Electronic Research Archive

ERA, 32(6): 4074–4095.
DOI: 10.3934/era.2024183
Received: 01 April 2024
Revised: 04 June 2024
Accepted: 18 June 2024
Published: 25 June 2024

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

Research article

Multi-modal adaptive feature extraction for early-stage weak fault diagnosis in bearings

Zhenzhong Xu[1],†, Xu Chen[2],†, Linchao Yang[3], Jiangtao Xu[1],* and Shenghan Zhou[2]

[1] College of Aerospace and Civil Engineering, Harbin Engineering University, Harbin 150001, China
[2] School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China
[3] School of Economics and Management, North China Electric Power University, Beijing 102206, China
† These two authors contributed equally to this work.

* **Correspondence:** Email: hit_xjt@163.com.

Abstract: We present a novel multi-modal adaptive feature extraction algorithm considering both time-domain and frequency-domain modalities (AFETF), coupled with a Bidirectional Long Short-Term Memory (Bi-LSTM) network based on the Grey Wolf Optimizer (GWO) for early-stage weak fault diagnosis in bearings. Singular Value Decomposition (SVD) was employed for noise reduction, while Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) was utilized for signal decomposition, facilitating further signal processing. AFETF algorithm proposed in this paper was employed to extract weak fault features. The adaptive diagnostic process was further enhanced using Bi-LSTM network optimized with GWO, ensuring objectivity in the hyperparameter optimization. The proposed method was validated for datasets containing weak faults with a 0.2 mm crack and strong faults with a 0.4 mm crack, demonstrating its effectiveness in early-stage fault detection.

Keywords: bearings; weak fault; time domain; frequency domain; AFETF

1. Introduction

Bearings are critical components in rotating machinery and are the primary cause of system faults. In the field of machine health monitoring, diagnosing faults in bearings is crucial. It is estimated that 45–55% of equipment faults are due to issues with bearings [1]. For example, bearings are critical components whose reliable operation is essential for wind turbines. In severe cases, bearing faults can result in downtime, damage, injuries, and significant economic losses [2]. The early detection of weak faults in bearings is paramount for ensuring continuous energy production and preventing potential disruptions or hazards. However, diagnosing faults in bearings poses formidable challenges due to the complexity of operating environments and the subtlety of fault symptoms [3]. Traditional fault diagnosis methods based on signal analysis often fall short in accurately capturing weak fault signatures amidst intricate vibration patterns and environmental noise [4]. Therefore, to guarantee the steady functioning of rotating machinery, the development of efficient fault diagnosis models and the use of advanced intelligent algorithms are crucial [5].

Signal-based methods, a common approach for fault diagnosis [6], are extensively employed in bearing fault diagnosis. Sensors collect time-domain signals, representing vibration signals as data waveforms evolving over time. Time-domain analysis assesses rolling bearing vibration signals using various indexes like mean square value, variance, and kurtosis [7] to determine the occurrence of faults. Nevertheless, the operating environment of the bearing is frequently complicated, and external vibration sources affect the precision of time-domain index computations, causing it difficult to pinpoint the exact location or kind of bearing fault and resulting in potential misjudgments. Consequently, researchers frequently employ Fourier Transform (FT) [8,9] to convert the original signal into the frequency domain, providing a better representation of the bearing's periodic and frequency characteristics. Nevertheless, in real-world situations, the vibration signals produced by spindle equipment operations are not smooth, rendering it difficult to resolve the problem with straightforward time or frequency domain analyses. Consequently, time-frequency analysis techniques for non-linear and non-smooth data decomposition have gained prominence in recent fault diagnosis research. Traditional methods for handling non-linear and non-smooth sequences, such as Wavelet Transform (WT) [10,11], Variational Mode Decomposition (VMD) [12,13], Empirical Mode Decomposition (EMD) [14,15], Ensemble Empirical Mode Decomposition (EEMD) [16,17], and Complementary Ensemble Empirical Mode Decomposition (CEEMD) [18], are widely employed for denoising and feature extraction in bearing fault diagnosis. Yang et al. [19] proposed an IMF prediction method which added all the prediction results of IMF decomposed by EMD, the nonlinear response of cracked rotor can be predicted. The IMF prediction is more advantageous than the Lyapunov prediction method through validation. Ming et al. [20] proposed an optimal wavelet filtering method which is developed by maximizing the L-kurtosis through the genetic algorithm to extract the bearing weak repetitive transients and validated the feasibility of this method. Guo et al. [21] proposed an iterative morphological difference product wavelet and achieved higher diagnosis accuracy. However, when it comes to selecting the Intrinsic Mode Functions (IMF) post-decomposition, the majority of approaches rely on human experience or consider only one aspect, either from the time or frequency domain perspective. This process is often insufficiently comprehensive, making it challenging to accurately identify the IMF that encapsulates the most informative details about early-stage weak faults.

In the realm of deep learning, bearing fault diagnosis methods based on models such as Convolutional Neural Networks (CNN) [22], Long Short-Term Memory (LSTM) [23], Bi-LSTM [24], and Deep Belief Networks (DBN) have gained widespread adoption. Li et al. [25] proposed a connection network constructed by resonance-based sparse signal decomposition and broad learning system without the need for deep architecture for rolling element bearing weak fault and verified the effectiveness of this method. Zhou et al. [26] proposed a novel end-to-end deep network-based sparse denoising framework based on a model-data-collaborative linkage framework for bearing weak fault diagnosis. Jin et al. [27] proposed a weak fault diagnosis method for train axle box bearing based on parameter optimization VMD and improved Deep Belief Network (DBN) for train axle box bearing. Zhong et al. [28] proposed exponential excitation networks, which improves diagnosing performance by amplifying important information, that is, the discriminative information between normal and weak fault conditions. For deep learning models, the choice of hyperparameters directly affects the performance of the model; however, the hyperparameter optimization process often relies on the accumulation of professional knowledge and long-term experience. Manual tuning is cumbersome, costly, and inefficient, and the performance of neural network models is limited by human experience. Therefore, automatically finding the optimal parameters and thus determining the optimal model is essential for efficiently improving the model performance.

To enhance the effectiveness of fault diagnosis using deep learning models, numerous studies have explored the integration of optimization algorithms to improve model performance. Commonly employed algorithms include the particle swarm algorithm [29], cuckoo search algorithm [30], among others. Additionally, to prevent the overlooking of optimal solutions, Gao et al. [31] introduced an adaptive convolutional neural network based on Nesterov momentum for rolling bearing fault diagnosis. The traditional momentum method is replaced by Nesterov momentum, and an adaptive rule dynamically adjusting the learning rate is incorporated to enhance the network's generalization ability. Li et al. [32] devised an automated hyperparameter selection algorithm for motor bearing fault diagnosis. It exhibits superior noise immunity and achieves 100% model accuracy, a notable improvement over other deep learning method.

Early detection of weak faults is crucial for rotating machinery operation, preventing downtime, reducing maintenance costs, extending equipment lifespan, and enhancing safety. However, early weak fault signals in rotating machinery bearings exhibit non-linearity and non-stationarity, posing challenges for effective fault feature extraction. Additionally, traditional diagnostic models often require manual parameter tuning, introducing subjectivity and impacting diagnostic performance. We address these challenges by proposing the adaptive feature extraction algorithm, considering both time-domain and frequency-domain modalities (AFETF), coupled with a GWO-optimized Bi-LSTM algorithm for early weak fault diagnosis in rotating machinery bearings. The effectiveness of the proposed algorithm is demonstrated through comparative analysis. The major contributions are:

1) Utilization of SVD for signal denoising and CEEMDAN for adaptive signal decomposition to obtain Intrinsic Mode Functions (IMFs), beneficial for extracting weak fault features under noisy conditions.

2) Application of the AFETF algorithm to process IMFs, integrating time-domain and frequency-domain modalities information to enhance weak fault signal characterization.

3) Adoption of the GWO-optimized Bi-LSTM algorithm for bearing weak fault diagnosis, improving robustness, and generalization capability by automating hyperparameter optimization.

2. Theoretical background

2.1. SVD noise reduction

Singular Value Decomposition (SVD) is a commonly used matrix decomposition method in mathematics [33,34], which has been widely used in signal processing, noise control, and data dimensionality reduction in recent years [35]. In decomposing a matrix using SVD, a matrix of order is assumed, where all the elements belong to the actual or complex domain, and there exists a decomposition equation:

$$Q = U \sum V^T, \tag{1}$$

where U: $a * a$ unitary matrix, V: $b * b$ unitary matrix, satisfies $UU^T = I, VV^T = I, T$: transpose, $\sum$: semi-positive definite diagonal matrix, where the elements are zero except the diagonal elements.

