Sound & Vibration 2025, 59(5), 3354.
doi: https://doi.org/10.59400/sv3354

Article

DRSEL: A dual branch feature level ensemble learning framework based on multi sensor data fusion for fault diagnosis of rotating machines

Yuan Zhuang[1,2,3], Wei Qu[4], Junjie Ding[3,5], Minling Pan[4,*], Jiahua Su[1]

[1] College of Mechanical, Naval Architecture & Ocean Engineering, Beibu Gulf University, Qinzhou 535011, China
[2] Guangxi Key Laboratory of Precision Navigation Technology and Application, Guilin University of Electronic Technology, Guilin 541004, China
[3] Shenzhen Junrongyao Precision Hardware Products Co., Ltd., Shenzhen 518000, China
[4] College of Electronic and Information Engineering, Beibu Gulf University, Qinzhou 535011, China
[5] School of Information Science, Guangdong University of Finance & Economics, Guangzhou 510320, China
*** Corresponding author:** Minling Pan, pan_minling@bbgu.edu.cn

CITATION

Zhuang Y, Qu W Ding J, et al. DRSEL: A dual branch feature level ensemble learning framework based on multi sensor data fusion for fault diagnosis of rotating machines. Sound & Vibration. 2025; 59(5): 3354. https://doi.org/10.59400/sv3354

ARTICLE INFO

Received: 9 July 2025
Revised: 4 August 2025
Accepted: 24 August 2025
Available online: 1 September 2025

Abstract: This paper proposes a feature-level ensemble learning framework for fault diagnosis of rotating machinery based on multi-sensor data fusion, aiming to address the inherent limitations of conventional single-model or single-sensor approaches in capturing complex and variable fault characteristics. Firstly, an improved ensemble learner is designed by integrating residual networks with Swin Transformer blocks, enabling the extraction of multi-scale, hierarchical, and complementary features from multi-sensor vibration signals. This design allows the model to capture local fine-grained patterns while simultaneously learning long-range dependencies, thus achieving a more comprehensive and discriminative feature representation. Then, the feature vectors produced by multiple base learners are concatenated to perform feature-level fusion, which effectively leverages the complementary information across heterogeneous sensors and significantly enhances diagnostic accuracy, robustness, and stability. Finally, the proposed framework is validated through two real-world industrial case studies involving a train bogie gearbox and an induction motor, covering diverse operating speeds, load conditions, and fault types. Experimental results demonstrate that the method achieves a fault diagnosis accuracy exceeding 96.88%, markedly outperforming traditional single-model approaches and conventional fusion strategies. Moreover, the framework exhibits strong generalization ability under variable working conditions. These findings highlight its practical applicability in industrial scenarios and underline its potential to support the development of intelligent and reliable predictive maintenance systems.

Keywords: multi-sensor data fusion; feature-level ensemble learning; fault diagnosis; train bogie gearbox; induction motor

1. Introduction

Rotating machinery is extensively applied in various fields, e.g., equipment manufacturing, rail transportation, and energy and power systems [1]. It constitutes the core components of numerous critical devices and serves as the foundation of modern industrial development. Common types of rotating machinery include bearings, gears, and motors, which form the driving, transmission, and execution systems of the majority of industrial equipment. Therefore, ensuring the safe and stable operation of rotating machinery is essential for maintaining the continuity of industrial production and enhancing equipment operational efficiency [2]. Failures in rotating machinery

can result in a wide range of consequences: minor failures may lead to decreased efficiency, energy waste, and increased operational costs, whereas severe failures may cause production accidents, significant economic losses, and even endanger human safety. However, the accurate diagnosis of faults in rotating machinery under industrial conditions is restricted by several challenges. These include: (1) fault characteristics vary considerably across different types of rotating machinery; (2) most existing methods rely on single-sensor data, which provide limited information, are susceptible to interference, and lead to poor diagnostic accuracy and robustness; (3) most models with a single network architecture lack sufficient generalization ability and are not suitable for complex operating conditions and multiple fault types [3]. Consequently, it is of great significance to develop a fault diagnosis method based on multi-sensor signals and ensemble learning, which exhibits broad applicability and strong adaptability to real-world industrial scenarios.

In recent years, the rapid development of deep learning has brought significant transformation to the field of fault diagnosis [4–8]. Leveraging large-scale data and high-performance computing environments, various data-driven fault diagnosis methods based on deep learning have achieved remarkable results in multiple practical scenarios, gradually becoming a research hotspot and development trend in intelligent fault diagnosis. Among these methods, residual learning, Transformer architectures, and their various variants have attracted considerable attention. Qin et al. proposed a dynamic wide-kernel residual network with adaptive symmetric loss for bearing fault diagnosis [9]. Hou et al. introduced a bearing fault diagnosis method based on joint feature extraction using Transformer and ResNet [10]. Xiao et al. developed a Bayesian variational Transformer (Bayesformer) for diagnosing faults in rotating machinery [11]. Lv et al. presented a fault diagnosis approach for bearings that integrates adaptive feature mode decomposition with Transformer networks [12]. Yan et al. proposed LiConvFormer, a lightweight fault diagnosis framework based on separable multi-scale convolution and walk-based self-attention mechanisms [13]. Although these approaches have achieved promising results, they primarily focus on bearing faults and have not demonstrated their effectiveness in diagnosing faults in more structurally complex rotating machinery, e.g., gearboxes and motors. Moreover, methods relying on a single type of network model have certain limitations. For example, Transformer-based models tend to suffer from overfitting when the available training data for fault diagnosis is limited.

Unlike single-component systems, e.g. bearings, industrial equipment often consists of complex multi-coupled systems, e.g., motors and gearboxes, which exhibit diverse fault types. As a result, relying solely on vibration signals from a single sensor leads to insufficient multidimensional information and limited fault feature representation, thereby constraining the model's performance in diagnosing faults in complex systems. The development of data fusion and intelligent sensing technologies has offered promising solutions to this issue. Research on fault diagnosis based on multi-sensor fusion aims to enhance diagnostic performance by integrating multi-source information, achieving feature complementarity, and improving feature expressiveness [14–16]. Among these studies, Wang et al. proposed a multi-sensor

fusion-based intelligent fault diagnosis method for rolling bearings using Variational Mode Decomposition (VMD) and an ultra-lightweight GoogLeNet architecture [17]. He introduced a zero-sample learning model for fault diagnosis of unknown compound faults in train bearings, which utilizes fused vibration and acoustic signals based on label feature vector generation to improve diagnostic accuracy [18]. Zhang et al. proposed a cross-domain fusion network of multimodal data for gearbox fault diagnosis under varying operating conditions [19]. To address the issue of noise interference, Qiu et al. developed a multimodal fusion fault diagnosis approach based on a multi-scale stacked denoising autoencoder and a dual-branch feature fusion network [20]. Xu et al. introduced a multimodal multi-sensor feature fusion algorithm using spiking neural networks [21]. These approaches construct a more comprehensive state representation by fusing multi-sensor data, effectively improving the accuracy and adaptability of fault identification.

Compared to conventional models with simple architectures, ensemble learning effectively enhances the generalization capability and robustness of diagnostic systems by integrating the strengths of multiple base models [22, 23]. This approach demonstrates superior performance, particularly in handling high-dimensional, complex, and multi-condition datasets, and has become one of the critical directions in the development of intelligent diagnostics. In the context of fault diagnosis using ensemble learning, Tong et al. proposed a multi-branch ensemble framework based on multi-scale convolutional neural networks (MSCNNs) as the base learners, which enables the fusion of multi-sensor information for bearing fault diagnosis [24]. Ye et al. introduced a multi-sensor information fusion deep ensemble learning network (MIFDELN) for bearing fault diagnosis [25]. Fu et al. developed a robust fault diagnosis method for bearings by combining ensemble learning with adaptive weight selection [26]. You et al. proposed a fault diagnosis strategy for air conditioning systems based on ensemble learning, incorporating generative adversarial networks (GANs) to enhance the learning of the diagnostic model [27]. Xiao et al. presented a domain-extended meta-ensemble learning approach for the diagnosis of faults in bearings and gearboxes [28]. These studies demonstrate that ensemble learning methods incorporating multi-sensor data have shown excellent performance in the fault diagnosis of components, e.g., bearings. However, in contrast to the popularity of bearing-related studies, similar diagnostic approaches for more complex systems, e.g., motors and gearboxes, remain relatively underexplored.

In order to address the aforementioned issues and bridge the gap in existing research, this paper proposes an ensemble learning-based fault diagnosis framework using multi-sensor data. The base learners in the proposed framework consist of ResNet-18 and Swin Transformer, where each base learner is responsible for feature extraction from a single sensor. This design enables efficient feature extraction and comprehensive information fusion from multiple sensors, thereby improving diagnostic accuracy and robustness. The main contributions of this work are as follows:

(1) A dual-branch ensemble learner is developed as the base learner, composed of ResNet-18 and Swin Transformer. This architecture is designed to enhance global contextual modeling while maintaining sensitivity to local details, thus

enabling more effective mining of key fault features from heterogeneous multi-source data.

(2) A feature-level ensemble learning fault diagnosis framework named DRSEL based on multi-sensor data fusion is proposed. In this framework, each base learner is assigned to a specific sensor for feature extraction and learning. The feature vectors extracted from all base learners are concatenated and passed through a multi-layer perceptron (MLP) and a fully connected layer to perform final fault state classification.

(3) The effectiveness, robustness, and generalization ability of the proposed framework for fault diagnosis in rotating machinery are validated through two case studies: 1 based on a bogie gearbox and another based on an electric motor.

