Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bayesian-Optimized VMD and Attention-Based CNN- BiLSTM for Early Gearbox Fault Diagnosis

Hongfei Yao

yaohf520@126.com

Qiqihar University

Yufei Kang

Qiqihar University

Jian Ma

Qiqihar University

Dian Liu

Qiqihar University

Hao LI

Qiqihar University

Article

Keywords: Gearbox fault diagnosis, Bayesian optimization, variational mode decomposition, convolutional neural network, bidirectional long short-term memory, attention mechanism, deep learning

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10632675/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Bayesian-Optimized VMD and Attention-Based CNN-BiLSTM for Early Gearbox Fault Diagnosis

Hongfei YAO[1]*, Yufei KANG [1]*, Jian MA[1], Dian LIU[1], Hao LI[1]

[1] School of Mechanical Engineering, Qiqihar University, Qiqihar 161006, China
* Corresponding author: HONGFEI YAO (e-mail: yaohf520@126.com); YUFEI KANG (e-mail: 740458851@qq.com);

Abstract

Gearboxes, critical components in rotating machinery, require timely incipient fault diagnosis to prevent catastrophic failures and ensure operational reliability. However, traditional methods struggle to detect weak fault signatures amid strong background noise and variable working conditions. This study proposes an integrated intelligent diagnosis framework combining Bayesian-optimized Variational Mode Decomposition (BO-VMD) with an attention-based CNN-BiLSTM model. BO-VMD adaptively optimizes VMD decomposition parameters through kurtosis difference maximization, enabling effective separation of fault-related features from noisy backgrounds. The CNN-BiLSTM model hierarchically extracts spatial-temporal representations, while the attention mechanism enhances the weight distribution of key fault features. Validated on the Tsinghua University gearbox dataset under nine working conditions, the framework achieves 90.0% average accuracy for incipient fault classification, outperforming standalone CNN and BiLSTM models, particularly under time-varying operational conditions. This study provides a robust and scalable solution for condition-based maintenance and intelligent monitoring of critical industrial mechanical systems.

Keywords: Gearbox fault diagnosis; Bayesian optimization; variational mode decomposition; convolutional neural network; bidirectional long short-term memory; attention mechanism; deep learning

1. Introduction

As the core of mechanical transmission systems, gearboxes undertake the key func-tions of power transmission and speed regulation, and are widely used in important fields of the national economy such as new energy power generation and rail transit [1-3]. When operating under high load, variable speed and strong impact conditions for a long time, gearboxes are prone to faults such as tooth surface wear and tooth root cracks. Among them, early diagnosis of minor faults smaller than 0.3 mm is the key to preventing equip-ment failure [4-5]. Statistics show that over 60% of faults in rotating machinery are related to gearboxes, and 45% of serious accidents result from minor faults not being detected in time, with direct economic losses exceeding 10 billion yuan annually [6]. Therefore, con-ducting research on the diagnosis technology for minor faults in gearboxes to achieve ear-ly warning and precise identification of faults has significant theoretical and engineering significance.

Gearbox fault diagnosis technology has evolved from traditional signal processing methods to intelligent diagnosis methods. Traditional methods take Fourier transform, wavelet transform, etc. as the core and achieve fault identification by extracting time-domain and frequency-domain features [7-8]. Although wavelet energy–based feature extraction combined with support vector machine has been widely used in bearing fault detection, it still faces difficulties in extracting subtle fault features from high-dimensional signals and exhibits poor anti-noise performance in actual engineering scenarios [9]. Since the rise of deep learning technology, intelligent diagnosis methods based on neural networks have been widely used due to their strong feature learning capabilities. The CNN diagnosis method proposed by Yang et al. can directly extract features from raw vibration signals and performs well under stable working conditions. However, this method fails to make full use of temporal information and lacks adaptability across different working conditions [10]. Wang et al. adopted LSTM networks to capture signal timing dependen-cies and enhance the processing capability of non-stationary signals, but they ignore local spatial features, and the accuracy of identifying minor faults needs to be improved [11]. To balance spatial and temporal feature extraction, the CNN-BILSTM hybrid model has be-come a research hotspot [12-13]. It integrates the advantages of local feature extraction of CNN and the bidirectional temporal modeling ability of BiLSTM, which can more com-prehensively mine fault features. The CNN-BiLSTM diagnostic method proposed by Chen et al [14]. has an accuracy rate of 85% under fixed working conditions, but it does

