Sound & Vibration 2025, 59(6), 3617.
doi: https://doi.org/10.59400/sv3617

Article

An incremental intelligent fault diagnosis method for marine diesel engines based on CNN-Transformer and cosine similarity

Yingying Wu, Yongjian Wang*, Hangxi Cai, Guoqiang Li, Xin Wei

Marine Engineering Institute, Jimei University, Xiamen 361021, China

*** Corresponding author:** Yongjian Wang, 2201624259@qq.com

CITATION

Wu Y, Wang Y, Cai H, et al. An incremental intelligent fault diagnosis method for marine diesel engines based on CNN-Transformer and cosine similarity. Sound & Vibration. 2025; 59(6): 3617. https://doi.org/10.59400/sv3617

ARTICLE INFO

Received: 21 October 2025
Revised: 15 November 2025
Accepted: 22 November 2025
Available online: 1 December 2025

Abstract: This paper proposes an incremental intelligent fault diagnosis method for marine diesel engines based on a Convolutional Neural Network (CNN)-Transformer architecture and cosine similarity. The method is designed to address critical limitations of conventional supervised diagnostic frameworks, including heavy reliance on labeled data, weak cross-condition generalization, and the inability to identify new or evolving fault types. The model first employs CNN to extract local temporal features from vibration signals and then uses a Transformer to learn high-level semantic representations of fault attributes. During the incremental learning phase, known fault classes—such as exhaust valve failures—are used to train the model. In the testing phase, the model calculates the cosine similarity between feature embeddings of unseen samples and the prototypes of known classes in the attribute space to determine their classification or novelty. This mechanism enables effective identification of both known and novel faults, including those in cylinder liners and piston rings, without requiring prior labeled data for the latter. Experimental results demonstrate that the proposed approach achieves superior classification accuracy, robustness, and adaptability compared to traditional supervised methods, offering a scalable and generalizable solution for intelligent marine diesel engine fault diagnostics.

Keywords: marine diesel engine; fault diagnosis; incremental fault diagnosis; vibration signals; CNN-Transformer

1. Introduction

Marine diesel engines serve as the core of ship propulsion systems, where reliability is critical to operational safety [1]. However, due to harsh working environments and complex mechanical stresses, components such as valves and cylinder liner–piston ring (CLPR) assemblies are prone to faults like leakage and excessive wear. Efficient and accurate fault diagnosis is thus essential to minimize maintenance costs and prevent severe failures. Traditional supervised learning-based diagnostic methods require extensive labeled datasets, making them inadequate for identifying emerging or rare faults, particularly in open-set conditions where unknown fault types may occur [2–4]. Additionally, the nonlinear and high-dimensional characteristics [5] of vibration signals impose higher demands on model performance.

Before the widespread adoption of deep learning techniques, damage identification in mechanical and structural systems was mainly achieved through signal processing and physics-based feature extraction methods, such as wavelet-based feature analysis and modal flexibility-based damage indicators. These approaches

have demonstrated effectiveness in specific scenarios; however, they often rely on handcrafted features and prior domain knowledge, limiting their adaptability to complex and non-stationary operating conditions commonly encountered in marine diesel engines.

In a broader engineering context, deep learning techniques have been widely applied in the field of structural health monitoring (SHM), with bridge health monitoring serving as a representative application. Existing studies have shown that convolutional neural networks, recurrent neural networks, and attention-based models can effectively extract damage-sensitive features from vibration and response signals under complex operating environments. These advances in bridge SHM demonstrate the strong capability of deep learning in handling non-stationary signals, noise interference, and limited prior knowledge. Such characteristics are highly relevant to engine fault diagnosis problems, where operating conditions are complex, and fault patterns are diverse. Over the past decade, deep learning (DL)-based methods have achieved remarkable progress in intelligent fault diagnosis, driven by the development of deep neural network architectures capable of modeling complex relationships between monitoring data and machine health conditions [6–8]. These methods typically rely on large-scale labeled datasets to train accurate diagnostic models and have demonstrated strong performance in closed-set conditions. To address open-set fault scenarios, some approaches incorporate unknown fault detection modules to extend generalization capabilities [9]. However, such models generally lack the capacity for online adaptation, as they cannot autonomously update their knowledge base after deployment.

To overcome this limitation, class-incremental learning [10] (CIL) has emerged as a promising paradigm, enabling models to integrate newly observed fault types while retaining previously acquired diagnostic knowledge [11, 12]. A key challenge in CIL is catastrophic forgetting, where learning from new data leads to the degradation of prior knowledge [13, 14]. This issue arises from the intrinsic stability-plasticity dilemma in neural networks: acquiring new information requires plasticity, whereas retaining old knowledge demands stability [15]. Addressing this trade-off is essential for building robust, continually evolving diagnostic systems suitable for real-world industrial applications.

Wang et al. [6] proposed a graph continual learning network for machinery fault diagnosis, which integrates new fault type detection and class-incremental learning by using a GCN-based model to identify unseen classes and automatically update the model without catastrophic forgetting. Zhang et al. [16] proposed a multivariable ensemble-based incremental support vector machine (MEISVM) method for intelligent fault diagnosis of roller bearings, which effectively detects compound and varying-severity faults using vibration signals and enables incremental learning from new data. Shi et al. [17] proposed a cross-domain class-incremental broad network (CDCIBN) to address fault diagnosis under variable operating conditions, incorporating a domain-adaptation loss function and class-incremental learning mechanism to achieve robust performance without replaying old data. Fu et al. [18] proposed a broad auto-encoder-based intelligent fault diagnosis framework with

incremental learning capabilities, where sample-incremental and class-incremental strategies are developed to enable continuous model updating without retraining for newly arriving fault samples and fault modes.

Although these methods have achieved significant progress, they are predominantly applied to rotating machinery or bearings under controlled laboratory settings, with limited fault types and stable operating conditions. However, marine diesel engines operate in complex and highly variable environments, presenting more diverse and evolving fault patterns. This highlights a critical gap in the current research: existing incremental learning techniques lack validation in safety-critical and domain-specific applications such as marine propulsion systems. Therefore, it is imperative to develop tailored incremental diagnostic strategies that can ensure reliable performance and continual adaptation in real-world maritime scenarios.

To address this challenge, incremental learning and zero-shot learning have emerged as promising paradigms for intelligent fault diagnosis. Incremental learning aims to enable models to continuously learn new fault knowledge without retraining from scratch, while zero-shot learning focuses on identifying unknown fault types based on attribute descriptions and similarity relationships. Motivated by these concepts, this paper proposes an incremental intelligent fault diagnosis method for marine diesel engines based on a CNN-Transformer architecture and cosine similarity. The proposed method establishes an attribute-based representation of fault features and enables effective identification of both known and unknown fault categories.

The main contributions of this paper can be summarized as follows:

(1) It combines CNNs for localized feature extraction with Transformers for global contextual understanding, enabling robust fault representation from complex vibration signals;
(2) It introduces a cosine similarity-driven inference strategy for zero-shot recognition of unseen fault types—such as cylinder liner–piston ring (CLPR) faults—without requiring additional labeled samples;
(3) It supports continual model updates from streaming fault data, effectively mitigating catastrophic forgetting and enhancing generalization across varying operating conditions.

Experimental results verify the method' s superiority over conventional supervised models in both classification accuracy and cross-condition robustness. This work presents a scalable, real-time diagnostic solution that advances the development of intelligent and autonomous monitoring systems for marine propulsion platforms.

2. Approach

2.1. An incremental learning fault diagnosis method based on CNN-Transformer model and cosine distance

In the fault diagnosis of marine diesel engine vibration signals, a CNN-Transformer-based model is designed for the attribute association characteristics of different fault classes. The model combines the powerful feature extraction capability of one-dimensional convolutional neural network (1DCNN) and the efficient capability

of Transformer to capture long-range dependencies and global context understanding, which is able to efficiently learn the vibration signal features and characterise the fault attributes, and provide strong support for identifying unknown fault types using the effective features and attributes of the known fault classes under zero-sample learning.

Convolutional neural networks (CNNs) are effective in extracting localized temporal features from vibration signals, whereas they are limited in modeling long-range dependencies and global relationships. In contrast, Transformer-based models excel at capturing global semantic information through self-attention mechanisms but lack the capability to efficiently extract low-level local features when applied alone. By combining CNN and Transformer, the proposed framework integrates the strengths of both architectures, enabling joint learning of local temporal characteristics and global semantic representations, which is particularly beneficial for zero-shot fault diagnosis.