To overcome these challenges, namely, the limited fault feature representation of single-sensor data, the weak robustness under variable operating conditions, and the limited generalization capability of single-model architectures. This paper proposes a novel dual-branch ensemble learning framework named DRSEL. The remainder of the paper is organized as follows: Section 2 reviews the theoretical foundations of the core models employed, including ResNet-18 and Swin Transformer; Section 3 elaborates on the proposed DRSEL framework, including data preprocessing, model architecture, and the feature-level fusion strategy; Section 4 presents comprehensive experimental evaluations on two representative datasets to validate the effectiveness and robustness of the proposed method; Section 5 discusses parameter sensitivity and ablation studies to assess the contribution of each component; and Section 6 concludes the paper and outlines potential directions for future research.

2. Theoretical background

2.1. ResNet

ResNet was proposed by He et al. [29] from Microsoft Research in 2015. It is primarily composed of multiple residual blocks. Let x denote the input vector of a residual block, and $F(\cdot)$ represent the residual mapping function. The component $H(x) = F(x) + x$ convolution on the input x followed by a ReLU activation function. The term w_i denotes the layer weights of the residual block, as shown in Equation (1). ResNet has demonstrated superior performance in deep learning tasks and has achieved promising results in the field of mechanical fault diagnosis.

$$H(x) = F(x, \{w_i\}) + x \tag{1}$$

2.2. Swin transformer

The Swin Transformer employs a hierarchical architecture analogous to that of convolutional neural networks (CNNs) and is composed of four sequential stages [30]. **Figure 1** depicts the overall structure of the Swin Transformer network as well as the internal configuration of the Swin Transformer Block. As illustrated, the input image is initially divided into non-overlapping patches via a patch partitioning operation. These patches are subsequently processed by a linear embedding layer to align the

channel dimensions. Beginning from Stage 2, each subsequent stage incorporates a patch merging operation to progressively downsample the feature map resolution. Specifically, in the first patch merging layer, features from neighboring patches are concatenated along the channel dimension. The output resolutions of the four stages are denoted as $H/4 \times W/4$, $H/8 \times W/8$, $H/16 \times W/16$, $H/32 \times W/32$ respectively, and their spatial resolutions are consistent with those in typical CNN architectures. Consequently, this hierarchical structure enables the backbone network to be readily adapted for various visual tasks in existing methods.

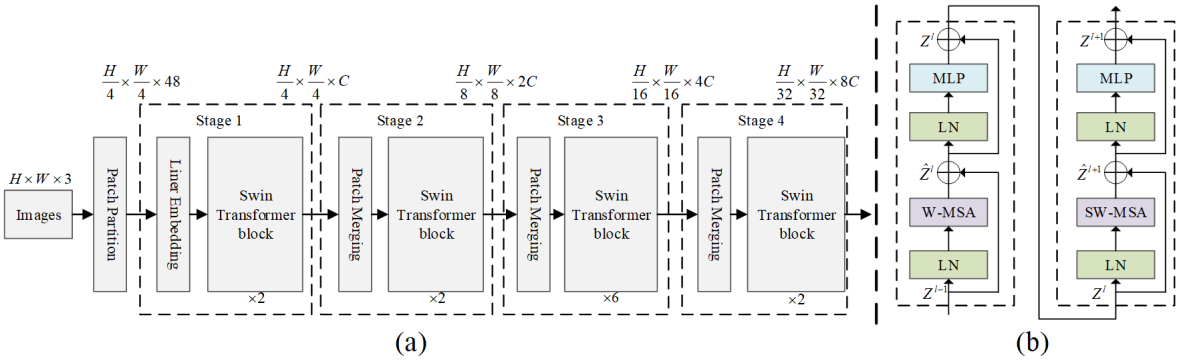

Figure 1. Structure of **(a)** Swin Transformer and **(b)** Swin Transformer block.

The Swin Transformer block initially partitions the image into non-overlapping windows and applies window-based multi-head self-attention (W-MSA) to compute self-attention within each local window, thereby achieving linear computational complexity. While this approach effectively reduces computational overhead, it inherently limits interactions across different windows. To overcome this drawback, the shifted window partitioning strategy, referred to as shifted window multi-head self-attention (SW-MSA), is introduced. This mechanism enables information exchange between adjacent windows and significantly enlarges the receptive field. The computational formulation of the Swin Transformer block is expressed as follows:

$$\begin{aligned} \widehat{Z}^{l} &= W - MSA(LN(Z^{l-1})) + Z^{l-1}, \\ Z'^{l} &= MLP(LN(\widehat{Z}^{l})) + \widehat{Z}^{l}, \\ \widehat{Z}^{l+1} &= SW - MSA(LN(Z'^{l})) + Z'^{l}, \\ Z^{l+1} &= MLP(LN(\widehat{Z}^{l+1})) + \widehat{Z}^{l+1}. \end{aligned} \tag{2}$$

Here, $\widehat{Z}^{l}$ and Z'^{l} represent the feature outputs of the W-MSA module and the MLP module, respectively. W-MSA and SW-MSA refer to the conventional layer and the shifted window multi-head self-attention layer, respectively.

3. Proposed method

3.1. Processing of multi sensor data

Multisensor signal fusion can enhance classification performance by integrating data from different physical dimensions, achieving information complementarity and redundancy. In the proposed method, both the Swin Transformer and ResNet-18

networks are incorporated. Since the Swin Transformer requires 2D image inputs, and the ResNet-18 performs better with 2D image signals than with 1D time-series data, 2D time-frequency images are used as the unified input format. Therefore, during the preprocessing phase, the continuous wavelet transform (CWT) will be used to convert the 1D time-series signals into 2D time-frequency signals.

CWT as an efficient time-frequency analysis tool, possesses excellent time-frequency localization properties. It can perform multi-scale decomposition of non-stationary signals, preserving time information while extracting frequency variation features. This makes it particularly suitable for processing complex signals such as motor faults and bearing impacts, and it has been widely applied. The preprocessing steps employed in this paper are as follows:

Step 1: Since the sampling frequencies of different sensors may vary, the original signals with higher sampling frequencies are downsampled using an anti-aliasing filter. After downsampling, each signal will have the same number of data points.

Step 2: Multi-sensor signals are sampled with a length of 2048 and a step size of 2048, with no overlapping sampling. The sampling length of 2048 is set primarily considering the length of the dataset and the time for 1 full rotation of the measured component (In Case 1, the total available data length is only 640,000 points, and the data length of Case 2 is 420,000 points). This ensures both the coverage of a complete cycle and an adequate number of samples. Furthermore, too short a sampling length would reduce the time-frequency resolution after CWT and increase edge effects, which could affect the completeness and accuracy of fault feature extraction.

Step 3: Each segment of the signal from every channel is normalized within a range of 0 to 1. Then, the CWT is applied to the samples. The wavelet basis used is the Complex Morlet Wavelet, which is a combination of a Gaussian envelope and a sine wave, offering good time-frequency localization properties. It is especially suitable for processing non-stationary signals and is widely used in time-frequency analysis and fault diagnosis. The expression for the wavelet is as follows. The scale range is set to (1, 6). This results in the time-frequency feature map for each sample, as shown in formula (4).

$$\psi(t) = \exp\left(-\frac{t^2}{2\sigma^2}\right) \cdot \exp(i\omega_0 t) \tag{3}$$

$$M(W, H, C) = \int_{-\infty}^{+\infty} x(t) \cdot \psi^* \left(\frac{t-b}{a}\right) dt \tag{4}$$

Here, $M(W, H, C)$ represents the obtained time-frequency feature map, while *W, H, C* correspond to the width, height, and number of channels of the matrix, respectively. The number of channels is set to 3. σ controls the width of the Gaussian envelope, ω_0 is the center frequency, which controls the frequency of the wavelet. $\psi(t)$ represents the Complex Morlet Wavelet basis, which captures the local features of the signal, a denotes the scale, and b is the translation number. t represents time, consistent with the definition in the continuous wavelet transform where $\psi(t)$ is analyzed in the time domain.

The CWT scale range is set to (1, 6) based on the frequency characteristics of

the acquired signals and the properties of the Morlet wavelet. Given the sampling frequency of 42 and 64 kHz for Cases 1 and 2, this scale range roughly corresponds to a frequency band of 1.9–13 kHz and 2.7–16 kHz, respectively. This band effectively covers the mid-to-high frequency components where fault-related features such as impacts and transients are most prominent in motor and bearing signals. Additionally, this choice offers a good trade-off between time and frequency resolution while maintaining computational efficiency.

Step 4: The obtained time-frequency feature map is resized to dimensions (224, 224) for the width and height to ensure an appropriate input size of (224, 224, 3).

Finally, the resulting input feature map is divided into training, validation, and testing sets. The division ratio will be determined based on the specific requirements of the task.

3.2. DRSEL and base learner

The base learner of DRSEL consists of ResNet-18 and Swin Transformer, forming a dual-branch feature-level ensemble learning module with strong generalization and feature extraction capabilities, shown in **Figure 2**. ResNet-18 is lightweight in structure and can efficiently extract local spatial features, making it suitable for modeling low-level visual information such as texture and edges. Swin Transformer, on the other hand, leverages a window-based multi-head self-attention mechanism to model global dependencies in images and is adept at capturing long-range correlations and structural information. The feature-level fusion of the 2 enables the network to fully exploit both local and global features, thereby enhancing its robustness and discriminative ability under complex backgrounds and variable operating conditions. Moreover, Transformer-based architectures typically rely heavily on large-scale datasets, whereas ResNet demonstrates stable performance in small-sample scenarios. The combination of both networks improves the adaptability of the proposed method to limited data conditions. During the forward propagation process, each batch is configured with input samples of size (32, 224, 224, 3), where 32 denotes the batch size, 224 represents the image height and width, and 3 indicates the number of channels. After passing through the 2 branches, each sample yields two 128-dimensional feature vectors extracted by ResNet-18 and Swin Transformer, respectively. These vectors are then concatenated to form a 256-dimensional fused feature vector for each sample.