2.2. CEEMDAN signal decomposition

The bearing vibration signal is non-smooth and non-linear whereas the Wavelet Transform (WT) and the Empirical Modal Decomposition (EMD) are commonly used methods, the former will produce different effects when dealing with different basis functions, which relies more on people's subjective experience. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is a further optimized algorithm based on the EMD, which can decompose the signal into several IMFs of a smooth single mode and a cosine term according to the characteristics of the signal itself [36].

Add N times Gaussian white noise $z_i(t)(t = 1,2,\dots,N)$ to the original signal $y(t)$. This is shown in the following equation:

$$y_i(t) = y(t) + z_i(t). \tag{2}$$

Each original signal $y(t)$ with Gaussian white noise added is decomposed using EMD to obtain a number of IMFs. Assuming that there are M IMFs, and the overall averaging is calculated as shown in the following equation:

$$\overline{c_m(t)} = \frac{1}{M}\sum_{i=1}^{M} c_{m,i}(t). \tag{3}$$

According to the Eq (3), the first order IMF can be calculated as shown below:

$$c_1 = \overline{c_1(t)} = \frac{1}{M}\sum_{i=1}^{M} c_{1,i}(t). \tag{4}$$

The residual can be calculated from c_1 as:

$$h_1(t) = y(t) - c_1. \tag{5}$$

Define $y_{Ek}(\cdot)$ as the k th IMF after EMD decomposition, $p_k(t)$ denotes the signal residual after the k th addition of noise, and A denotes the amplitude of the added noise. Gaussian white noise is added on top of the residual from the first stage of decomposition, and the EMD decomposition is continued for the processed signal residual. The IMFs generated from the first stage residuals are averaged together to obtain the second stage IMF c_2:

$$p_k(t) = h_k(t) + Ay_{EK}\big(z_i(t)\big), \tag{6}$$

$$c_2 = \frac{1}{M}\sum_{i=1}^{M} y_{E1}(p_1(t)). \tag{7}$$

Repeat the steps above for each subsequent stage of the residual to finally obtain the *m* th signal residual, as shown in the following equation:

$$h_m(t) = h_{m-1}(t) - c_m. \tag{8}$$

Multiple EMD decompositions are performed on the signal until the signal residuals cannot be further decomposed to complete the CEEMDAN process, and finally, after M decompositions, all CEEMDAN decompositions of the signal are as follows:

$$y(t) = \sum_{m=1}^{M} IMF_m + h_M(t). \tag{9}$$

2.3. *Grey Wolf Optimization algorithm*

The Grey Wolf Optimization algorithm (GWO) is based on the collaboration mechanism of grey wolves and simulates a wolf pack hunting to achieve the goal. Compared with other intelligent optimization algorithms, it has the characteristics of solid convergence, few parameters, easiness to implement, and can achieve the global optimal solution [37]. A certain number of grey wolves are divided into $\alpha,\ \beta,\ \delta$ and ω according to their duties. α is the sole leader and is responsible for the group decision-making of the wolf pack; β is the auxiliary, assisting the leader in managing the wolf pack and also acting as the replacement of the leader; δ needs to follow the orders of α and β, and is responsible for scouting, hunting, and guarding, etc.; and ω is the lowest level of the wolf pack, which is responsible for following the orders of the superiors and maintaining the equilibrium of the pack races.

The steps of GWO include searching, tracking, encircling, hunting, and attacking. The principle of the algorithm is that, in the encirclement, each wolf represents a potential solution, and $\alpha,\ \beta,\ \delta, \omega$ gradually decreases according to their levels of priority. According to the positions of the first three wolves, the other individuals calculate the distances and constantly approach them to update the position to complete the search of optimality.

In the Q-dimensional space, a single wolf is denoted as $W_i = (w_{i,1}, w_{i,2}, \dots, w_{i,q})$, and the constituent wolves are denoted as $W_i = (W_1, W_2, \dots, W_N)$. The calculation is shown in the following equations:

$$D_m = C_k \cdot W_m - W_{ik}(t), m = \alpha, \beta, \delta; i = 1,2,3 \dots, n, \tag{10}$$

$$C_k = 2r_1, k = 1,2,3, \tag{11}$$

$$W'_{ik} = W_m - A_k \cdot D_m, \tag{12}$$

$$W_{i(t+1)} = \frac{1}{3} \sum W'_{ik}, \tag{13}$$

$$A_k = 2ar_2 - a, k = 1,2,3, \tag{14}$$

where $a = 2 - t/t_{max}$, a: convergence factor, which decreases linearly with the number of iterations and takes values in the range [0,2], D_m: distance of grey wolf W_i from wolves $\alpha,\ \beta,\ \delta$ at iteration t, C_k: coefficient vector, A_k: coefficient vector, r_1: random values of [0,1], r_2: random values of [0,1], W'_{ik}: vector of the grey wolf's movement as it approaches $\alpha,\ \beta,\ \delta, W_{i(t+1)}$: move value of the grey wolf W_i for the next iteration.

2.4. *Bi-LSTM algorithm*

Bidirectional Long Short-Term Memory (Bi-LSTM) is composed of forward LSTM and backward LSTM. Bi-LSTM is connected to one output for forward and backward propagation, respectively, but forward and backward propagations are not associated. The hidden layers of both propagations are independent of each other, and their confidential information layers can be propagated in both directions. The composition of Bi-LSTM makes itself more suitable for diagnosing early-stage weak faults of bearings by fully using the information of timing features [38].

The LSTM consists of a forget gate f_t, an input gate i_t, a cell state c_t and an output gate o_t, and σ is an activation function-sigmoid. The composition of LSTM is shown in Figure 1.

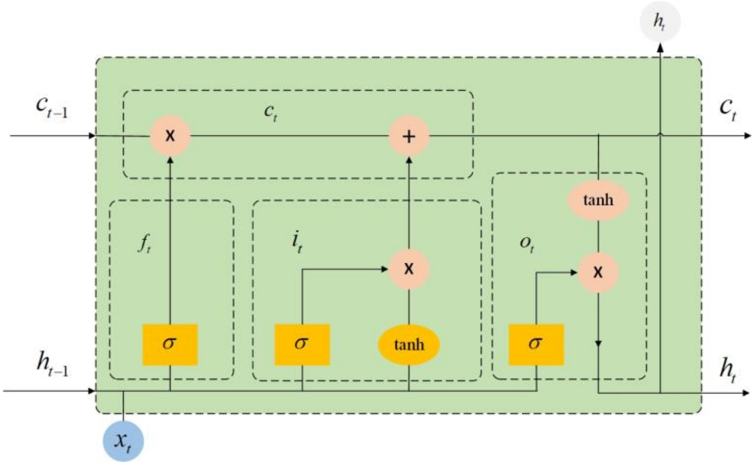


Figure 1. The composition of LSTM.

Forward LSTM operates memory cells using input gates, output gates, and forget gates. In this computational process, the forget gate decides to remember or forget the hidden state of the previous layer according to the probability. The error term propagation of the backward LSTM is concentrated in the direction along the time axis and in the direction of the previous layer of the network.

During the actual training process of LSTM, each output is compared with the corresponding labels to update the weight parameters of the network. Each label is coded using one-hot coding, and its coded result needs to match the output of softmax, which takes the value range of [0,1]. Using this function allows the LSTM to judge the probability of the labels. The value of the softmax function is calculated as shown in the following equation:

$$S_i = \frac{e^{z^i}}{\sum_{j=1}^{J} e^{z^j}}, \tag{15}$$

where Z^i: the i th element of the array Z, J: number of categories of the array Z, e: natural number.

In the softmax function, the cross entropy is used as the loss function in this paper, and its expression is shown in the following equation:

$$L = \sum_{k=1}^{n} y_i log\ (S_i), \tag{16}$$

where y_i: category of the i th data, S_i: probability of a category for the i th data, n: total number of categories.