In this framework, the convolutional neural network (CNN) is employed to extract localized temporal features from high-frequency vibration signals, such as impact characteristics and periodic patterns, while the Transformer is utilized to capture long-range dependencies and global semantic relationships among fault attributes. Compared with CNN-only architectures that lack global contextual modeling and Transformer-only architectures that are sensitive to raw signal noise and computationally expensive, the proposed CNN-Transformer hybrid structure achieves a more balanced representation capability and improved robustness for zero-shot fault diagnosis.

The structure of the CNN-Transformer model is shown in **Figure 1**. Based on the functional division, the model can be divided into four modules: data preprocessing module, feature extraction module, attribute learning module, and recognition and classification module.

(1) Data preprocessing module: the vibration data collected from known faults of the exhaust valve assembly of the marine diesel engine is divided into a training set, and the vibration data of other types of faults (such as cylinder liner-piston ring faults) is divided into a test set. The data set is preprocessed using data enhancement and normalisation to optimise the data quality and improve the model generalisation capability.

(2) Feature extraction module: features are extracted from all faulty vibration samples of the diesel engine by the encoder of the CNN neural network, and the training is continued until the model converges. In the training phase, the label information of unknown category faults is not used and only aligned with known category fault vibration samples to ensure consistent data dimensions. The CNN encoder was combined with a Softmax layer to adapt the model to the training set. After the training is completed, the parameters of the trained CNN encoder layers are saved. In the testing phase, the saved parameters are loaded.

(3) Attribute learning module: the Transformer attribute learner is used for predicting the attributes of the samples and deriving their labels. That is, the fault features of known categories are input to multiple attribute learners in the attribute learning network for training, and the model with the best performance and its parameters are finally saved. At this stage, the labelling information of the unknown category

faults remains unused, and only the unknown category samples are aligned with the known category samples to ensure the validity of the subsequent zero-sample classification. Afterward, the trained attribute learner is used for predicting the attributes of the unknown category fault features and generating the corresponding attribute vectors.

(4) Identification and classification module: the obtained attribute vectors of the unknown category fault samples are labelled with the known category fault attributes, and the cosine distance is calculated to complete the identification and classification of the unknown category fault samples.

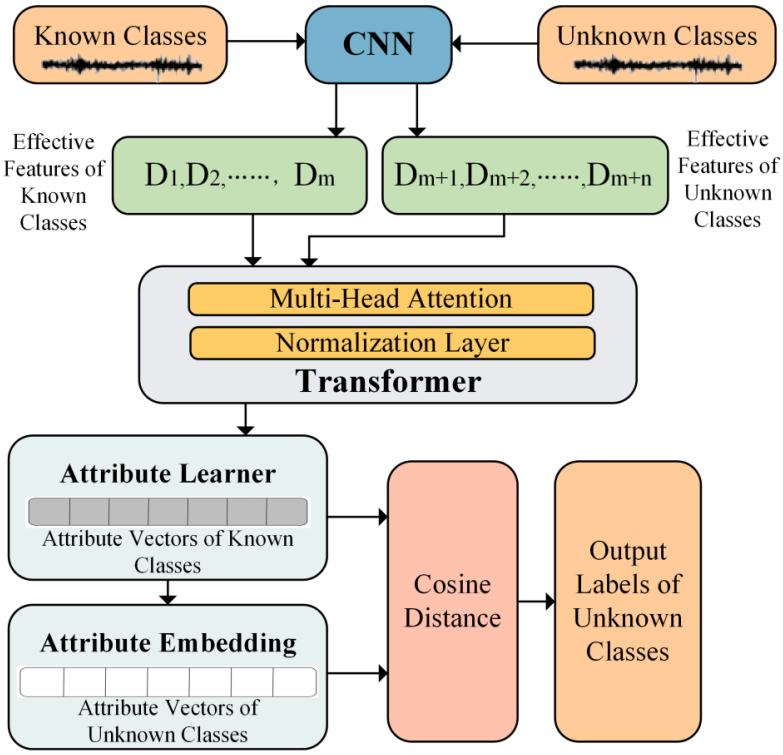


Figure 1. Schematic of CNN-Transformer model architecture.

2.2. Convolutional neural network-based feature learning

CNN (Convolutional Neural Network) [19] has an outstanding feature extraction capability, and the commonly used convolution operations are one-dimensional convolution and two-dimensional convolution. One-dimensional convolution performs well with one-dimensional sequence data, while two-dimensional convolution specializes in processing image data [20]. One-dimensional convolutional neural network (1DCNN) is a deep learning network model specialized in processing time series data and one-dimensional signals, which is widely used in the field of vibration signal analysis, and its advantage lies in the sliding operation of the input signals through a one-dimensional convolutional kernel, extracting the features from the local region and forming the feature mapping. The core structure of 1DCNN consists of a convolutional layer, a batch normalization layer, an activation layer, a pooling layer, and a fully connected layer composition, among others, as shown in **Figure 2**.

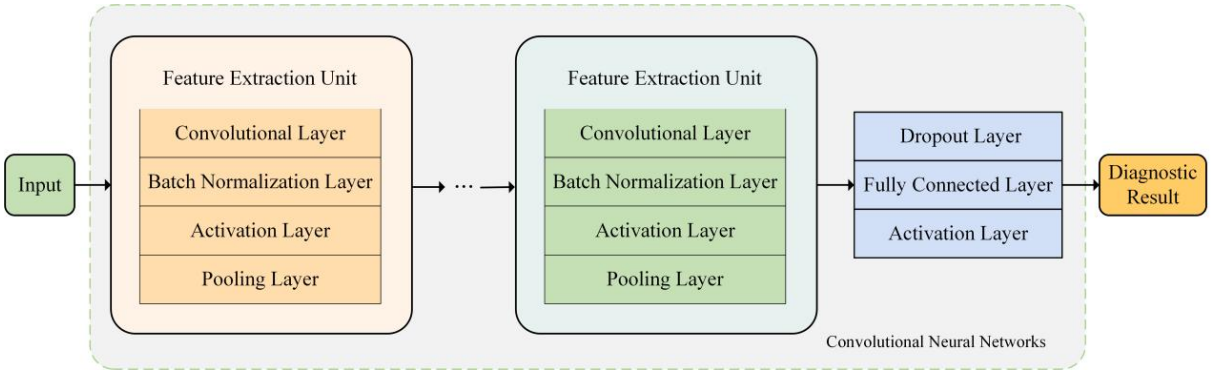


Figure 2. Basic structure of Convolutional Neural Networks (CNN).

The convolutional layer can significantly reduce the model complexity through the parameter sharing mechanism, and at the same time use the local connection mechanism to extract the local features of the signal; the activation function (e.g., LReLU) introduces the nonlinear ability and improves the ability of the model to express the complex features; the pooling layer reduces the risk of overfitting and improves the computational efficiency by lowering the data dimensionality; and the fully connected layer spreads the mapping of the extracted features and passes it to the classifiers to achieve fault classification. In fault diagnosis, 1DCNN is capable of modeling and identifying subtle changes in frequency and amplitude characteristics in vibration signals, providing highly discriminative characterization of different fault modes; moreover, its shallower network structure and fewer parametric quantities enable it to have a good performance in processing high-frequency data such as vibration signals.

One-dimensional convolutional neural networks are suitable for dealing with the temporal characteristics of vibration signals and can efficiently extract local features of fault data. Feature learning of vibration signals includes the following steps:

(1) Input preprocessing: the input signal is normalized and noise reduced, and the sampling rate is unified to ensure the consistency of the time domain and frequency domain characteristics.

