Article:
Yin, Y., Huang, D., Qin, N. et al. (2025) A Stacked Generalized Zero-Shot Learning Framework for Fault Diagnosis of High-Speed Train Bogies With Data Imbalance. IEEE Transactions on Instrumentation and Measurement, 74. 3536714. ISSN: 0018-9456

https://doi.org/10.1109/tim.2025.3565027

A Stacked Generalized Zero-Shot Learning Framework for Fault Diagnosis of High-Speed Train Bogies with Data Imbalance

Yirui Yin, Deqing Huang, *Senior Member*, *IEEE*, Na Qin, *Member*, *IEEE*, Kang Li, *Senior Member*, *IEEE*, Tianwei Wang, and Ronghua Zong

***Abstract*—In fault diagnosis for high-speed trains (HSTs), data-driven methods are becoming increasingly popular. However, diagnosing unknown faults remains challenging due to limitations of existing GZSL methods, such as prediction confidence shifts from data imbalance, and domain shift between known and unknown classes, making it an open problem in engineering applications. This paper addresses this issue by proposing a stacked generalized zero-shot learning (Stacked-GZSL) framework which combines attribute-based GZSL strategies and generative model-based GZSL strategies. The former leverages the correlations among data attributes and categories to tackle feature distribution differences across classes, while the latter addresses weight drift issue caused by the imbalance within known data and between known and unknown classes. Then, stacked model techniques are employed to resolve compatibility issues and further enhance accuracy. Experimental results confirm that the proposed framework can effectively address the challenge of unknown compound faults in train bogies under data imbalance. The framework achieves over 90% accuracy for known faults and 70% for unknown faults, outperforming single GZSL strategies and other classic ensemble models.**

***Index Terms*—Bogie fault diagnosis, generalized zero-shot learning, data imbalance, unknown compound fault, deep learning**

NOMENCLATURE

AD	Anti-yaw damper.
AS	Air spring.
BN	Batch normalization.
CN60	China national 60kg/m tracks.
Conv1d	A one-dimensional convolution layer.
D	Discriminator.
F	Filters.
G	Generator.
GAN	Generative adversarial network.
GZSL	Generalized zero-shot learning.
HST	High-speed train.
K	Kernel size.
LD	Lateral damper.
MD	Mahalanobis distance.
ResBlock	Residual blocks in the residual network.
S	Stride.
VAE	Variational autoencoder.
WGAN	Wasserstein generative adversarial network.
WGAN-GP	Wasserstein generative adversarial network with gradient penalty.
ZSL	Zero-shot learning.
C	Common set of attribute labels.
$c(y)$	Attribute label combination for a fault class.
$\mathcal{L}$	Loss.
S	Total set of visible fault class samples.
x	Visible class fault sample.
X	Set of visible class fault samples.
$\tilde{x}$	Generated fault sample.
$\hat{x}$	Interpolated data.
y	Fault category.
Y^S	Set of visible class fault categories.
z	Random Gaussian noise.
Z	Set of random Gaussian noise.

This work was supported by Natural Science Foundation of China (62173279, U1934221), Sichuan Science and Technology Program under Grant (2022YFG0247, 2021JDJQ0012) and Central University Basic Research Fund of China (2682021ZTPY027). Corresponding author: Na Qin, qinna@swjtu.edu.cn.

Y. Yin, D. Huang, N. Qin, T. Wang, and R. Zong are with the Institute of Systems Science and Technology, School of Electrical Engineering, Southwest Jiaotong University, Chengdu 611756, China. K. Li is with the School of Electronic and Electrical Engineering, University of Leeds, LS9 6ST, Leeds, UK. (E-mail: 2022200363@my.swjtu.edu.cn; elehd@swjtu.edu.cn; qinna@swjtu.edu.cn; K.Li@leeds.ac.uk; wtw@my.swjtu.edu.cn; zongrhua@163.com;)

I. INTRODUCTION

With the acceleration of urbanization and economic globalization, high-speed trains (HSTs) are playing an increasingly important role in enhancing regional connectivity and promoting economic development. However, the high-speed nature of HSTs imposes extremely high requirements on their reliability and safety, highlighting the importance of research in high-speed train fault diagnosis.

The bogies, as the critical components of HSTs, are responsible for steering the wheelsets and the train suspension systems, directly affecting the stability of train operation and passenger comfort. Fig. 1 shows position of bogies in a HST. A fault in the bogie may lead to unstable train operations, and even result in derailment and other safety incidents, making its maintenance work crucial. Efficient maintenance ensures safety, extends equipment life, and reduces costs [1][2].

The key damping components used in HST bogies to mitigate vibrations between the wheelsets and the car body include air springs (AS), secondary lateral dampers (LD), and anti-yaw dampers (AD) [3]. Their positions are illustrated in Fig. 2. The AS uses air as an elastic medium to support the car body's weight and absorb track vibrations and shocks. The LD

Fig. 1: Position of bogies in a high-speed train.

reduces lateral swaying and vibrations, stabilizing the car body and reducing lateral forces from uneven tracks or curves. The AD prevents the "hunting" phenomenon, damping longitudinal vibrations to enhance stability, especially during high-speed or long-distance travels [4]. The AS is made of rubber or synthetic materials, which can degrade, resulting in reduced elasticity, cracking, or leakage, causing failures. The LD and AD may also see the development of cracks and oil leaks, leading to a loss in the damping effectiveness [5].

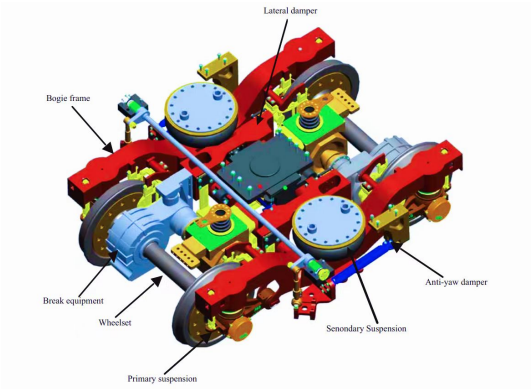


Fig. 2: Main components of the bogie.

Research on the fault diagnosis of bogies is extensive and can generally be divided into two main categories: model-based and data-driven approaches [6]. The former relies on the physical model of the system, requiring in-depth expertise and an understanding of system behavior. Sakellariou et al. [7] developed a Functional Model Based Method for fault detection and magnitude estimation in railway vehicle suspensions using stochastic functional models and statistical decision tools. Due to the high complexity and multivariable coupling of the bogie system, modeling is extremely challenging. It requires significant computational resources and time. Additionally, the varying operating conditions demand repeated validation and adjustment, which reduces its generality and flexibility compared to data-driven methods.

With the increased availability of operational data and the enhancement of computational power, data-driven fault diagnosis methods are increasingly popular due to their excellent performance.

Despite many merits offered by data-driven methods, they also face significant challenges in practical applications [8]. The effectiveness of data-driven methods largely depends on the quality of the input data. Missing values or errors can significantly degrade fault diagnosis accuracy. Normal operation data typically outnumber fault samples by a large margin. This imbalance can biases the training model towards the majority class, leading to high false alarm rates and missed detections [9][10].

Compound faults in bogies, though less frequent than single failures, can still occur and simultaneously disrupt multiple system functions, leading to severe consequences. For instance, a failure in the AS can alter the car body's dynamic response, affecting both ADs' and LDs' load and performance. Therefore, diagnosis of compound faults is crucial for enhancing the stability, safety, and economy of HSTs and optimizing maintenance strategies [11][12]. However, a significant issue arises as recorded fault data is much less than normal operation data, and data for compound faults is even rarer. Therefore, these faults are often classified as unknown. Despite this, research on diagnosing unknown bogie faults remains limited.

For the diagnosis of unknown faults, zero-shot learning (ZSL) is a promising solution emerged in recent years [13]. ZSL is a machine learning strategy for recognizing categories without specific sample encounters. Moreover, most practical applications require the identification of both seen and unseen fault classes, rather than solely focusing on unknown classes as in ZSL. This necessitates the use of generalized zero-shot learning (GZSL). GZSL is mainly divided into three categories: attribute-based, embedding-based, and generative model-based methods [14]. ZSL was first introduced to fault diagnosis for thermal power devices using an expert semantic knowledge base [15][16][17]. Currently, most ZSL based fault diagnosis approaches are centered around attribute prediction. In the context of fault diagnosis, "attributes" typically refer to domain-specific features such as fault location, type, or system parameters, which can be directly mapped to physical characteristics for interpretability.

But to date, most ZSL and GZSL methods in engineering applications still suffer from seen class overfitting, where the model heavily favors seen classes during inference due to their availability during training. This stems from three main issues: domain shifts due to feature distribution mismatch, training data bias that causes overconfidence in seen classes, and incomplete evaluation protocols. Although generative and calibration methods offer partial solutions, achieving balanced generalization remains an open challenge, especially under real-world conditions where fault distribution within seen classes is also highly imbalanced.