not en-hance the features of minor faults, and its performance drops significantly in scenarios with low signal-to-noise ratio.

In summary, the diagnosis of minor faults in gearboxes faces three major challenges: First, the characteristic signals of minor faults are weak and can be easily masked by oper-ating noise, environmental interference, and normal vibration signals, resulting in an ex-tremely low signal-to-noise ratio [15]; Second, in industrial scenarios, gearboxes are often under conditions of variable speed and variable load, and the fault features are time-varying and non-stationary, increasing the difficulty of feature extraction [16]; Third-ly, the characteristics of different minor faults vary little, and the features change dynami-cally during fault evolution, making it difficult for traditional methods to achieve precise identification [17].

This paper focuses on the key technical issues of minor fault diagnosis of gearboxes and proposes the diagnostic method of BO-VMD-CNN-BiLSTM. The main research con-tents are as follows:

1) Design a Bayesian optimization VMD algorithm, aiming to maximize the kurtosis difference of modal components. Adaptively optimize the penalty factor α and the number of modes K of VMD to achieve enhanced extraction of minor fault features.

2) Construct a multi-dimensional feature matrix to extract time-domain and frequen-cy-domain statistical features from the key modal components after VMD decomposition, providing rich feature inputs for deep learning models;

3) Propose the CNN-BiLSTM model based on the attention mechanism, strengthen the weight distribution of key fault features, and improve the model's recognition ability for minor faults;

4) Conduct multi-condition micro-fault diagnosis experiments on the gearbox dataset of Tsinghua University to verify the effectiveness and robustness of the proposed method.

2. Bayesian optimization VMD algorithm(BO-VMD)

2.1. The principle of variational Mode Decomposition (VMD)

VMD is an adaptive modal decomposition method based on variational optimization. Its core idea is to decompose the original signal into K modal components with a center frequency by constructing a variational problem, and each modal component satisfies the characteristic of limited bandwidth [18].

2.1.1. Construction of VMD variational problems

Suppose the original vibration signal is$x(t)$, decompose it into K modal components $\{u_k(t)\}_{k=1}^{K}$,The analytical signal of each modal component can be obtained through Hilbert transform, and its expression is:

$$z_k(t) = u_k(t) + j\hat{u}_k(t) = A_k(t)e^{j(j\omega_k t+\phi_k(t))} \tag{1}$$

Among them, the Hilbert transform where$\hat{u}_k(t)$ is$u_k(t)$,$A_k(t)$,$\omega_k(t)$,$\phi_k(t)$ are the instantaneous amplitudes, instantaneous frequencies, and instantaneous phases of the modal components, respectively [19].

The variational problem of VMD can be expressed as:

$$\begin{cases} \min\{u_k\},\{\omega_k\} \sum_{k=1}^{K} \left\| \partial_t \left[(u_k(t) + j\hat{u}_k(t))e^{-j\omega_k t} \right] \right\|_2^2 \\ s.t. \sum_{k=1}^{K} u_k(t) = x(t) \end{cases} \tag{2}$$

Here,∂trepresents the partial derivative with respect to time$t$, $\| \bullet \|_2^2$ represents the squared L_2norm, and$\omega_k$is the central frequency of the modal component [20].