(2) Convolution operation: local features (mutation points, periodic features, etc.) of the vibration signal are extracted through multiple one-dimensional convolution layers. Let the input signal be $x \in R^{t\times 1}$. After the convolution operation of the l layer, the feature mapping is expressed as:

$$h^{(l)} = \mathrm{Re}LU(W^{(l)} * h^{(l-1)} + b^{(l)}) \tag{1}$$

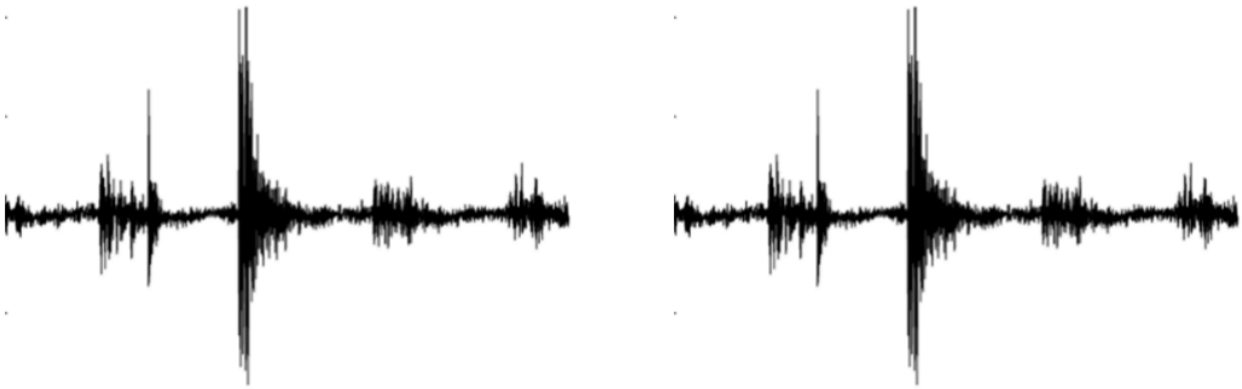
Where $W^{(l)}$ and $b^{(l)}$ are convolution kernel weights and bias, respectively, and * denotes the convolution operation, ReLU is the activation function.

(3) Pooling and dimensionality reduction: the feature layer is dimensionality reduced using a maximum pooling layer to retain the main features to reduce the computational complexity.

(4) After multi-layer convolution and pooling operations, the output high-dimensional feature mapping is expanded into a one-dimensional vector $f(x) \in R^{d}$, which is used as the feature vector of the signal for subsequent attribute description learning

by the attribute learner (Transformer).

The convolution operation in CNN uses a convolution kernel (also known as a filter) to slide over the input signal to extract the local information into a feature mapping, as shown in **Figure 3**. As can be seen in the figure, using the signal through the convolutional layer retains the salient features of the original waveform, but its data size has changed, i.e., fewer data points characterize the original information. This study uses CNN to complete the zero-sample representation learning, i.e., extracting useful information from the signal data and using the feature vector with reduced information size as the input data for the Transformer attribute learner, which can significantly improve the training efficiency and generalization performance of the model.



(a) Original time-domain waveform of the input signal. **(b)** Time-domain waveform extracted by the first convolutional layer.

Figure 3. Time-domain of features extracted by convolutional layers.

2.3. Transformer-based attribute learning

The Transformer [21] model is a deep learning architecture based on an attention mechanism, and the unique self-attention mechanism [22] and the ability to process information about different locations in a sequence in parallel give it a clear advantage in efficiently capturing long-distance dependencies and contextual information. The core of the model consists of two parts: encoder and decoder, as shown in **Figure 4**.

The encoder mainly consists of multiple stacked sub-layers, including a multi-head attention mechanism and a feed-forward neural network, and each sub-layer improves the training efficiency and stability through residual connection and normalisation operation. The input signal is positionally encoded and summed with the embedding vectors, and the global features are extracted by the encoder to capture the dependencies between the elements in the input sequence. The decoder structure is similar to the encoder, but with an additional masked multi-head attention mechanism for handling the autoregressive task. The decoder generates the final target sequence by receiving the output features from the encoder as well as the historical output from the current decoding.

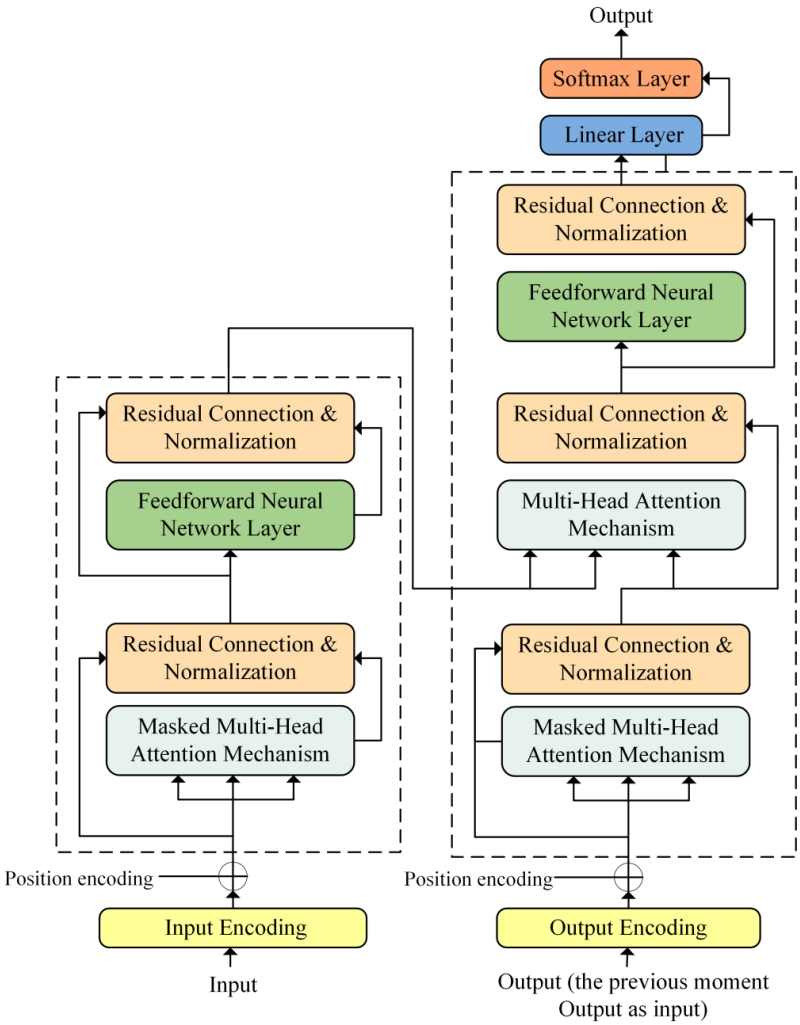


Figure 4. Transformer network architecture.

The multi-head attention mechanism is the core module of Transformer, which achieves multi-dimensional feature capture capability of the input data by calculating the weighted values of different heads, and finally, generates the prediction results through the output of the decoder at the linear and Softmax layers. Transformer has a powerful global feature capture capability, good parallel computing capability, and long sequence modelling performance, which can efficiently perform semantic modelling of signal features to generate attribute features. Its core components include a multi-head self-attention mechanism and a feed-forward neural network:

(1) Input Embedding: The feature f(x) extracted by the CNN is mapped into a fixed dimension embedding vector $z_0 \in R^{d \times D}$, where D denotes the input dimension of the Transformer.

(2) Multi-Headed Attention Mechanism: global dependencies are captured by calculating the similarity between input features. With the multi-head mechanism, the model is able to capture global context information in different dimensions in parallel. For the input sequence z, the attention weight is calculated by the following equation:

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

Where Q is the query vector, K is the key vector, V is the value vector, d_k is the number of columns of the Q, K matrix, and W_i^Q, W_i^K, and W_i^V are the weight matrices of Q, K, V, respectively.

(3) Feedforward neural network: the attention output is further processed by the feedforward neural network to form the attribute description vector. The feedforward network contains two fully connected layers and activation function:

$$FFN(x) = \text{ReLU}(W_1 x + b_1)W_2 + b_2 \tag{4}$$

Where W_1, W_2, b_1, b_2 are learnable parameters.

(4) Attribute output: the final output is the attribute feature vector a(y)∈R^D of the fault category, the elements of the R- vector are real numbers, and the dimension of the D- attribute vector, i.e., the dimension of the attribute space indicates the embedding of the fault category in the attribute space.

2.4. Attribute learner and cosine distance

Incremental learning is a machine learning method that does not need to provide labelled samples for all categories, and is especially suitable for scenarios where all fault category labelled data are not available in advance. The core idea is to achieve recognition of unknown category samples by learning attribute information of known categories and embedding the attributes into a high-dimensional space. The overall attribute learning and matching process is illustrated in **Figure 5**.