To address these challenges in bogie fault diagnosis, we propose a novel approach to effectively integrate different GZSL approaches [18] as each method offers unique advantages in addressing specific issues. To mitigate the domain shifts between seen and unseen classes, attribute-based strategies [19] provide strong interpretability and generalization, effectively narrowing the feature distribution gap across categories. To address the imbalance within seen classes and the inherent bias toward them, we introduce a generative distribution balancing strategy [20] to reduce the model's tendency to overfit seen classes. Integrating these methods leverages their complementary strengths to boost recognition performance for both seen and unseen classes.

However, integrating these approaches into a unified framework presents several challenges. The most critical challenge lies in the structural differences among GZSL methods, includ-

ing discrepancies in feature space representations, prediction formats, and output probabilities, which hinder direct fusion. To overcome this, we propose a stacking-based fusion strategy, where a meta-model is designed to combine the outputs of different models at the feature level, substantially improving the compatibility and robustness of the overall framework. The main contributions in this paper can be summarized as follows:

1) A previously overlooked yet practically challenging issue in the fault diagnosis for bogie systems is addressed: the exploration of unknown compound faults, while also considering the imbalance within the known classes.

2) A novel vibration-optimized GAN loss function is proposed to enable robust synthetic sample generation for low-dimensional bogie systems, which significantly addresses the issue of inadequate calibration of generative models in real-world HST maintenance scenarios.

3) A dual-strategy GZSL is developed, in which attribute-based and model-based approaches are synergistically integrated to simultaneously mitigate domain shift and training data bias, thereby bridging theoretical GZSL advancements with practical bogie fault diagnosis requirements.

4) A stacked meta-model is designed to resolve the compatibility challenges when different GZSL methods are integrated, via feature-level fusion, substantially mitigating the limitations of individual approaches while enhancing diagnostic accuracy and robustness.

The remainder of the paper is organized as follows: Section II briefly introduces the preliminaries and the proposed framework. Section III details the experimental data and parameter settings. Section IV presents and analyzes the experimental results. Finally, Section V summarizes the work and suggests a few topics for future research.

II. Methodologies

The overall proposed framework mainly consists of three parts as shown in Fig. 3, integrating two ZSL strategies to enhance the accuracy and generalizability for bogie fault diagnosis.

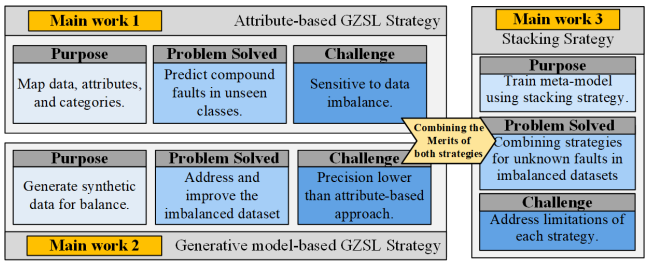


Fig. 3: Three parts of the overall framework.

The attribute-based GZSL strategy focuses on establishing a ternary mapping relationship between data, attributes, and categories to predict compound faults in unseen classes, utilizing attribute information from known classes. Although this strategy is straightforward and generally performs well, its effectiveness can be significantly influenced by data imbalance. Conversely, the generative model-based GZSL strategy aims to improve predictive accuracy for unknown class faults by generating synthetic data to balance the dataset, offering commendable generalization capabilities but potentially lacking the precision of the attribute-based approach. The overall framework employs a stacking strategy to train a meta-model using the predictive results of both strategies, effectively harnessing their advantages to compensate for their respective limitations and enhance overall diagnostic performance.

A. Attribute-Based GZSL Strategy

GZSL leverages auxiliary information to achieve cross-modal knowledge transfer, enabling the inference of information about unseen classes. In 2009, Larochelle et al. [21] introduced a concept similar to ZSL. In the same year, Palatucci et al. [22] proposed the concept of ZSL, using semantic output codes to bridge the gap between seen and unseen classes.

Attribute-based GZSL is a method that allows classification of categories for which no data has been seen during the training phase [23]. This approach relies on attribute descriptions of categories, such as shape, size, and color, which are shared across different categories.

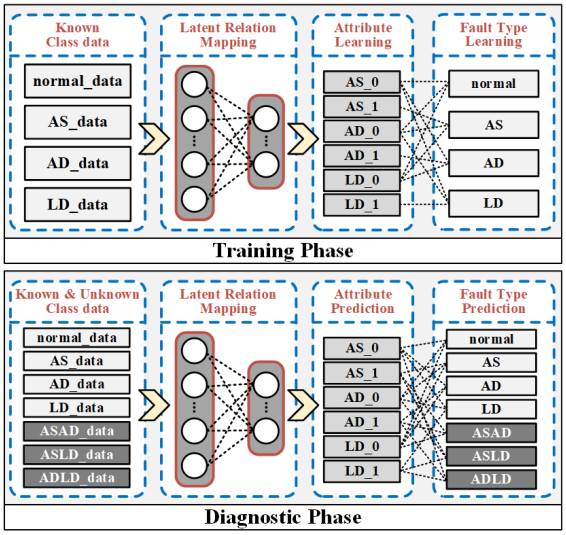


Fig. 4: Workflow of the designed attribute-based GZSL method.

Fig. 4 illustrates the working principle of the attribute-based GZSL approach for bogie fault diagnosis. The data in the figure is divided into normal data and fault data. For example, AD_data represents a single fault occurring at the AD, while ASAD_data represents a compound fault occurring at both the AS and AD. First, a set of attributes are defined. For bogies, these attributes are related to the fault locations. Each single fault is labeled in the dataset, and supervised learning is used to train the model to learn the mapping from attributes to features. During the training phase, the attribute vectors for each category are embedded, establishing a triadic relationship among data, attributes, and categories. In the testing phase, for a sample from an unseen or seen category, the model can identify its attributes and match these attributes with the known category attribute vectors to determine the most probable category.

During the mapping phase of the model, the learning model employs a residual network. Compared to a one-dimensional convolutional neural network [24], the introduction of residual connections can address the problems of gradient vanishing and explosion in deep networks [25]. Residual blocks facilitate faster information flow, allowing the model to reach optimization with fewer training iterations and thereby improving training efficiency. Fig. 5 shows the two types of residual blocks.

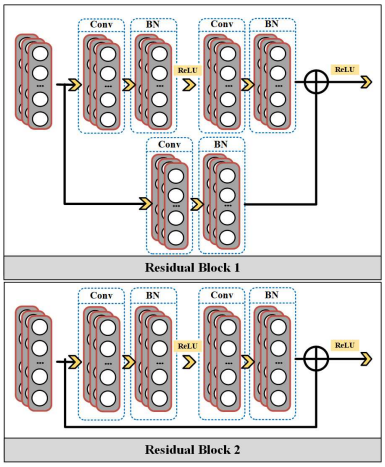


Fig. 5: The two types of designed residual blocks.

In summary, this method offers strong generalization ability and high interpretability. By relying on the generic descriptions of attributes rather than specific category samples, the model can generalize better to new categories. Additionally, using attributes enhances the interpretability of the decision-making process, as decisions are based on understandable features.

However, data imbalance can significantly impact this attribute-based diagnosis approach. When one category has more samples than others, the model tends to learn and predict better for the categories with more abundant samples. This imbalance can lead to insufficient and inaccurate learning of attribute features, ultimately resulting in biased category predictions.

B. Generative Model-Based GZSL Strategy

The generative model-based GZSL strategy primarily relies on generative models such as GANs [26] and VAEs [27][28]. These models can learn data distributions and generate new samples similar to real data. When addressing data sample imbalance, generative models can generate known class samples to balance the dataset and create unknown compound fault data. This approach effectively transforms the GZSL problem into a supervised classification problem [29].

Fig. 6 illustrates the working principle of the designed generative model-based GZSL approach for bogie fault diagnosis. In the training phase, known fault classes are constructed based on attribute labels. These attribute label sets are combined with random noise to train the generative model. In the GAN, the discriminator distinguishes between generated and real known class data, determining if the generated data is fake and, if

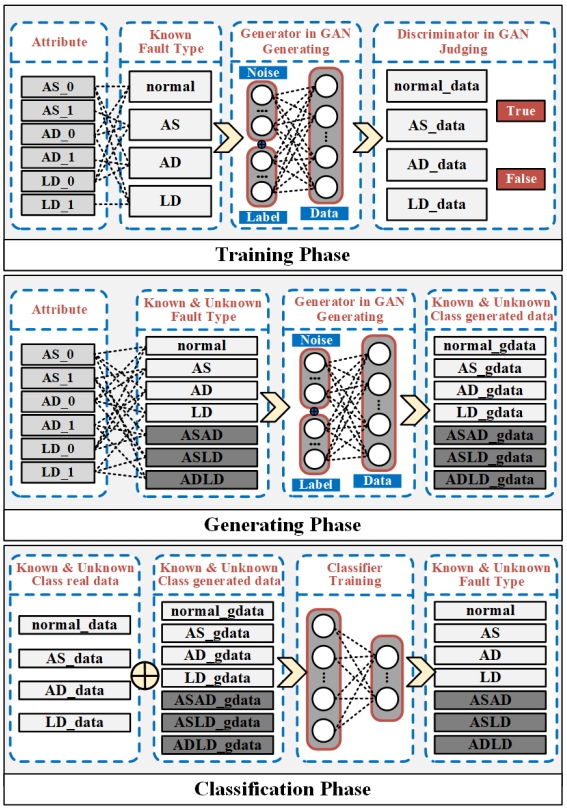


Fig. 6: Workflow of the designed generative model-based GZSL method.

plausible, identifying its category. This training iterates until the GAN can generate realistic data.

