Mathematical Biosciences and Engineering

MBE, 21(12): 7688–7706.
DOI: 10.3934/mbe.2024338
Received: 17 September 2024
Revised: 16 November 2024
Accepted: 22 November 2024
Published: 4 December 2024

https://www.aimspress.com/journal/MBE

Research article

Research on bearing fault diagnosis based on a multimodal method

Hao Chen[1,2,*], Shengjie Li[1], Xi Lu[1,2], Qiong Zhang[1,2], Jixining Zhu[1] and Jiaxin Lu[1]

[1] School of Information Engineering, Nantong Institute of Technology, Nantong 226002, Jiangsu, China
[2] Division of Information and Communication Convergence Engineering, Mokwon University, Daejeon 35349, Korea

* **Correspondence:** Email: chenhao@ntit.edu.cn.

Abstract: As an essential component of mechanical systems, bearing fault diagnosis is crucial to ensure the safe operation of the equipment. However, vibration data from bearings often exhibit non-stationary and nonlinear features, which complicates fault diagnosis. To address this challenge, this paper introduces a novel multi-scale time-frequency and statistical features fusion model (MTSF-FM). Specifically, the method first employs continuous wavelet transform to generate time-frequency images, capturing local and global features of the signal at different scales. Contrast enhancement techniques are then used to improve the visual quality of these images. Next, features are extracted from the time-frequency images using a visual geometry group network to obtain deep features of image modalities. In parallel, 13 key features are extracted from the original vibration data in the time-frequency domain. Convolutional neural networks are then employed for deep feature extraction. Experimental results demonstrate that MTSF-FM achieves accuracies of 98.5% and 95.1% on two public datasets. These findings highlight the effectiveness of MTSF-FM in analyzing complex vibration data and propose a novel method for bearing fault diagnosis.

Keywords: bearing fault diagnosis; non-stationary; nonlinear; time-frequency; deep feature

1. Introduction

Mechanical equipment is the cornerstone of modern industry, and its normal operation is critical to the productivity and safety of various sectors. However, the increasing complexity of industrial systems and the high cost of data collection pose significant challenges for fault diagnosis [1]. As key

components of mechanical equipment, bearings directly impact the performance and lifespan of the equipment due to their operational condition [2]. However, despite significant advances in bearing fault diagnosis technologies, the inherent diversity and complexity of bearing data continue to challenge accurate diagnosis [3]. In particular, the non-stationary and nonlinear characteristics of vibration signals further complicate feature extraction, making it essential to develop advanced techniques for effective fault diagnosis [4].

When a bearing is defective, various abnormal signals such as vibration, temperature, and current are generated [5]. Among them, vibration signals are widely used in bearing fault diagnosis due to their low collection cost and high sensitivity to small fault signs [6]. Vibration data typically exhibit non-stationary and nonlinear features, making the signal features extremely complex. Therefore, the key to improving the accuracy of bearing fault diagnosis lies in effectively extracting these complex vibration signals.

Analysis methods based on vibration signals mainly include signal processing and machine learning techniques [7,8]. These methods provide powerful technical support for the accurate diagnosis of bearing faults. Signal processing techniques extract key fault features by processing and analyzing vibration signals in detail. To achieve this, signal denoising and filtering techniques are commonly used to reduce background noise and interference, making fault features more prominent. During feature extraction, considering the non-stationary and nonlinear features of bearing fault signals, it is challenging to fully capture the complex fault features by relying solely on time-domain or frequency-domain analysis. Therefore, time-frequency analysis methods such as the short-time Fourier transform and wavelet transform are widely used. These methods can simultaneously analyze both time and frequency domain features of the signal, revealing its local features [9,10]. Tao et al. [11] proposed an unsupervised fault diagnosis method based on wavelet packet decomposition and reconstruction. They extracted the energy eigenvectors of bearing vibration signals to generate two-dimensional time-frequency images, thus visualizing the fault features. Similarly, Che et al. [12] used time-domain and frequency-domain metrics to pre-train bearing vibration signals, effectively achieving fault diagnosis.

With the development of data processing technology, machine learning methods are gradually being introduced into the field of bearing fault diagnosis. Methods such as autoencoders [13] and flow-aware networks [14] have made significant progress in processing complex data. These techniques optimize feature extraction and model performance, offering new insights for machinery fault diagnosis. Deep learning, a subset of machine learning, has shown great potential in this domain due to its superior ability in feature extraction and nonlinear mapping [15–18]. In particular, the emergence of convolutional neural networks (CNNs) has greatly advanced this field. Janssens first applied CNNs for bearing fault diagnosis in 2016 [19]. Subsequently, researchers have proposed various methods to improve the performance of CNNs [20–22]. For example, Choudhary et al. [23] used LeNet-5 and an artificial neural network to diagnose faults in rotating machinery bearings and validated the effectiveness of LeNet-5. Chen et al. [24] inputted the original vibration signal into a CNN, and the network used two convolution kernels of different sizes to extract features at different frequencies. These features were then fed into the long short-term memory (LSTM) for fault recognition. Han et al. [25] used the time-domain image of bearing vibration data as input, extracted features through a CNN, and ultimately achieved bearing state recognition using a support vector machine (SVM). Considering that the continuous wavelet transform (CWT) is effective in analyzing the non-stationary and nonlinear features in bearing vibration data, Fu et al. [26] first applied CWT to the bearing vibration signal to obtain the transformed time-domain signal. Then, they extracted temporal features

of the signal using bidirectional long short-term memory (BiLSTM), while spatial features were extracted using CNN. Song et al. [27] proposed the CNN-BiLSTM model to address challenges such as the insufficient utilization of temporal features and the high cost of parameter tuning. They used particle swarm optimization to optimize the training hyperparameters of the network, significantly improving fault diagnosis performance. Dong et al. [28] proposed a self-attention-enhanced CNN and introduced empirical wavelet transform to decompose the original signal into three frequency components. This strategy enables the model to capture key information in the signal more efficiently. Additionally, challenges such as small data issues and label noise further complicate fault diagnosis tasks [29]. To address these challenges, Wang et al. [30] proposed an enhanced transformer with asymmetric loss function for few-shot fault diagnosis with noisy labels, significantly improving diagnostic accuracy. Zhang et al. [31] introduced a federated learning framework to tackle fault diagnosis problems arising from data privacy concerns and variations in data distribution. Although these studies have made significant progress in feature extraction, network structure design, model optimization, and handling small data samples, the feature representation of a single modality still limits the ability of the model to comprehensively diagnose complex failure modes. Furthermore, while increasing the complexity of the network structure can improve model performance, it also leads to a significant increase in computational complexity.

In summary, researchers significantly improved the accuracy of fault diagnosis through continuous optimization of signal processing technologies and machine learning algorithms. However, the inherent non-stationary, nonlinear features of vibration data still pose a challenge to traditional single-modality representation methods. CNNs are effective at capturing the local features and hierarchical structure of vibration data; however, when confronted with complex and variable failure modes, the extraction method relying only on time-frequency features may still be insufficient. To address these challenges, a novel multi-scale time-frequency and statistical features fusion model (MTSF-FM) is proposed. The main contributions of this paper are as follows:

1) Multi-scale time-frequency feature extraction: using CWT to generate time-frequency images and enhance feature recognition through contrast enhancement.