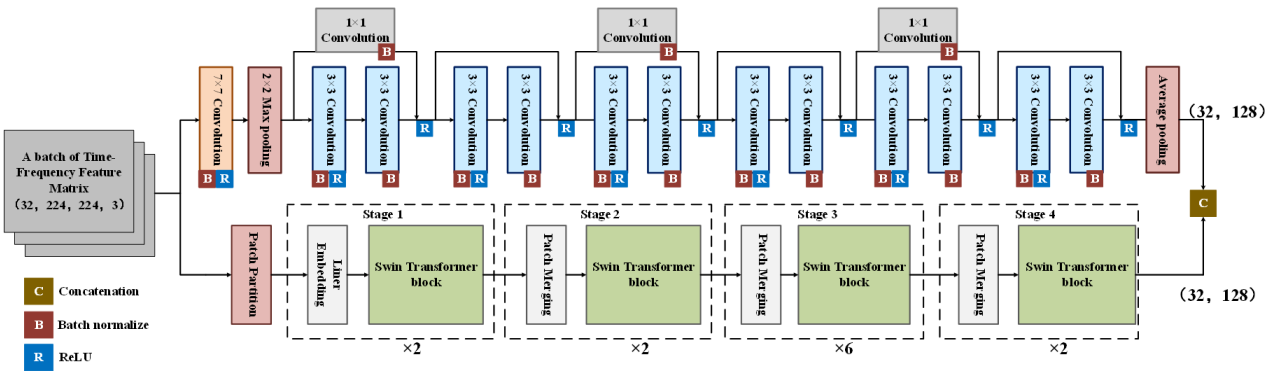

Figure 2. Structure of base learner for DRSEL.

DRSEL is composed of two or more base learners (in this study, only two are utilized), and the number of learners can be extended according to the number of sensors, as shown in **Figure 3**. After a batch of multi-sensor samples is input into their corresponding base learners, the data of each sample from each sensor is mapped to a 256-dimensional feature vector. Consequently, the output of 1 batch is of size (32, 256), where 32 denotes the batch size and 256 represents the feature vector of a single sample. These outputs are then concatenated along the feature dimension, resulting in a representation of size (32, 512), which is subsequently fed into an MLP. The MLP comprises 2 linear layers: the first layer takes an input of size 512 and outputs a 1024-dimensional vector. This layer is followed by batch normalization to stabilize feature distribution, reduce overfitting, and accelerate training convergence. A ReLU activation function is applied to introduce non-linearity, thereby enhancing noise robustness and feature expressiveness. The second linear layer reduces the feature dimension to 256. Finally, a fully connected layer is used to perform classification.

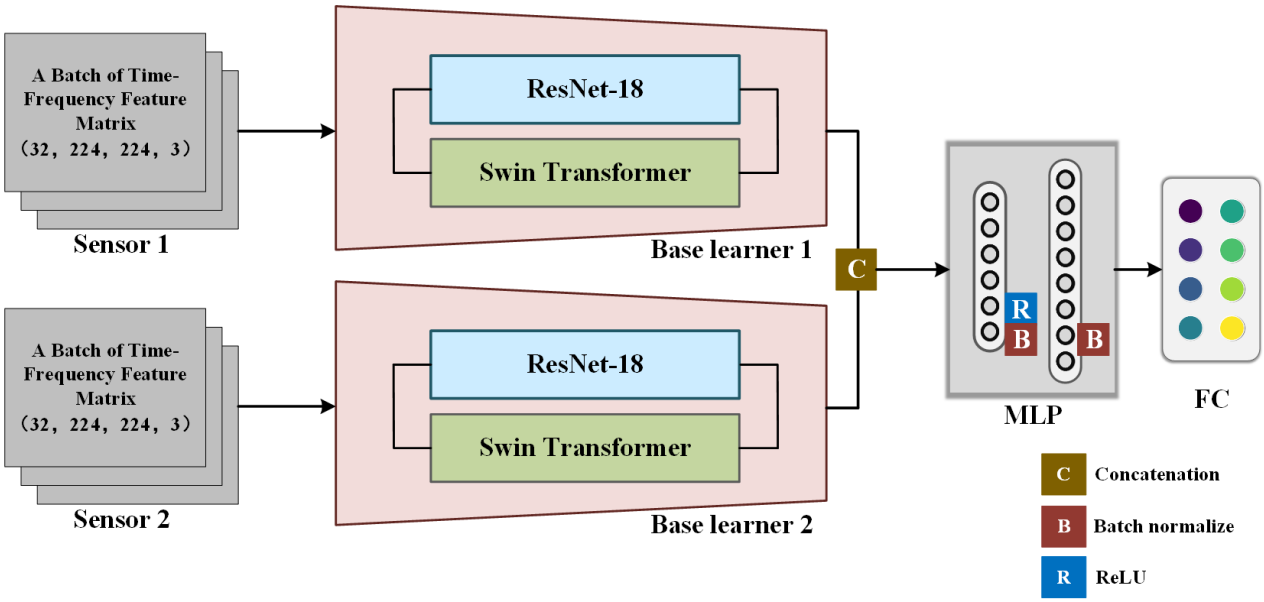

Figure 3. Structure of DRSEL.

3.3. Overall framework

The architecture of the proposed DRSEL framework is depicted in **Figure 4**. In the initial stage, multi-sensor signals are subjected to a series of preprocessing operations, including signal segmentation and transformation via continuous wavelet transform. Subsequently, the processed data are partitioned into training, validation, and test subsets based on the specific requirements of the diagnostic task. The second stage involves model training and parameter optimization using the training and validation sets. In the final stage, the diagnostic capability of the proposed framework is thoroughly evaluated and analyzed on the test set.

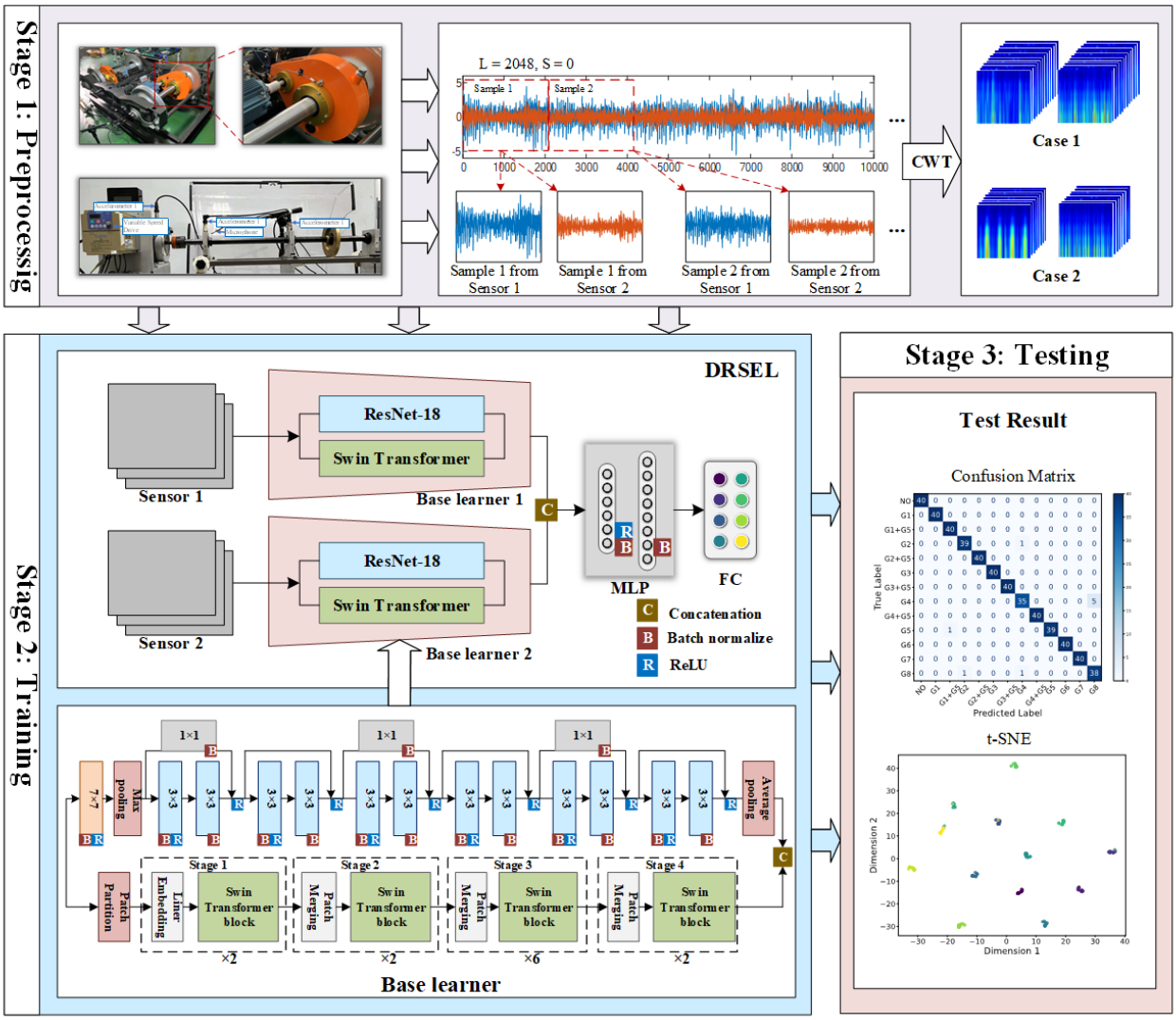

Figure 4. Overall flowchart of the proposed fault diagnosis method.