3. Proposed methods

3.1. Adaptive feature extraction considering both time-domain and frequency-domain modalities

CEEMDAN decomposes the vibration signal according to the frequency. However, there are pseudo-components in the decomposition process, so it is very critical to filter out the IMF that contains the wealthiest information about early-stage weak faults, which directly affects the establishment of the fault diagnosis model and ultimately affects the accuracy of weak fault diagnosis. The time-domain characteristics of the vibration signal are analyzed according to the time history of the signal to obtain its statistical characteristics. Its processing will not cause losses of signal or distortions, and it is not affected by the influence of the measurement points, which renders itself rather suitable for weak fault diagnosis.

We propose an adaptive extraction algorithm of weak fault features considering time-frequency modalities (AFETF) for envelope spectrum analysis. We select the kurtosis, mean square error, and sample entropy as the parameter indexes in the time domain dimension. Kurtosis is more sensitive to shock pulses and pulse-like fault signals, which makes itself applicable for surfacing damage faults and diagnosing early weak faults [39]. The mean square error can reflect the degree of data dispersion and is an essential indicator of accuracy. The early-stage weak faults will be enhanced when the faults appear and change the mean square deviation [40]. The sample entropy can reflect the complexity of the time series, and at the same time, it is less dependent on the data and it is high in consistency [41]. These three indexes are more sensitive to early weak faults.

Kurtosis is affected by the number and the size of signal x waveform shocks, and the formula for calculating the kurtosis value is shown below:

$$K = \frac{E(x-\mu)^2}{\sigma^4}, \tag{17}$$

where E: expectation, μ: signal mean, σ: signal mean variance.

The formula for calculating the mean square error is shown in the following equation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \tag{18}$$

where N: number of sampling points.

The sample entropy equation is shown below:

$$SampEn(m, r, N) = -l\,n\left[\frac{A^m(r)}{B^m(r)}\right], \tag{19}$$

where $B^m(r)$: the probability of matching to m points under the distance restriction r, $A^m(r)$: the probability of matching to $m+1$ points.

The specific equation of the metric proposed in this paper used to quantify information of the early-stage weak faults contained in IMFs is shown as follow:

$$X_i = \frac{K_i}{K_{max}} + \frac{\sigma_i}{\sigma_{max}} + \frac{S_i}{S_{max}}, \tag{20}$$

where K_i: the kurtosis of i th IMF, K_{max}: the max kurtosis of IMFs, σ_i: the mean square error of i th IMF, σ_{max}: the max mean square error of IMFs, S_i: the sample entropy of i th IMF, S_{max}: the max sample entropy of IMFs.

Although time domain modality analysis is suitable for early weak fault diagnosis, its precision must be improved. Frequency domain modality analysis can make up well for this defect. In this paper, the theoretical characteristic frequency of each fault is calculated according to the bearing type, size parameter, and rotational speed. After CEEMDAN has been used to decompose the original signal into several IMFs, assuming that L IMFs are obtained, their envelope signals are obtained using the Hilbert variation. Then, the spectral transformation is performed to obtain the envelope spectrum of the IMF. Given the manufacture and assembly of bearings and the influence of the operating environment, there is often a particular gap between the actual fault characteristic frequency of bearings and the theoretical fault characteristic frequency, and the fluctuation range is usually within 5%. The number of fault characteristic frequencies and that of their multiplicative frequencies are obtained in the envelope spectrum. We define the fault characteristic frequency screening function, as shown in the following equation:

$$F_i = Number_m(W_i(f)), f \in [f' - \gamma f', f' + \gamma f'], i = 1,2, \dots, L, \tag{21}$$

where f': the theoretical characteristic frequency of fault, f: the range of characteristic frequency for screening, $W_i(f)$: the envelope spectrum of the i th IMF component, γ: the screening factor within 5%, m: the number of selected spectral amplitudes, $Number$: the number of spectra in the first m frequencies with the highest spectral amplitude that match the screening range.

After obtaining F_i for each IMF, its percentage is calculated as shown in the following equation:

$$Y_i = \frac{F_i}{\sum_{i=1}^{L} F_i}. \tag{22}$$

The index of measuring fault information is obtained in the dimension of time domain, and the number of corresponding fault characteristic frequencies and their multiples are obtained in the dimension of frequency domain. Based on the time-frequency characteristics above, it defines Z as a parameter index which can achieve the most abundant information of early-stage weak faults, and its calculation formula is shown as the following equation:

$$Z_i = max(X_i * Y_i), i = 1,2, \dots, L. \tag{23}$$

3.2. Adaptive classification algorithm: GWO-BiLSTM

LSTM is improved by introducing a flexible and self-control self-loop design on the RNN [42]. LSTM can only propagate in one direction, and in the process of weak fault diagnosis. LSTM misses the part of the weak fault feature information because the weak fault itself is not apparent. Bi-LSTM can realize the bi-directional propagation to extract information of weak fault feature to the maximum extent to improve the accuracy of the model. In weak fault diagnosis, by using Bi-LSTM, the number of neurons in the hidden layer of the model, the initial learning rate, and the regularization coefficient need to be specified subjectively, which cannot guarantee the optimal solution, thus affecting the accuracy of the model [43]. Therefore, we propose a classification algorithm for GWO-BiLSTM, which uses GWO to optimize and adjust the hyperparameters of Bi-LSTM to obtain the optimal solution of the model parameters, thus improving the accuracy of the model.

To improve the accuracy of the Bi-LSTM model, this paper uses the GWO algorithm to optimize the hyperparameters of the Bi-LSTM, including the number of neurons in the hidden layer, the learning rate, and the regularization coefficient. The obtained results of model classification must be infinitely close to the actual results to get the minimum objective function. The fitness function is defined as shown in the following equation:

$$fitness = -\sum_{S=1}^{S}(y' - y)^2, \tag{24}$$

where y': the real class, y: the actual class, S: the length of training.

The flow of the optimized classification algorithm based on GWO-BiLSTM is shown in Figure 2.

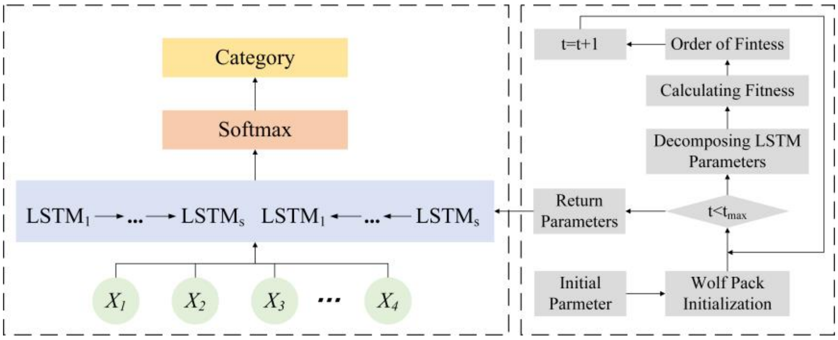


Figure 2. GWO-BiLSTM overall architecture.

First, the number of hidden layer neurons, the learning rate, and the regularization coefficient of the Bi-LSTM are initialized. The random distribution function is used to initialize the grey wolf position, decompose the parameters of the LSTM network, including the weights, etc., and substitute them into the network to iteratively calculate the fitness value until the end of the iteration. The best W_α is returned as the optimal solution. Inputting it into the Bi-LSTM network and inputting the dataset for training can generate the final classification result.

3.3. Adaptive weak fault diagnosis model

The process of early-stage weak fault diagnosis for bearings proposed in this paper is shown in Figure 3.

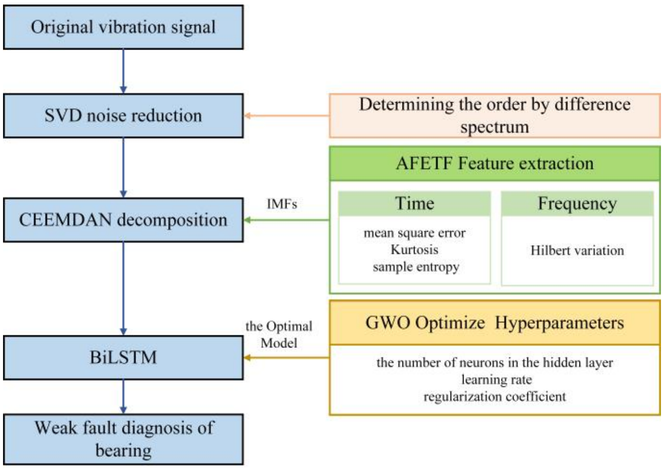


Figure 3. Flowchart of the proposed method.