2) Comprehensive feature fusion: key features from time, frequency, and time-frequency domains are fused with multi-scale time-frequency images using CNN.

3) Multimodal deep network: a multimodal network is designed to combine data from multiple modalities, overcoming the limitations of single-modality methods.

The rest of the paper is organized as follows: Section 2 describes the structure of MTSF-FM in detail. Section 3 describes the experimental setup, evaluation metrics, comparison experiments, and ablation experiments and discusses the results. Section 4 contains a conclusion.

2. Methodology

The fusion of statistical and time-frequency features is a critical aspect of the MTSF-FM. Figure 1 illustrates the MTSF-FM structure. The network is divided into two branches, each of which extracts features for different types of input data. In the first branch, after data segmentation, a vector set containing thirteen statistical features is computed to capture the overall features of the vibration signal. These statistical features (such as mean, standard deviation, and kurtosis) are then processed through an embedding layer, which converts them into fixed-length vector representations suitable for further analysis. These vectorized statistical features are passed through a CNN to extract hierarchical

representations of the features via convolution and pooling operations.

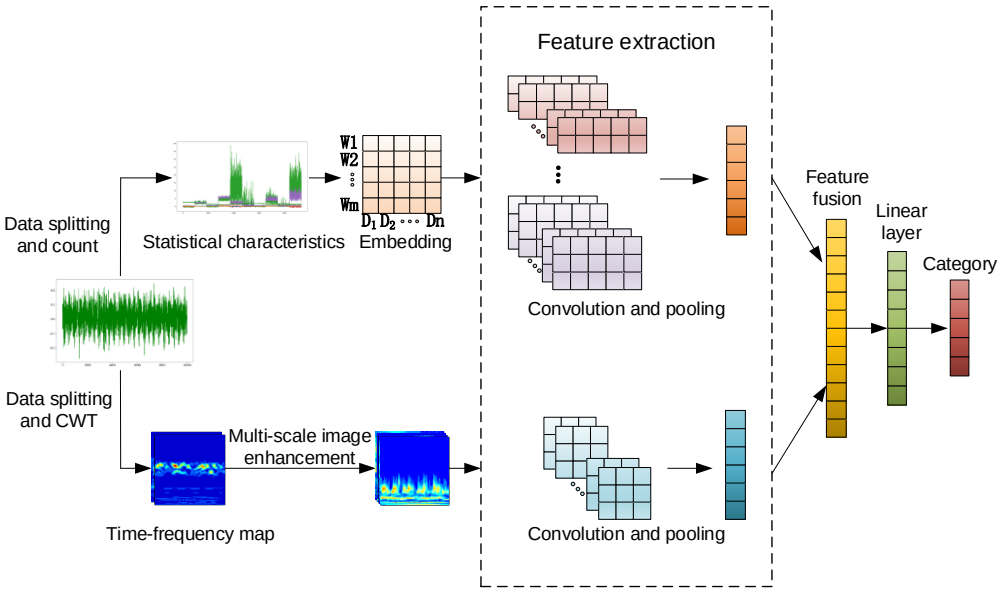

Figure 1. MTSF-FM structure.

In the second branch, the vibration signal is first processed using CWT to capture multi-scale time-frequency features, which highlight local variations in both frequency and time. These features are then visualized as time-frequency images. To improve the clarity of these images, a contrast enhancement technique is applied, making key features more visible. The enhanced images are subsequently passed through a visual geometry group (VGG) network, where convolutional and pooling layers are used to extract deep features. This allows the network to capture important multi-scale information at various resolutions. These features are then fused in a fusion layer, where the statistical and time-frequency features are concatenated into a single multimodal feature vector. This fusion strategy integrates complementary information from both feature types, improving the ability of the model to represent the vibration signal and enhancing fault diagnosis accuracy.

Finally, the fused feature vector is processed through a fully connected layer to predict the final class. This fully connected layer consolidates the multimodal feature information and provides the model with the capability to make the final fault classification.

2.1. Time-frequency feature extraction

To fully capture the time-frequency features of the vibration signals, the CWT is applied to convert the one-dimensional vibration signals into two-dimensional time-frequency images. First, the original vibration signal is sampled, and then the transformation is performed using the complex Morlet wavelet, as shown in Eq (1).

$$\psi(t) = \pi^{-1/4} e^{i\omega_0 t} e - e^{-t^2/2} \tag{1}$$

where ω_0 is the frequency parameter and i is the imaginary unit. To perform the CWT on signals at

different scales and obtain the wavelet coefficients, $CWT(a,\tau)$ is expressed as in Eq (2).

$$CWT(a,\tau) = \int_{-\infty}^{\infty} x(t)\,\psi^*(\frac{t-\tau}{a})dt \tag{2}$$

where $x(t)$ is the original vibration signal, a is the scale parameter, τ is the translation parameter, and $\psi^*(t)$ is the complex conjugate of the wavelet function $\psi(t)$. Next, the modulus is calculated from the wavelet coefficients, which reflect the energy distribution of the signal across different times and frequencies. The modulus is computed as shown in Eq (3).

$$Energy(a,\tau) = |CWT(a,\tau)| \tag{3}$$

Finally, the calculated energy distribution data are used to generate time-frequency images.

To fully capture the signal features at different scales, a multi-scale feature strategy is proposed. Specifically, the strategy applies wavelet transforms on two sets of scale ranges to generate two different sets of time-frequency images $CWT_{\text{scale}_j}(a,\tau), for\ j = 1,2.$

To optimize the quality of the time-frequency images and enhance their representation, a linear contrast stretching method is applied to improve the image contrast. Suppose each pixel value in the image is I, and the stretched I' is given by Eq (4).

$$I' = \alpha \cdot I + \beta \tag{4}$$

where $\alpha = 1.5$ is the gain factor and β is the offset.

2.2. Statistical feature extraction

In feature selection, this paper considers the complexity and non-stationarity of the signals, ensuring that the selected features can comprehensively reflect the features of the fault signals. Thirteen features are extracted from the original data, covering the time domain, the frequency domain, and the time-frequency domain. Specifically, the time-domain features include mean, standard deviation, kurtosis, skewness, peak-to-peak, root mean square, signal-to-noise ratio, and variance. The frequency-domain features include spectral centroid, spectral bandwidth, and dominant frequency, while the time-frequency domain features include wavelet energy and wavelet entropy. The selection of these features is based on the studies by Dhamande et al. [32] and Altaf et al. [33]. The former emphasizes the importance of time-domain and frequency-domain features in bearing fault diagnosis, while the latter proposes a vibration signal fault diagnosis method based on statistical features. These studies demonstrate the effectiveness of selecting these features in improving diagnostic accuracy.