In the generating phase, attribute labels are combined to represent both known single faults and unknown compound faults. These combinations, along with random Gaussian noise, are used as inputs to the GAN to generate data. This process produces data for all categories of both known single faults and unknown compound faults. By generating data for all fault categories and combining them with available real fault data, this method better addresses issues of data imbalance and the absence of unknown fault data in the classification phase. As a result, the fault classifier achieves improved classification performance.

For the generation phase, this study employs an improved GAN network. The primary generative models are GAN and VAE. GANs exhibit superior conditional control and diversity in generated data [30], while VAEs are less prone to mode collapse and easier to train, though they generally produce lower quality data compared to GANs [31]. To address the issue of mode collapse often encountered with GANs, a variety of optimizations methods have been proposed to enhance the stability and performance of GANs.

For example, [32] introduces the Wasserstein GAN (WGAN), which uses the Wasserstein distance as a new loss function. A K-Lipschitz discriminator and the maximization of Wasserstein distance address known GAN issues, including training stability and sample quality. In WGAN-GP, a gradient penalty term is introduced to penalize the expected

gradient norm of the discriminator's outputs when it exceeds a threshold [33]. This smooths the discriminator's gradient, preventing gradient explosion or vanishing, thus improving training stability and sample quality. By combining the WGAN loss with a classification loss to enforce distinctiveness in generated features, the proposed f-CLSWGAN framework has the potential to significantly enhance the performance of ZSL [34]. The optimization function for f-CLSWGAN is defined as follows:

$$\min_G \max_D \mathcal{L}_{WGAN} + \beta \mathcal{L}_{CLS}, \tag{1}$$

where β is a hyperparameter used to balance the two losses. $\mathcal{L}_{WGAN}$ is the loss function of WGAN, and $\mathcal{L}_{CLS}$ is the classification loss function for ZSL.

For training the visible samples $S = \{(x, y, c(y)) | \, x \in X, \; y \in Y^S, c(y) \in C\}$, where x represents a set of sample data from dataset X, y represents a visible fault class from the visible fault class set Y^S, and $c(y)$ represents the attribute label combination for a fault class. C denotes the common set of attribute labels. Given the training data S from known categories, the objective is to learn a conditional generator G and a discriminator D.

The loss $\mathcal{L}_{WGAN}$ is computed as follows:

$$\begin{aligned} \min_G \max_D \; \mathcal{L}_{WGAN} &= \mathbb{E}\left[D\left(x, c(y)\right)\right] \\ &- \mathbb{E}\left[D\left(\tilde{x}, c(y)\right)\right] - \lambda \mathbb{E}\left[\left(\|\nabla_{\hat{x}} D\left(\hat{x}, c(y)\right)\|_2 - 1\right)^2\right], \end{aligned} \tag{2}$$

where the symbol $\mathbb{E}$ denotes the expectation in probability theory and statistics. It represents the theoretical weighted average of all possible values a random variable can assume, with the weights determined by the probability distribution of these values. Additionally, $\tilde{x} = G(z, c(y))$, where $z \in Z$ is random Gaussian noise sampled from the set Z, and $c(y) \in C$ represents attribute labels taken as inputs. λ is the penalty coefficient, and $\hat{x}$ is the interpolated data defined as $\hat{x} = \alpha x + (1 - \alpha)\tilde{x}$ with $\alpha \sim U(0, 1)$. This denotes that α is randomly sampled from a uniform distribution U. The loss $\mathcal{L}_{CLS}$ is computed as follows:

$$\mathcal{L}_{CLS} = -\mathbb{E}_{\tilde{x} \sim p_{\tilde{x}}}\left[\log P\left(y|\tilde{x}; \theta\right)\right], \tag{3}$$

where y is the class of known data, and $P\left(y|\tilde{x}; \theta\right)$ represents the probability that the generated data $\tilde{x}$ is correctly classified as class y.

For the convolutional feature extraction of the generated data, the proposed method aims to ensure better performance for GZSL by making the features exhibit high inter-class distinctiveness and high intra-class similarity. The former is ensured by the classification loss in f-CLSWGAN, which guarantees high discriminability of features between different classes. The latter is achieved through our innovative design utilizing an outlier elimination method to maintain high similarity within the same class. The structure of proposed network f-CLSWGAN with Mahalanobis distance (f-CLSWGAN-MD) is shown in the Fig. 7.

Assuming that data not satisfying intra-class similarity are considered as outliers, a threshold is set such that if the feature

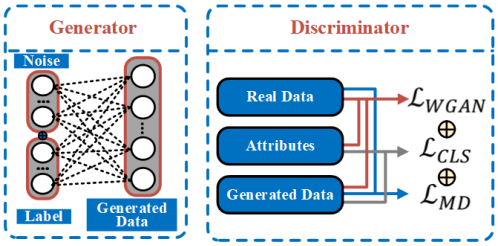


Fig. 7: Loss components of the proposed f-CLSWGAN-MD network.

similarity between the generated data and the original data is below this threshold, it is treated as an outlier for regularization. The Mahalanobis distance [35] is used to measure the similarity. The intra-class similarity loss is defined as follows:

$$\mathcal{L}_{MD} = \mathbb{E}\left[\max\left(0, d\left(\tilde{x}\right) - \tau\right)\right], \tag{4}$$

where τ is the threshold:

$$\tau = percentile\left(d\left(\tilde{x}\right), q\right), \tag{5}$$

where the term q represents the percentile. The Mahalanobis distances between the generated data and the mean vector are sorted, and the value at the q-th percentile is chosen as the threshold τ. For each generated sample $\tilde{x}$, the Mahalanobis distance $d\left(\tilde{x}\right)$ between it and the mean vector μ is calculated:

$$d\left(\tilde{x}\right) = \sqrt{\left(\tilde{x} - \mu\right)^T S^{-1} \left(\tilde{x} - \mu\right)}, \tag{6}$$

where S^{-1} is the inverse of the covariance matrix.

By introducing $\mathcal{L}_{MD}$, the intra-class similarity can be significantly improved. Given this consideration, the optimization function for f-CLSWGAN-MD can now be redefined as follows:

$$\min_G \max_D \mathcal{L}_{WGAN} + \beta \mathcal{L}_{CLS} + \gamma \mathcal{L}_{MD}, \tag{7}$$

where β and γ are hyperparameters used to balance the three types of losses.

One-dimensional data often exhibit strong temporal or local correlations, but traditional methods struggle with inter-class overlap and intra-class outliers. By integrating Mahalanobis distance and an innovative outlier elimination method, the approach ensures intra-class similarity, inter-class distinctiveness, and effectively filters noise. It also captures both local and global features of one-dimensional data, enhancing the quality of generated data for downstream classifiers.

With this enhancement, the generative model-based GZSL strategy is able to extend the class coverage by generating data for unseen categories, thereby extending the model's recognition capability and their generalization ability.

However, this strategy heavily depends on the quality of the generated data. If the generated data significantly deviates from the reality, the overall model performance will be significantly impacted.

C. Stacking Strategy

The stacking strategy, commonly referred to as Ensemble Learning in the field of machine learning [36] is a method that enhances prediction accuracy and stability by combining multiple models.

The concept of Stacking was first introduced in [37]. It was initially used to minimize generalization error by training a meta-model to combine the predictions of multiple base models. In [38], this concept was further expanded by using regression models as meta-models. It is demonstrated that combining different types of base learners, such as decision trees [39] and neural networks, significantly improves model performance. The most recent development includes its application to deep neural networks. For instance, shallow neural networks can be used as base models, while deep neural networks serve as meta-models. This approach has been successfully applied in image classification, speech recognition, and natural language processing [40].

Fig. 8 illustrates the designed Stacked framework tailored to this specific fault diagnosis problem:

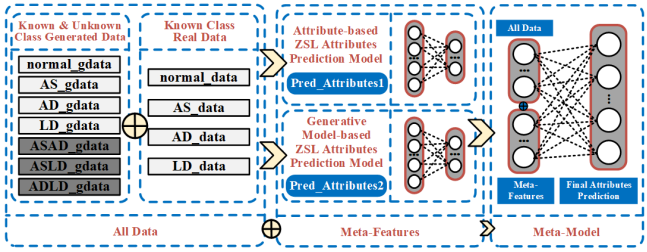


Fig. 8: Workflow of the designed stacked framework.