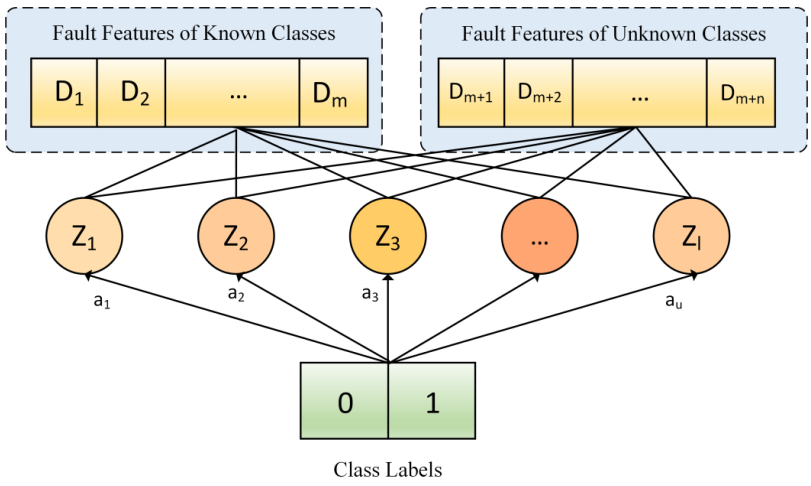


Figure 5. The attribute learning and matching process.

After feature extraction, each vibration sample is represented as a d-dimensional attribute vector in the learned attribute space. For each known fault class, a class prototype is constructed by calculating the mean attribute vector of all training samples

belonging to that class. These prototypes serve as semantic representations of known fault categories.

Attribute-Based Learner (AL) is the core module for increasing learning, which is used to map the attribute information of fault categories into a high-dimensional attribute space [23].

In this study, the dimensionality of the attribute vector D is determined by the output dimension of the Transformer encoder. Each fault sample is represented as a D-dimensional attribute embedding in the learned attribute space. During inference, unknown fault samples are matched with known fault classes using a prototype-based strategy, where the class prototype is defined as the mean attribute vector of each known class.

This high-dimensional attribute space can be viewed as an abstract feature space where each dimension represents a fault characteristic. By learning the characteristics of known fault categories, the attribute learner is able to generate attribute features matching faults for unknown fault categories.

The attribute learner helps the model to quickly process the unknown class fault data without prior learning, i.e., let the training data set for the known fault class be $D_S = \{(x_i^S, y_i^S)\}$, where x_i^S is the vibration signal sample and y_i^S is the label of the known class. The unknown category test set is $D_U = \{x_j^U\}$, and the corresponding set of labels is y_j^U. The feature vector $f(x_i^s) = \{D_1, D_2, ..., D_m\}$ of the vibration signal is extracted by a Convolutional Neural Network (CNN), and the attribute vector $a(y_i^s)$ is generated by a Transformer. Subsequently, these attribute vectors are mapped into the same high-dimensional attribute space. The feature vector of the unknown fault class data is $f_{\varnothing}(x_{ki}^t) = \{D_{m+1}, D_{m+2}, ..., D_{m+n}\}$ and $Z = \{Z_1, Z_2, ..., Z_d\}$ is the attribute vector. The known class features are used in the training phase to train the attribute learner $a_1, a_2, ..., a_u$. In the testing phase, the attribute learner $a_1, a_2, ..., a_u$ is trained.

In the testing phase, the trained attribute learner is first used to generate the attribute vectors of the test set, i.e., the attribute representations of the unknown class fault samples. In this process, the attribute vectors of the known class data and the unknown class fault data are mapped into a high-dimensional vector space, which facilitates similarity or difference metrics. After feature extraction, each vibration sample is represented as a D-dimensional attribute vector in the learned attribute space. For each known fault class, a class prototype is constructed by calculating the mean attribute vector of all training samples belonging to that class. The class prototype serves as the semantic center of the corresponding fault category and is defined as:

$$c = \frac{1}{N}\sum_{i=1}^{N} z_i \tag{5}$$

Where z_i denotes the attribute vector of the i-th training sample and N is the number of samples in the known fault class.

To characterize the dispersion of attribute vectors within each known fault class, a class radius is further defined. The class radius represents the maximum distance between the attribute vectors and the corresponding class prototype, and it is calculated

as:

$$\mathrm{r} = \max \|\mathrm{z}_i - \mathrm{c}\| \tag{6}$$

Where r denotes the class radius. This radius provides a boundary describing the distribution range of known fault samples in the attribute space. During the testing phase, the attribute vectors of unknown fault samples are first generated by the trained attribute learner. The similarity between the attribute vector $\mathrm{z}_{\mathrm{ki}}^{\mathrm{t}}$ of an unknown fault sample and the class prototype c of a known fault class is then measured using cosine distance, which is defined as:

$$CosineDistance = \frac{Z_{test} \cdot c}{\| z_{test} \| \| c \|} \tag{7}$$

Cosine distance is selected as the similarity metric due to its scale-invariant property, which makes it particularly suitable for vibration-based feature representations where amplitude variations frequently occur under different operating conditions. In contrast, Euclidean distance is sensitive to magnitude scaling, and Mahalanobis distance requires reliable covariance estimation, which becomes unstable when the number of unknown fault samples is limited. Therefore, cosine similarity provides a more robust and interpretable criterion for attribute-based zero-shot fault recognition in this study.

The cosine distance is used as a metric to complete the classification and identification of known fault classes. When the distance is greater than the radius component r_i^s of the known class data, it is decided that it does not belong to the known data, and the label '0' is output; vice versa, the label '1' is output, which completes the classification and identification of the unknown fault class.

3. Experimental platform construction and signal acquisition

3.1. Experimental platform construction

Based on the research objectives of this study, a marine diesel engine fault diagnosis experimental platform was constructed to enable controlled and repeatable vibration signal acquisition. The diesel engine used in the experiments was a QC-50BW-CY model, with a rated power of 56 kW and a rated speed of 1500 r/min. The engine had a bore and stroke of 105 mm and 125 mm, respectively, and was configured with four cylinders. A WHI224 generator set was employed, and a water-resistance load tank was used to provide adjustable loading conditions.

For vibration signal acquisition, a DH5922 data acquisition instrument and the DHDAS testing and analysis system (Donghua Testing Technology Co., Ltd.) were used. The vibration acceleration sensors had a maximum measurable acceleration of 50 g (m/s^2), which is sufficient to capture fault-induced vibration characteristics. The overall experimental platform configuration and the vibration sensor layout for the cylinder liner–piston ring assembly and exhaust valve fault tests are shown in **Figures 6** and **7**.

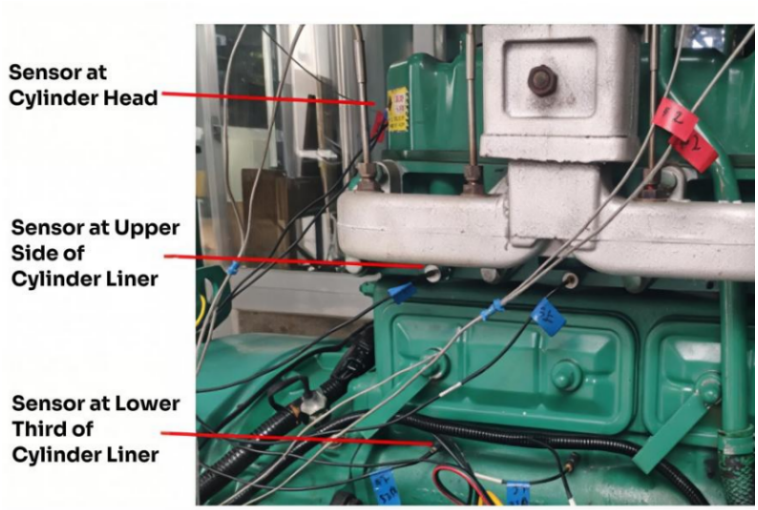


Figure 6. Vibration sensor layout for cylinder liner–piston assembly faults.

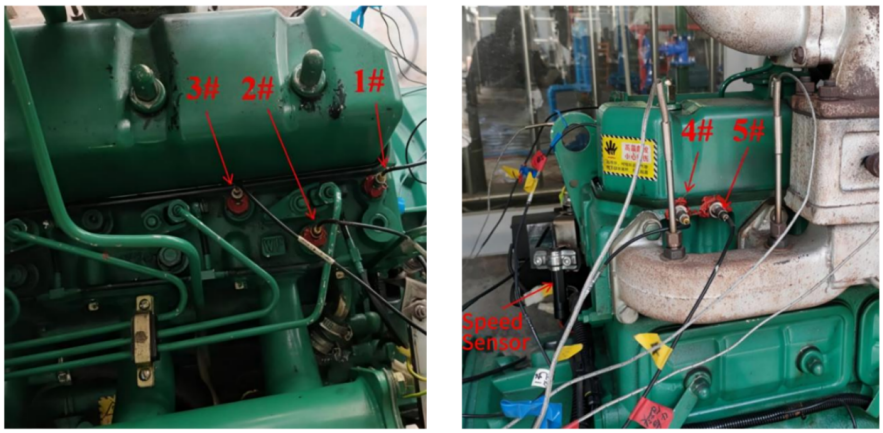


Figure 7. Exhaust valve failure test vibration sensor layout.