2.1.2. Solution to variational problems

To solve the above variational problem, the Lagrange multiplier$\lambda(t)$and the penalty factor$\alpha$are introduced to transform the constrained variational problem into an unconstrained variational problem:

$$\begin{aligned} &L(\{u_k\},\{\omega_k\},\lambda) \\ &= \alpha \sum_{k=1}^{K} \left\| \partial_t \left[(u_k(t) + j\hat{u}_k(t))e^{-j\omega_k t} \right] \right\|_2^2 + \left\| x(t) - \sum_{k=1}^{K} u_k(t) \right\|_2^2 + \left\langle \lambda(t), x(t) - \sum_{k=1}^{K} u_k(t) \right\rangle \end{aligned} \tag{3}$$

The alternating direction multiplier method Alternating Direction Method of Multipliers (ADMM) is adopted to iteratively solve this variational problem, and the update formula of the modal components and the center frequency is obtained:

Modal component update:

$$\hat{u}_k^{n+1}(\omega)=\frac{\hat{x}(\omega)-\sum_{i\neq k}\hat{u}_i^n(\omega)+\frac{\hat{\lambda}^n(\omega)}{2}}{1+2\alpha(\omega-\omega_k^n)^2} \tag{4}$$

Central frequency update

$$\omega_k^{n++1}=\frac{\int_0^\infty \omega\left|\hat{u}_k^{n+1}(\omega)\right|^2 d\omega}{\int_0^\infty \left|\hat{u}_k^{n+1}(\omega)\right|^2 d\omega} \tag{5}$$

Lagrange multiplier update:

$$\hat{\lambda}^{n+1}(\omega)=\hat{\lambda}^n(\omega)+\tau(\hat{x}(\omega)-\sum_{k=1}^K\hat{u}_k^{n+1}(\omega)) \tag{6}$$

where $\hat{x}(\omega)$,$\hat{u}_k(\omega)$, and $\hat{\lambda}(\omega)$ represent the Fourier transforms of $x(t)$, $u_k(t)$, and $\lambda(t)$, respectively; τ is the iteration step size, and n is the number of iterations.

When the iteration meets the convergence condition

$$\frac{\sum_{k=1}^K\left\|\hat{u}_k^{n+1}(\omega)-\hat{u}_k^n(\omega)\right\|_2^2}{\left\|\hat{u}_k^n(\omega)\right\|_2^2}<tol \tag{7}$$

When$tol$is the convergence threshold, in this paper, when $1e-7$ taken, the iteration stops and reaches the final K modal components [21].

2.2. Bayesian optimization principle

Bayesian optimization is a global optimization algorithm based on probabilistic models. Its core is to regress and fit the posterior distribution of the objective function through the Gaussian Process (GP), and then use the acquisition function to balance the ability to explore unknown regions and utilize the existing optimal information. Efficiently search for the global optimal hyperparameter combination with very few evaluations [22], and its complete iterative process is as follows:

1) Initialize sample points: Select a small number of initial hyperparameter combinations as sample points, calculate the objective function values corresponding to each sample, and complete the algorithm initialization;

2) Gaussian process regression modeling: Taking the initial sample points as input and the objective function values as output, the posterior distribution of the sample points is calculated through Gaussian process regression, and a probability proxy model of the objective function is constructed.

3) Selection of iteration points for acquisition functions: Based on the posterior distribution of the surrogate model, acquisition functions such as expectation boosting (EI), confidence upper bound (UCB), and probability boosting (PI) are used to calculate the acquisition function values of all candidate hyperparameters. The point with the largest acquisition function value is selected as the next iteration point to be evaluated.

4) Iterative update and termination judgment: Evaluate the objective function values of the new iteration points, add them to the sample set, and update the Gaussian process surrogate model; If the current result meets the preset stop conditions (such as the upper limit of the number of iterations, the convergence threshold of the objective function), the global optimal hyperparameter is output. If not met, repeat steps 2 to 4 until the stop condition is satisfied. The iterative process of Bayesian optimization is shown in Figure1.