We begin by employing SVD to reduce noise in the original bearing signal. The order of the SVD is then determined using the difference spectrum. Next, CEEMDAN decomposes the signal into multiple IMFs. We utilize AFETF algorithm proposed in this paper to screen the IMF, which contains the richest weak fault information. Finally, GWO is employed to optimize the hyperparameters of Bi-LSTM and train the model for final diagnosis.

4. Experiments and results analysis

To verify the effectiveness of the proposed method, a simulation test of bearing fault is performed in the test rig shown in Figure 4. The test rig consists of an electric motor, an inertia wheel for applying radial load, a coupling, a conveyor belt drive mechanism, a conveyor belt, a crank-connecting rod mechanism, a gearbox with the bearing, and a reciprocating mechanism with a spring. The bearing selected for this test is a deep groove ball bearing (MB ER-10K), and the rotational frequency of the shaft is defined as f_r. Two acceleration meters as shown in Figure 4, are PCB Model 608A11. The sensitivity of these acceleration meters is 10.2 (mV/ms^{-2}) / 100 (mv/g). The acquisition card model is NI 9234. According to the formula for calculating the fault frequency of the bearing, the characteristic fault frequencies of the outer race, inner race and ball are $f_o = 3.052f_r$, $f_i = 4.948f_r$, and $f_b = 3.984f_r$, respectively.

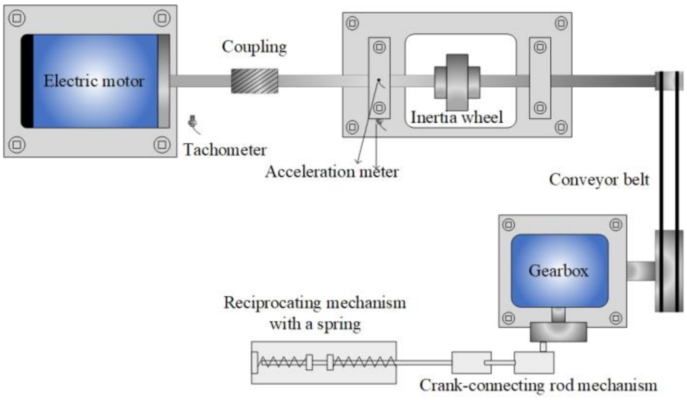


Figure 4. Schematic diagram of bearing test rig.

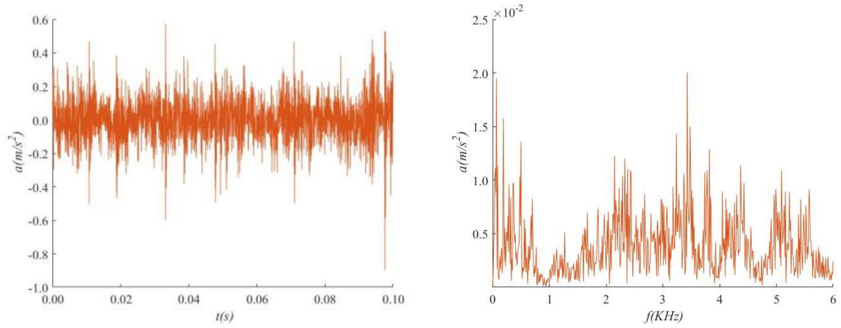
Vibration tests were conducted on bearings featuring damage to the outer race, inner race, and steel ball components. Localized cracks with a width and a height of 0.2 and 0.4 mm were etched on the outer race grooves, inner race grooves, and steel ball surfaces, respectively. The sampling frequency f_s of this test is 25.6 kHz and the motor speed is 900 r/min, corresponding to a rotational frequency f_r of 15 Hz.

Table 1 displays the characteristic fault frequency of each component under real measurements.

Table 1. Actual characteristic fault frequency of individual components (Hz).

Component	Outer race	Inner race	Ball
Measured rotation frequency	14.33	14.76	14.12
Characteristic fault frequency	43.75	73.03	56.25

Taking the outer race weak fault with a 0.2 mm crack vibration signal as an example, its time domain and frequency domain are shown in Figure 5.



(a) Time domain within 0.1 seconds. (b) Frequency domain within 6000 Hz.

Figure 5. Original signal of the outer race fault with a 0.2 mm crack.

Figure 5(a) reveals prominent shock characteristics in the time-domain plot, while Figure 5(b) illustrates a frequency domain plot with a poorly smoothed signal waveform and numerous mutation points, indicating significant background noise. Extracting fault features in this scenario without noise reduction is challenging.

We initially employ SVD for noise reduction of fault signal data to eliminate background noises. However, selecting the appropriate order is crucial, as different singular value orders yield vastly different effects on noise reduction. The distribution of singular values is depicted in Figure 6, where the singular value gradually decreases with increasing order.

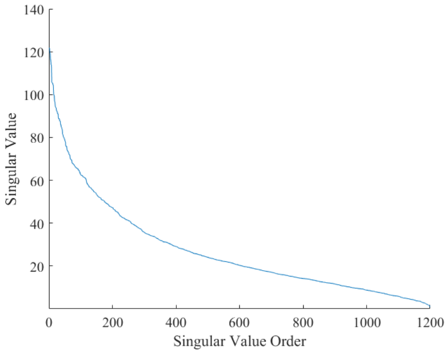


Figure 6. Distribution of singular values.

Figure 7 displays the singular value difference spectrum. At a singular value order of 8, the maximum difference value reaches 8.157, indicating a clearer distinction between the signal and noise. Hence, a singular value order of 8 is selected for noise reduction in this study.

Figure 8 illustrates the time domain comparison before and after noise reduction, indicating a reduced signal amplitude and a more defined boundary contour. Similarly, Figure 9 displays smoother frequency domains post-noise reduction, with fewer mutation points. Irrelevant noise bands are suppressed, preserving valuable signal bands and enhancing signal amplitude. SVD noise reduction better represents the characteristics of the original signal.

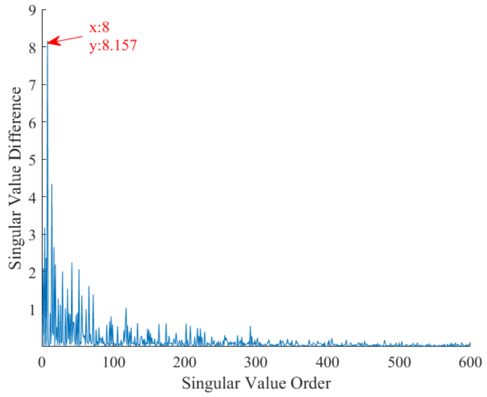
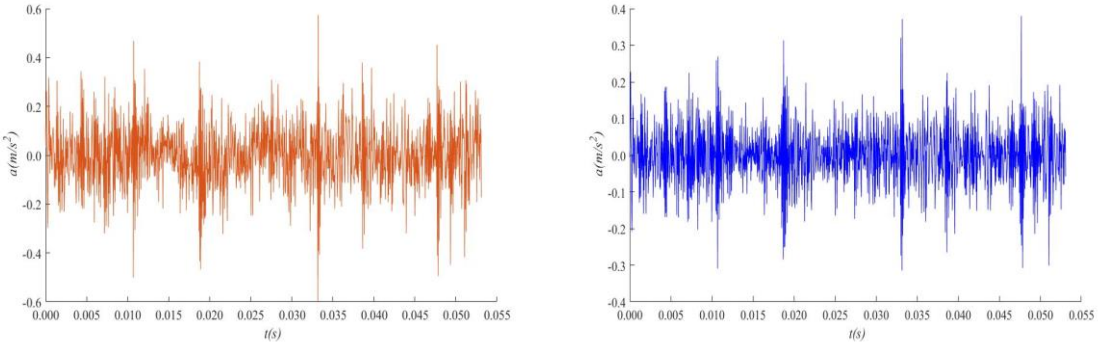
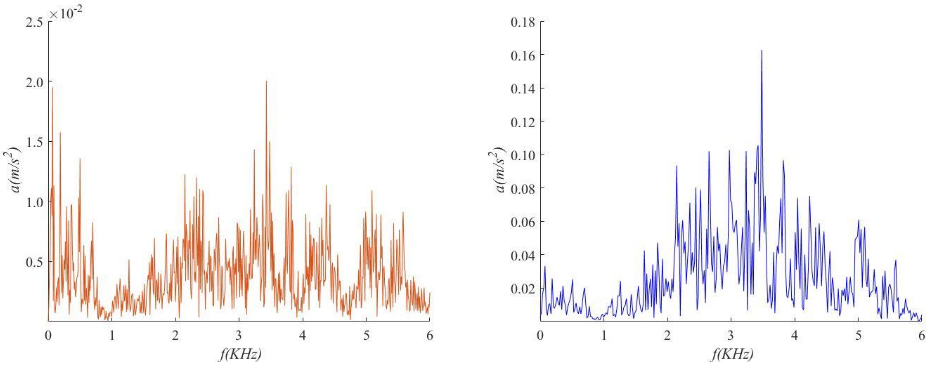


Figure 7. Singular value difference spectrum.