Time-domain features reflect the overall trend and fluctuation of the signal. For example, the mean, standard deviation, kurtosis, and skewness help identify the underlying fluctuation patterns of a signal, while the signal-to-noise ratio helps assess the signal quality and determine whether it is affected by noise. Frequency-domain features can identify specific frequency components in the fault signal by analyzing the spectral information of the signal. For example, spectral centroid and spectral bandwidth reflect the distribution features of the signal energy, and dominant frequency indicates the main vibration frequency in the signal. Time-frequency features combine information from both time and frequency dimensions. By analyzing the energy distribution of a signal at different times and frequencies, transient frequency changes and local features in the signal can be captured. In bearing

fault diagnosis, fault signals are often accompanied by rapidly changing frequency components, which are difficult to adequately capture by traditional time or frequency domain analysis. Wavelet energy and entropy can capture these transient changes in both time and frequency domains, improving fault diagnosis accuracy.

As bearing fault diagnosis faces problems such as signal complexity and non-stationarity, a single modality feature often cannot fully reflect all the information in the signal. Therefore, the integrated time-domain, frequency-domain, and time-frequency features can comprehensively characterize the signal from multiple perspectives and improve the accuracy of fault diagnosis.

2.3. Multimodal feature fusion

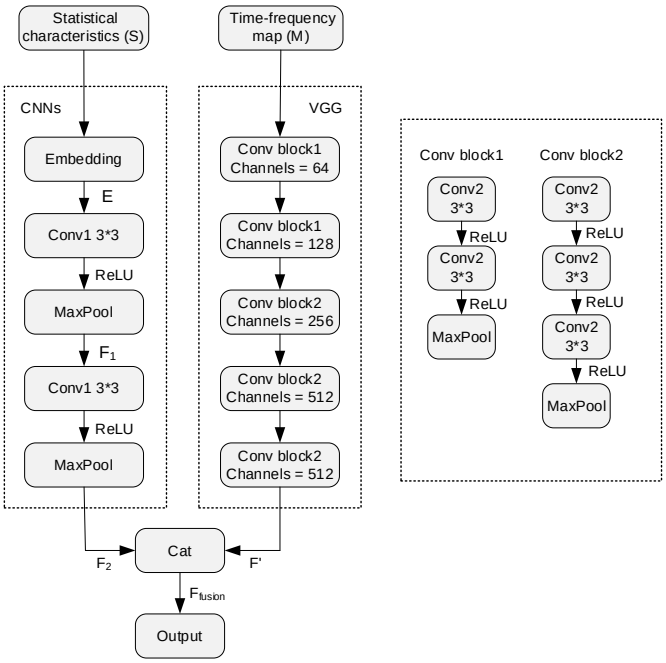

Figure 2. Two-branch network structure.

Figure 2 illustrates the two-branch network structure proposed in this paper. The first branch is dedicated to statistical feature extraction. The statistical features are vectorized and represented through the embedding layer and then passed into CNNs for feature extraction. The CNNs include two sets of convolutional and pooling layers. Suppose S is the input statistical feature; the vector obtained after the embedding layer is derived based on Eq (5).

$$E = Embedding(S) \tag{5}$$

The output from the first set of convolution and pooling layers is given by Eq (6).

$$F_1 = Maxpool(ReLU(Conv(E))) \tag{6}$$

where $Maxpool$ denotes the maximum pooling operation, $ReLU$ denotes the rectified linear unit, and $Conv$ denotes the convolution operation. The output from the second set of convolution and pooling layers is given by Eq (7).

$$F_2 = Maxpool(ReLU(Conv(F_1))) \tag{7}$$

The second branch is the time-frequency image feature extraction network. Specifically, this VGG includes five convolutional blocks. Convolutional block 1 includes two convolutional layers and a pooling layer, and convolutional block 2 contains three convolutional layers and a pooling layer. Suppose M is the input feature map; the output after processing by convolution block 1 can be expressed by Eq (8).

$$F_{channel=c} = Maxpool(ReLU(Conv(ReLU(Conv(M))))) \tag{8}$$

where c denotes the number of output channels in the convolutional layer. The structure of convolutional block 2 is expanded compared to convolutional block 1, specifically by an additional convolutional layer. Therefore, more convolution operations are required when implementing convolution block 2. The feature map obtained after passing through the five convolutional blocks is given by Eq (9).

$$F' = ConvBlock(F_{channel=c}, F'_{channel=c}) \tag{9}$$

where $ConvBlock$ denotes the convolution block operation, and the values of c are 64, 128, 256, 512, and 512. Finally, the features extracted by the CNNs and VGG are fused, and the resulting features are shown in Eq (10).

$$F_{fusion} = Cat(F_2, F') \tag{10}$$

where Cat denotes the concatenation operation of the features.

3. Experiments and discussion

Table 1. Parameter setting.

Parameter	Value
Window size	512
Window step	256
Learning rate	0.00001
Epoch	100
Batch size	32
Optimizer	Adaptive moment estimation
Loss function	Cross entropy loss

This paper validates the effectiveness of MTSF-FM using two public datasets. The experimental setup includes the Ubuntu operating system, two NVIDIA RTX 4090 GPUs (24 GB of video memory), and CUDA version 11.7. The parameter settings are provided in Table 1.

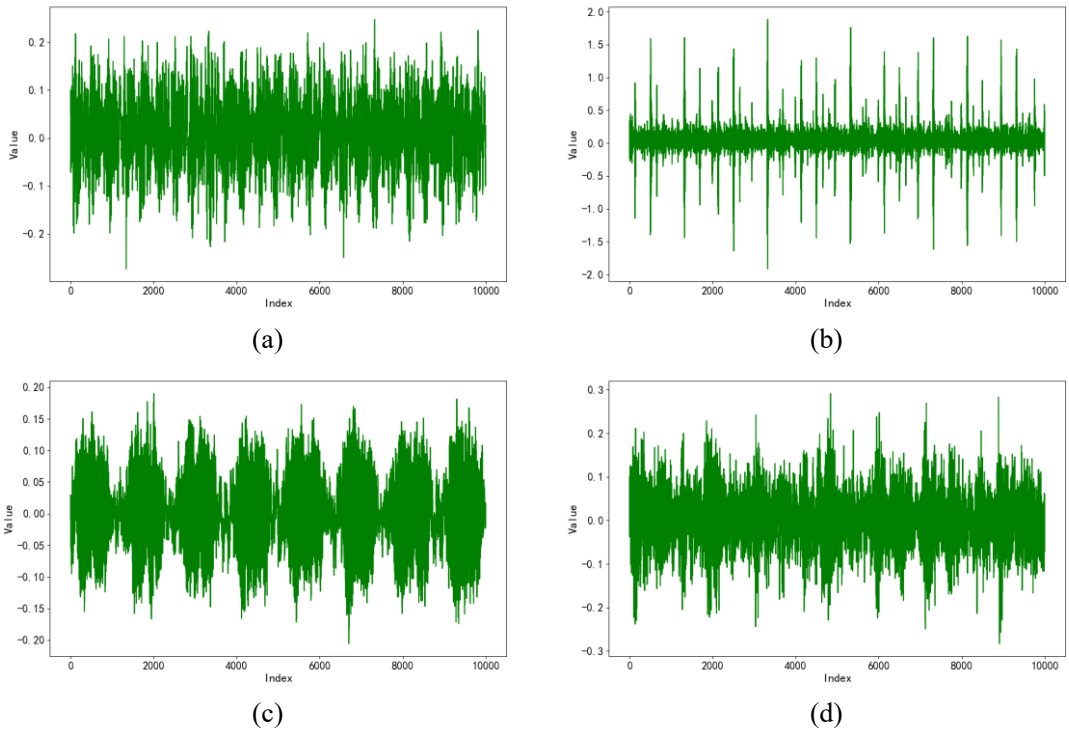

Figure 3. Bearing vibration data timing diagram. (a) Normal. (b) Inner race fault. (c) Normal. (d) Inner race fault.