The proposed stacking strategy first uses generated data to compensate for missing and imbalanced data, creating a balanced and complete dataset. Two GZSL strategies are employed for attribute prediction, yielding two predicted attributes used as meta-features. These meta-features, along with the dataset, are used to train the meta-model, resulting in the final attribute label prediction.

The core of this strategy lies in its ability to improve prediction accuracy by combining the strengths of different models. For unknown composite fault prediction under imbalanced data conditions, ensemble learning effectively merges the strengths of the two GZSL strategies, outperforming any single base model. Additionally, it reduces overfitting, as individual complex models tend to overfit the training data, especially when data is limited or has too many features. Aggregating multiple models, especially simpler ones, smooths out prediction outliers and improves generalization on unseen data. Lastly, it increases robustness, as each base model has its own strengths in handling different types of noise and outliers. Model aggregation enhances the overall model's tolerance to a variety of noise and outliers.

III. Experimental Setup

This section introduces the experimental setup, including the data source, parameter settings in th experiments and for the stacked model.

A. Data source

Collecting high-frequency data for faulty bogies is nearly impossible due to their rare failure rate and associated safety and economic risks. To address this, mechanical faults were simulated using SIMPACK, a reliable dynamic tool widely used in rail transit. Models from the State Key Laboratory of Traction Power in Southwest Jiaotong University follow realistic dynamics. Simulation data, validated in [3] and [12], are extensively used for bogie fault diagnosis and algorithm evaluation. The simulation configuration of the vehicle is shown in Fig. 9.

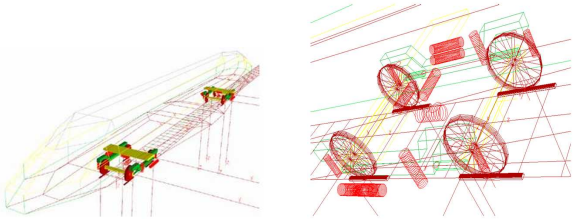


Fig. 9: Simulation configuration of HST bogie.

The vehicle model mainly consists of one car body, two bogies, and four sets of wheel pairs, with a total of 62 degrees of freedom. The model uses LMA economy-type treads, CN60 tracks, and the Wuhan-Guangzhou railway track spectrum as excitations. To enhance authenticity, track excitations and route spectra from real-world environments were incorporated, minimizing the gap between simulated and real data. Combined with train operation dynamics, this approach generated high-quality vibration signals. The sampling frequency was set to 243 Hz. According to [41], the two-channel optimal fault mode representation was chosen as the signal channel. Normal operation data, as well as single fault data for AS, AD, and LD, are collected and encoded simply as 0 and 1. Each type of data is collected for a duration of 220 seconds, a simple data cleaning procedure was conducted to remove invalid data and the first 4 seconds of data. 52,488 data samples were collected for each component, with each data sample having a dimension of (243, 2).

To generate the data imbalance condition, four groups of data with different imbalance conditions are designed, as summarized in Table I.

TABLE I: Experimental Groups for Imbalance Data

Number	Normal	AS	LD	AD
1. Data Balance	52488	52488	52488	52488
2. Fault Imbalance Scenario 1	52488	10500	10500	10500
3. Fault Imbalance Scenario 2	52488	5240	5240	5240
4. Practical Engineering Fault Imbalance	52488	5240	20900	20900

The four groups of data include the Data Balance Group, Fault Imbalance Scenario 1 Group (with an approximate 5:1 ratio for normal data and single fault data), Fault Imbalance Scenario 2 Group (with an approximate 10:1 ratio for normal data and single fault data), and the Practical Engineering Fault Scenario Group, where the ratio was set to approximately 10:1:4:4. The rationale for the final group is that in HST

bogies, AS is primarily used to buffer vertical vibrations, resulting in a relatively lower fault frequency, possibly accounting for 10-20% of the total fault rate. AD and LD dampers are used to suppress hunting motion and lateral vibrations of the train, and their fault rates are relatively higher, approximately accounting for 40-45% of the total fault rate. These estimates are based on the general fault characteristics of each component in the HST bogie system and industry experience.

B. Parameter settings for different strategies

1. Parameter Settings for Attribute-based GZSL

Residual networks are used to extract deep fault data features, with the network configuration shown in Table II. The data batch size was set to 60. The Adam optimizer, with an initial learning rate of the default 1e-3, was used in conjunction with the ReduceLROnPlateau callback function for learning rate adjustment. The number of epochs with no improvement before reducing the learning rate was set to 10. The learning rate reduction factor was set to 0.1, and the minimum improvement required to consider as an improvement was 0.0001. No cooldown period was set. The total number of iterations was 200.

TABLE II: Simplified Configuration of ResNet Architecture

Layer	Input Size	Parameters	Output Size	Connected to
C1_BN_R	(bz,243,2)	K=3,S=2,F=64	(bz,122,64)	
MP1	(bz,122,64)		(bz,61,64)	C1_BN_R
A1_C1_BN_R	(bz,61,64)	K=3,S=1	(bz,61,64)	MP1
A1_C2_BN	(bz,61,64)	K=3,S=1	(bz,61,64)	A1_C1_BN_R
A1_C	(bz,61,64)	K=1,S=1	(bz,61,64)	C1_BN_R
A1_SC_R	(bz,61,64)		(bz,61,64)	A1_C2_BN, A1_C
C2	(bz,61,64)	K=3,S=2,F=128	(bz,31,128)	A1_SC_R
B1_C1_BN_R	(bz,31,128)	K=3,S=1	(bz,31,128)	C2
B1_C2_BN	(bz,31,128)	K=3,S=1	(bz,31,128)	B1_C1_BN_R
B1_SC_R	(bz,31,128)		(bz,31,128)	B1_C2_BN, C2
C3	(bz,31,128)	K=3,S=2,F=256	(bz,16,256)	B1_SC_R
B2_C1_BN_R	(bz,16,256)	K=3,S=1	(bz,16,256)	C3
B2_C2_BN	(bz,16,256)	K=3,S=1	(bz,16,256)	B2_C1_BN_R
B2_SC_R	(bz,16,256)		(bz,16,256)	B2_C2_BN, C3
C4	(bz,16,256)	K=3,S=2,F=512	(bz,8,512)	B2_SC_R
B3_C1_BN_R	(bz,8,512)	K=3,S=1	(bz,8,512)	C4
B3_C2_BN	(bz,8,512)	K=3,S=1	(bz,8,512)	B3_C1_BN_R
B3_SC_R	(bz,8,512)		(bz,8,512)	B3_C2_BN, C4
AAP1	(bz,8,512)	Output=1	(bz,512)	B3_SC_R
DS	(bz,512)		(bz,6)	AAP1

bz: batch size.
K: Kernel size.
S: Stride.
F: Filters.

In the table, “C” represents a one-dimensional convolution layer, “BN” stands for batch normalization, “A1” , “B1”, “B2” and “B3” stand for two types of residual blocks “A” and “B” in the residual network and “AAP” represents adaptive average pooling. “DS” stands for dense layer with sigmoid function. The ReLU activation function, denoted as “R,” was selected for the residual network because it can mitigate the vanishing gradient problem and its sparse activation improves the efficiency of the residual network.

2. Parameter Settings for Generative Model-based GZSL

The GAN network includes a generator and a discriminator. Their respective network architectures are shown in Tables III and IV. The update frequency of the generator and the discriminator was set to 1:5. Both models were optimized using the RMSprop optimizer with a learning rate of 0.00005 to maintain stability during the training process. The hyperparameter β controls the gradient penalty weight in the loss function and was set to 8. The hyperparameter γ controls the regularization of discrete values in the loss function and was set to 4. A total of 25,000 training epochs were conducted. The specific ablation experiments for the hyperparameter settings will be presented later. The network ultimately connects to a linear layer, utilizing the sigmoid activation function to predict the probability of each attribute. This results in an attribute set corresponding to the final attribute category.

TABLE III: detailed configuration of the generator

Layer name	Input Size	Detailed parameters	Output size
Noise_multiply_Lable	(bz,6)&(bz,6)		(bz,6)
Dense1	(Bz,6)	activation=‘relu’	(bz,18)
Reshape	(bz,18)		(bz,2,9)
UpSampling1d1	(bz,2,9)	Size = 3	(bz,6,9)
Conv1d_BN_LeakyReLU1	(bz,6,9)	F=27,K=2,S=3	(bz,2,27)
UpSampling1d2	(bz,2,27)	Size = 3	(bz,6,27)
Conv1d_BN_LeakyReLU2	(bz,6,27)	F=81,K=2,S=3	(bz,2,81)
UpSampling1d3	(bz,2,81)	Size = 3	(bz,6,81)
Conv1d_BN_LeakyReLU3	(bz,6,81)	F=243,K=2,S=3	(bz,2,243)
Permute	(bz,2,243)		(bz,243,2)

bz: batch size.
F: Filters.
K: Kernel size.
S: Stride.