(a) Original signal. (b) Denoised signal.

Figure 8. Time domain comparison.



(a) Original signal. (b) Denoised signal.

Figure 9. Frequency domain comparison.

Following noise reduction, the signal undergoes CEEMDAN decomposition, yielding 12 IMFs. These decomposition results are illustrated in Figure 10.

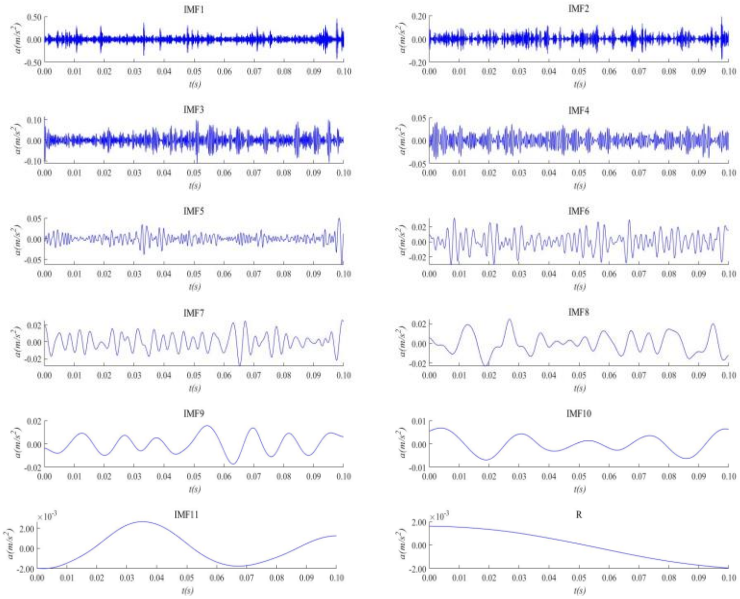


Figure 10. The IMFs obtained from CEEMDAN decomposition. Note R represents the residual.

CEEMDAN decomposes the signal into 12 IMFs based on frequency, with the final IMF representing the residual. Throughout the decomposition process, a reduction in the richness of information within each IMF becomes evident.

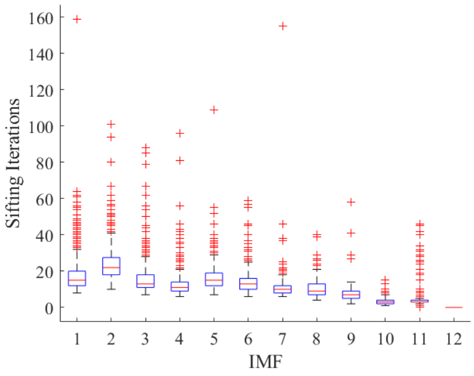


Figure 11. The sifting iterations boxplot of CEEMDAN mode.

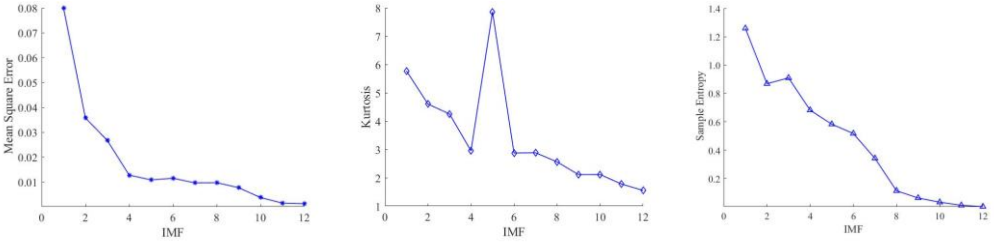
Figure 11 illustrates the sifting iterations for each IMF during CEEMDAN decomposition. The maximum number of sifting iterations for IMF1 is 159, indicating the rapid signal decomposition capability of CEEMDAN [44].

We propose a AFETF algorithm, which can help filter the IMF containing the richest fault information for diagnosis of early-stage weak faults.

Figure 12 presents the calculation results of mean square error, kurtosis, and sample entropy for each IMF in the time domain.

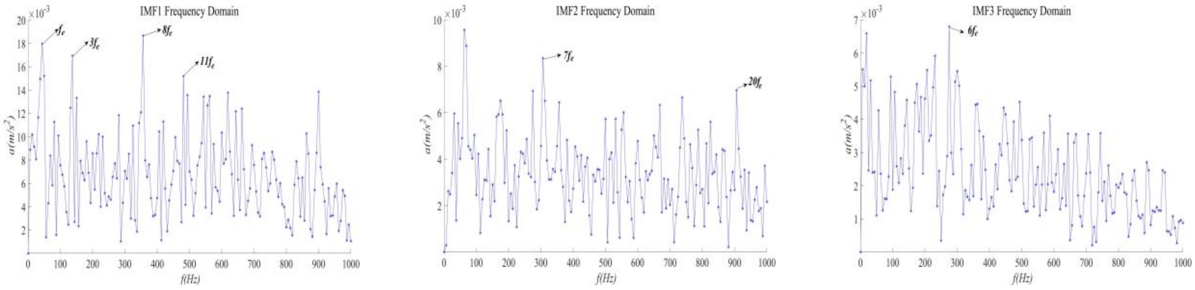
Figure 12 indicates that the first five IMFs contain the most significant amount of information. Setting L as 5, γ as 0.05, and m as 5, and according to Eq (21), the envelope spectrum of the first five

IMFs are plotted so as to find the number of times the fault frequency that will be multiplied.

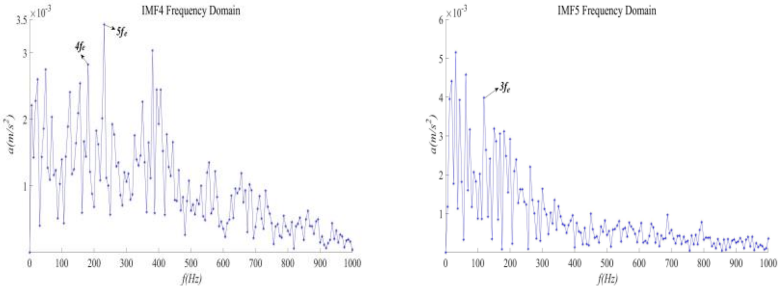


(a) Mean square error. (b) Kurtosis. (c) Sample entropy.

Figure 12. IMF time domain index.



(a) IMF1. (b) IMF2. (c) IMF3.



(d) IMF4. (e) IMF5.

Figure 13. The characteristic frequency of the outer race fault and its frequency multiplier in the IMF frequency domain.

Figure 13 shows that among the eigenfrequencies of the first five highest amplitudes and their multiples, there are four for IMF1, two for IMF2, one for IMF3, two for IMF4, and one for IMF5.

The fault information contained in each IMF is calculated according to Eq (23) and the results are shown in Figure 14.

According to Figure 14, it can be learned that according to the adaptive feature extraction method proposed in this paper, IMF1 contains the most abundant fault information, so IMF1 is chosen as the feature of weak fault extracted from the outer race.

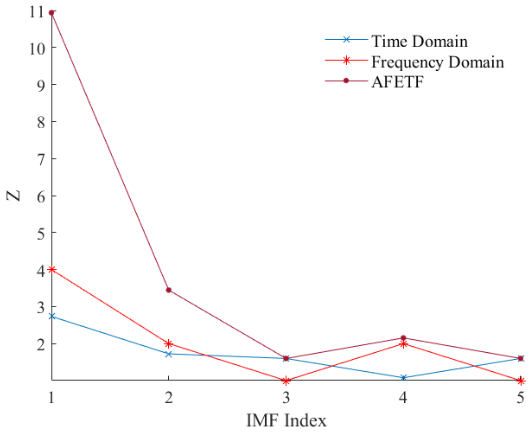


Figure 14. Fault information index.