Case Western Reserve University (CWRU) [34] dataset: This dataset contains bearing vibration data from various failure modes and operating conditions. The data are collected using accelerometers placed on the drive end, fan end, and base of the motor casing. The data from these three different locations reflect different physical phenomena, providing complementary information for fault diagnosis. The dataset includes normal (N), inner race fault (IRF), ball fault (BF), and outer race fault (ORF) categories. For this paper, the drive end bearing fault data are sampled at 12 kHz, with fault diameters of 7, 14, and 21 mils. The experimental dataset consists of 10 sets: 1 set of normal data and 9 sets of fault data.

Spectra Quest (SQ) [35] dataset: This dataset is obtained from the SQ experimental platform at Xi'an Jiaotong University. The experiment uses a comprehensive mechanical fault simulation test bed to simulate motor bearing ORF and IRF. Vibration signals of varying severity (mild, moderate, and severe) are collected at three rotational frequencies. For this paper, data corresponding to three different fault severities at the same rotational frequency are analyzed. The experimental dataset consists of a total of 7 sets: 1 set of normal data and 6 sets of fault data. Figure 3 shows the bearing vibration signals under N and IRF conditions.

3.1. Evaluation metrics

To comprehensively evaluate the performance of MTSF-FM in bearing fault diagnosis, accuracy, precision, recall, and F1 score [20] are used for evaluation. Suppose TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives. Accuracy, precision, recall, and F1 score are calculated as shown in Eq (11) to (14).

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{11}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{12}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{13}$$

$$F1score = \frac{2 * Rec\,all * Pre\,cision}{Rec\,all + Pre\,cision} \tag{14}$$

3.2. CWRU experiment

The CR dataset contains 10 sets of bearing data with 119,808 rows in each set. The data is segmented using a sliding window with a window size of 512 and a step size of 256, resulting in 466 segments per set. Therefore, the total number of segments for all 10 sets is 4660. Thirteen statistical features are computed for each data segment, forming a feature matrix of size (4660, 13). Additionally, 4660 time-frequency images are generated by applying CWT to each of the two data segments. The dataset is split into training, validation, and test sets. The training set contains 3260 samples, the validation set contains 936 samples, and the test set contains 464 samples.

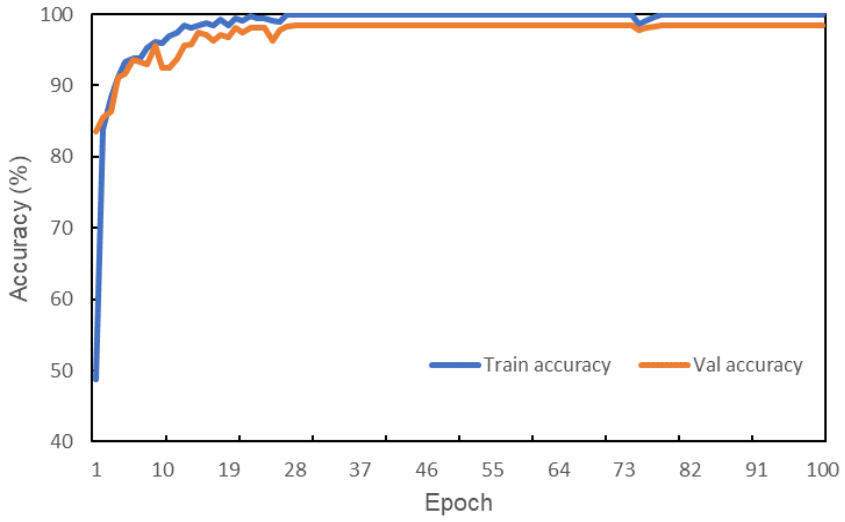

Figure 4. Accuracy trend during training on the CWRU dataset.

Figure 4 shows the accuracy trends of the training and validation sets during the training process. After ten epochs, both training and validation accuracy reaches approximately 95%. As the number of epochs increases, training accuracy approaches 100%, while validation accuracy approaches 99%. The overall trend indicates that MTSF-FM learns effectively and generalizes well on the validation set.

Table 2. Experimental results on the CWRU dataset.

Category	Precision (%)	Recall (%)	F1 score (%)
N	100.00	96.97	98.46
7-IRF	100.00	100.00	100.00
7-BF	93.88	95.83	94.85
7-ORF	100.00	100.00	100.00
14-IRF	100.00	97.92	98.95
14-BF	94.00	97.92	95.92
14-ORF	100.00	100.00	100.00
21-IRF	100.00	97.92	98.95
21-BF	97.92	97.92	97.92
21-ORF	100.00	100.00	100.00
Average	98.6	98.4	98.5

Table 2 presents the experimental results of MTSF-FM on the CWRU test set, including precision, recall, and F1 scores for 10 categories. The F1 score ranges from 94.85% to 100.00%, with an average F1 score of 98.5%. Additionally, the precision, recall, and F1 score for categories 7-IRF, 7-ORF, 14-ORF, and 21-ORF reach 100%, indicating that the MTSF-FM can completely and accurately identify these faults. Even for the more challenging 7-BF category, the F1 score reaches 94.85%. These results demonstrate that MTSF-FM achieves balanced and excellent classification performance across different fault types.