TABLE IV: detailed configuration of the discriminator

Layer name	Input Size	Detailed parameters	Output size
Label_embedding	(bz,6)		(bz,243,6)
Data_concatenate_label	(bz,243,6)		(bz,243,8)
Conv1d_LeakyReLU1	(bz,243,8)	F=81,K=3,S=3	(bz,81,81)
Conv1d_LeakyReLU2	(bz,81,81)	F=27,K=3,S=3	(bz,27,27)
Conv1d_LeakyReLU3	(bz,27,27)	F=9,K=3,S=3	(bz,9,9)
Conv1d_LeakyReLU4	(bz,9,9)	F=3,K=3,S=3	(bz,3,3)
Flatten	(bz,3,3)		(bz,9)
Dense	(bz,9)		(bz,1)

bz: batch size.
F: Filters.
K: Kernel size.
S: Stride.

In the tables, “Conv1d” represents a one-dimensional convolution layer, “BN” stands for batch normalization, and “UpSampling1d” denotes a one-dimensional upsampling layer. The network utilizes LeakyReLU to improve gradient flow within the network, as opposed to ReLU, which may suffer from the “dead neuron” problem. The GAN handles inputs and data distributions with negative values. LeakyReLU provides a small gradient for negative inputs, helping to stabilize the training process and achieve better convergence.

3. Parameter Settings for the Stacked Model

The network architecture settings for the Stacked ensemble of the two diagnostic models are shown in Table V.

TABLE V: detailed configuration of the stacked model

Layer name	Input Size	Output size
Data_pred1_pred2_concatenate	(bz,243*2)&(bz,6)&(bz,6)	(bz,498)
Dense1_relu	(bz,498)	(bz,128)
Dense2_relu	(bz,128)	(bz,64)
Dense_sigmoid	(bz,64)	(bz,6)

bz: batch size.

The model accepts three different inputs. The input data of shape (bz, 243, 2) is flattened to (bz, 486) and concatenated

with the attribute label results predicted by the two base models. The Adam optimizer was used with an initial default learning rate of 0.001 for adjustment.

IV. Experimental Results and Analysis

This section primarily discusses the experimental results for bogie fault diagnosis. In order to validate the effectiveness of the designed framework, ablation experiments and comparative experiments were also conducted. This section presents the results of each part of the experiments and summarize the key findings.

A. Experimental results

Based on the groups set in Table I, experiments were conducted to explore the results under different data imbalance conditions, as shown in Table VI. The experimental results include the recognition accuracy based on the attribute-based GZSL strategy, the recognition accuracy based on the generative model-based GZSL strategy, and the overall recognition accuracy of the proposed method based on the Stacked model. In addition, for the proposed method, the diagnostic accuracy of known classes under normal and single-fault conditions and the recognition accuracy of unknown classes under compound fault conditions are listed for each group of data. For the GZSL problem, the harmonic mean of the class-average top-1 accuracy of known and unknown classes is often chosen as the evaluation metric. Therefore, the harmonic mean of the proposed method is also provided in the Table VI. Fig. 10 shows the confusion matrices for the recognition of each category in the four groups of experiments using the proposed method.

From the first two groups of data, it can be observed that under data-balanced conditions and when the data imbalance is not pronounced (Groups 1 and 2), the diagnostic accuracy of the proposed model based on the Stacking ensemble is slightly lower than that of the attribute-based GZSL method. However, in the case of significant data imbalance (Group 3), the generative model-based GZSL demonstrates more robust diagnostic performance compared to the attribute-based GZSL, which suffers from model drift due to the influence of sample quantity. Overall, the proposed Stacking ensemble method exhibits good recognition accuracy across all groups, particularly under data imbalance conditions, highlighting the advantages of ensemble learning.

The diagnostic accuracy for known class faults in the four groups is consistently high, around 90%. Except for the extremely imbalanced Group 3, the accuracy for unknown classes in each group exceeds 70%. The harmonic mean of recognition accuracy for the imbalanced data groups (Groups 2 and 4) also exceeds 80%, and even for the extremely imbalanced Group 3, the harmonic mean of recognition accuracy reaches 74%.

Additionally, the performance of the generative model-based GZSL method needs to be analyzed separately. Fig. 11 presents a comparative analysis of real and generated time-domain data.

Direct assessment of generation performance in the time domain is challenging due to the data's high non-linearity, which even t-SNE struggles to clarify. To address this, a convolutional classifier was trained on real data to extract high-dimensional features from the penultimate layer. Both real and generated data were input into the classifier, and the extracted features were visualized using t-SNE for dimensionality reduction. The resulting clustering in Fig. 12 highlights whether the generated data aligns with the original data distribution.

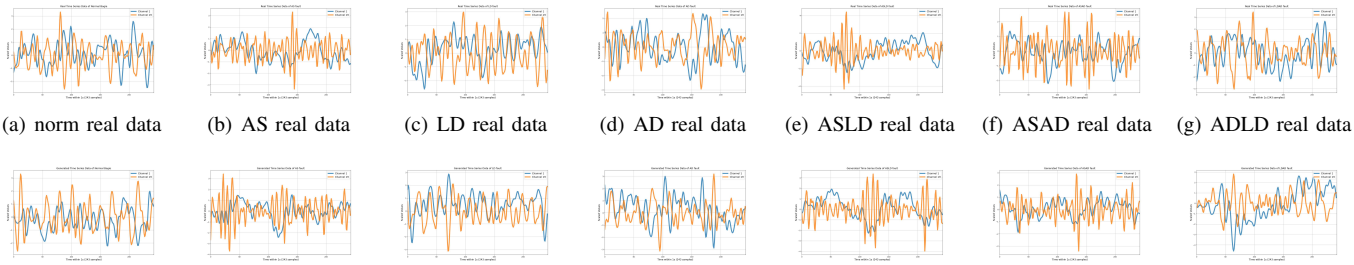
From the figure, distinct clusters with high intra-class cohesion indicate strong inter-class separability in the penultimate

TABLE VI: The experimental results for four groups

Group Number	Attribute-Based GZSL Diagnostic Accuracy	Generative Model-Based GZSL Diagnostic Accuracy	Overall Fault Diagnosis Accuracy Based on Stacking Ensemble Model	Known Class Diagnostic Accuracy	Unknown Class Diagnostic Accuracy	Harmonic Mean
1	92.34%	90.01%	92.22%	99.40%	82.67%	90.27%
2	89.79%	86.55%	89.67%	96.62%	80.40%	87.77%
3	73.63%	75.74%	77.74%	88.11%	63.91%	74.09%
4	82.93%	80.48%	83.88%	92.04%	73.00%	81.42%

(a) Confusion matrix of data balance group. (b) Confusion matrix of fault imbalance scenario 1 group. (c) Confusion matrix of fault imbalance scenario 2 group. (d) Confusion matrix of practical engineering fault imbalance group.

Fig. 10: Confusion matrix results for 4 group.



(a) norm real data (b) AS real data (c) LD real data (d) AD real data (e) ASLD real data (f) ASAD real data (g) ADLD real data

(h) norm generated data (i) AS generated data (j) LD generated data (k) AD generated data (l) ASLD generated data (m) ASAD generated data (n) ADLD generated data

Fig. 11: Comparison of real and generated data across different datasets. Top row: real data. Bottom row: generated data.

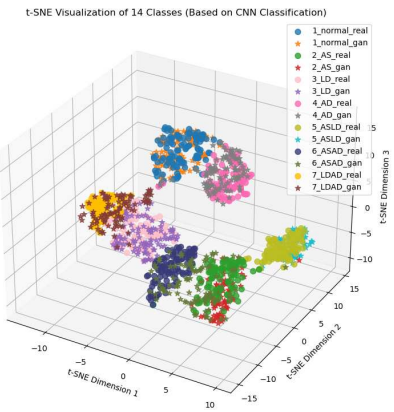
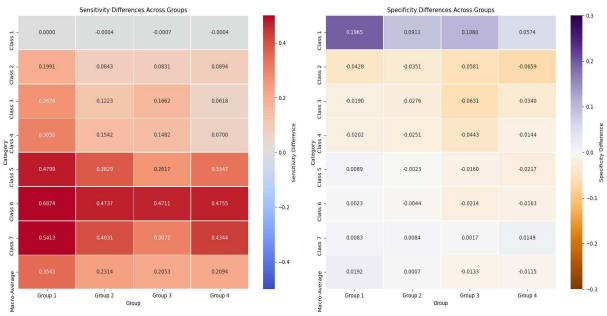
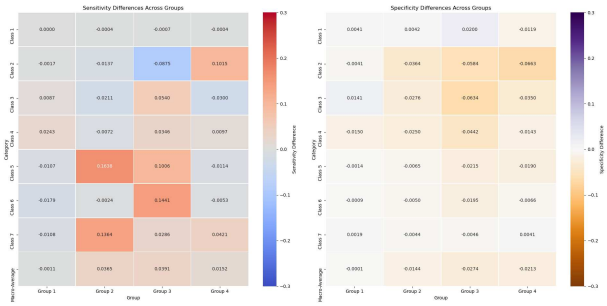


Fig. 12: T-SNE plots of real data and generated data.