Initially, the weak fault with a 0.2 mm crack signal is sliced into segments of 0.1 seconds each. Using AFETF proposed in this paper, the optimal IMF is selected and labeled to create a dataset of 1140 × 2560. Subsequently, 30% of the data from each category is extracted as the test set. Next, the strong fault with a 0.4 mm crack signal is processed similarly to form the same 1140 × 2560 dataset.

As an example, for weak fault with a 0.2 mm crack, the IMFs selected for the three types of fault signals are shown in Figure 15.

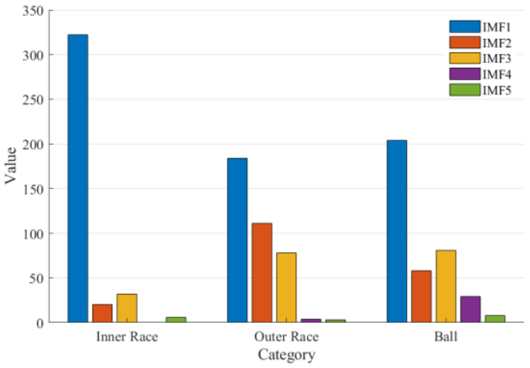
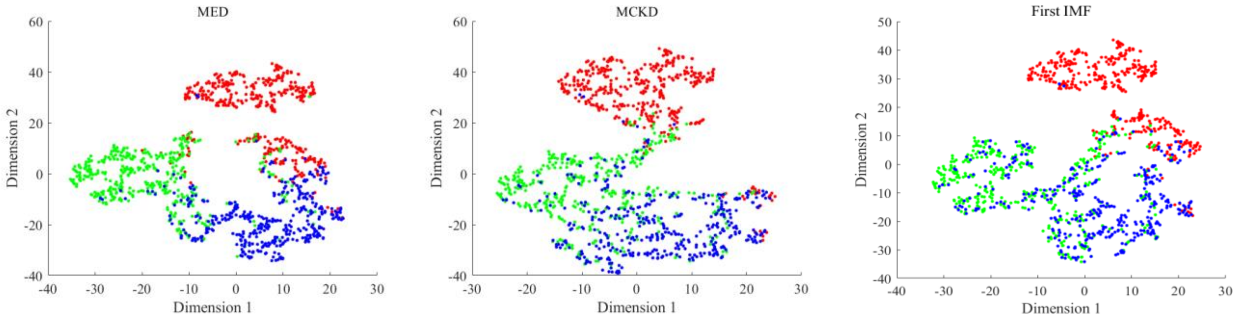


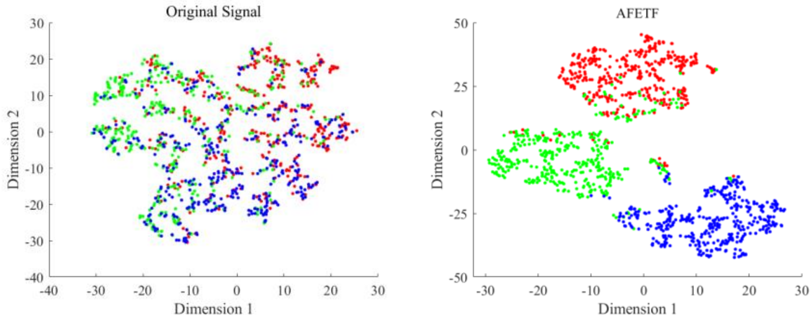
Figure 15. IMF selection results for three types of faults.

From Figure 15, it can be seen that according to the AFETF algorithm of this paper, IMF1, IMF2, and IMF3 are the major features to be selected for faults from the outer race, IMF1 for faults from the inner race, and IMF1, IMF2, and IMF3 for the ball faults.

To assess the effectiveness of the proposed method, the weak fault with a 0.2 mm crack is taken as an example. First, noise reduction is performed using SVD, followed by signal decomposition with CEEMDAN. Subsequently, feature extraction is conducted using Minimum Entropy Deconvolution (MED), Maximum Correlated Kurtosis Deconvolution (MCKD), and AFETF proposed in this paper. Additionally, the original signal and the first IMF selected after signal decomposition are included as features for comprehensive comparison. These features are visualized in 2D using t-sne, as depicted in Figure 16.



(a) MED. (b) MCKD. (c) First IMF.



(d) Original signal. (e) AFETF.

Figure 16. Different fault features visualization by t-sne under various feature extraction methods. Note red indicates ball fault features, green indicates inner race fault features, and blue indicates outer race fault features.

According to Figure 16, it can be found that the three features extracted by the AFETF algorithm proposed in this paper have more distinct boundaries.

Taking weak fault with a 0.2 mm crack diagnosis as an example, after feature extraction using the different feature extraction algorithms mentioned above, the GWO-BiLSTM algorithm is used for classification, and the curve of its train accuracy is shown in Figure 17.

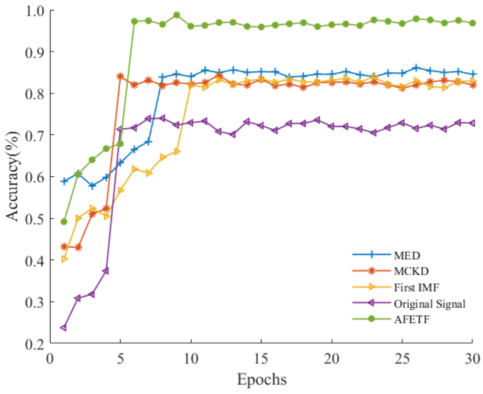
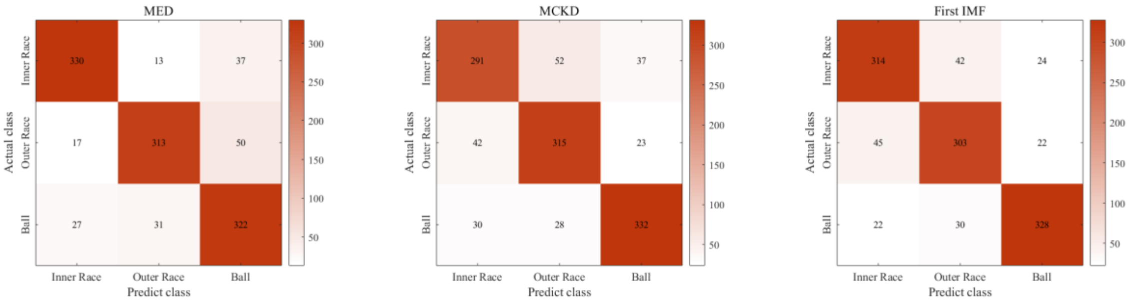
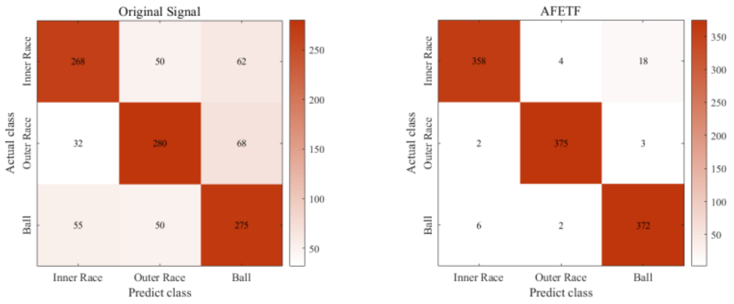


Figure 17. Train accuracy curve.

The GWO-BiLSTM classification results of confusion matrixes are shown in Figure 18.

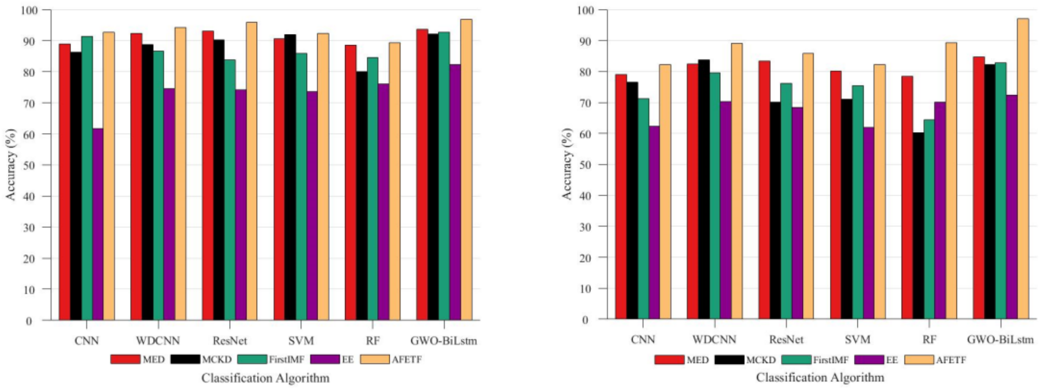


(a) MED. (b) MCKD. (c) First IMF.