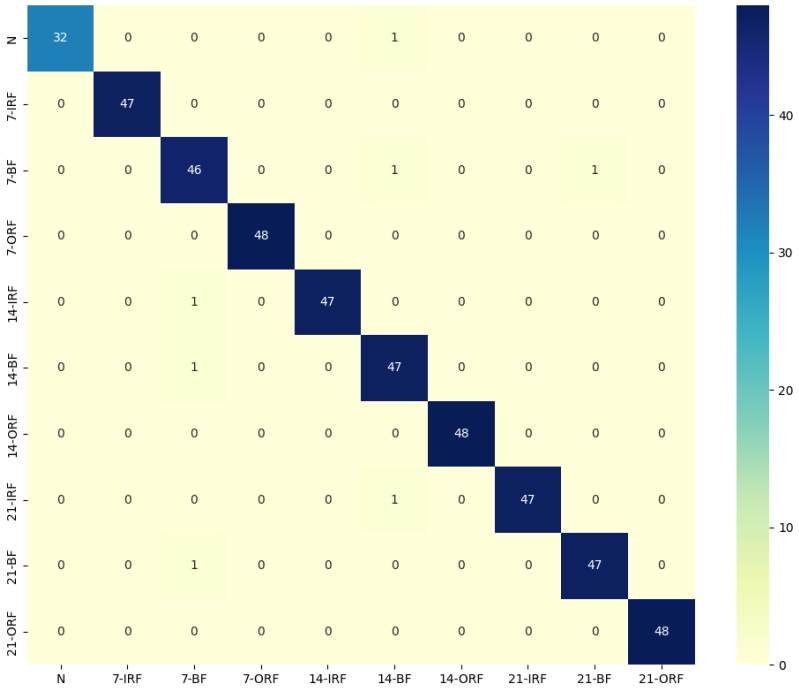

Figure 5. Confusion matrix on the CWRU dataset.

Figure 5 shows the confusion matrix of MTSF-FM on the CWRU test set. The diagonal elements represent correctly classified instances, while the off-diagonal elements indicate misclassified instances. The matrix shows that most samples are correctly categorized. For example, category N

has 32 correctly classified samples, with only 1 misclassified as category 14-BF. Similarly, all instances of category 14-ORF are predicted correctly. These results further highlight the high accuracy of MTSF-FM in fault diagnosis.

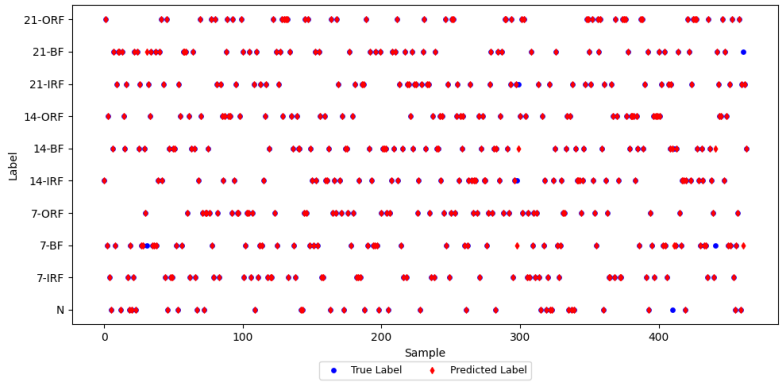

Figure 6. Comparison results of predicted and true values in the CWRU test set.

Figure 6 compares the true and predicted labels of each sample on the CWRU test set. Most samples show complete overlap between the true and predicted labels, indicating that MTSF-FM accurately classifies most samples. A few samples show inconsistencies in their predictions, mainly concentrated on specific instances, but the overall error is minimal. These results further validate the effectiveness of MTSF-FM in the classification task, supporting the findings in Table 2 and the confusion matrix.

3.3. SQ experiment

The dataset contains seven sets of bearing data, each with approximately 384,000 rows. Using the sliding window, each set is segmented into 1498 segments, resulting in a total of 10,486 segments across the seven datasets. Additionally, 10,486 time-frequency images are generated by applying CWT to each data segment. The dataset is divided into training, validation, and test sets, with 7340 samples in the training set, 2098 in the validation set, and 1048 in the test set.

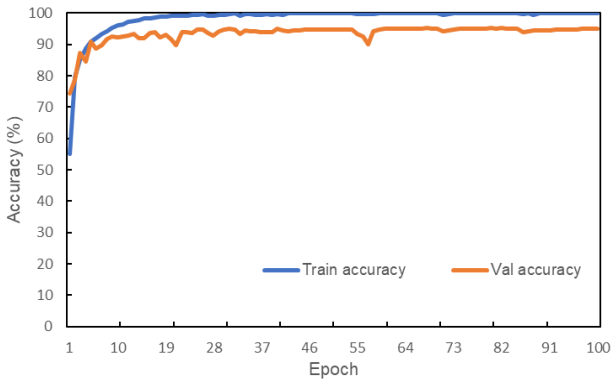

Figure 7. Accuracy trend during training on the SQ dataset.

Figure 7 shows the accuracy trend of MTSF-FM during training on the SQ dataset. As the number of epochs increases, accuracy steadily improves, indicating the effective learning ability of the model. By epoch 84, the accuracy on the validation set reaches 95.1%, highlighting a strong performance and generalization ability of the model on the SQ dataset.

Table 3. Experimental results on the SQ dataset.

Category	Precision (%)	Recall (%)	F1 score (%)
IRF1	90.51	94.08	92.26
IRF2	92.62	90.79	91.69
IRF3	97.32	95.39	96.35
ORF1	95.95	93.42	94.67
ORF2	93.63	96.71	95.15
ORF3	97.35	96.71	97.03
N	99.26	99.26	99.26
Average	95.2	95.2	95.2

Table 3 presents the experimental results of MTSF-FM on the SQ test set. The model performs excellently in detecting category N, achieving precision, recall, and F1 score above 99%. It also maintains high precision, recall, and F1 scores for most fault categories, generally above 90%. However, the F1 score for the IRF2 category is slightly lower at 91.69%, indicating that further optimization is needed for this category.

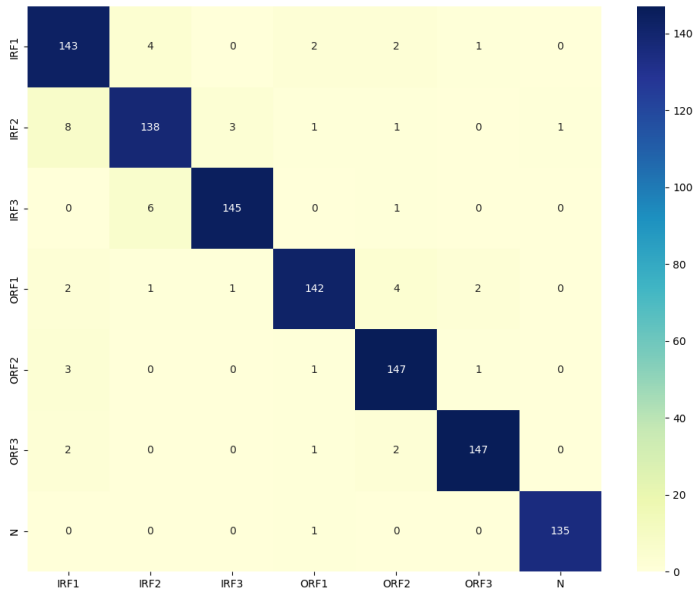

Figure 8. Confusion matrix on the SQ dataset.

Figure 8 shows the confusion matrix of MTSF-FM on the SQ test set. The model performs excellently in most categories. For category N, 135 samples are correctly identified, with only 1 misclassification. Misclassifications occur due to the similarity of features among the categories IRF1, IRF2, and ORF1.