(a) sensitivity and specificity differences of original data distribution



(b) sensitivity and specificity differences of balanced data distribution

Fig. 13: Impact of data balancing on sensitivity and specificity across groups using GZSL strategy.

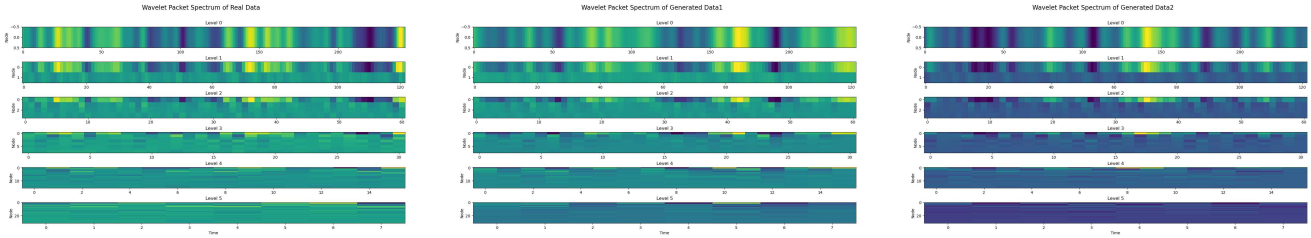
layer's features. Most categories show overlap between real and generated data, suggesting that the generated samples effectively capture real data features. However, in some categories, such as ASAD, the generated data features overlap with the distribution of AS, revealing cross-distribution characteristics.

In addition, sensitivity and specificity experiments were conducted to compare the changes before and after data balancing. Since the study focuses on unknown fault types, to prevent data imbalance from affecting the evaluation of method performance, two assessments were performed using the GZSL strategy based on generative models: The first evaluation maintained the original data distribution to compare the sensitivity and specificity before and after data generation. The second utilized a balanced dataset to better evaluate the performance of the GZSL strategy based on generative models. The detailed results are shown in the Fig. 13.

For sensitivity, the model maintains stable sensitivity for normal classes across both original and balanced test sets. The GZSL strategy significantly enhances sensitivity for unknown faults, especially in the original distribution, though improvements are less noticeable in the balanced test set. This indicates that data augmentation strengthens the model's ability to learn unknown faults but slightly weakens performance on known faults. Sensitivity for Class 2 decreases, likely due to issues with augmented data quality, causing feature confusion with other classes. Overall, the strategy improves recognition of minority classes.

For specificity, while improving sensitivity, the GZSL strategy slightly reduces specificity, particularly for negative samples in single and unknown fault classes. Specificity for unknown faults, such as Classes 6 and 7, drops due to feature overlap introduced by augmentation, affecting the model's ability to distinguish negative samples. This decline is more pronounced in the original distribution test set, where the model leans toward positive class recognition. In the balanced test set, the impact is more consistent, but improvements for unknown and single fault classes are more evident in the original distribution.

To validate GAN-generated samples' coverage of unseen class attributes, controlled experiments were conducted with:



(a) Wavelet packet spectrum of real data. (b) Wavelet packet spectrum of generated data using proposed f-CLSWGAN-MD. (c) Wavelet packet spectrum of generated data using WGAN.

Fig. 14: Wavelet packet spectrum of real and generated data.

Experimental group (Real known classes data and GAN-generated unknown-class data), Control 1 (Real known classes data only) and Control 2 (Real known and unknown classes data). Experiment 1 compared the results of the experimental group with Control 1, while Experiment 2 compared the experimental group with Control 2. Testing used real known and unknown class samples, with 20 replicates under identical conditions (matched data sizes). Statistical analysis results were shown in Table VII.

TABLE VII: Group Performance Comparisons with Effect Sizes

	Mean Difference	p-value	95% Confidence Interval	Cohen's d
1	27.9330	2.5971e-25	[25.6876, 30.1784]	7.9637
2	-3.5560	2.9881e-06	[-4.8709, -2.2411]	-1.7313

Note:
Mean Difference: The absolute difference in accuracy (%) between compared groups.
p-value: Probability of observing the result under the null hypothesis (no difference); values < 0.05 indicate statistical significance.
95% CI: Confidence interval for the mean difference; non-overlap with zero supports significance.
Cohen's d: Standardized effect size; $|d| \geq 0.8$ suggests a large effect.

The statistically significant p-values confirm that these differences are unlikely to be due to random variation. The experimental group achieved a 27.93% accuracy improvement over Control 1, demonstrating the effectiveness of GAN-generated samples in mitigating complete unknown-class recognition failure. Meanwhile, the 3.56% deficit relative to Control 2, with its narrow confidence interval, indicates GAN samples closely approximate real unknown-class distributions.

B. Ablation study

In the designed generative model-based GZSL method, several hyperparameter settings are involved. To explore the optimal performance, ablation experiments were conducted and the results are presented in this section. To validate the generative capability of using MD for loss regularization, we first compare the spectrum of the generated data with that of the real data. Using the unknown class compound fault ASAD as an example, the wavelet packet analysis spectrum of the actual fault data from sensor 19 is compared with the corresponding generated data presented in Fig. 14. Additionally, comparisons are made between the data generated by wGAN models, both with and without outlier regularization using Mahalanobis distance.

By comparing the three wavelet packet spectrum plots, it is evident that the proposed f-CLSWGAN-MD method generates data with spectral characteristics that are closer to the real data. The quality of the generated data is superior to that produced by the WGAN. The spectral characteristics across various frequency bands highly match those of the real data, indicating that the generated data is of high quality and can effectively simulate the features of the real data. Therefore, using MD to regularize the loss results in better fidelity and consistency in the generated data.

To investigate the impact of different hyperparameter ratios on the convergence of GAN's data generation loss, various hyperparameter values were set as shown in Table VIII. A total of 25,000 training iterations were conducted, with the discriminator's loss and the loss between generated data and real data recorded every 50 iterations.

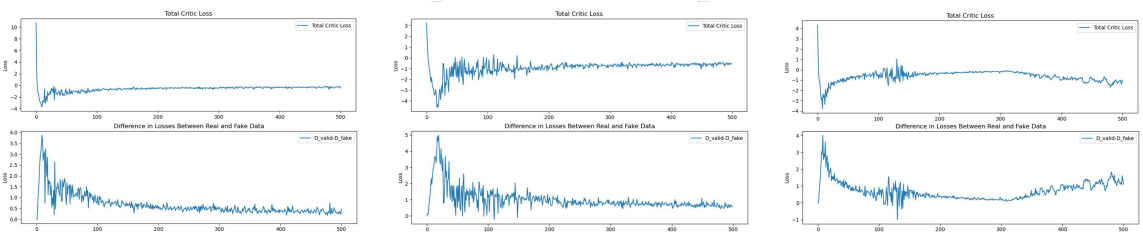
TABLE VIII: Settings for Five Groups of Hyperparameters

Group Number	Hyperparameter Configuration Set
1	$\boldsymbol{\beta = 8, \gamma = 4}$
2	$\beta = 12, \gamma = 4$
3	$\beta = 4, \gamma = 4$
4	$\beta = 8, \gamma = 8$
5	$\beta = 8, \gamma = 1$

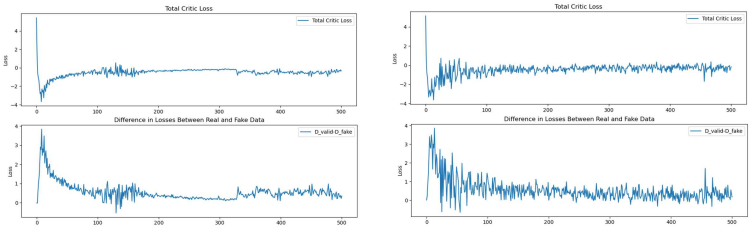
The Fig. 15 compares five sets of trends for two types of losses in the discriminator and the differences between generated and real data. The results indicate that the convergence effect of the proportions set in this paper is the best, with minimal fluctuations. If the value of γ is too large, it may lead to non-convergence of the results, while an excessively large value of β may cause significant fluctuations.

C. Comparative experiments

For the task of mining deep mapping relationships among attributes, features, and data, two types of residual network blocks are designed as shown in Fig. 5. To evaluate their performance, we conducted a comparative experiment with 7 experimental groups, as detailed in Table IX. These were tested against the four environmental scenarios listed in Table I, yielding accuracies as shown in the Fig. 16.



(a) Loss Trends for Group 1. (b) Loss Trends for Group 2. (c) Loss Trends for Group 3.



(d) Loss Trends for Group 4. (e) Loss Trends for Group 5.

Fig. 15: Trends of two types of losses for discriminator and differences between generated and real data.