(d) Original Signal. (e) AFETF.

Figure 18. GWO-BiLSTM fault diagnosis confusion matrix based on different feature extraction methods.



(a) Strong fault with a 0.4 mm crack. (b) Weak fault with a 0.2 mm crack.

Figure 19. Comparison of diagnostic accuracy for strong and weak faults under different feature extraction methods and diagnostic models.

The features are extracted using MED, MCKD, the first IMF, the original signal, and AFETF algorithm proposed in this paper, and then classified using CNN, Wide and Deep Convolutional Neural Network (WDCNN), Residual Neural Network (ResNet), Support Vector Machine (SVM), Random Forest (RF), and GWO-BiLSTM algorithms, respectively, adopted in this paper. The classification results for the weak fault with a 0.2 mm crack and for the strong fault with a 0.4 mm crack are shown in Figure 19.

The accuracy of strong fault with a 0.4 mm crack diagnosis is shown in Table 2.

Table 2. Diagnosis accuracy in the case of strong fault with a 0.4 mm crack.

Methods	MED	MCKD	First IMF	Original signal	AFETF
CNN	88.88	86.35	91.33	61.76	92.65
WDCNN	92.36	88.74	86.63	74.52	94.26
ResNet	93.12	90.32	83.75	74.26	95.98
SVM	90.67	91.93	85.96	73.68	92.27
RF	88.49	80.06	84.53	76.12	89.34
GWO-BiLSTM	93.64	92.25	92.81	82.32	96.82

The diagnosis accuracy in the case of weak fault with a 0.2 mm crack is shown in Table 3.

Table 3. Diagnosis accuracy in the case of weak fault with a 0.2 mm crack.

Methods	MED	MCKD	First IMF	Original signal	AFETF
CNN	78.98	76.63	71.23	62.26	82.26
WDCNN	82.43	83.74	79.63	70.25	89.03
ResNet	83.36	70.12	76.25	68.36	85.94
SVM	80.17	71.13	75.36	61.98	82.20
RF	78.46	60.26	64.47	70.12	89.27
GWO-BiLSTM	84.65	82.28	82.89	72.20	96.93

According to Figure 19, Tables 2 and 3, regarding feature extraction, the diagnosis algorithms are unified using GWO-BiLSTM. AFETF algorithm proposed in this paper improves the accuracy respectively by 3.18, 4.57, 4.01, and 14.5% in the diagnosis of strong fault with a 0.4 mm crack. In the diagnosis of weak fault with a 0.2 mm crack, the accuracy is respectively improved by 12.28, 14.65, 14.04, and 24.73%. Regarding diagnosis algorithms, approaches of feature extraction are unified using AFETF algorithm proposed in this paper. In the diagnosis of strong fault with a 0.4 mm crack, compared with other algorithms, the accuracy of GWO-BiLSTM proposed in this paper is respectively improved by 4.17, 2.56, 0.84, 4.55, and 7.48%. In the diagnosis of weak fault with a 0.2 mm crack, the accuracy is respectively improved by 14.67, 7.9, 10.99, 14.73, and 7.66%.

5. Discussion and conclusions

We present an adaptive feature extraction and diagnosis method for detecting early-stage weak faults in bearings. Following SVD noise reduction, CEEMDAN decomposes the signal into modal components. Subsequently, the proposed algorithm, AFETF extracts fault features rich in information. GWO-BiLSTM, adept at identifying timing features, conduct adaptive fault diagnosis. Experimental results demonstrate superior diagnostic accuracy for both 0.4 and 0.2 mm faults compared to traditional methods. While effective for both strong and weak faults, the method exhibits particularly significant accuracy improvement for weak faults. This study underscores the suitability of the proposed diagnosis method for detecting early-stage weak faults in bearings.

We focus on the early-stage weak fault diagnosis of bearings under constant speed conditions. Future work will extend this research to weak fault diagnosis under variable speed conditions.

Use of AI tools declaration

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

Acknowledgments

This research was funded by the National Natural Science Foundation of China under Grant 11372080.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. F. Jia, Y. G. Lei, J. Lin, X. Zhou, N. Lu, Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data, *Mech. Syst. Signal Process.*, **72** (2016), 303–315. https://doi.org/10.1016/j.ymssp.2015.10.025
2. Y. J. Zhou, X. Y. Long, M. W. Sun, Z. Q. Chen, Bearing fault diagnosis based on Gramian angular field and DenseNet, *Math. Biosci. Eng.*, **19** (2022), 14086–14101. https://doi.org/10.3934/mbe.2022656
3. Z. W. Shang, C. L. Pan, Y. Yu, F. Liu, M. S. Gao, Weak local fault diagnosis of gearboxes based on adaptive inertia factor particle swarm independent component analysis, *Insight Nondestr. Test. Cond. Monit.*, **65** (2023), 415–422. https://doi.org/10.1784/insi.2023.65.8.415
4. W. Cui, G. Y. Meng, A. M. Wang, X. E. Zhang, J. Ding, Application of rotating machinery fault diagnosis based on deep learning, *Shock Vib.*, **2021** (2021), 3083190. https://doi.org/10.1155/2021/3083190
5. Y. H. Zhang, T. T. Zhou, X. F. Huang, L. C. Cao, Q. Zhou, Fault diagnosis of rotating machinery based on recurrent neural networks, *Measurement*, **171** (2021), 108774. https://doi.org/10.1016/j.measurement.2020.108774
6. Q. S. Wang, Z. C. Sun, Y. M. Zhu, C. H. Song, D. Li, Intelligent fault diagnosis algorithm of rolling bearing based on optimization algorithm fusion convolutional neural network, *Math. Biosci. Eng.*, **20** (2023), 19963–19982. https://doi.org/10.3934/mbe.2023884
7. Z. X. Wei, Y. X. Wang, S. L. He, J. D. Bao, A novel intelligent method for bearing fault diagnosis based on affinity propagation clustering and adaptive feature selection, *Knowledge-Based Syst.*, **116** (2017), 1–12. https://doi.org/10.1016/j.knosys.2016.10.022