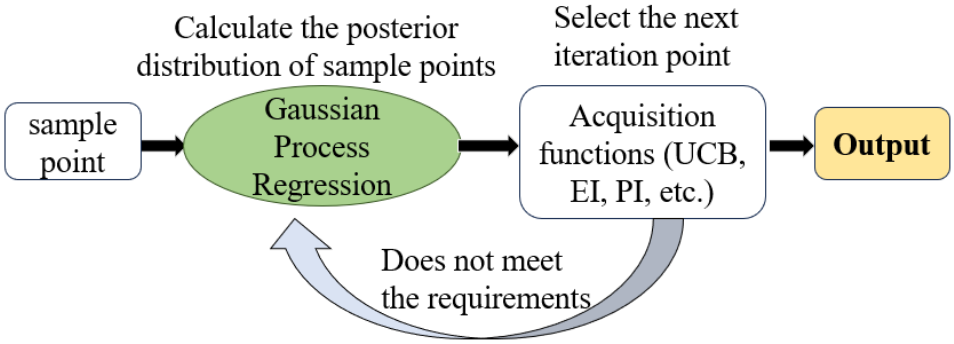


Figure 1. shows the flow chart of the Bayesian optimization algorithm

2.2.1. Objective function design

To achieve the effective separation of minor fault features, the objective function is to maximize the kurtosis difference of each modal component after VMD decomposition. Kurtosis, as an indicator for measuring the peak characteristics of signals, can effectively reflect the impact features of minor faults. It is defined as:

$$K(u_k) = \frac{1}{N}\sum_{i=1}^{N}(u_k(i) - \mu_{uk})^4 / \sigma_{u_k}^4 \tag{8}$$

Among them, μ_{uk}, σ_{u_k} are the mean and standard deviation of modal component $u_k(t)$ respectively, and N is the signal length.

The objective function is designed as:

$$f(\alpha, K) = \max_{1\le k\le K} K(u_k) - \min_{1\le k\le K} K(u_k) \tag{9}$$

When K exceeds the range of [4, 15] or α exceeds the range of [3000,10000], a penalty is applied to the objective function:

$$f(\alpha, K) = \begin{cases} \max_{1\le k\le K} K(u_k) - \min_{1\le k\le K} K(u_k) & 4 \le \mathrm{K} \le 15, 3000 \le \alpha \le 10000 \\ -1e6 & otherwise \end{cases} \tag{10}$$

2.2.2. Selection of acquisition function

The Expected Improvement (EI) is adopted as the acquisition function, which is defined as:

$$EI(\theta) = E[max(f(\theta) - f^*, 0)] \tag{11}$$

Among them, $\theta = (\alpha, K)$ is the combination of hyperparameters, and f^* is the currently observed optimal objective function value.

Based on the Gaussian distribution assumption, the analytical solution of EI is:

$$EI(\theta) = \begin{cases} (\mu(\theta) - f^*)\Phi(\frac{\mu(\theta)-f^*}{\sigma(\theta)}) + \sigma(\theta)\phi(\frac{\mu(\theta)-f^*}{\sigma(\theta)}) & \sigma(\theta) > 0 \\ 0 & \sigma(\theta) = 0 \end{cases} \tag{12}$$

Here, $\mu(\theta)$ and $\sigma(\theta)$ are the mean estimation and standard deviation estimation of the objective function at point θ in the Gaussian process model, respectively; $\Phi(\square)$ and $\phi(\square)$ are the cumulative distribution function and probability density function of the standard normal distribution, respectively.

2.3. The BO-VMD algorithm process

The core idea of BO-VMD (Bayesian Optimized Variational Mode Decomposition) is to adaptively search for the key hyperparameters of variational mode decomposition through the Bayesian optimization algorithm, that is, the optimal combination of mode component number K and penalty factor α, to replace the traditional manual trial-and-error method. Then, the obtained optimal parameters are used to perform VMD decomposition on the signal to improve the decomposition accuracy and fault feature extraction ability.