TABLE IX: Experimental groups setup for different networks and structures for attribute-based GZSL

Group Number	different networks and different structures
1	**Designed Dual Residual Block**
2	Single A Residual Block
3	Single B Residual Block
4	Conv1d Network
5	Dense Network
6	Support Vector Machines
7	Decision Tree

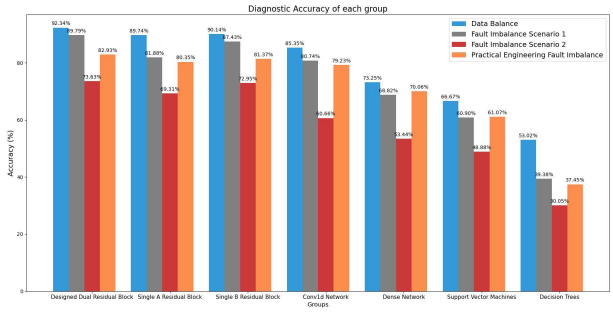


Fig. 16: Diagnostic accuracy of 5 groups under 4 conditions.

The results clearly demonstrate that the designed dual-residual-block attribute prediction model achieves the highest accuracy across all four scenarios, with 82.93% under practical fault imbalance conditions. Compared to other methods, such as Single A and Single B residual blocks, which show performance drops in imbalanced scenarios, and traditional models like SVM and decision trees, which fall to 48.88% and 30.05% respectively, the dual-residual-block model demonstrates superior robustness and feature-mining capability. These results highlight its effectiveness in addressing both balanced and imbalanced fault diagnosis tasks.

Besides the GAN network, other models like VAE and Diffusion [42] are also capable of generating data. Therefore, two additional experimental groups were set up. Since clustering all categories together could interfere with the data and lead to an excessive number of categories, this experiment focuses solely on exploring the performance of the three types of generators. The experiment uses an unknown composite fault ASAD as an example. Dimensionality reduction feature extraction and t-SNE visualization of the generative models reveal the clustering effects of the generated data compared to real data, as shown in Fig. 17.

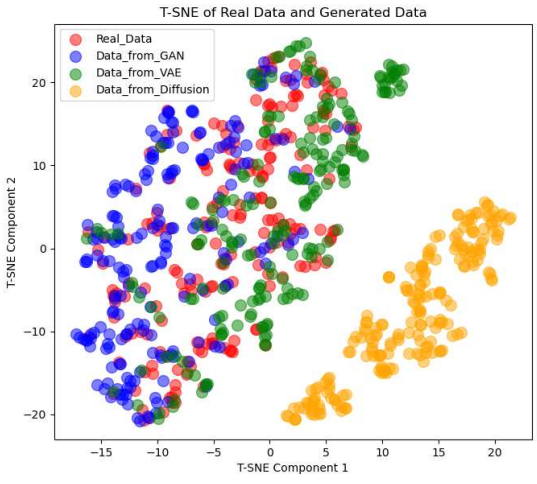


Fig. 17: t-SNE Plots of real data and data Generated by designed GAN, VAE and diffusion network.

It can be observed that the data points generated by the GAN are closely aligned with the distribution of real data points, covering a wide area and separated naturally between

clusters, demonstrating high diversity and robust generative capabilities. In contrast, data points generated by the VAE are relatively concentrated in the central area, indicating a compact distribution. This may suggest that the VAE-generated data lacks diversity and may exhibit some degree of blurring or distortion. The poorest performance is observed with the Diffusion model, whose clusters are more segregated. This indicates that the Diffusion model may not have adequately learned the intrinsic structure of the data for this specific task, resulting in greater variability in the generated data. Although the Diffusion model has a great potential in generating high-quality and diverse images, its training process is usually slower and highly sensitive to parameters and training techniques, performing poorly in generating asymmetrical low-dimensional data.

Additionally, the overall diagnostic accuracy based on the three generative models for GZSL is also listed in the Table X.

TABLE X: Diagnostic Accuracy of Each Group

Group Number	Diagnostic Accuracy of each group			
	Data Balance	Fault Imbalance Scenario 1	Fault Imbalance Scenario 2	Practical Engineering Fault Imbalance
1. Designed GAN	**90.01%**	**86.55%**	**75.74%**	**80.48%**
2. VAE	87.42%	81.81%	67.95%	75.59%
3. Diffusion	64.76%	60.03%	57.21%	59.83%

It is evident that the GAN designed here significantly outperforms the other two generative models in terms of recognition efficiency, showcasing strong adaptability and robustness. However, GANs may face challenges such as mode collapse during training, requiring careful optimization to ensure data quality. While the VAE benefits from stable training, its lack of diversity in generated data makes it unsuitable for GZSL tasks. On the other hand, diffusion models, despite their novelty, solid mathematical foundation, and strong interpretability, show limitations in generating low-dimensional data. Their longer training times and high sensitivity to parameters further restrict their practicality in real-world tasks. Overall, the designed GAN demonstrates higher accuracy and stability in generating capabilities for unknown compound faults, making it the most suitable model for practical GZSL applications.

A number of ensemble models have been proposed and include training as a meta-model, random forests, voting, and others. To verify the diagnostic accuracy of the model, comparative experiments are conducted as listed in Table XI. The table shows the diagnostic accuracy of four ensemble models under four conditions. To more intuitively display the performance of each model, Fig. 18 presents their performance radar charts.

The Stacked GZSL framework demonstrates significant advantages over Random Forests and traditional ensemble methods in handling imbalanced data and unknown faults. Unlike hard or weighted voting, which fail to adaptively combine model strengths and are often limited by the performance of weaker models, the stacking approach integrates diverse

TABLE XI: Diagnostic Accuracy of Each Group

Group Number	Diagnostic Accuracy of each group			
	Data Balance	Fault Imbalance Scenario 1	Fault Imbalance Scenario 2	Practical Engineering Fault Imbalance
1. Stacking Method	**92.22%**	**89.67%**	**77.74%**	**83.88%**
2. Majority voting	90.81%	85.37%	72.46%	78.25%
3. Weighted Voting	91.95%	88.20%	73.85%	80.11%
4. Random Forest	92.78%	89.37%	74.65%	81.22%

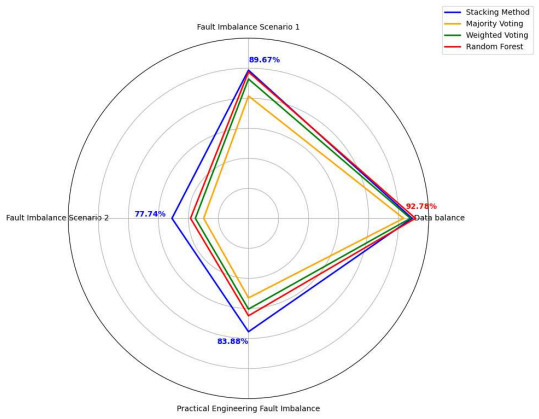


Fig. 18: Performance radar charts of four ensemble models under four conditions.

models more effectively, resulting in superior accuracy and robustness, particularly in challenging fault scenarios.

While Random Forests excel in balanced data conditions and can even outperform the proposed method in such cases, their simpler structure and reliance on decision tree ensembles make them more suitable for initial model construction and rapid iteration. However, their performance may be limited in more complex or imbalanced scenarios compared to the proposed Stacked GZSL framework. Moreover, since Random Forests are limited to decision tree ensembles, their flexibility is relatively low.

Traditional ensemble methods like hard and weighted voting are simple and computationally efficient but lack adaptability, as they cannot dynamically adjust model weights. Consequently, their performance often lags behind stacked methods in complex or imbalanced tasks.

The stacked GZSL method effectively combines the strengths of base models and optimizes predictions for unknown classes, demonstrating high accuracy, robustness, and adaptability, especially under data imbalance. While it incurs higher computational costs and relies on base model performance, it consistently outperforms other methods in diagnosing known and unknown faults, making it the preferred choice for practical GZSL applications.

V. Conclusion

In addressing the issues of data imbalance and unknown compound fault diagnosis, this paper has proposed a Stacked-GZSL framework, which has shown to achieve efficient fault

diagnosis through the following innovations. Firstly, it introduces the GZSL strategy, utilizing attribute concept combinations to predict unknown compound faults. Secondly, it improves the loss function of the wGAN network to address the imbalance of single fault data, enhancing data generation quality. Moreover, it combines two GZSL strategies based on generative models and attribute prediction models. Lastly, by employing stacking model techniques, the advantages of both strategies are retained, resulting in more accurate overall predictions.

Despite these advancements, certain challenges remain. The framework's performance heavily depends on the quality and completeness of the attribute information, as poorly defined or weakly correlated attributes can adversely affect diagnostic accuracy. The generative model-based strategy also relies on the quality and stability of synthetic data, which can be compromised by issues such as mode collapse or suboptimal hyperparameter settings. Furthermore, while the stacked model improves performance, it introduces higher training complexity, requiring significant computational resources and more intricate debugging efforts. Additionally, the method is currently limited to offline scenarios, with computational complexity and the need for rapid adaptation posing challenges for real-time implementation. Lastly, although the framework demonstrates promising results, its generalization capability to unknown classes may be constrained under extreme data imbalance or when attribute correlations are insufficient.