4. Experiment and result analysis

This section aims to validate the effectiveness of DRSEL in rotating machinery fault diagnosis through 2 case studies. The fault data employed in these cases are drawn from the University of Ottawa's induction motor fault dataset and the BJTU-RAO dataset from Beijing Jiaotong University, which encompasses common motor faults as well as gear and bearing faults occurring in train axle boxes.

Due to the length constraints of the datasets, each sample in both cases is fixed at 2048 data points without overlapping sampling. After the continuous wavelet transform (CWT), all samples are converted into time-frequency feature maps with dimensions of 224 × 224. For each fault category, 200 samples are selected, with 120 used for training, 40 for validation, and 40 for testing. The ratio of dataset splitting is referred to literature [10,16]. Given the limited size of the training set, data augmentation, e.g., random horizontal flipping, random rotation, and random vertical flipping, is employed to enhance the diversity of the training data. During the training process, the number of epochs is uniformly set to 30. The optimizer employed is ADAM, and the loss function is categorical cross-entropy. The initial learning rate is set to 0.001 and is adjusted using

a cosine annealing schedule. Selection of these hypermeters was guided by literatures [10,16], the validation of these values can be referred to in the subsequent experiments.

The hardware configuration used in this study includes an Intel i7-13700KF CPU and an NVIDIA RTX 4080 GPU. All experiments are implemented in Python using the PyTorch framework. Each experiment is conducted 5 times, and the reported results represent the average of these 5 independent runs.

The comparative methods include 5 widely adopted supervised learning approaches, as detailed below:

Swin Transformer: A supervised learning method that introduces local window attention and a shifted window mechanism to effectively model the hierarchical structure and local features of images.

ResNet-18: A supervised learning method that mitigates the vanishing gradient problem by incorporating deep residual connections, thereby enhancing fault classification performance.

ShuffleNet v2: A lightweight neural network architecture that optimizes computational efficiency through channel grouping and channel shuffling mechanisms while maintaining strong feature representation capabilities.

SqueezeNet: A parameter-efficient convolutional network that employs "squeeze-and-expand" modules to significantly reduce the number of parameters while preserving discriminative power.

AlexNet: A classical convolutional neural network that enhances nonlinear representational capacity through mechanisms including ReLU activation, local response normalization, and dropout.

4.1. Case 1

To validate the effectiveness of DRSEL in rotating machinery fault diagnosis, the reduction gearbox data from the BJTU-RAO dataset was employed as Case 1 for testing [31]. The experimental platform simulating the fault conditions of the metro train bogie transmission system is illustrated in **Figure 5**. This platform is a scaled-down (1:2) and simplified version of an actual metro bogie. The single power transmission chain in the platform consists of a motor, a reduction gearbox, and an axle box. The transmission chain is driven by a 3-phase asynchronous AC motor, with the motor speed regulated by a frequency converter. The load is applied through a hydraulic system. The gearbox uses helical gears, where the driving gear has 16 teeth and the driven gear has 107 teeth. The supporting bearing model for the driving gear is HRB 32305. The dataset includes 13 operational conditions, comprising 8 single-fault conditions and 4 compound-fault conditions. Furthermore, the experiments were conducted under 3 different rotational speed conditions: 20 Hz, 40 Hz, and 60 Hz, as summarized in the **Table 1** below.

4.1.1. Performance comparison and analysis

Figure 6 illustrates the performance comparison of 6 methods during the training process under the operating condition of 40 Hz rotational speed. It can be observed that DRSEL, ResNet-18, and ShufflenNet v2 exhibit similar convergence speeds, with their training accuracy reaching 100% after approximately 11 epochs. In contrast, the remaining 3 methods approach stability only around the 30th epoch, and their accuracy

remains around 85%. However, it is noteworthy that when evaluating validation accuracy, DRSEL ultimately stabilizes at approximately 98%, while ResNet-18 and ShufflenNet v2 stabilize at around 93% and 84%, respectively. These results indicate that, among all the methods, DRSEL demonstrates superior feature extraction and fitting capabilities, enabling it to effectively learn patterns and regularities within the training data.

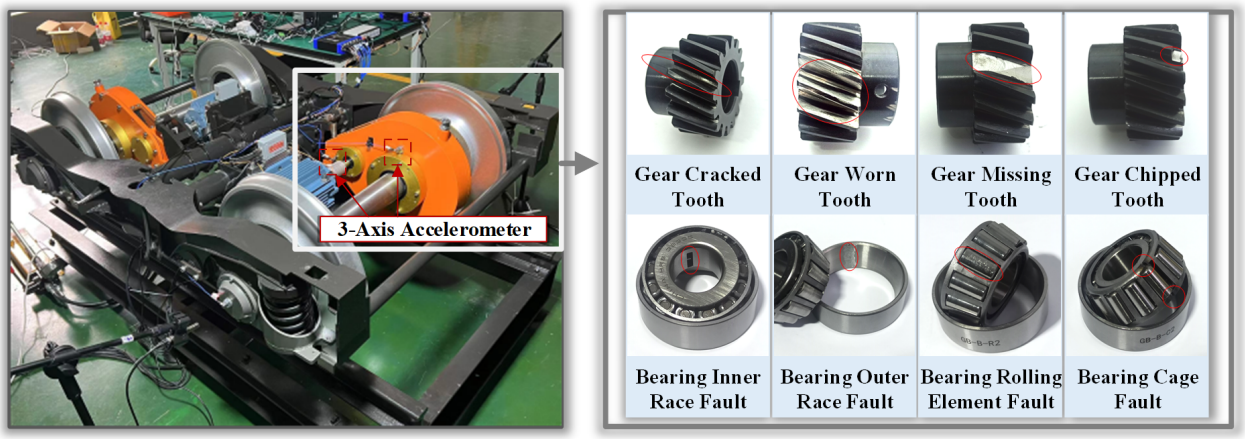

Figure 5. BJTU-RAO bogie fault simulation test bench and gearbox fault types.

Table 1. The description of different fault types and labels.

Heath states	Fault type	Labels	Working condition
Normal	Single Component	G0	
Gear Cracked Tooth	Single Component	G1	
Gear Worn Tooth	Single Component	G2	
Gear Missing Tooth	Single Component	G3	
Gear Chipped Tooth	Single Component	G4	
Bearing Inner Race Fault	Single Component	G5	
Bearing Outer Race Fault	Single Component	G6	20Hz, 40Hz, 60Hz rotating speed
Bearing Rolling Element Fault	Single Component	G7	
Bearing Cage Fault	Single Component	G8	
Gear Cracked Tooth+ Bearing Inner Race Fault	Compound Fault	G1+G5	
Gear Worn Tooth+ Bearing Inner Race Fault	Compound Fault	G2+G5	
Gear Missing Tooth+ Bearing Inner Race Fault	Compound Fault	G3+G5	
Gear Chipped Tooth+ Bearing Inner Race Fault	Compound Fault	G4+G5	

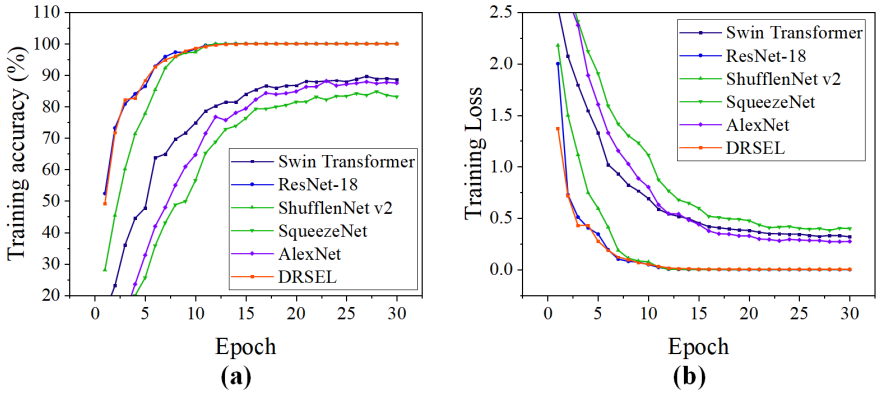

Figure 6. Training **(a)** accuracy and **(b)** loss curve.

To further evaluate the performance of the proposed method, two key metrics—test accuracy and F1 score were calculated, as shown in **Table 2**. DRSEL achieved the highest test accuracy, outperforming the second-best method, ResNet-18, by approximately 4.84%. Although both ResNet-18 and ShufflenNet v2 converged rapidly on the training set, they failed to achieve high classification accuracy on the test set, likely due to weak generalization capabilities and the occurrence of overfitting. In contrast, DRSEL achieved a test accuracy of 98.26%, further validating its strong feature extraction and fitting capabilities, as well as its excellent classification performance.

Table 2. Performance comparison under 40Hz rotating speed.

Methods	Swin transformer	ResNet-18	ShufflenNet v2	SqueezeNet	AlexNet	DRSEL
Accuracy (%)	84.81	93.42	84.23	83.50	82.69	98.27
F1 Score	0.847	0.934	0.842	0.845	0.823	0.983

To gain an intuitive understanding of the performance of the 6 methods under 3 different rotational speeds, tests were conducted for all 6 methods, and the results are shown in **Figure 7**. It can be observed that DRSEL achieved test accuracies of 96.88%, 98.27%, and 98.15% at the 3 different rotational speeds, outperforming the other 5 methods. In contrast, the other 5 methods exhibited test accuracies below 85% at all 3 rotational speeds, with the exception of ResNet-18. This may be due to the following reasons: (1) compared to multi-sensor data, local or single-dimensional information may lead to missing key features, thereby limiting the comprehensiveness and accuracy of fault identification and weakening the anti-interference capability; (2) compound faults cause the fault signals to overlap, resulting in mixed features. If the model's robustness is weak, this will lead to a decline in feature extraction and classification accuracy. In this context, DRSEL, which integrates multi-sensor data and ensemble learning, offers significant advantages.