The specific process is shown in Figure 2

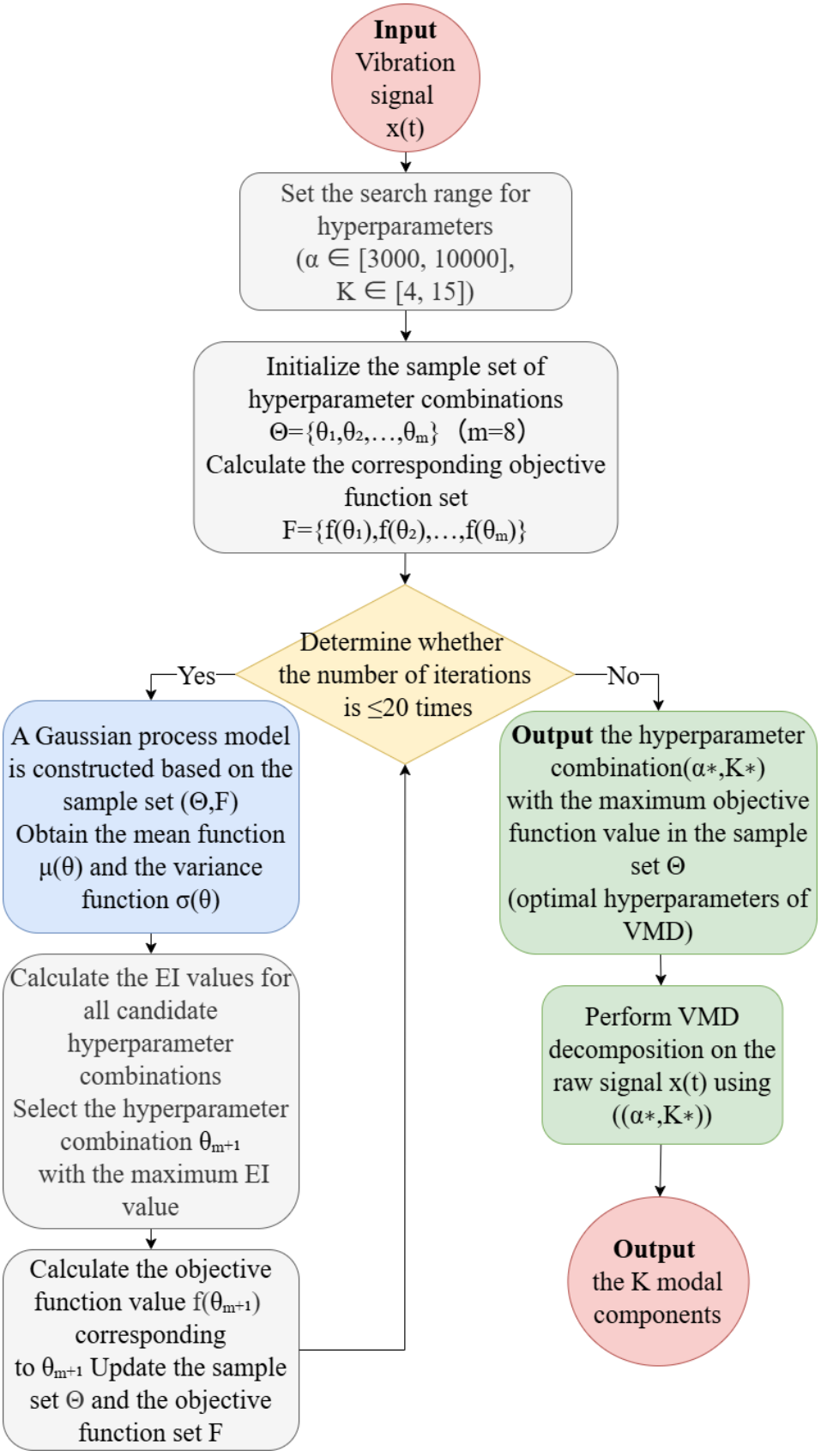


Figure 2 Flowchart of the BO-VMD algorithm

3. CNN-BiLSTM hybrid model

3.1. Overall architecture of the model

The CNN-BiLSTM hybrid model consists of a feature extraction layer, an attention layer and a classification layer. The overall architecture is shown in Figure 2. The feature extraction layer consists of the CNN module and the BiLSTM module. The CNN module extracts the local spatial features of the feature matrix, while the BiLSTM module captures the bidirectional temporal dependency of the feature sequence. The attention layer allocates weights to the feature vectors output by BiLSTM to enhance the key fault features. The classification layer predicts fault categories through a fully connected network.[23]