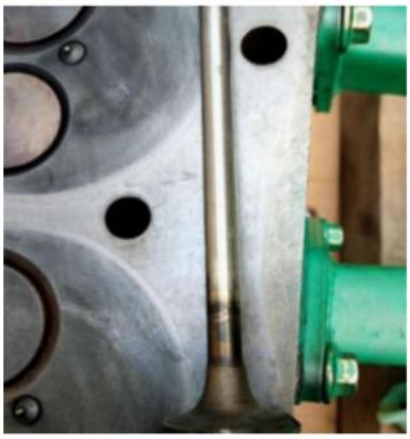
Using this experimental platform, three typical exhaust valve assembly faults—pitting corrosion, burning, and excessive clearance—were artificially introduced and treated as known fault classes. Vibration data collected under these exhaust valve fault conditions, together with data from normal operating conditions, were used as known-class samples for model training. In addition, three cylinder liner-piston ring assembly (CLPRs) fault conditions—broken piston ring, 0.25 mm wear, and 0.5 mm wear—were introduced and treated as unknown fault classes. Vibration data corresponding to CLPR's faults were collected exclusively as unknown-class samples and used only during the testing phase.

3.2. Exhaust valve fault setting and signal acquisition

To obtain vibration data under exhaust valve fault conditions, the No. 4 cylinder of the marine diesel engine was selected as the experimental fault cylinder. Exhaust valve pitting corrosion faults were simulated by introducing punching points on the exhaust valve seat, while exhaust valve burning faults were simulated by cutting notches on the sealing cone surface of the exhaust valve. Excessive exhaust valve clearance faults were simulated by increasing the valve clearance to 0.7 mm, whereas the normal allowable clearance should not exceed 0.35 mm. The exhaust valve fault simulation process is

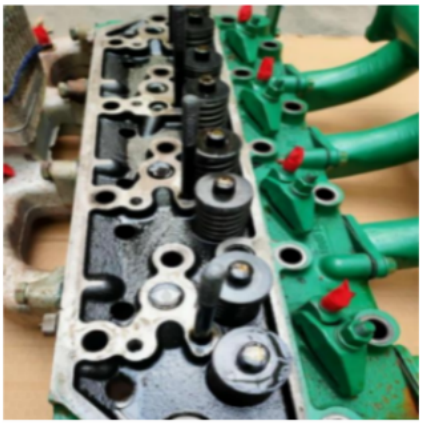
illustrated in **Figure 8**.

(a) Pitting corrosion of exhaust valve seat.



(b) Burning of exhaust valve spindle.

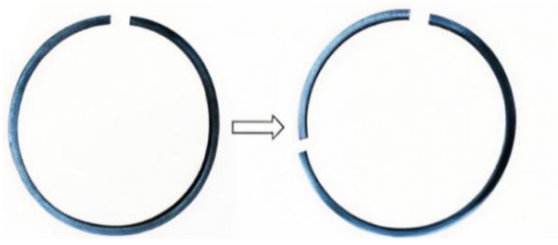
(c) Excessive clearance of exhaust valve.



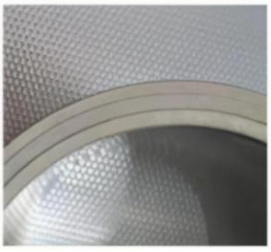
(d) Cylinder head valve assembly of diesel engine.

Figure 8. Set of diesel engine exhaust valve fault state.

Using the same experimental platform, vibration data were also collected from the No. 1 cylinder under normal operating conditions, as well as under three CLPR fault conditions, including cylinder liner wear of 0.25 mm, cylinder liner wear of 0.5 mm, and a broken piston ring (one channel), as shown in **Figure 9**. In total, vibration signals corresponding to four operating states were obtained.



(a) Fracture of first piston ring.



(b) Excessive wear of cylinder liner.

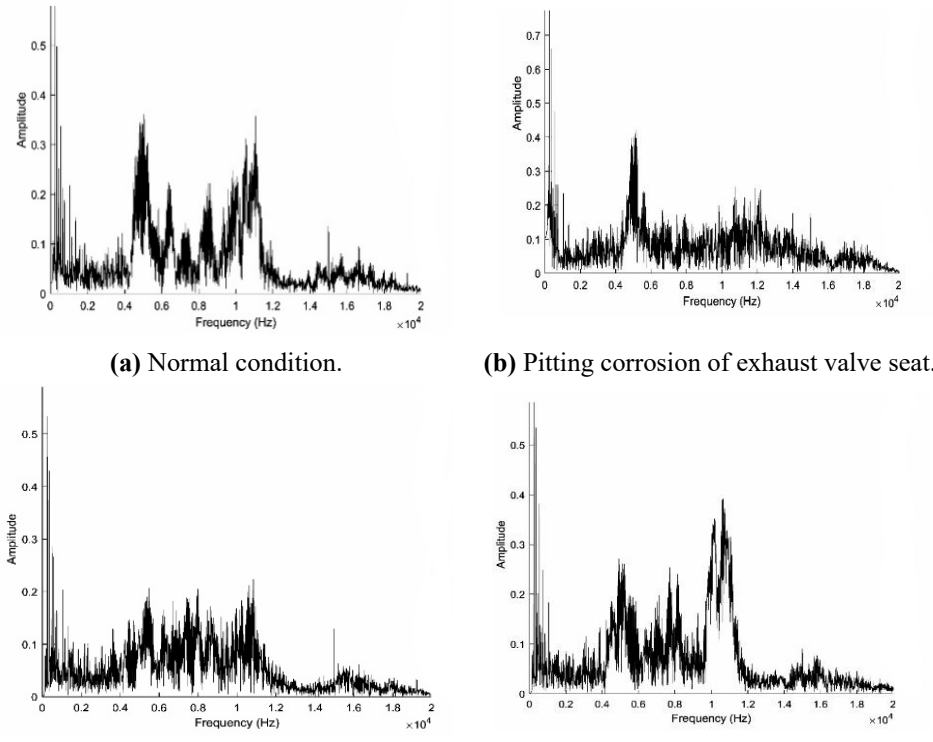
Figure 9. *Cont.*

(c) Excessive wear of cylinder liner (+0.5 mm). **(d)** Cylinder wall damage condition.

Figure 9. CLPRs faults set and its occur damage.

Vibration sensors were installed at three measurement locations as described in Section 3.1, and vibration signals were collected under different fault states and load conditions. The sampling frequency was set to 51.2 kHz to ensure sufficient temporal resolution for fault feature extraction. All experiments were conducted at the rated operating conditions of 1500 r/min and 56 kW. Vibration data corresponding to three exhaust valve fault states and the normal state were measured under these conditions.