Therefore, future work will focus on addressing these limitations and adapting the framework for real-time fault diagnosis. Key directions include developing incremental learning methods based on fault information, improving the quality of attribute information, enhancing the stability of generative models, and optimizing the stacked model to reduce computational complexity. Strengthening the framework's generalization under extreme data imbalance and weak attribute correlations will also be prioritized. Leveraging remote transmission technologies and edge computing frameworks will help bridge the gap between offline analysis and real-time monitoring.

REFERENCES

[1] C. Zhang, H. Kordestani, and M. Shadabfar, "A combined review of vibration control strategies for high-speed trains and railway infrastructures: Challenges and solutions," *Journal of Low Frequency Noise, Vibration and Active Control*, vol. 42, no. 1, pp. 272–291, 2023.

[2] L. Chen, N. Qin, X. Dai, and D. Huang, "Fault diagnosis of high-speed train bogie based on capsule network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6203–6211, 2020.

[3] D. Huang, S. Li, N. Qin, and Y. Zhang, "Fault diagnosis of high-speed train bogie based on the improved-ceemdan and 1-d cnn algorithms," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[4] W. Zhai, K. Wei, X. Song, and M. Shao, "Experimental investigation into ground vibrations induced by very high speed trains on a non-ballasted track," *Soil Dynamics and Earthquake Engineering*, vol. 72, pp. 24–36, 2015.

[5] Y. Zang, W. Shangguan, B. Cai, H. Wang, and M. G. Pecht, "Methods for fault diagnosis of high-speed railways: A review," *Proceedings of the institution of mechanical engineers, part O: journal of risk and reliability*, vol. 233, no. 5, pp. 908–922, 2019.

[6] M. Mansouri, M.-F. Harkat, H. N. Nounou, and M. N. Nounou, *Data-driven and model-based methods for fault detection and diagnosis*. Elsevier, 2020.

[7] J. Sakellariou, K. Petsounis, and S. Fassois, "On board fault detection and identification in railway vehicle suspensions via a functional model based method," in *Proceedings of the International Conference on Noise and Vibration Engineering (ISMA 2002), Leuven, Belgium*, pp. 16–18, 2002.

[8] H. Chen, B. Jiang, S. X. Ding, and B. Huang, "Data-driven fault diagnosis for traction systems in high-speed trains: A survey, challenges, and perspectives," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1700–1716, 2020.

[9] T. Zhang, J. Chen, F. Li, K. Zhang, H. Lv, S. He, and E. Xu, "Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions," *ISA transactions*, vol. 119, pp. 152–171, 2022.

[10] Y. Geng, Z. Wang, L. Jia, Y. Qin, and X. Chen, "Bogie fault diagnosis under variable operating conditions based on fast kurtogram and deep residual learning towards imbalanced data," *Measurement*, vol. 166, p. 108191, 2020.

[11] W. Zhao, Y. Liu, B. Wang, and Y. Liao, "Composite fault mechanism and vibration characteristics of high-speed train axle-box bearings," *International Journal of Mechanical System Dynamics*, 2024.

[12] N. Qin, B. Wu, D. Huang, and Y. Zhang, "Stepwise adaptive convolutional network for fault diagnosis of high-speed train bogie under variant running speeds," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 8389–8398, 2022.

[13] X. Sun, J. Gu, and H. Sun, "Research progress of zero-shot learning," *Applied Intelligence*, vol. 51, pp. 3600–3614, 2021.

[14] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.

[15] 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.

[16] X. Chen, C. Zhao, and J. Ding, "Pyramid-type zero-shot learning model with multi-granularity hierarchical attributes for industrial fault diagnosis," *Reliability Engineering & System Safety*, vol. 240, p. 109591, 2023.

[17] J. Zhao, J. Yue, and C. Zhao, "Learning to better see the unseen: Broad-deep mixed anti-forgetting framework for incremental zero-shot fault diagnosis," *arXiv preprint arXiv:2403.13845*, 2024.

[18] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot learning of object categories," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 01, pp. 1–1, 2013.

[19] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2021–2030, 2017.

[20] G.-S. Xie, Z. Zhang, H. Xiong, L. Shao, and X. Li, "Towards zero-shot learning: A brief review and an attention-based embedding network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1181–1197, 2022.

[21] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks.," *Journal of machine learning research*, vol. 10, no. 1, 2009.

[22] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," *Advances in neural information processing systems*, vol. 22, 2009.

[23] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.

[24] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1d convolutional neural networks and applications: A survey," *Mechanical systems and signal processing*, vol. 151, p. 107398, 2021.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[28] M. B. Sariyildiz and R. G. Cinbis, "Gradient matching generative networks for zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2168–2178, 2019.

[29] M. R. Vyas, H. Venkateswara, and S. Panchanathan, "Leveraging seen and unseen semantic relationships for generative zero-shot learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 70–86,

Springer, 2020.
[30] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 4, pp. 3313–3332, 2021.
[31] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," *arXiv preprint arXiv:2008.12595*, 2020.
[32] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, pp. 214–223, PMLR, 2017.
[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.
[34] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5542–5551, 2018.
[35] G. J. McLachlan, "Mahalanobis distance," *Resonance*, vol. 4, no. 6, pp. 20–26, 1999.
[36] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.
[37] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.
[38] L. Breiman, "Stacked regressions," *Machine learning*, vol. 24, pp. 49–64, 1996.
[39] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32.
[40] J. Martinez-Gil, "A comprehensive review of stacking methods for semantic similarity measurement," *Machine Learning with Applications*, vol. 10, p. 100423, 2022.
[41] L. Su, L. Ma, N. Qin, D. Huang, and A. Kemp, "Fault diagnosis of high-speed train bogie based on spectrogram and multi-channel voting," in *2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS)*, pp. 22–26, IEEE, 2018.
[42] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*, pp. 8162–8171, PMLR, 2021.

Yirui Yin received the B.S. degree from the East China Jiaotong University, Nanchang, China, in 2022. She is currently studying for the Master's degree in the School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China.

Her research interests lie in the areas of fault diagnosis and deep learning.

Deqing Huang received the B.S. and Ph.D. degrees with a major of applied mathematics from the Mathematical College, Sichuan University, Chengdu, China, in 2002 and 2007, respectively. He attended the Department of Electrical and Computer Engineering (ECE), National University of Singapore (NUS), Singapore, in 2006, where he received the second Ph.D. degree with a major in control engineering in 2011. From January 2010 to February 2013, he was a Research Fellow in the Department of Electrical and Computer Engineering of NUS. From March 2013 to January 2016, he was a Research Associate with the Department of Aeronautics, Imperial College London, London, U.K. In January 2016, he joined the Department of Electronic and Information Engineering, Southwest Jiaotong University, Chengdu, China as professor and department head.

His research interests lie in the areas of modern control theory, artificial intelligence and fault diagnosis as well as robotics.

Na Qin received the B.S. degree in electrical technology from the School of Electrical Engineering, Zhengzhou University, Zhengzhou, China, in 2000, and the M.S. degree in electric power system and automation and Ph.D. degree in electrical engineering from the School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China, in 2003 and 2014, respectively. She is currently a professor with the School of Electrical Engineering, Southwest Jiaotong University.

Her research interests include intelligent information processing, fault diagnosis, pattern recognition, federated learning and intelligent systems.

Kang Li received the B.Sc. degree in Industrial Automation from Xiangtan University, Hunan, China, in 1989, the M.Sc. degree in Control Theory and Applications from Harbin Institute of Technology, Harbin, China, in 1992, and the Ph.D. degree in Control Theory and Applications from Shanghai Jiaotong University, Shanghai, China, in 1995. He was also awarded the D.Sc. degree in Science from Queen's University Belfast, UK, in 2015.

Between 1995 and 2018, he worked at Shanghai Jiaotong University, Delft University of Technology and Queen's University Belfast as a research fellow. Between 2002 and 2018, he was a Lecturer, a Senior Lecturer (2007), a Reader (2009) and a Chair Professor (2011) with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, U.K. He currently holds the Chair of Smart Energy Systems at the University of Leeds, UK.

His research interests cover nonlinear system modeling, identification, and control, with applications to smart energy systems, microgrids, and decarbonization of transport systems and manufacturing processes.

Tianwei Wang received his B.S. and M.S.degree in Electronic Information Engineering and Control science and Engineering from Southwest Jiaotong University in 2021 and 2024, repectively. He is currently working towards a Ph.D. degree in Electrical Engineering at Southwest Jiaotong University, Chengdu, China.

His research interests include fault diagnosis, fault tolerant control of rectifier and intelligent information processing.

Ronghua Zong received the B.S. degree from the Sichuan Agricultural University, Ya'an, China, in 2022. She is currently studying for the Master's degree in the School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China.

Her research interests lie in the areas of fault diagnosis and deep learning.