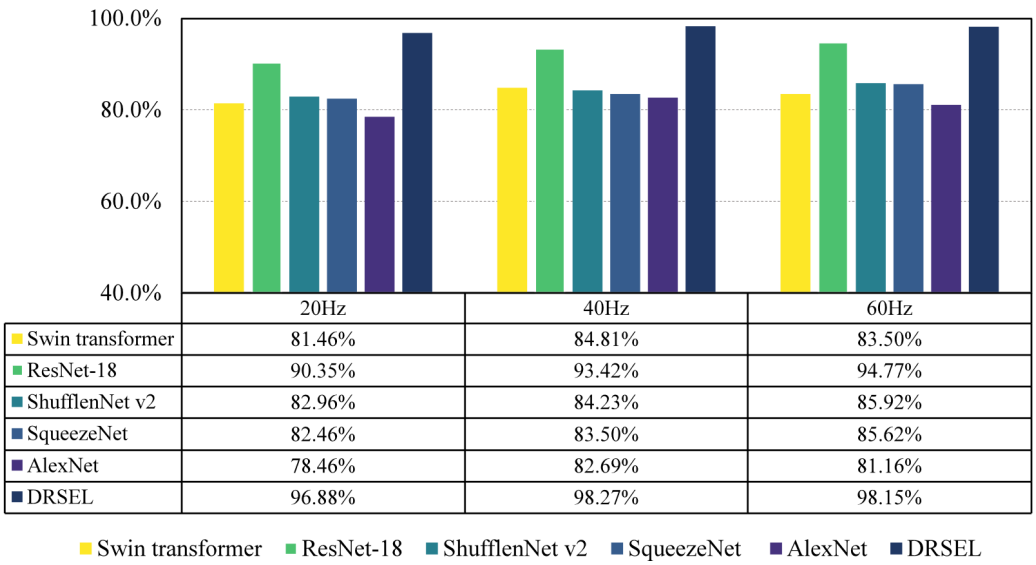

Figure 7. Comparison of experiment results under different rotating speeds for Case 1.

4.1.2. Measurement of generalization ability

In this test, the focus is on evaluating the model's generalization capability. Since the BJTU-RAO dataset does not include data under varying rotational speeds, data from the 20Hz, 40Hz, and 60Hz speeds were mixed to form data under varying speeds, and then randomly shuffled to simulate conditions with fluctuating rotational speeds. The 6 methods were then tested, and the results are shown in **Figure 8**. It can be observed that AlexNet, SqueezeNet, and ShufflenNet v2 exhibited a significant decrease in accuracy compared to the single rotational speed tests, while DRSEL maintained an accuracy of 97.46%, with no noticeable decline. This indicates that DRSEL has a stronger generalization ability compared to these methods. This is because DRSEL uses 2 learners to extract and integrate fault features from different sensors, achieving information complementarity at the feature level, which enhances its representational power. Compared to a single feature source, DRSEL is capable of more comprehensively modeling fault features, improving the model's ability to identify complex operating conditions and compound faults, thereby further enhancing its generalization performance and diagnostic accuracy.

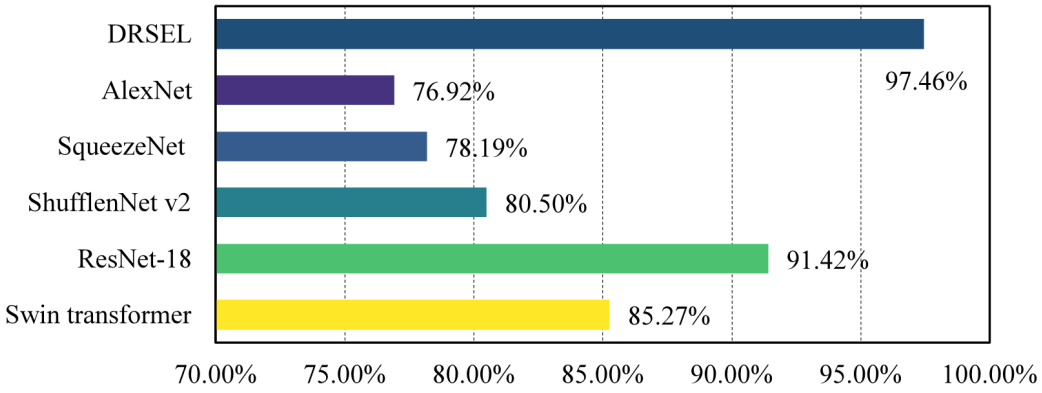

Figure 8. Comparison of experiment results under variable rotating speed for Case 1.

4.1.3. Visualization analysis

Figures 9 and **10** present the confusion matrices, offering an intuitive depiction of each model's classification performance across various fault types. The 2D confusion matrix visualizations under the 40 Hz rotational speed condition demonstrate that the proposed method consistently exhibits higher average diagonal values compared to the other approaches. This indicates a stronger capability in distinguishing fault categories and achieving superior classification accuracy. To visually demonstrate the classification performance of the proposed method, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied to analyze the diagnostic results of the concentrated model. The results for all 6 methods are shown in the figure. It is observed that, except for ResNet-18 and DRSEL, the clustering maps for the other methods exhibit significant overlap, making it difficult to delineate the classification boundaries, especially for G4 and G8. On the other hand, DRSEL clearly separates nearly all categories, further confirming its feasibility and effectiveness.

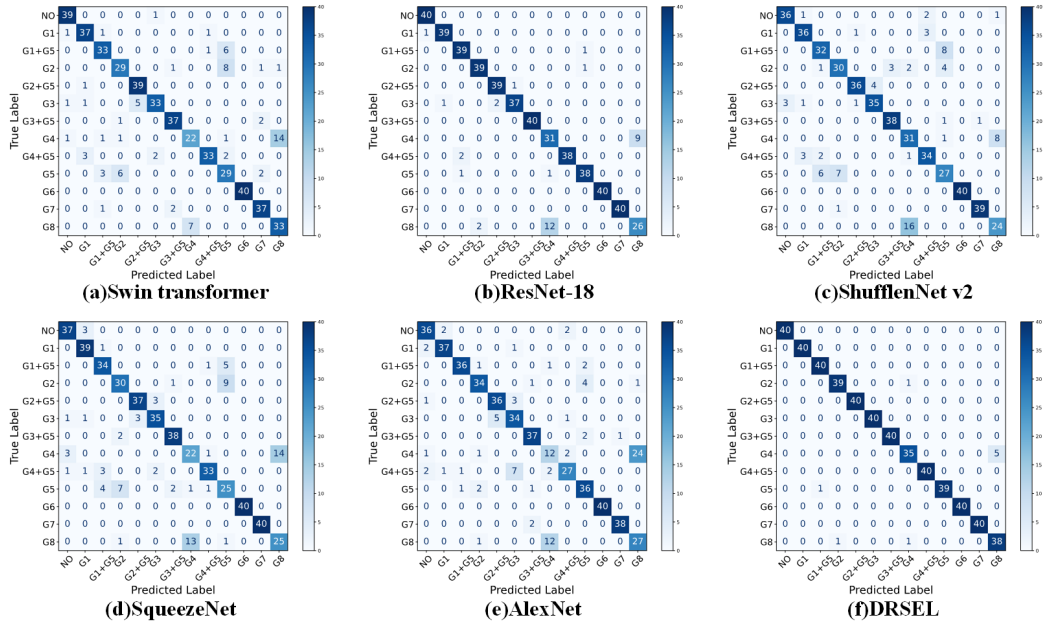

Figure 9. The confusion matrices under 40Hz rotating speed.

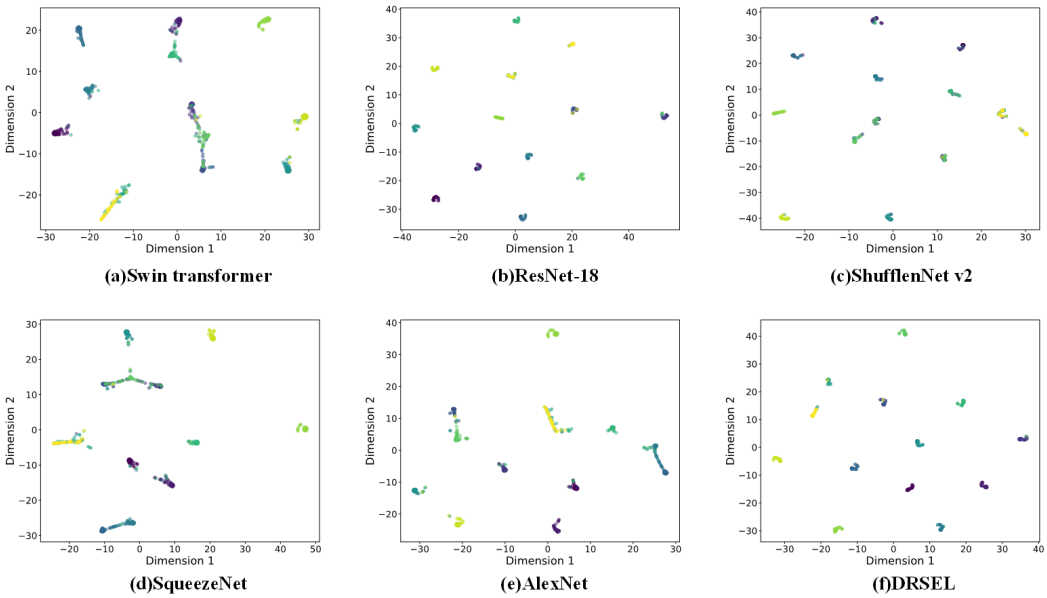

Figure 10. The t-SNE results under a 40Hz rotating speed.