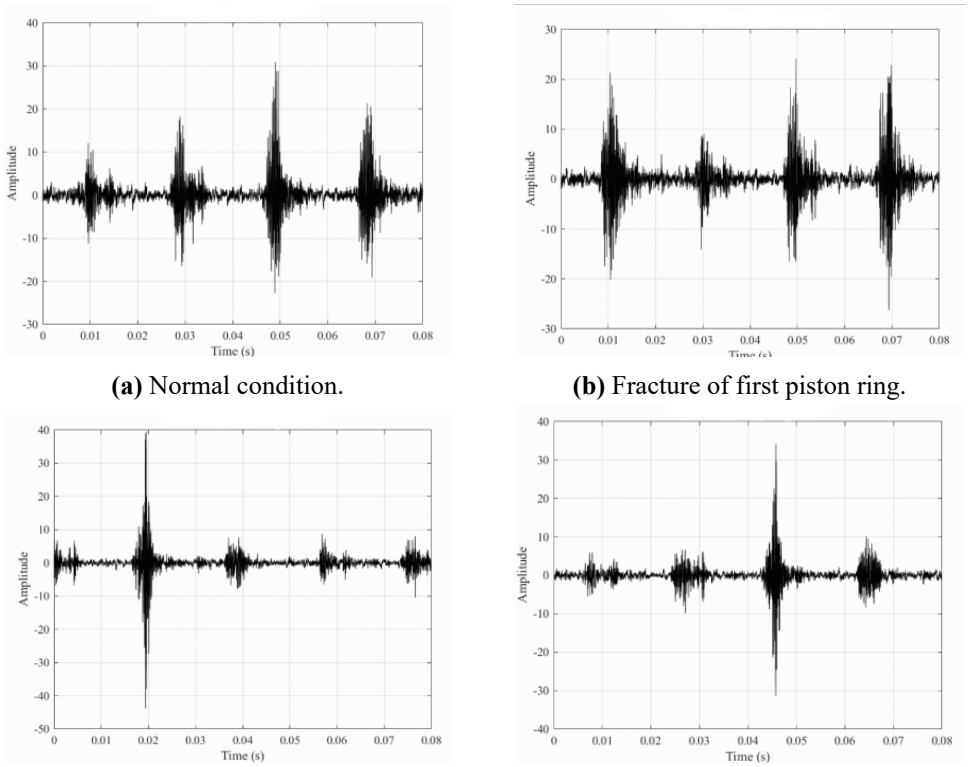
Figure 10 shows the vibration signals collected by the vibration sensors under a 75% load condition at the top position of the No. 1 cylinder liner. **Figure 11** presents the corresponding time–frequency representations obtained from the averaged vibration signals over ten working cycles under the same load condition.



(a) Normal condition. **(b)** Pitting corrosion of exhaust valve seat.

(c) Burning of exhaust valve spindle. **(d)** Excessive clearance of exhaust valve.

Figure 10. The frequency domain diagram of vibration signals for the exhaust valve in four states at Sensor Point 1#.



(a) Normal condition. **(b)** Fracture of first piston ring.

(c) Excessive wear of cylinder liner. **(d)** Excessive wear of cylinder liner (+0.5 mm).

Figure 11. Time-frequency signal mean value of ten working cycles of NO.1 cylinder line and piston-rings at four state.

To ensure the validity of the experimental evaluation, the vibration data corresponding to exhaust valve faults and normal conditions were used exclusively as known-class samples for model training, while the vibration data associated with CLPRs faults were reserved only for testing. No samples from unknown fault classes were involved in the training process, which avoids data leakage between training and testing stages.

All experimental data were collected from a single diesel engine test platform. Although this may introduce potential correlations between samples, the strict separation of known and unknown fault data ensures a fair evaluation of the zero-shot fault diagnosis capability of the proposed method.

4. Fault diagnosis results and discussion

4.1. Data set pre-processing and preparation

In order to ensure the accuracy and comparability of the experimental data, the collected vibration signals are pre-processed with steps such as denoising, normalisation, and signal segmentation. During data acquisition, vibration signals may inevitably be affected by environmental noise and occasional outliers caused by sensor disturbances or transient operating fluctuations. To mitigate these effects, standard data preprocessing procedures were applied prior to model training and testing. Specifically, the raw vibration signals were first subjected to normalization to eliminate amplitude-scale differences across samples.

In addition, signal segments containing abnormal spikes or incomplete acquisition

were removed during segmentation to avoid the influence of outliers. No manual feature selection or fault-specific filtering was introduced, ensuring that the original fault-related characteristics of the vibration signals were preserved. After preprocessing, the vibration signals are segmented into fixed-length time series segments (4096 sampling points per segment) for input to the deep learning model for training and testing. Signal segmentation is based on rotational speed synchronisation information, which ensures that each sample corresponds to a complete duty cycle and guarantees sample independence and consistency.

The incremental learning task is to classify unknown fault samples without having knowledge of the unknown fault samples. In this context, the training set and the test set are required to have no overlapping samples, and the model should avoid touching the samples in the test set during the training phase, so as not to be able to generalize the model to the unknown category well. Since the marine diesel engine state sample identification is achieved by attribute description, an attribute semantic space describing the marine diesel engine state needs to be constructed (see **Table 1**).

Table 1. Description of diesel engine status attributes.

Attribute category	Attribute description	Number of samples
Normal Condition of Diesel Engine	Normal Condition	25,000
Fault Condition of Exhaust Valve Assembly	Pitting Corrosion	25,000
	Burning	25,000
	Excessive Clearance	25,000
Fault Condition of Cylinder Liner–Piston Ring Assembly	Broken Piston Ring	2500
	Wear (0.25 mm)	2500
	Wear (0.5 mm)	2500

As can be seen from **Table 1**, this experimental study uses the three fault characteristic attributes corresponding to the three faults of pitting corrosion, burnout, and excessive clearance of the exhaust valve assembly of a marine diesel engine for describing the fault classes of the diesel engine exhaust valve assembly. The three fault feature attributes corresponding to broken ring, worn 0.25 mm, 0.5 mm faults of cylinder liner-piston ring sets (CLPRs) of diesel engine are used for describing the diesel engine cylinder liner-piston ring set fault class, the dataset performs the preprocessing strategy of overlap sampling and normalisation, the overlap sampling step is 512, and the length of the samples are all 4096. The normal state class of the diesel engine and the faults of the exhaust valve assemblies are the known class training set, and CLPRs faults are the unknown class test set, and the training set $\cap$ test set = $\varnothing$.

The CNN-Transformer model batch size is set to 128, the learning rate is 0.001, and the loss function is the cross-entropy loss function. The detailed hyperparameter configuration of the proposed CNN-Transformer model, including the number of convolutional layers, kernel sizes, and pooling strategies. Transformer depth, attention heads, and fully connected layer dimensions are summarized in **Table 2**.

Table 2. Parameters of the CNN-Transformer model.

No.	Layer type	Output size	Hyperparameters	Number of parameters
1	Input Layer	128 × 1 × 4096	---	---
2	Convolutional Layer 1	128 × 64 × 4096	Kernel Size: (128, 1); Number of Kernels: 64; Stride: 1; Activation: LReLU	8448
3	Pooling Layer 1	128 × 64 × 2048	Pool Size: (2, 1); Stride: 2	---
4	Convolutional Layer 2	64 × 128 × 2048	Kernel Size: (64, 1); Number of Kernels: 128; Stride: 1; Activation: LReLU	16,512
5	Pooling Layer 2	64 × 128 × 1024	Pool Size: (2, 1); Stride: 2	---
6	Convolutional Layer 3	32 × 256 × 256	Kernel Size: (32, 1); Number of Kernels: 256; Stride: 1; Activation: LReLU	33,024
7	Pooling Layer 3	32 × 256 × 128	Pool Size: (2, 1); Stride: 2	---
8	Convolutional Layer 4	512 × 16 × 64	Kernel Size: (16, 1); Number of Kernels: 512; Stride: 1; Activation: LReLU	66,048
9	Pooling Layer 4	512 × 16 × 32	Pool Size: (2, 1); Stride: 2	---
10	Transformer Layer	512 × 8 × 512	Number of Heads: 8; Hidden Units: 512; Number of Layers: 6	---
11	Fully Connected Layer	1024 × 1	Activation: LReLU	4,194,304
12	Output Layer	4 × 1	Softmax	---

To ensure the validity and fairness of the incremental fault diagnosis experiments, the vibration dataset was strictly partitioned into training and testing sets with no sample overlap. The vibration signals corresponding to the normal condition and the exhaust valve assembly faults (including pitting corrosion, burning, and excessive clearance) were used exclusively as known-class samples for model training. In contrast, the vibration signals associated with cylinder liner–piston ring assembly (CLPRs) faults (including broken piston ring, 0.25 mm wear, and 0.5 mm wear) were treated as unknown-class samples and were only used during the testing phase.

Specifically, all samples in the training set and the test set were generated from different signal segments, and no data segment from the unknown fault classes was accessed during model training. This strict separation ensures that the proposed method performs true zero-shot fault diagnosis rather than supervised or semi-supervised classification. In addition, to evaluate the robustness of the proposed approach under data imbalance conditions, the training set contained 25,000 samples for each known fault class, while the test set contained 2500 samples for each unknown fault class. During the training stage, a subset of the known-class samples was randomly selected as a validation set to monitor model convergence and prevent overfitting, while the unknown-class samples were never involved in either training or validation.