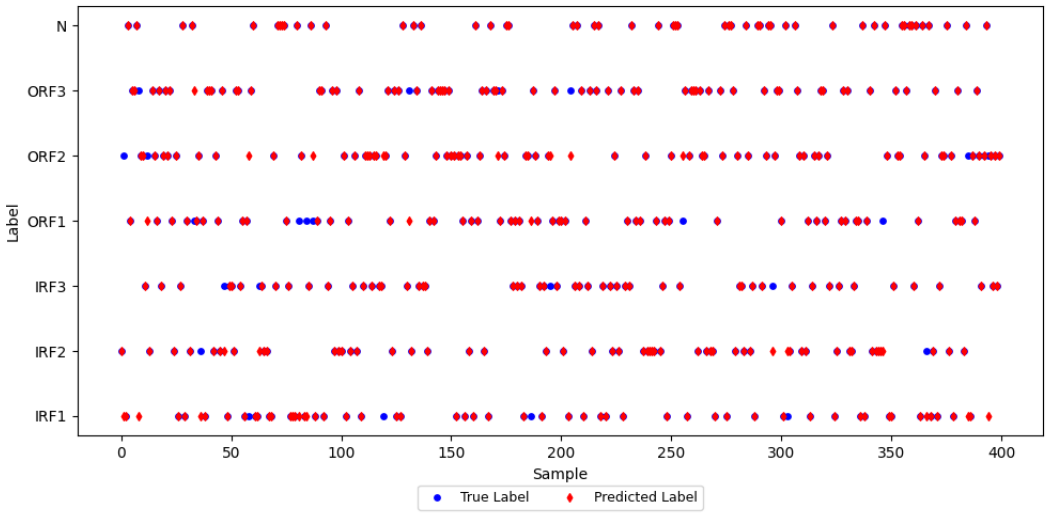

Figure 9. Comparison results of predicted and true values in the SQ test set.

Figure 9 shows the classification comparison results of MTSF-FM on the SQ test set. The blue and red dots in the graph mostly overlap, indicating a high consistency between predicted and true values. However, a few samples in the middle region show biased predictions, highlighting the limitations of the model in classifying certain samples.

3.4. Comparison experiments

Table 4. Comparison experiments result on the CWRU dataset.

Model	Accuracy (%)	Average F1 score (%)
SVM [25]	89.3	89.1
LSTM [16]	93.8	93.7
Transformer [22]	92.7	92.7
LeNet-5 [23]	93.5	93.4
MTSF-FM	98.5	98.5

Table 4 compares the fault diagnosis performance of various models on the CWRU dataset, including SVM, LSTM, Transformer, LeNet-5, and MTSF-FM. MTSF-FM achieves 98.5% in both accuracy and average F1 score, significantly outperforming the other models. Specifically, SVM performs relatively poorly, with 9.2% lower accuracy than MTSF-FM. This demonstrates that SVM has more obvious limitations when dealing with complex and multidimensional bearing fault data. LSTM outperforms SVM in the bearing fault diagnosis with its excellent timing capture capability but still lags MTSF-FM. Despite its powerful self-attention mechanism, Transformer does not fully leverage its advantages on the CWRU dataset, performing slightly worse than LSTM. LeNet-5 effectively extracts features through convolutional layers, achieving an accuracy of 93.5%. In summary, MTSF-FM performs the best with its multimodal fusion strategy.

Table 5. Comparison experiments result on the SQ dataset.

Model	Accuracy (%)	Average F1 score (%)
SVM	88.8	88.9
LSTM	80.8	80.3
Transformer	86.1	85.9
LeNet-5	90.6	90.7
MTSF-FM	95.1	95.2

Table 5 presents the experimental results of several models on the SQ dataset. Among them, MTSF-FM maintains the best performance with an accuracy of 95.1% and an average F1 score of 95.2%. SVM performs well on the SQ dataset, achieving nearly 90% accuracy and average F1 score. This suggests that the data distribution and feature dimensions of the SQ dataset are beneficial to the SVM, enhancing its classification performance. Meanwhile, LSTM performs worse, with lower accuracy and average F1 score compared to SVM and LeNet-5, indicating that it does not fully capture the temporal properties of the data. The performance of the Transformer is average and fails to outperform LeNet-5. LeNet-5, on the other hand, demonstrates the power of CNNs in feature extraction. Even when dealing with different types of datasets, it maintains high classification accuracy, achieving 90.6%. However, compared to MTSF-FM, LeNet-5 is still 4.5% lower in accuracy.

In summary, MTSF-FM demonstrates excellent fault diagnosis performance on both datasets with its multimodal fusion strategy. The results also highlight how different datasets influence model performance, underscoring the importance of selecting an appropriate model based on dataset characteristics in practical applications. In general, when dealing with diverse and complex data, single modality information is insufficient to fully capture fault features. By fusing information from multiple modalities, MTSF-FM effectively overcomes this limitation, providing strong support for research in bearing fault diagnosis.

3.5. Ablation experiments

Table 6. Ablation experiments result on the CWRU dataset.

Module		Accuracy (%)	Average F1 score (%)
Time-frequency features	Statistical features		
✓		96.2	96.2
	✓	93.8	93.5
✓	✓	98.5	98.5

Table 6 demonstrates the ablation experiments result on the CWRU test set. The results demonstrate the complementary roles of time-frequency and statistical features in bearing fault diagnosis. When only time-frequency features are used, the model achieves an accuracy of 96.2%, showing their effectiveness in capturing critical bearing failure information. However, with the addition of statistical features, performance improves significantly, with accuracy rising to 98.5%. This validates the auxiliary role of statistical features in fault diagnosis and emphasizes the importance of multimodal feature fusion for improving model accuracy. Meanwhile, the average F1 score increases from 96.2% to 98.5%, indicating that the model strikes a better balance between precision and recall.

Table 7. Ablation experiments result on the SQ dataset.

Module		Accuracy (%)	Average F1 score (%)
Time-frequency features	Statistical features		
✓		91.2	91.1
	✓	75.0	74.8
✓	✓	95.1	95.2

Table 7 demonstrates the ablation experiments result on the SQ test set. Compared to the CWRU dataset, the experimental results on the SQ dataset exhibit more significant differences. When using only time-frequency features, the model achieves an accuracy of 91.2%. In contrast, when using only statistical features, the accuracy drops to 75.0%, highlighting the limitations of statistical features on this dataset. However, when time-frequency and statistical features are combined, model performance significantly improves, with accuracy increasing to 95.1%. This represents an improvement of 3.9% over time-frequency features alone and 20.1% over statistical features alone. These results underscore the importance of time-frequency features on the SQ dataset and reaffirm the effectiveness of the multimodal feature fusion strategy. Additionally, the improvement in the average F1 score reflects the enhanced stability of the model in fault recognition.

3.6. Discussion

The findings of this paper emphasize the critical importance of multimodal fusion strategies in improving bearing fault diagnosis accuracy. The performance of SVM, LSTM, Transformer, and LeNet-5 is compared across two different datasets.