4.2. Case 2

This case study validates the effectiveness of DRSEL for electric motor fault diagnosis using the University of Ottawa constant and variable speed electric motor vibration and acoustic fault signature dataset [32]. The experimental setup is shown in **Figure 11**. The motor model used is the Marathon Electric D396, with bearing model 6205. The dataset includes 3-channel vibration signals (623C01) and a single-channel acoustic signal (130F20), with a sampling frequency of 42,000 Hz. The dataset design includes 4 different motor speeds for each fault type, covering both constant and variable speed conditions. Data collection involved a total of 8 motors of the same model, including 1 healthy motor and 7 faulty motors. These motors correspond to 2 types of electrical faults (voltage unbalance and single-phase operation, stator winding

faults), 4 types of mechanical faults (rotor unbalance, rotor misalignment, bowed axis, bearing faults), and 1 type of electrical-mechanical coupled fault (broken bar fault). The experiment was conducted under 3 different constant speed conditions (15Hz, 30Hz, 45Hz) and 2 variable speed conditions, namely the uniform acceleration condition (15Hz to 45Hz) and the uniform deceleration condition (45Hz to 15Hz), as shown in **Table 3**.

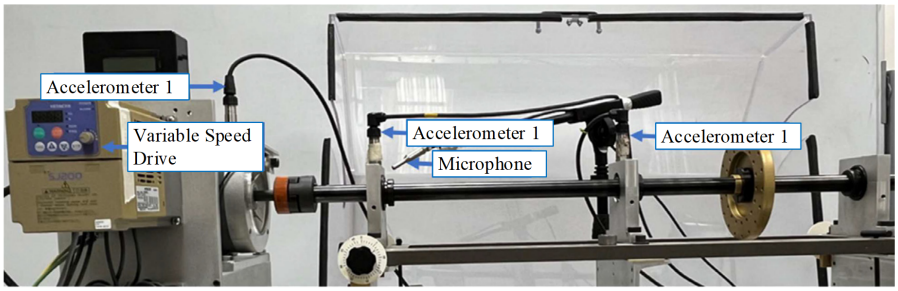

Figure 11. Experiment system for the OU motor dataset.

Table 3. The description of different fault types and labels.

Heath states	Fault type	Labels	Working condition
Normal	\	NO	15Hz, 30Hz, 45Hz rotating speed
Rotor Unbalance	Mechanical	RU	
Rotor Misalignment	Mechanical	RM	
Stator Winding Fault	Mechanical	SW	
Voltage Unbalance and Single Phasing	Electrical	VU	
Bowed Axis	Mechanical	BA	
Rotor Bar Broken	Electromechanical Coupling	RBB	
Bearing Faults	Mechanical	BF	

4.2.1. Performance comparison and analysis

Figure 12 records the performance of the 6 methods during the training process under the 45Hz operating condition. Although all 6 methods achieve 100% training accuracy after 30 epochs, it is evident that DRSEL converges the fastest, reaching 100% accuracy within just 6 epochs. This is further corroborated by the loss curve shown in (b), which also highlights the rapid convergence of DRSEL. This once again demonstrates its strong feature extraction and fitting capabilities, enabling it to efficiently learn the patterns and regularities within the training data.

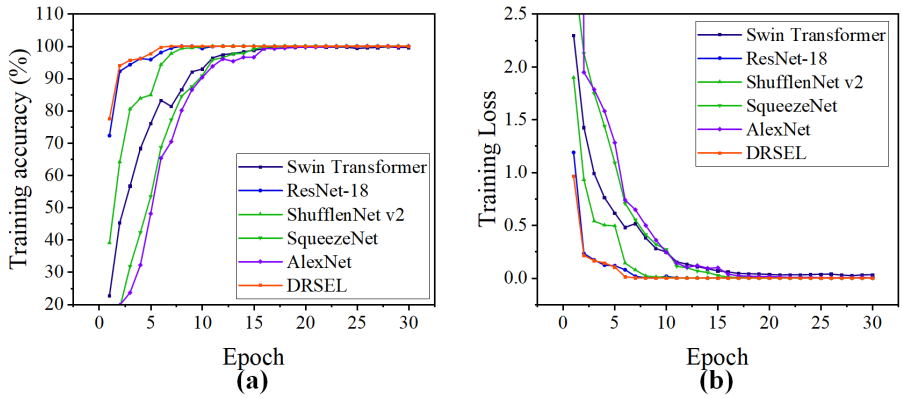

Figure 12. Training **(a)** accuracy and **(b)** loss curve.

Meanwhile, in order to make a more accurate comparison of the performance of the 6 methods, the methods were tested, and the testing accuracy and F1 scores were recorded, as shown in **Table 4**. In this case, all methods performed better than in Case 1, likely due to the presence of compound faults in Case 1. DRSEL achieved 100% accuracy in all 5 tests, further demonstrating its stability and exceptional feature extraction ability.

Table 4. Performance comparison under 45Hz rotating speed.

Methods	Swin transformer	ResNet-18	ShufflenNet v2	SqueezeNet	AlexNet	DRSEL
Accuracy (%)	96.88	99.38	97.19	97.81	92.81	100
F1 Score	0.969	0.993	0.972	0.978	0.928	1

Furthermore, the 6 methods were tested under 3 different constant speed conditions, and the results are shown in **Figure 13**. Although all 6 methods achieved over 90% accuracy in all tests, which may be attributed to the clear differences in features between different faults, it is noteworthy that DRSEL achieved 100% accuracy in all 15 tests across the 3 different speeds. This further validates its stability and exceptional performance. This can be attributed to (1) its use of multi-sensor input, integrating vibration and acoustic signals to complement multi-modal features, and (2) the ensemble learning approach that combines dual model feature information, enhancing feature representation and discriminability, thereby improving feature extraction capability.

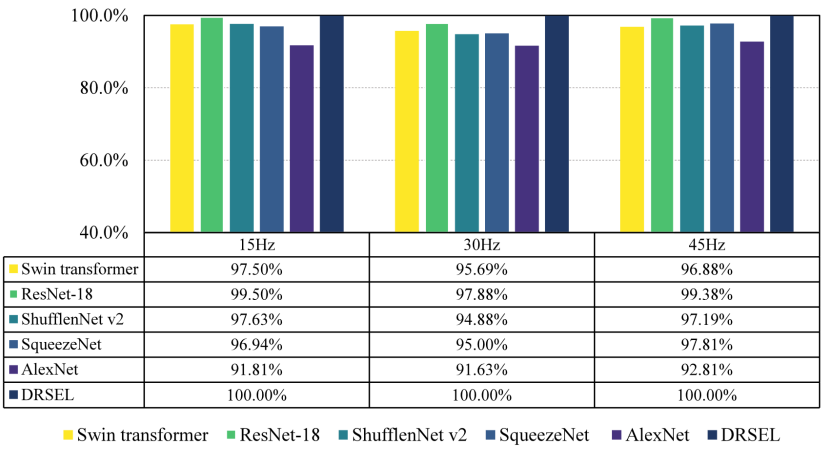

Figure 13. Comparison of experiment results under different rotating speeds for Case 2.

4.2.2. Measurement of generalization ability

To further test the model's generalization ability, the fault diagnosis performance of the motor under variable speed operation was evaluated in the 15Hz-45Hz ramp-up and 45Hz-15Hz ramp-down conditions. The data for these 2 conditions were split into training, validation, and test sets in a 3:1:1 ratio, ensuring that each set included both the ramp-up and ramp-down processes. The 6 methods were then tested, and the results are shown in **Figure 14**. It is evident that, compared to constant speed conditions, the accuracy of all 6 methods decreased under variable speed conditions. Among them, SqueezeNet showed the largest drop, likely due to its lightweight architecture that uses a large number of 1 × 1 convolutions to significantly reduce the number of parameters.

This may have compromised its ability to extract deep semantic information, causing it to struggle in conditions with significant variations in operating speed and complex feature differences, resulting in insufficient discriminative information capture. On the other hand, DRSEL remained stable at 99.88% and 99.94%, demonstrating that its hierarchical, multi-scale representation, combined with deep local features, helps build richer feature expressions for multimodal information. This makes DRSEL more adaptable to the diversity and uncertainty in complex operating conditions.

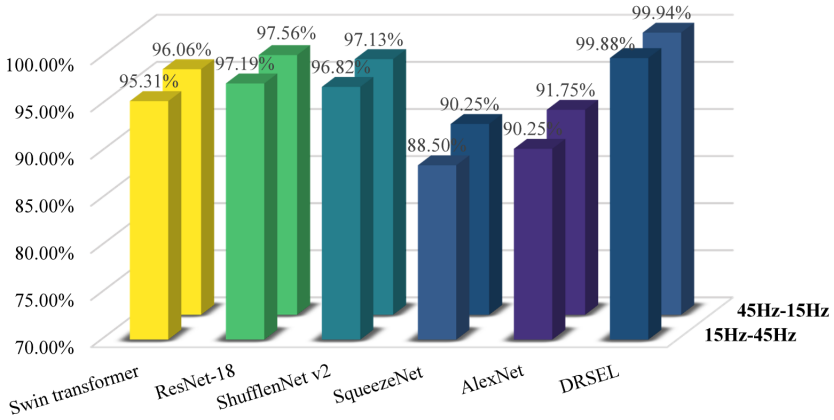

Figure 14. Comparison of experiment results under variable rotating speed for Case 2.