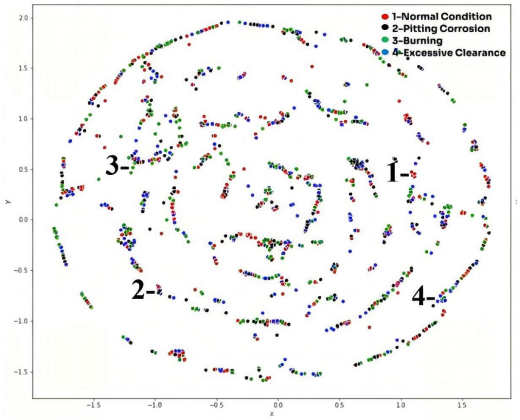
4.2. Visual analysis of CNN feature extraction effect

Based on the CNN-Transformer model and zero-sample learning fault diagnosis mechanism, take the experimental condition of serial number 1 in **Table 3** as an example, feature extraction is carried out on the vibration samples of marine diesel engine exhaust valve fault (training set) and cylinder liner-piston group fault (test set) through CNN neural network encoder, and the CNN extracted features are visualised using t-SNE downscaling, and the result is as shown in **Figure 12**. In the figure, whether it is the training set or the test set, the clustering of the features extracted by CNN becomes more concentrated, and the original linearly indivisible discrete data is embedded into the nonlinearly divisible space, and the CNN completes the effective extraction of the features initially with its strong nonlinear expressive ability.

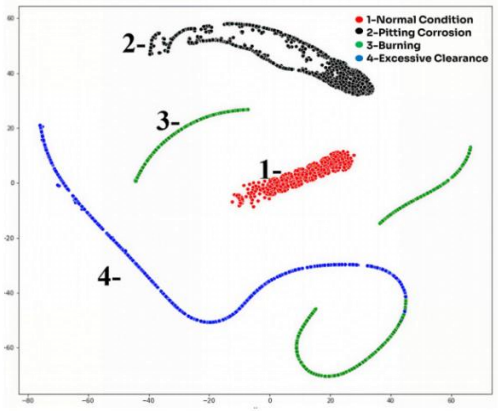
Table 3. Ten sampling prediction results.

No.	Number of training samples	Number of testing samples	Accuracy
1	10,180	970	82.35%
2	7510	710	78.91%
3	12,030	980	86.14%
4	14,500	1480	76.43%
5	15,150	1640	73.13%
6	16,200	2080	80.64%
7	11,270	820	91.98%
8	11,510	1700	81.02%
9	16,450	1680	77.38%
10	13,790	1080	83.62%

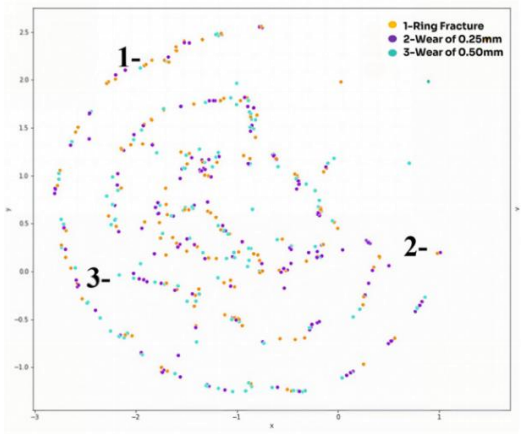
Note: Training set: vibration data of exhaust valve faults; Testing set: vibration data of cylinder liner–piston ring set (CLPRs).



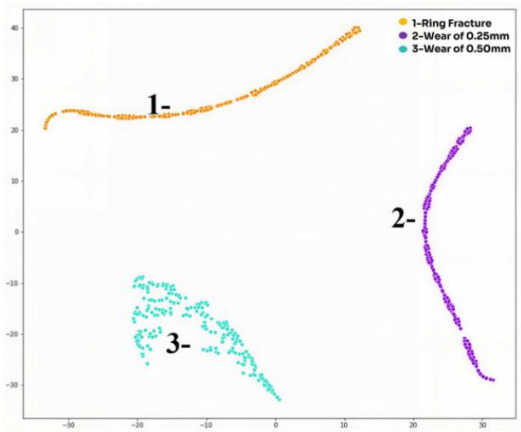
(a) Feature extraction of the known class samples by first convolutional layer.



(b) Feature extraction of the known class samples by fourth convolutional layer.



(c) Feature extraction of the unknown class samples by first convolutional layer.



(d) Feature extraction of the unknown class samples by fourth convolutional layer.

Figure 12. Visualization of known fault types in exhaust valve faults and CNN feature extraction results.

4.3. Fault diagnosis results

4.3.1. Visualization and analysis of fault diagnosis results

Known classes of effective features extracted by CNN from one-dimensional time-series signals are used to train the Transformer attribute learner, which saves the optimal inter-layer parameter fixation learner. Then, the attribute learner embeds the

effective features output from the feature extraction network into the high-dimensional attribute space, and finally, based on the output attribute vector, the cosine distance to the known class prototype is calculated, and the results of the test set samples are outputted with a threshold rating of whether the radius component is exceeded. To ensure the validity and authenticity of the experimental results, this study uses random sampling of data from the training set of vibration data of exhaust valve failures, and the test set of vibration data of CLPRs without playback to construct new training and test sets. This strategy aims to evaluate the learning performance of the model in the face of unbalanced samples. As an example, the specific experimental results and prediction data for 10 random samples are shown in **Table 3**.

The average accuracy of the 10 experiments is 81.16%, with a maximum of 91.98%, indicating that the CNN-Transformer model is able to perform the zero-sample classification task. To analyse the fault diagnosis and learning effect of the marine diesel engine based on the CNN-Transformer model and incremental learning, t-SNE is used to visualize the feature attributes of the training set of exhaust valve fault vibration data, and the test set of vibration data of CLPRs in Group 1 in **Table 3**, as shown in **Figure 13**.

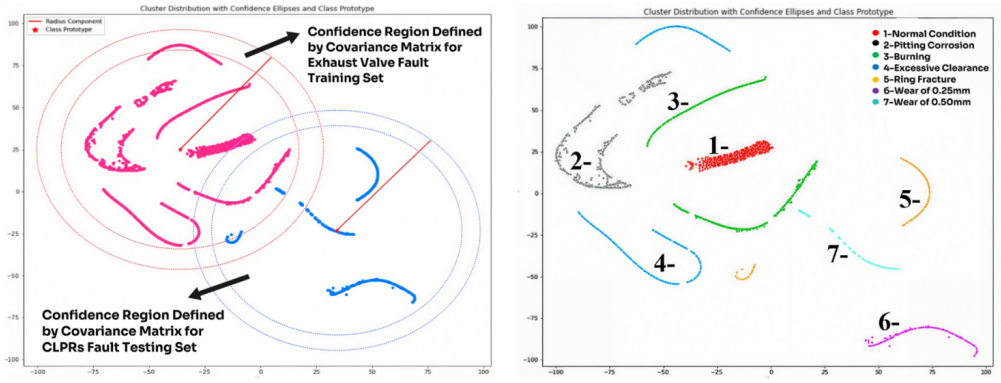


Figure 13. Visualization of fault diagnosis results for marine diesel engine based on CNN-Transformer model and zero-shot learning.

From the figure, it can be seen that the known classes of effective features extracted using CNN are successfully mapped into the same space after the Transformer attributes are learnt for the two classes of data effective features in different feature spaces. From the confidence region defined by the covariance matrix, the distance between some of the test set data, and the training set category prototypes is smaller than the radius of the covariance matrix of the known classes, indicating that some of the test set data cannot be completely recognised. The accuracy of the overall experimental test is shown in **Table 3**.

4.3.2. Comparison with other diagnostic methods

To further demonstrate the effectiveness of the proposed method, comparative experiments were conducted using the same vibration dataset under identical training and testing conditions. A conventional one-dimensional convolutional neural network (1D-CNN) with a Softmax classifier was implemented as a baseline supervised learning model. In addition, a CNN-based model without the Transformer attribute learner was evaluated to investigate the contribution of global semantic modeling.

The experimental results indicate that the baseline 1D-CNN model exhibits satisfactory performance when diagnosing known fault categories; however, it fails to correctly identify unknown fault types due to the absence of attribute-based learning. By contrast, the proposed CNN–Transformer incremental learning framework demonstrates superior capability in zero-shot fault diagnosis, highlighting its advantage in handling unseen fault conditions.

