, and T. Li, "Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved transformer," *Engineering Applications of Artificial Intelligence*, vol. 124, p. 106507, 2023.

[4] Z. Hu, H. Zhao, L. Yao, and J. Peng, "Semantic-consistent embedding for zero-shot fault diagnosis," *IEEE Transactions on Industrial Informatics*, 2022.

[5] L. Feng and C. Zhao, "Fault description based attribute transfer for zero-sample industrial fault diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 1852–1862, 2020.

[6] X. Chen, B. Zhang, C. Zhao, J. Ding, and W. Wang, "From coarse to fine: Hierarchical zero-shot fault diagnosis with multi-grained attributes," *IEEE Transactions on Fuzzy Systems*, 2024.

[7] Z. Li, K. Liu, M. Lin, D. Xin, H. Tang, and G. Wu, "A zero-sample state evaluation model for valve-side bushing of uhv converter transformer oriented to digital twin under attribute analysis," *IET Generation, Transmission & Distribution*, vol. 17, no. 5, pp. 1123–1134, 2023.

[8] R. Zhang, X. Bai, L. Pan, Z. Dong, and R. Song, "Zero-small sample classification method with model structure self-optimization and its application in capability evaluation," *Applied Intelligence*, vol. 52, no. 5, pp. 5696–5717, 2022.

[9] Y. Xu, K. Feng, X. Yan, X. Sheng, B. Sun, Z. Liu, and R. Yan, "Cross-modal fusion convolutional neural networks with online soft label training strategy for mechanical fault diagnosis," *IEEE Transactions on Industrial Informatics*, 2023.

[10] S. Zhang, Z. Sun, M. Wang, J. Long, Y. Bai, and C. Li, "Deep fuzzy echo state networks for machinery fault diagnosis," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 7, pp. 1205–1218, 2019.

[11] X. Xu, D. Cao, Y. Zhou, and J. Gao, "Application of neural network algorithm in fault diagnosis of mechanical intelligence," *Mechanical Systems and Signal Processing*, vol. 141, p. 106625, 2020.

[12] E.-J. Pérez-Pérez, F.-R. López-Estrada, V. Puig, G. Valencia-Palomo, and I. Santos-Ruiz, "Fault diagnosis in wind turbines based on anfis and takagi–sugeno interval observers," *Expert systems with applications*, vol. 206, p. 117698, 2022.

[13] H. Xue, D. Ding, Z. Zhang, M. Wu, and H. Wang, "A fuzzy system of operation safety assessment using multimodel linkage and multistage collaboration for in-wheel motor," *IEEE Transactions on Fuzzy Systems*, vol. 30, DOI 10.1109/TFUZZ.2021.3052092, no. 4, pp. 999–1013, 2022.

[14] H.-J. Rong, P. P. Angelov, X. Gu, and J.-M. Bai, "Stability of evolving fuzzy systems based on data clouds," *IEEE Transactions on Fuzzy Systems*, vol. 26, DOI 10.1109/TFUZZ.2018.2793258, no. 5, pp. 2774–2784, 2018.

[15] F. Xiao, Z. Cao, and A. Jolfaei, "A novel conflict measurement in decision-making and its application in fault diagnosis," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 1, pp. 186–197, 2020.

[16] Y. Zheng, Z. Xu, and X. Wang, "The fusion of deep learning and fuzzy systems: A state-of-the-art survey," *IEEE Transactions on Fuzzy Systems*, vol. 30, DOI 10.1109/TFUZZ.2021.3062899, no. 8, pp. 2783–2799, 2022.

[17] P. Wan, D. Sun, M. Zhao, and S. Huang, "Multistability for almost-periodic solutions of takagiâsugeno fuzzy neural networks with nonmonotonic discontinuous activation functions and time-varying delays," *IEEE Transactions on Fuzzy Systems*, vol. 29, DOI 10.1109/TFUZZ.2019.2955886, no. 2, pp. 400–414, 2021.

[18] M. Ali, M. Adnan, M. Tariq, and H. V. Poor, "Load forecasting through estimated parametrized based fuzzy inference system in smart grids," *IEEE Transactions on Fuzzy Systems*, vol. 29, DOI 10.1109/TFUZZ.2020.2986982, no. 1, pp. 156–165, 2021.

[19] H. H. Y. Saâad, N. A. M. Isa, and M. M. Ahmed, "A structural evolving approach for fuzzy systems," *IEEE Transactions on Fuzzy Systems*, vol. 28, DOI 10.1109/TFUZZ.2019.2904928, no. 2, pp. 273–287, 2020.

[20] D. Wang, J. Huang, and Y. Xu, "Matrix representations of the inverse problem in the graph model for conflict resolution with fuzzy preference," *Applied Soft Computing*, vol. 147, p. 110786, 2023.

[21] A. Chibani, M. Chadli, P. Shi, and N. B. Braiek, "Fuzzy fault detection filter design for tâs fuzzy systems in the finite-frequency domain," *IEEE Transactions on Fuzzy Systems*, vol. 25, DOI 10.1109/TFUZZ.2016.2593921, no. 5, pp. 1051–1061, 2017.