The strength of MTSF-FM lies in its ability to simultaneously integrate time-frequency and statistical features, effectively capturing local and global fault data features. This method fully leverages complementary information from different data modalities, overcoming the limitations of relying on a single modality for feature representation. Further ablation experiments confirm this; while time-frequency features alone achieve high diagnostic accuracy, combining them with statistical features leads to a significant improvement in system performance. This demonstrates the synergy between the two modalities in fault diagnosis.

The performance of the LSTM and Transformer in the experiments demonstrates their sensitivity to the inherent features of the data. The sequence modeling capabilities of LSTM are not fully realized in the SQ dataset, as the temporal patterns in the data do not align with the expectations of LSTM. Similarly, despite its powerful self-attention mechanism, Transformer does not outperform LeNet-5 on both datasets. On the SQ dataset, SVM outperforms both Transformer and LSTM, suggesting that the underlying assumptions and decision boundaries of SVM align better with the feature space and distribution of this dataset. However, the relatively poor performance of SVM on the CWRU dataset reflects a mismatch between the feature space and distribution and the underlying assumptions of the model. This finding emphasizes the importance of flexibility in adapting the modeling strategy to the features of the dataset. The strong performance of LeNet-5 on both datasets demonstrates the robustness and versatility of CNNs in extracting salient features from complex signals. Its ability to maintain high accuracy across different feature environments highlights the potential of CNNs in bearing fault diagnosis.

Although this study focuses on mechanical fault diagnosis, the design methodology of MTSF-

FM is also applicable to fault detection in other fields. For example, MTSF-FM could potentially be effective in the fields of aero-engine fault detection and automobile electronic control system diagnosis. Applying the model in these areas will help validate its ability to generalize to different operating environments and fault modes. Future research will explore the generalization of the model in these fields to further validate its effectiveness in complex industrial scenarios.

Despite the excellent performance of the MTSF-FM on vibration signals in this study, there are still challenges in applying it to different mechanical equipment or non-vibration datasets (such as acoustic signals and thermal imaging data). Different equipment operating conditions, fault types, and sensor configurations may affect feature extraction and model adaptation. For example, high noise levels or non-stationary signals in these datasets could impair the quality of time-frequency features or statistical features, resulting in lower diagnostic accuracy.

Additionally, MTSF-FM is mainly validated on smaller datasets, and how to maintain the accuracy and efficiency of the model as the dataset size increases still needs to be addressed. In practical industrial applications, improving the computational efficiency and response speed of the model to meet real-time diagnostic demands is another aspect that requires further optimization. Furthermore, although MTSF-FM has proven effective through the multimodal data fusion strategy, future research should focus on finding more efficient ways to fuse features from different sources and dynamically adjust the model structure to handle various types of signals and datasets.

4. Conclusions

This paper proposes a multimodal feature extraction model, MTSF-FM, which integrates CWT and contrast enhancement techniques to improve the richness and visibility of vibration signal features. By combining statistical features from the time, frequency, and time-frequency domains with multi-scale time-frequency image features, MTSF-FM captures a comprehensive set of local and global signal features. Furthermore, the model leverages CNN for deep feature extraction and fusion, enabling the effective integration of diverse information sources. This multimodal method significantly overcomes the limitations of traditional single-modality methods, leading to enhanced diagnostic accuracy in bearing fault detection.

The experimental results validate the effectiveness of MTSF-FM, achieving an accuracy of 98.5% on the CWRU dataset and 95.1% on the SQ dataset.

In the future, we will optimize the multimodal data fusion method to enhance the ability of the model to handle large-scale datasets and real-time diagnostic tasks. Additionally, we will explore the application of MTSF-FM to different mechanical systems and non-vibration datasets. Finally, we will focus on improving the scalability of the model to better adapt to complex working conditions and variable data.

Data availability statement

CWRU dataset: The data that support the findings of this study are openly available in [https://engineering.case.edu/bearingdatacenter/download-data-file].

SQ dataset: The data that support the findings of this study are openly available in [https://github.com/sliu7102/SQ-dataset-with-variable-speed-for-fault-diagnosis].

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. C. Li, S. Li, Y. Feng, K. Gryllias, F. Gu, M. Pecht, Small data challenges for intelligent prognostics and health management: A review, *Artif. Intell. Rev.*, **57** (2024), 214. https://doi.org/10.1007/s10462-024-10820-4
2. X. Chen, R. Yang, Y. Xue, M. Huang, R. Ferrero, Z. Wang, Deep transfer learning for bearing fault diagnosis: A systematic review since 2016, *IEEE Trans. Instrum. Meas.*, **72** (2023), 1–21. https://doi.org/10.1109/TIM.2023.3244237
3. G. J. Jang, M. C. Noh, S. S. Kim, C. S. Shin, S. S. Lee, C. J. Lee, Vibration data feature extraction and deep learning-based preprocessing method for highly accurate motor fault diagnosis, *J. Comput. Des. Eng.*, **10** (2023), 204–220. https://doi.org/10.1093/jcde/qwac128
4. Y. Xue, C. Wen, Z. Wang, W. Liu, G. Chen, A novel framework for motor bearing fault diagnosis based on multi-transformation domain and multi-source data, *Knowledge-Based Syst.*, **283** (2024), 111205. https://doi.org/10.1016/j.knosys.2023.111205
5. M. Sohaib, M. J. Kim, Fault diagnosis of rotary machine bearings under inconsistent working conditions, *IEEE Trans. Instrum. Meas.*, **69** (2019), 3334–3347. https://doi.org/10.1109/TIM.2019.2933342
6. J. Pacheco-Chérrez, A. J. Fortoul-Díaz, F. Cortés-Santacruz, M. L. Aloso-Valerdi, I. D. Ibarra-Zarate, Bearing fault detection with vibration and acoustic signals: Comparison among different machine leaning classification methods, *Eng. Fail. Anal.*, **139** (2022), 106515. https://doi.org/10.1016/j.engfailanal.2022.106515
7. J. Jiao, M. Zhao, J. Lin, K. Liang, A comprehensive review on convolutional neural network in machine fault diagnosis, *Neurocomputing*, **417** (2020), 36–63. https://doi.org/10.1016/j.neucom.2020.07.088
8. H. Wang, Z. Liu, D. Peng, Z. Cheng, Attention-guided joint learning CNN with noise robustness for bearing fault diagnosis and vibration signal denoising, *ISA Trans.*, **128** (2022), 470–484. https://doi.org/10.1016/j.isatra.2021.11.028
9. N. Liu, T. Schumacher, Y. Li, L. Xu, B. Wang, Damage detection in reinforced concrete member using local time-frequency transform applied to vibration measurements, *Buildings*, **13** (2023), 148. https://doi.org/10.3390/buildings13010148
10. P. Zhou, S. Chen, Q. He, D. Wang, Z. Peng, Rotating machinery fault-induced vibration signal modulation effects: A review with mechanisms, extraction methods and applications for diagnosis, *Mech. Syst. Signal Process.*, **200** (2023), 110489. https://doi.org/10.1016/j.ymssp.2023.110489
11. H. Tao, J. Qiu, Y. Chen, V. Stojanovic, L. Cheng, Unsupervised cross-domain rolling bearing fault diagnosis based on time-frequency information fusion, *J. Franklin Inst.*, **360** (2023), 1454–1477. https://doi.org/10.1016/j.jfranklin.2022.11.004