4.3.3. Limitations and future work

It should be noted that the experimental validation in this study is conducted on a single marine diesel engine platform under fixed rated operating conditions. The current investigation primarily serves as a proof-of-concept to verify the feasibility of the proposed CNN–Transformer-based incremental fault diagnosis framework. In future work, the proposed method will be extended to multi-engine platforms, variable load and speed conditions, and more complex marine environments to further enhance its generalization capability and industrial applicability.

5. Conclusion

To address the limitations of traditional supervised learning methods for intelligent fault diagnosis of marine diesel engines, which require large amounts of labeled fault data and are restricted to predefined fault categories with limited generalization capability, an incremental learning-based fault diagnosis method is proposed in this study. The proposed approach integrates a CNN-Transformer architecture with an attribute-based incremental learning framework. During the training phase, vibration data corresponding to exhaust valve faults are used as known fault samples to establish a mapping between fault features and their semantic attributes. During the testing phase, the trained model is directly applied to diagnose previously unseen fault categories without retraining.

Experimental results demonstrate that the proposed method achieves an average diagnostic accuracy of 81.16% and a maximum accuracy of 91.98% for unknown fault categories under zero-shot learning conditions. These results indicate that the combination of attribute-based incremental learning and the CNN-Transformer architecture is capable of effectively identifying unknown fault types in marine diesel engines. It should be noted that the number of samples for unknown fault classes is significantly smaller than that of known fault classes, which reflects realistic industrial scenarios but may influence diagnostic performance in practical applications.

Nevertheless, several limitations of this study should be acknowledged. The experimental data were collected from a single diesel engine platform under fixed operating conditions, and the diversity of fault types and operating scenarios remains limited. In addition, the imbalance between known and unknown fault samples may affect the robustness of the diagnostic results. Furthermore, zero-shot fault diagnosis inherently faces challenges due to the absence of labeled data for unseen faults. Future work will focus on validating the proposed method under multi-platform conditions, variable operating parameters, and more complex fault scenarios, as well as exploring multimodal information fusion to further enhance diagnostic robustness and

generalization capability.

Author contributions: Conceptualization, YW (Yongjian Wang) and YW (Yingying Wu); methodology, YW (Yongjian Wang); software, YW (Yingying Wu); validation, YW (Yingying Wu), YW (Yongjian Wang), and GL; formal analysis, YW (Yingying Wu); investigation, YW (Yingying Wu); resources, HC; data curation, YW (Yingying Wu); writing—original draft preparation, YW (Yingying Wu); writing—review and editing, GL; visualization, YW (Yingying Wu); supervision, YW (Yongjian Wang); project administration, YW (Yongjian Wang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 52372361, 52475104). The APC was funded by the same grants.

Institutional review board statement: Not applicable.

Informed consent statement: Not applicable.

Data availability statement: Not applicable.

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

References

1. Wang L. Research on Health State Assessment and Fault Diagnosis of Marine Diesel Engine Based on Performance Parameters [Master's Thesis]. Harbin Engineering University; 2023. (in Chinese)
2. Wang C, Lu N, Cheng Y, et al. A Data-Driven Aero-Engine Degradation Prognostic Strategy. IEEE Transactions on Cybernetics. 2021; 51(3): 1531–1541. doi: 10.1109/TCYB.2019.2938244
3. Wang X, Jiang B, Ding SX, et al. Extended Relevance Vector Machine-Based Remaining Useful Life Prediction for DC-Link Capacitor in High-Speed Train. IEEE Transactions on Cybernetics. 2022; 52(9): 9746–9755. doi: 10.1109/TCYB.2020.3035796
4. Chen C, Shi J, Shen M, et al. Pseudo-Label Guided Sparse Deep Belief Network Learning Method for Fault Diagnosis of Radar Critical Components. IEEE Transactions on Instrumentation and Measurement. 2023; 72: 1–12. doi: 10.1109/TIM.2023.3256474
5. Bai H, Zhan X, Yan H, et al. Research on Diesel Engine Fault Diagnosis Method Based on Stacked Sparse Autoencoder and Support Vector Machine. Electronics. 2022; 11(14): 2249. doi: 10.3390/electronics11142249
6. Wang S, Lei Y, Lu N, et al. Graph Continual Learning Network: An Incremental Intelligent Diagnosis Method of Machines for New Fault Detection. IEEE Transactions on Automation Science and Engineering. 2024; 1–11. doi: 10.1109/TASE.2024.3417208
7. Li X, Zhang W, Li X, et al. Partial Domain Adaptation in Remaining Useful Life Prediction With Incomplete Target Data. IEEE/ASME Transactions on Mechatronics. 2024; 29(3): 1903–1913. doi: 10.1109/TMECH.2023.3325538
8. Lee D, Lee JG, Choi M, et al. Multi-fidelity sub-label-guided transfer network with physically interpretable synthetic datasets for rotor fault diagnosis. Engineering Applications of Artificial Intelligence. 2025; 148: 110467. doi: 10.1016/j.engappai.2025.110467
9. Lei Y, Li N, Li X. Big Data-Driven Intelligent Fault Diagnosis and Prognosis for Mechanical Systems. Springer; 2023. doi: 10.1007/978-981-16-9131-7
10. Chen X, Li X, Yu S, et al. Dynamic Vision Enabled Contactless Cross-Domain Machine Fault Diagnosis with Neuromorphic Computing. IEEE/CAA Journal of Automatica Sinica. 2024; 11(3): 788–790. doi: 10.1109/JAS.2023.124107
11. Dong H, Tian Z, Spencer JW, et al. Bilevel Optimization of Sizing and Control Strategy of Hybrid Energy Storage System in Urban Rail Transit Considering Substation Operation Stability. IEEE Transactions on Transportation Electrification. 2024; 10(4): 10102–10114. doi: 10.1109/TTE.2024.3385821

12. Yang B, Xu S, Lei Y, et al. Multi-source transfer learning network to complement knowledge for intelligent diagnosis of machines with unseen faults. Mechanical Systems and Signal Processing. 2022; 162: 108095. doi: 10.1016/j.ymssp.2021.108095
13. Chen J, Wang G, Lv J, et al. Open-Set Classification for Signal Diagnosis of Machinery Sensor in Industrial Environment. IEEE Transactions on Industrial Informatics. 2023; 19(3): 2574–2584. doi: 10.1109/TII.2022.3169459
14. Meire M, Van Baelen Q, Ooijevaar T, et al. Constraint Guided Autoencoders to Enforce a Predefined Threshold on Anomaly Scores: An Application in Machine Condition Monitoring. Journal of Dynamics, Monitoring and Diagnostics. 2023; 2(2). doi: 10.37965/jdmd.2023.234
15. Zhong Z, Mi J, Zhao Y, et al. Coordinated Control of the Onboard and Wayside Energy Storage System of an Urban Rail Train Based on Rule Mining. Urban Rail Transit. 2024; 10(3): 232–247. doi: 10.1007/s40864-024-00223-7
16. Zhang X, Wang B, Chen X. Intelligent fault diagnosis of roller bearings with multivariable ensemble-based incremental support vector machine. Knowledge-Based Systems. 2015; 89: 56–85. doi: 10.1016/j.knosys.2015.06.017
17. Shi M, Ding C, Chang S, et al. Cross-Domain Class Incremental Broad Network for Continuous Diagnosis of Rotating Machinery Faults Under Variable Operating Conditions. IEEE Transactions on Industrial Informatics. 2024; 20(4): 6356–6368. doi: 10.1109/TII.2023.3345449
18. Fu Y, Cao H, Chen X, et al. Broad auto-encoder for machinery intelligent fault diagnosis with incremental fault samples and fault modes. Mechanical Systems and Signal Processing. 2022; 178: 109353. doi: 10.1016/j.ymssp.2022.109353
19. Li M, Zhu Z. Spatial-Temporal Fusion Graph Neural Networks for Traffic Flow Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence. 2021; 35(5): 4189–4196. doi: 10.1609/aaai.v35i5.16542
20. Wang S, Liu T, Luo K, et al. Identification of engine faults based on acoustic emission signals using a 1DCNN-ViT ensemble model. Measurement Science and Technology. 2023; 34(2): 024007. doi: 10.1088/1361-6501/aca041
21. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. arXiv preprint. 2017. doi: 10.48550/ARXIV.1706.03762
22. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing. 2021; 452: 48–62. doi: 10.1016/j.neucom.2021.03.091
23. Zhang YH, Shao F, Zhao XP, et al. Rolling bearing fault diagnosis based on multi-label zero-shot learning. Journal of Vibration and Shock. 2022; 41(11): 55–64. (in Chinese)