[22] M. Wang, G. Feng, J. Qiu, H. Yan, and H. Zhang, "Fault detection filtering design for discrete-time interval type-2 tâs fuzzy systems in finite frequency domain," *IEEE Transactions on Fuzzy Systems*, vol. 29, DOI 10.1109/TFUZZ.2020.3006576, no. 2, pp. 213–225, 2021.

[23] Y. Siddiqui, J. Thies, F. Ma, Q. Shan, M. Nießner, and A. Dai, "Retrievalfuse: Neural 3d scene reconstruction with a database," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12 568–12 577, 2021.

[24] Z. Wang, Y. Hao, T. Mu, O. Li, S. Wang, and X. He, "Bi-directional distribution alignment for transductive zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19 893–19 902, 2023.

[25] Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, and T. Harada, "Goal-oriented gaze estimation for zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3794–3803, 2021.

[26] D. Wang, Y. Li, L. Jia, Y. Song, and Y. Liu, "Novel three-stage feature fusion method of multimodal data for bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021.

[27] Z. Chen, J. Wu, C. Deng, X. Wang, and Y. Wang, "Deep attention relation network: A zero-shot learning method for bearing fault diagnosis under unknown domains," *IEEE Transactions on Reliability*, vol. 72, no. 1, pp. 79–89, 2022.

[28] S. Xing, Y. Lei, S. Wang, N. Lu, and N. Li, "A label description space embedded model for zero-shot intelligent diagnosis of mechanical compound faults," *Mechanical Systems and Signal Processing*, vol. 162, p. 108036, 2022.

[29] J. Xu, H. Zhang, L. Zhou, and Y. Fan, "Zero-shot compound fault diagnosis method based on semantic learning and discriminative features," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[30] H. Lv, J. Chen, T. Pan, and Z. Zhou, "Hybrid attribute conditional adversarial denoising autoencoder for zero-shot classification of mechanical intelligent fault diagnosis," *Applied Soft Computing*, vol. 95, p. 106577, 2020.

[31] J. Xu, L. Zhou, W. Zhao, Y. Fan, X. Ding, and X. Yuan, "Zero-shot learning for compound fault diagnosis of bearings," *Expert Systems with Applications*, vol. 190, p. 116197, 2022.

[32] H. Wang, Y. Kang, L. Yao, H. Wang, and Z. Gao, "Fault diagnosis and fault tolerant control for t–s fuzzy stochastic distribution systems subject to sensor and actuator faults," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 11, pp. 3561–3569, 2020.

[33] S. U. Jan, Y. D. Lee, and I. S. Koo, "A distributed sensor-fault detection and diagnosis framework using machine learning," *Information Sciences*, vol. 547, pp. 777–796, 2021.

[34] S. G. Kumbhar *et al.*, "An integrated approach of adaptive neuro-fuzzy inference system and dimension theory for diagnosis of rolling element bearing," *Measurement*, vol. 166, p. 108266, 2020.

[35] M. Yeganejou, S. Dick, and J. Miller, "Interpretable deep convolutional fuzzy classifier," *IEEE Transactions on Fuzzy Systems*, vol. 28, DOI 10.1109/TFUZZ.2019.2946520, no. 7, pp. 1407–1419, 2020.

[36] H. Yang, Y. Fu, and D. Wang, "Multi-anfis model based synchronous tracking control of high-speed electric multiple unit," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 3, pp. 1472–1484, 2017.

[37] M. Mehrasa, M. Babaie, A. Zafari, and K. Al-Haddad, "Passivity anfis-based control for an intelligent compact multilevel converter," *IEEE Transactions on Industrial Informatics*, vol. 17, DOI 10.1109/TII.2021.3049313, no. 8, pp. 5141–5151, 2021.

[38] Y. Cui, Y. Xu, R. Peng, and D. Wu, "Layer normalization for tsk fuzzy system optimization in regression problems," *IEEE Transactions on Fuzzy Systems*, vol. 31, DOI 10.1109/TFUZZ.2022.3185464, no. 1, pp. 254–264, 2023.

[39] G. Heydari, A. Gharaveisi, and M. Vali, "New formulation for representing higher order tsk fuzzy systems," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 4, pp. 854–864, 2015.

[40] W. Xie, H. Wu, Y. Tian, M. Bai, and L. Shen, "Triplet loss with multistage outlier suppression and class-pair margins for facial expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 690–703, 2021.

[41] L. Wang, Y. Liu, H. Di, C. Qin, G. Sun, and Y. Fu, "Semi-supervised dual relation learning for multi-label classification," *IEEE Transactions on Image Processing*, vol. 30, DOI 10.1109/TIP.2021.3122003, pp. 9125–9135, 2021.

[42] K. Parida, N. Matiyali, T. Guha, and G. Sharma, "Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3251–3260, 2020.

[43] P. Mazumder, P. Singh, K. K. Parida, and V. P. Namboodiri, "Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3090–3099, 2021.

[44] S. Narayan, A. Gupta, F. S. Khan, C. G. Snoek, and L. Shao, "Latent embedding feedback and discriminative features for zero-shot classification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pp. 479–495. Springer, 2020.

[45] M. Mancini, M. F. Naeem, Y. Xian, and Z. Akata, "Open world compositional zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5222–5230, 2021.