8. S. Q. Huang, J. D. Zheng, H. Y. Pan, J. Y. Tong, Order-statistic filtering fourier decomposition and its application to rolling bearing fault diagnosis, *J. Vib. Control*, **28** (2022), 1605–1620. https://doi.org/10.1177/1077546321997598
9. D. C. Zhu, G. Q. Liu, X. Y. Wu, B. L. Yin, An enhanced empirical Fourier decomposition method for bearing fault diagnosis, *Struct. Health Monit.*, **23** (2024), 903–923. https://doi.org/10.1177/14759217231178653
10. W. L. Fu, X. H. Jiang, B. L. Li, C. Tan, B. J. Chen, X. Y. Chen, Rolling bearing fault diagnosis based on 2D time-frequency images and data augmentation technique, *Meas. Sci. Technol.*, **34** (2023), 045005. https://doi.org/10.1088/1361-6501/acabdb
11. Z. J. Xie, D. Yu, C. S. Zhan, Q. C. Zhao, J. X. Wang, J. Q. Liu, et al., Ball screw fault diagnosis based on continuous wavelet transform and two-dimensional convolution neural network, *Meas. Control*, **56** (2023), 518–528. https://doi.org/10.1177/00202940221107620
12. V. Sharma, A. Parey, Extraction of weak fault transients using variational mode decomposition for fault diagnosis of gearbox under varying speed, *Eng. Fail. Anal.*, **107** (2020), 104204. https://doi.org/10.1016/j.engfailanal.2019.104204
13. Q. B. Lu, X. Q. Shen, X. J. Wang, M. Li, J. Li, M. Z. Zhang, Fault diagnosis of rolling bearing based on improved VMD and KNN, *Math. Probl. Eng.*, **2021** (2021), 2530315. https://doi.org/10.1155/2021/2530315
14. T. Wu, Fault diagnosis method of rolling bearing based on EMD-Hilbert envelope spectrum and BPNN, in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, **632** (2021), 052084. https://doi.org/10.1088/1755-1315/632/5/052084
15. P. K. Sahu, R. N. Rai, Fault diagnosis of rolling bearing based on an improved denoising technique using complete ensemble empirical mode decomposition and adaptive thresholding method, *J. Vib. Eng. Technol.*, **11** (2023), 513–535. https://doi.org/10.1007/s42417-022-00591-z
16. J. B. Hou, Y. X. Wu, H. Gong, A. S. Ahmad, L. Liu, A novel intelligent method for bearing fault diagnosis based on EEMD permutation entropy and gg clustering, *Appl. Sci.*, **10** (2020), 386. https://doi.org/10.3390/app10010386
17. A. Kumar, Y. Berrouche, R. Zimroz, G. Vashishtha, S. Chauhan, C. P. Gandhi, et al., Non-parametric Ensemble Empirical Mode Decomposition for extracting weak features to identify bearing defects, *Measurement*, **211** (2023), 112615. https://doi.org/10.1016/j.measurement.2023.112615
18. F. Z. Liu, J. W. Gao, H. B. Liu, The feature extraction and diagnosis of rolling bearing based on CEEMD and LDWPSO-PNN, *IEEE Access*, **8** (2020), 19810–19819. https://doi.org/10.1109/ACCESS.2020.2968843
19. Y. F. Yang, H. Chen, T. D. Jiang, Nonlinear response prediction of cracked rotor based on EMD, *J. Franklin Inst.*, **352** (2015), 3378–3393. https://doi.org/10.1016/j.jfranklin.2014.12.015
20. A. B. Ming, W. Zhang, C. Fu, Y. F. Yang, F. L. Chu, Y. J. Liu, L-kurtosis-based optimal wavelet filtering and its application to fault diagnosis of rolling element bearings, *J. Vib. Control*, **30** (2024), 1594–1603. https://doi.org/10.1177/10775463231165816
21. J. C. Guo, Q. B. He, D. Zhen, F. S. Gu, A. D. Ball, An iterative morphological difference product wavelet for weak fault feature extraction in rolling bearing fault diagnosis, *Struct. Health Monit.*, **22** (2023), 296–318. https://doi.org/10.1177/14759217221086314

22. S. Q. Zhou, L. P. Lin, C. Chen, W. B. Pan, X. C. Lou, Application of convolutional neural network in motor bearing fault diagnosis, *Comput. Intell. Neurosci.*, **2022** (2022), 923130. https://doi.org/10.1155/2022/9231305
23. S. J. Hao, F. X. Ge, Y. M. Li, J. Y. Jiang, Multisensor bearing fault diagnosis based on one-dimensional convolutional long short-term memory networks, *Measurement*, **159** (2020), 107802. https://doi.org/10.1016/j.measurement.2020.107802
24. Z. F. Xu, X. Mei, X. Y. Wang, M. N. Yue, J. T. Jin, Y. Yang, et al., Fault diagnosis of wind turbine bearing using a multi-scale convolutional neural network with bidirectional long short term memory and weighted majority voting for multi-sensors, *Renewable Energy*, **182** (2022), 615–626. https://doi.org/10.1016/j.renene.2021.10.024
25. X. C. Li, J. C. Wang, B. Zhang, Fault diagnosis of rolling element bearing weak fault based on sparse decomposition and broad learning network, *Trans. Inst. Meas. Control*, **42** (2020), 169–179. https://doi.org/10.1177/0142331219864820
26. X. Zhou, H. X. Zhou, G. R. Wen, X. Huang, Z. H. Lei, Z. F. Zhang, et al., A hybrid denoising model using deep learning and sparse representation with application in bearing weak fault diagnosis, *Measurement*, **189** (2022), 110633. https://doi.org/10.1016/j.measurement.2021.110633
27. Z. Z. Jin, D. Q. He, Z. X. Wei, Intelligent fault diagnosis of train axle box bearing based on parameter optimization VMD and improved DBN, *Eng. Appl. Artif. Intell.*, **110** (2022), 104713. https://doi.org/10.1016/j.engappai.2022.104713
28. B. H. Zhong, M. H. Zhao, S. S. Zhong, L. Lin, Y. J. Zhang, Deep exponential excitation networks: toward stronger attention mechanism for weak fault diagnosis, *Struct. Health Monit.*, 2024. https://doi.org/10.1177/14759217231217936
29. X. Liu, R. Q. Wu, R. G. Wang, F. Zhou, Z. F. Chen, N. H. Guo, Bearing fault diagnosis based on particle swarm optimization fusion convolutional neural network, *Front. Neurorob.*, **16** (2022), 1044965. https://doi.org/10.3389/fnbot.2022.1044965
30. M. H. Xiao, Y. B. Liao, P. Bartos, M. Filip, G. S. Geng, Z. W. Jiang, Fault diagnosis of rolling bearing based on back propagation neural network optimized by cuckoo search algorithm, *Multimedia Tools Appl.*, **81** (2022), 1567–1587. https://doi.org/10.1007/s11042-021-11556-x
31. S. Z. Gao, Z. M. Pei, Y. M. Zhang, T. C. Li, Bearing fault diagnosis based on adaptive convolutional neural network with nesterov momentum, *IEEE Sens. J.*, **21** (2021), 9268–9276. https://doi.org/10.1109/JSEN.2021.3050461
32. Z. Li, Y. Wang, J. N. Ma, Fault diagnosis of motor bearings based on a convolutional long short-term memory network of bayesian optimization, *IEEE Access*, **9** (2021), 97546–97556. https://doi.org/10.1109/ACCESS.2021.3093363
33. D. Kalman, A singularly valuable decomposition: The SVD of a matrix, *Coll. Math. J.*, **27** (1996), 2–23. https://doi.org/10.2307/2687269
34. J. F. Huang, L. L. Cui, Tensor singular spectrum decomposition: Multisensor denoising algorithm and application, *IEEE Trans. Instrum. Meas.*, **72** (2023), 1–15. https://doi.org/10.1109/TIM.2023.3249249
35. H. Li, T. Liu, X. Wu, Q. Chen, A bearing fault diagnosis method based on enhanced singular value decomposition, *IEEE Trans. Ind. Inf.*, **17** (2021), 3220–3230. https://doi.org/10.1109/TII.2020.3001376

36. D. Huang, S. Li, N. Qin, Y. Zhang, Fault diagnosis of high-speed train bogie based on the improved-CEEMDAN and 1-D CNN algorithms, *IEEE Trans. Instrum. Meas.*, **70** (2021), 3508811. https://doi.org/10.1109/TIM.2021.3062104
37. S. Mirjalili, S. M. Mirjalili, A. Lewis, Grey wolf optimizer, *Adv. Eng. Software*, **69** (2014), 46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007
38. T. Han, R. Y. Ma, J. G. Zheng, Combination bidirectional long short-term memory and capsule network for rotating machinery fault diagnosis, *Measurement*, **176** (2021), 109208. https://doi.org/10.1016/j.measurement.2021.109208
39. Y. H. Miao, M. Zhao, J. Lin, Y. G. Lei, Application of an improved maximum correlated kurtosis deconvolution method for fault diagnosis of rolling element bearings, *Mech. Syst. Signal Process.*, **92** (2017), 173–195. https://doi.org/10.1016/j.ymssp.2017.01.033
40. P. K. Kankar, S. C. Sharma, S. P. Harsha, Fault diagnosis of ball bearings using machine learning methods, *Expert Syst. Appl.*, **38** (2011), 1876–1886. https://doi.org/10.1016/j.eswa.2010.07.119
41. Z. Y. Wang, L. G. Yao, Y. W. Cai, Rolling bearing fault diagnosis using generalized refined composite multiscale sample entropy and optimized support vector machine, *Measurement*, **156** (2020), 107574. https://doi.org/10.1016/j.measurement.2020.107574
42. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
43. F. Q. Zou, H. F. Zhang, S. T. Sang, X. M. Li, W. Y. He, X. W. Liu, Bearing fault diagnosis based on combined multi-scale weighted entropy morphological filtering and bi-LSTM, *Appl. Intell.*, **51** (2021), 6647–6664. https://doi.org/10.1007/s10489-021-02229-1
44. M. E. Torres, M. A. Colominas, G. Schlotthauer, P. Flandrin, A complete ensemble empirical mode decomposition with adaptive noise, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague Congress Ctr, Prague, (2011), 4144–4147. https://doi.org/10.1109/ICASSP.2011.5947265