4.2.3. Visualization analysis

The confusion matrix and t-SNE clustering analysis shown in **Figures 15** and **16**, obtained under the 45Hz operating condition, clearly show that DRSEL achieves distinct separation for each fault category. In the t-SNE clustering map, the boundaries between categories are clear, and the clusters are tightly packed. In contrast, other methods show some misclassified categories, primarily concentrated around RU (rotor unbalance), RM (rotor misalignment), VU (voltage unbalance), and RBB (rotor bearing fault). This could be due to the fact that RU and RM are both rotor faults, and misalignment and imbalance share similar vibration characteristics. Similarly, VU and RBB represent electrical and electromechanical coupling faults, which exhibit weaker vibration features, leading to difficulties in distinguishing them from other faults.

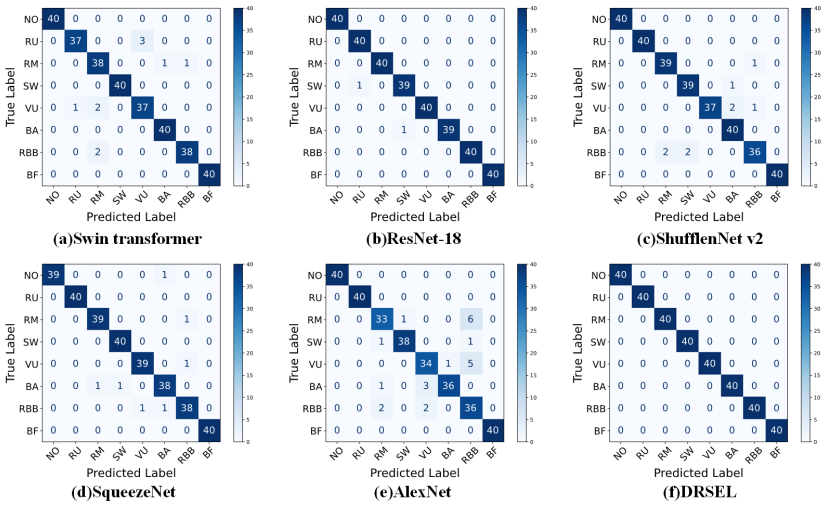

Figure 15. The confusion matrices under 45Hz rotating speed.

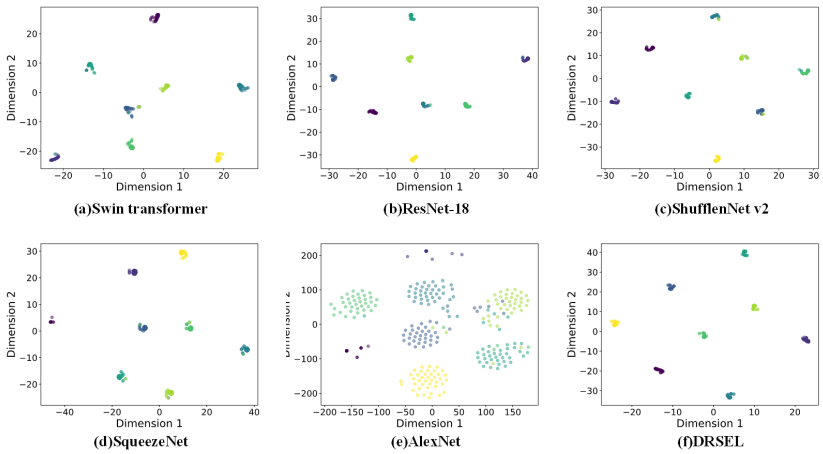

Figure 16. The t-SNE results under a 45Hz rotating speed.

5. Discussion

5.1. Sensitivity analysis on parameters

Parameter sensitivity analysis typically focuses on factors like the number of training iterations, the proportion of training to testing samples, and the value of the learning rate, to explore their impact on the final test accuracy of the model. Since the input size for the Swin Transformer is typically 224 × 224, this section does not investigate the effect of input sample size. The testing follows the same configuration as in Case 1, conducted at a 40Hz rotational speed, with the only changes being the number of training iterations and the size of the training sample, while the sample sizes for the test and validation sets and other parameters remain unchanged. Each experiment is conducted 5 times, and the reported results represent the average of these 5 independent runs.

Figure 17 illustrates the impact of varying training sample sizes and the number of iterations on the final test accuracy. It is evident that the accuracy increases as the number of training samples increases; however, when the sample size exceeds 120, i.e., when the ratio of training to testing samples exceeds 3:1, the improvement becomes marginal. Moreover, when the number of training samples decreases to 40, there is no significant decline in accuracy, further demonstrating the strong feature extraction capability and robustness of DRSEL. Additionally, variations in the number of training iterations have no substantial effect on DRSEL, as the model stabilizes and converges by the 12th epoch, meaning that increasing the number of iterations beyond this point does not impact DRSEL's performance.

Figure 18 illustrates the model's training accuracy, loss, and test accuracy under different learning rates. As shown in **Figure 18a,b**, smaller learning rates generally lead to faster convergence within fewer iterations. In contrast, when the learning rate is set to 0.01, the model fails to reach stability even after 30 epochs. However, as depicted in **Figure 18c**, a learning rate of 0.0001 results in a test accuracy of only 93.46%, indicating that the model converged too quickly and failed to sufficiently learn the underlying fault features. Therefore, considering both convergence behavior and generalization performance, a learning rate of 0.001 was selected as the optimal setting.

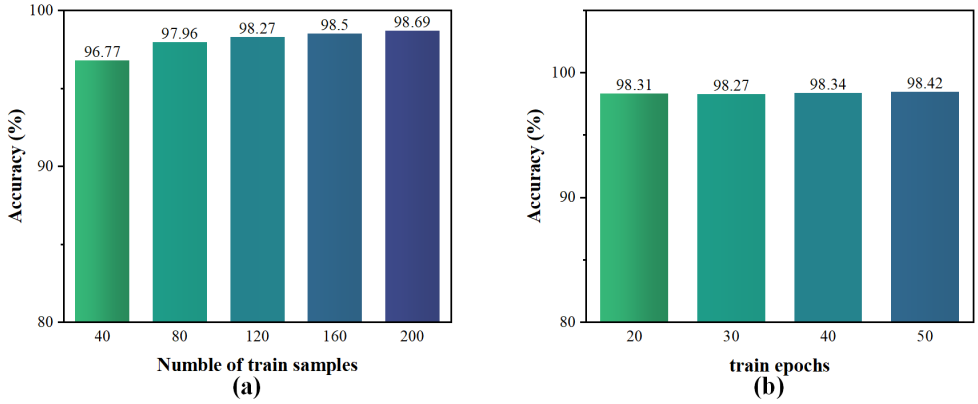

Figure 17. Sensitivity analysis on different **(a)** train sample numbers and **(b)** epochs.

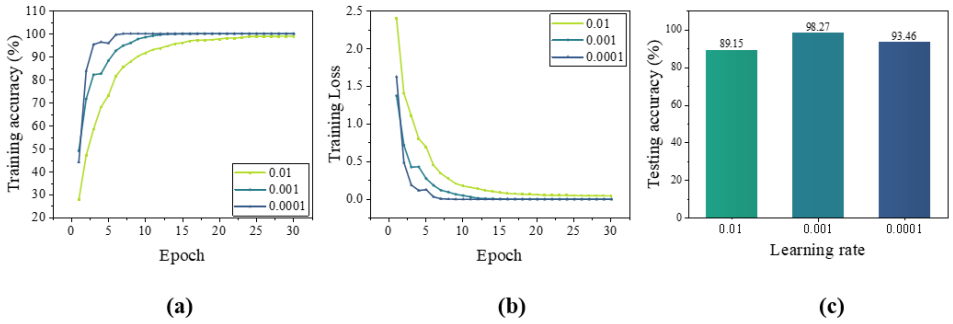

Figure 18. **(a)** training accuracy **(b)** training loss, and **(c)** testing accuracy under different learning rates.

5.2. Ablation study

The purpose of this test is to validate the necessity of each module in DRSEL and its contribution to performance enhancement. The test is conducted under the same conditions as in Case 1, at a 40Hz rotational speed, with all hyperparameters set identically to those used in Case 1. Additionally, DRSEL is decomposed into four initial models based on its structure, as detailed below:

Model 1: Swin Transformer

Model 2: ResNet-18

Model 3: Dual-branch Swin Transformer ensemble learning using multi-sensor signals, where each Swin Transformer branch outputs a 128-dimensional feature vector, which is then fused via concatenation and classified using MLP and fully connected layers.

Model 4: Dual-branch ResNet-18 ensemble learning using multi-sensor signals, where each ResNet-18 branch outputs a 128-dimensional feature vector, which is fused via concatenation and classified using MLP and fully connected layers.

Model 5: DRSEL

Moreover, each experiment is conducted 5 times, and the reported results represent the average of these 5 independent runs.

The results are shown in **Table 5**. It is evident that when using single-channel signals, neither Model 1 nor Model 2 achieved high accuracy. However, when using multi-sensor signals, Model 3 improved by 7.38% and 0.87% compared to Model 1, and Model 4 improved by 4.5% and 1.37% compared to Model 2. This demonstrates that

the use of multi-sensor signals as inputs, in combination with a feature-level ensemble learning framework, significantly enhances fault diagnosis accuracy. Furthermore, DRSEL's accuracy improved by 6.08% and 3.44% compared to Model 3 as well as 0.35% and 0.75% compared to Model 4, indicating that each module contributes to the enhancement of model performance, thus confirming their necessity.

Table 5. Ablation study results.

Models	Base learner		Ensemble framework	Case 1 (40Hz)	Case 2 (30Hz)
	Swin transformer	ResNet-18			
Model1	✓			84.81%	95.69%
Model2		✓		93.42%	97.88%
Model3	✓		✓	92.19%	96.56%
Model4		✓	✓	97.92%	99.25%
DRSEL	✓	✓	✓	98.27%	100%

The combination of ResNet-18 and Swin Transformer leverages the complementary strengths of convolutional networks and self-attention mechanisms. ResNet-18 excels at capturing local textures and edges, while Swin Transformer is proficient in modeling global dependencies through its hierarchical window-based attention. This heterogeneity fosters diverse feature representations and richer semantic understanding, leading to superior performance compared to homogeneous dual-branch models, which often suffer from feature redundancy. Furthermore, the hybrid architecture improves generalization by balancing local and global inductive biases, making it more robust across varying conditions.