12. C. Che, H. Wang, X. Ni, Q. Fu, Domain adaptive deep belief network for rolling bearing fault diagnosis, *Comput. Ind. Eng.*, **143** (2020), 106427. https://doi.org/10.1016/j.cie.2020.106427
13. K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, Y. Xu, Autoencoders and their applications in machine learning: A survey, *Artif. Intell. Rev.*, **57** (2024), 28. https://doi.org/10.1007/s10462-023-10662-6
14. S. Abdollahi, A. Deldari, H. Asadi, A. Montazerolghaem, S. M. Mazinani, Flow-aware forwarding in SDN datacenters using a knapsack-PSO-based solution, *IEEE Trans. Netw. Serv. Manage.*,**18** (2021), 2902–2914. https://doi.org/10.1109/TNSM.2021.3064974
15. S. Shen, H. Lu, M. Sadoughi, C. Hu, V. Nemani, A. Thelen, et al., A physics-informed deep learning approach for bearing fault detection, *Eng. Appl. Artif. Intell.*, **103** (2021), 104295. https://doi.org/10.1016/j.engappai.2021.104295
16. A. Khorram, M. Khalooei, M. Rezghi, End-to-end CNN+ LSTM deep learning approach for bearing fault diagnosis, *Appl. Intell.*, **51** (2021), 736–751. https://doi.org/10.1007/s10489-020-01859-1
17. Y. Zou, Y. Zhang, H. Mao, Fault diagnosis on the bearing of traction motor in high-speed trains based on deep learning, *Alexandria Eng. J.*, **60** (2021), 1209–1219. https://doi.org/10.1016/j.aej.2020.10.044
18. R. Zhu, Y. Chen, W. Peng, S. Z. Ye, Bayesian deep-learning for RUL prediction: An active learning perspective, *Reliab. Eng. Syst. Saf.*, **228** (2022), 108758. https://doi.org/10.1016/j.ress.2022.108758
19. Q. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, et al., Convolutional neural network based fault detection for rotating machinery, *J. Sound Vib.*, **377** (2016), 331–345. https://doi.org/10.1016/j.jsv.2016.05.027
20. D. Ruan, J. Wang, J. Yan, C. Gühmann, CNN parameter design based on fault signal analysis and its application in bearing fault diagnosis, *Adv. Eng. Inf.*, **55** (2023), 101877. https://doi.org/10.1016/j.aei.2023.101877
21. L. Jia, W. T. Chow, Y. Yuan, GTFE-Net: A gramian time frequency enhancement CNN for bearing fault diagnosis, *Eng. Appl. Artif. Intell.*, **119** (2023), 105794. https://doi.org/10.1016/j.engappai.2022.105794
22. Y. Hou, J. Wang, Z. Chen, J. Ma, T. Li, Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved Transformer, *Eng. Appl. Artif. Intell.*, **124** (2023), 106507. https://doi.org/10.1016/j.engappai.2023.106507
23. A. Choudhary, T. Mian, S. Fatima, Convolutional neural network based bearing fault diagnosis of rotating machine using thermal images, *Measurement*, **176** (2021), 109196. https://doi.org/10.1016/j.measurement.2021.109196
24. X. Chen, B. Zhang, D. Gao, Bearing fault diagnosis base on multi-scale CNN and LSTM model, *J. Intell. Manuf.*, **32** (2021), 971–987. https://doi.org/10.1007/s10845-020-01600-2
25. T. Han, L. Zhang, Z. Yin, C. A. Tan, Rolling bearing fault diagnosis with combined convolutional neural networks and support vector machine, *Measurement*, **177** (2021), 109022. https://doi.org/10.1016/j.measurement.2021.109022
26. G. Fu, Q. Wei, Y. Yang, C. Li, Bearing fault diagnosis based on CNN-BiLSTM and residual module, *Meas. Sci. Technol.*, **34** (2023), 125050. https://doi.org/10.1088/1361-6501/acf598

27. B. Song, Y. Liu, J. Fang, W. Liu, M. Zhong, X. Liu, An optimized CNN-BiLSTM network for bearing fault diagnosis under multiple working conditions with limited training samples, *Neurocomputing*, **574** (2024), 127284. https://doi.org/10.1016/j.neucom.2024.127284
28. Z. Dong, D. Zhao, L. Cui, An intelligent bearing fault diagnosis framework: One-dimensional improved self-attention-enhanced CNN and empirical wavelet transform, *Nonlinear Dyn.*, **112** (2024), 6439–6459. https://doi.org/10.1007/s11071-024-09389-y
29. C. Li, K. Luo, L. Yang, S. Li, H. Wang, X. Zhang, et al., A zero-shot fault detection method for UAV sensors based on a novel CVAE-GAN model, *IEEE Sens. J.*, **24** (2024), 23239–23254. https://doi.org/10.1109/JSEN.2024.3405630
30. H. Wang, C. Li, P. Ding, S. Li, T. Li, C. Liu, et al., A novel transformer-based few-shot learning method for intelligent fault diagnosis with noisy labels under varying working conditions, *Reliab. Eng. Syst. Saf.*, **251** (2024), 110400. https://doi.org/10.1016/j.ress.2024.110400
31. X. Zhang, C. Li, C. Han, S. Li, Y. Feng, H. Wang, et al., A personalized federated meta-learning method for intelligent and privacy-preserving fault diagnosis, *Adv. Eng. Inf.*, **62** (2024), 102781. https://doi.org/10.1016/j.aei.2024.102781
32. L. S. Dhamande, M. B. Chaudhari, Bearing fault diagnosis based on statistical feature extraction in time and frequency domain and neural network, *Int. J. Veh. Struct. Syst.*, **8** (2016), 229. https://doi.org/10.4273/ijvss.8.4.09
33. M. Altaf, T. Akram, M. A. Khan, M. Iqbal, M. Iqbal Ch, C. H. Hsu, A new statistical features based approach for bearing fault diagnosis using vibration signals, *Sensors*, **22** (2022), 2012. https://doi.org/10.3390/s22052012
34. A. W. Smith, B. R. Randall, Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study, *Mech. Syst. Signal Process.*, **64** (2015), 100–131. https://doi.org/10.1016/j.ymssp.2015.04.021
35. S. Liu, J. Chen, S. He, Z. Shi, Z. Zhou, Subspace network with shared representation learning for intelligent fault diagnosis of machine under speed transient conditions with few samples, *ISA Trans.*, **128** (2022), 531–544. https://doi.org/10.1016/j.isatra.2021.10.025

AIMS Press