6. Conclusion

This paper proposes an integrated learning fault diagnosis framework based on multi-sensor data fusion. The framework employs an ensemble learner composed of ResNet-18 and Swin Transformer networks as the base learners, aiming to enhance both local detail perception and global context modeling capabilities. It effectively explores key fault features within multi-source heterogeneous data while leveraging multi-sensor signals as inputs to achieve multi-scale feature extraction and multi-dimensional fusion for fault diagnosis. The proposed framework demonstrates excellent performance in 2 case studies, achieving fault diagnosis accuracy higher than 96.88% under various operating conditions, and over 97% accuracy under variable speed conditions, proving its strong robustness and outstanding generalization ability. The method proposed in this paper shows significant potential for practical applications. However, further validation in more diverse scenarios and operating conditions is required to ensure its broader deployability and more reliable results in fault diagnosis tasks.

Author contributions: Conceptualization, YZ; methodology, YZ; software, YZ; validation, MP and WQ; formal analysis, YZ; investigation, YZ; resources, MP; data curation, JD and Jiahua Su; writing—original draft preparation, YZ; writing—review and editing, YZ; visualization, JD; supervision, MP; project administration, MP; funding acquisition, MP. All authors have read and agreed to the published version

of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. He DQ, Wu JX, Jin ZZ, et al. AGFCN: A bearing fault diagnosis method for high-speed train bogie under complex working conditions. Reliability Engineering & System Safety. 2025; 258: 110907. doi: 10.1016/j.ress.2025.110907
2. He DQ, Zhao JY, Jin ZZ, et al. Prediction of bearing remaining useful life based on a two-stage updated digital twin. Advanced Engineering Informatics. 2025; 65: 103123. doi: 10.1016/j.aei.2025.103123
3. Tama BA, Vania M, Lee S, et al. Recent advances in the application of deep learning for fault diagnosis of rotating machinery using vibration signals. Artificial Intelligence Review. 2023; 56(5): 4667–4709. doi: 10.1007/s10462-022-10293-3
4. Zhu ZQ, Lei YB, Qi GQ, et al. A review of the application of deep learning in intelligent fault diagnosis of rotating machinery. Measurement. 2023; 206: 112346. doi: 10.1016/j.measurement.2022.112346
5. Qin YL, He DQ, Jin ZZ, et al. An improved deep learning algorithm for obstacle detection in complex rail transit environments. IEEE Sensors Journal. 2024; 24(3): 4011–4022. doi: 10.1109/jsen.2023.3340688
6. Sun HM, He DQ, Zhong JC, et al. Preventive maintenance optimization for key components of subway train bogie with consideration of failure risk. Engineering Failure Analysis. 2023; 154: 107634. doi: 10.1016/j.engfailanal.2023.107634
7. Yin XH, Mu ZQ, Cui QA, et al. Interpretable and spatio-distributed settlement related multimode process monitoring for Metro tunnels excavated by TBM. Advanced Engineering Informatics. 2025; 66: 103464. doi: 10.1016/j.aei.2025.103464
8. Lao ZP, He DQ, Wei ZX, et al. Intelligent fault diagnosis for rail transit switch machine based on adaptive feature selection and improved LightGBM. Engineering Failure Analysis. 2023; 148: 107219. doi: 10.1016/j.engfailanal.2023.107219
9. Qin GH, Zhang K, Lai XW, et al. An adaptive symmetric loss in dynamic wide-kernel ResNet for rotating machinery fault diagnosis under noisy labels. IEEE Transactions on Instrumentation and Measurement. 2024; 73: 3517512. doi: 10.1109/tim.2024.3375404
10. Hou SX, Lian A, Chu YD. Bearing fault diagnosis method using the joint feature extraction of Transformer and ResNet. Measurement Science and Technology. 2023; 34(7): 075108. doi: 10.1088/1361-6501/acc885
11. Xiao YM, Shao HD, Wang J, et al. Bayesian variational transformer: A generalizable model for rotating machinery fault diagnosis. Mechanical Systems and Signal Processing. 2024; 207: 110936. doi: 10.1016/j.ymssp.2023.110936
12. Lv J, Xiao QY, Zhai XD, et al. A high-performance rolling bearing fault diagnosis method based on adaptive feature mode decomposition and Transformer. Applied Acoustics. 2024; 224: 110156. doi: 10.1016/j.apacoust.2024.110156
13. Yan S, Shao HD, Wang J, et al. LiConvFormer: A lightweight fault diagnosis framework using separable multiscale convolution and broadcast self-attention. Expert Systems with Applications. 2024; 237: 121338. doi: 10.1016/j.eswa.2023.121338
14. Kibrete F, Woldemichael DE, Gebremedhen HS. Multi-sensor data fusion in intelligent fault diagnosis of rotating machines: A comprehensive review. Measurement. 2024; 232: 114658. doi: 10.1016/j.measurement.2024.114658
15. Lao ZP, He DQ, Jin ZZ, et al. Few-shot fault diagnosis of turnout switch machine based on semi-supervised weighted prototypical network. Knowledge-Based Systems. 2023; 274: 110634. doi: 10.1016/j.knosys.2023.110634
16. Zhuang Y, He DQ, Jin ZZ, et al. SWR2 Net: A fault diagnosis framework for rotating machine under limited samples and noise interference. Nonlinear Dynamics. 2025; 113(17): 22823–22852. doi: 10.1007/s11071-025-11302-0
17. Wang SQ, Feng ZG. Multi-sensor fusion rolling bearing intelligent fault diagnosis based on VMD and ultra-lightweight GoogLeNet in industrial environments. Digital Signal Processing. 2024; 145: 104306. doi: 10.1016/j.dsp.2023.104306
18. He DQ, Xu Y, Jin ZZ, et al. A zero-shot model for diagnosing unknown composite faults in train bearings based on label feature vector generated fault features. Applied Acoustics. 2025; 232: 110563. doi: 10.1016/j.apacoust.2025.110563
19. Zhang YC, Ding JL, Li YB, et al. Multi-modal data cross-domain fusion network for gearbox fault diagnosis under variable operating conditions. Engineering Applications of Artificial Intelligence. 2024; 133: 108236. doi: 10.1016/j.

engappai.2024.108236
20. Qiu Z, Fan SF, Liang HB, et al. Multimodal fusion fault diagnosis method under noise interference. Applied Acoustics. 2025; 228: 110301. doi: 10.1016/j.apacoust.2024.110301
21. Xu ZZ, Chen X, Xu JT. Multi-modal multi-sensor feature fusion spiking neural network algorithm for early bearing weak fault diagnosis. Engineering Applications of Artificial Intelligence. 2025; 141: 109845. doi: 10.1016/j.engappai.2024.109845
22. Mian Z, Deng XF, Dong XH, et al. A literature review of fault diagnosis based on ensemble learning. Engineering Applications of Artificial Intelligence. 2024; 127: 107357. doi: 10.1016/j.engappai.2023.107357
23. He YL, He DQ, Lao ZP, et al. Few-shot fault diagnosis of turnout switch machine based on flexible semi-supervised meta-learning network. Knowledge-Based Systems. 2024; 294: 111746. doi: 10.1016/j.knosys.2024.111746
24. Tong JY, Liu C, Bao JH, et al. A novel ensemble learning-based multisensor information fusion method for rolling bearing fault diagnosis. IEEE Transactions on Instrumentation and Measurement. 2023; 72: 9501712. doi: 10.1109/tim.2022.3225910
25. Ye MY, Yan XA, Jiang D, et al. MIFDELN: A multi-sensor information fusion deep ensemble learning network for diagnosing bearing faults in noisy scenarios. Knowledge-Based Systems. 2024; 284: 111294. doi: 10.1016/j.knosys.2023.111294
26. Fu GH, Wang XG, Liu YH, et al. A robust bearing fault diagnosis method based on ensemble learning with adaptive weight selection. Expert Systems with Applications. 2025; 269: 126420. doi: 10.1016/j.eswa.2025.126420
27. You YW, Tang JH, Guo M, et al. Ensemble learning based multi-fault diagnosis of air conditioning system. Energy and Buildings. 2024; 319: 114548. doi: 10.1016/j.enbuild.2024.114548
28. Xiao YM, Shao HD, Wang J, et al. Domain-augmented meta ensemble learning for mechanical fault diagnosis from heterogeneous source domains to unseen target domains. Expert Systems with Applications. 2025; 259: 125345. doi: 10.1016/j.eswa.2024.125345
29. He KM, Zhang XY, Ren SQ, et al. Deep Residual Learning for Image Recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 27–30 June 2016; Las Vegas, NV, USA. pp. 770–778. doi: 10.1109/cvpr.2016.90
30. Liu Z, Lin YT, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: Proceedings of the 18th IEEE/CVF International Conference on Computer Vision; 10–17 October 2021; Montreal, QC, Canada. pp. 9992–10002. doi: 10.1109/iccv48922.2021.00986
31. Ding A, Qin Y, Wang B, et al. Evolvable graph neural network for system-level incremental fault diagnosis of train transmission systems. Mechanical Systems and Signal Processing. 2024; 210: 111175. doi: 10.1016/j.ymssp.2024.111175
32. Sehri M, Dumond P. University of Ottawa constant and variable speed electric motor vibration and acoustic fault signature dataset. Data in Brief. 2024; 53: 110144. doi: 10.1016/j.dib.2024.110144