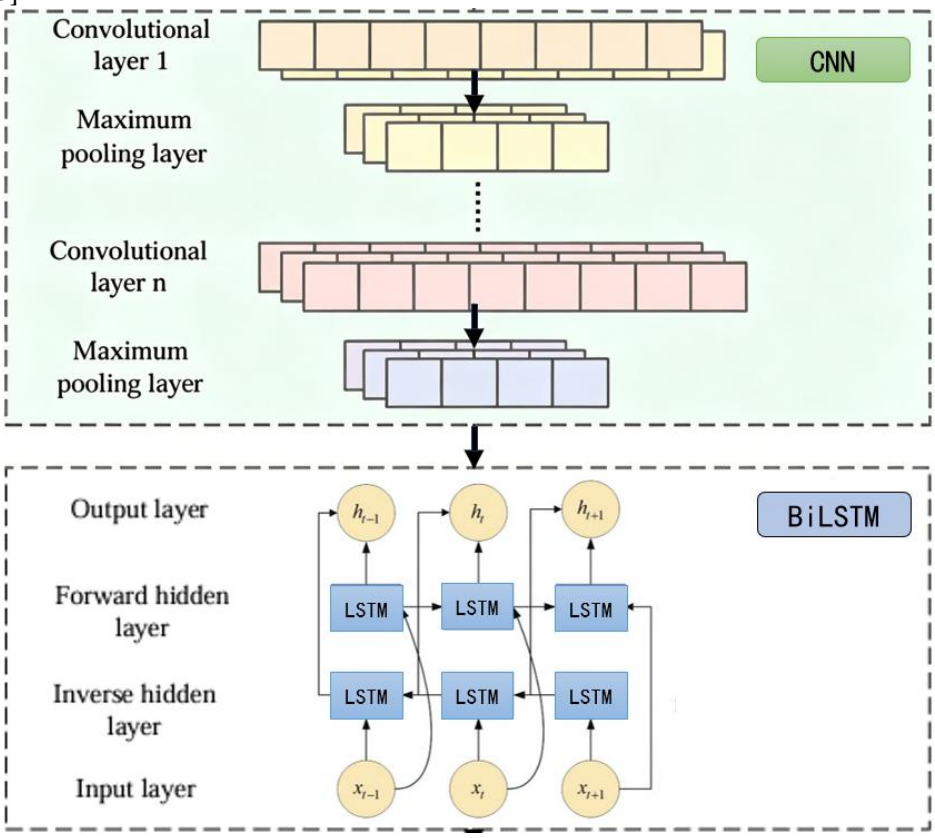


Figure 3 Overall architecture of the CNN-BiLSTM model

3.2. CNN module

The CNN module consists of three convolutional layers, a batch normalization layer, a ReLU activation function and a Max pooling layer [24]. The specific structure is as follows:

First convolutional layer: Input channels 1, output channels 64, convolutional kernel size 3×3, fill 1, step size 1; Batch normalization layer: Accelerates the convergence of model training and enhances the generalization ability of the model; ReLU activation function: Introducing nonlinearity to enhance the model's feature expression ability, with an expression of $\sigma(x) = max(0, x)$. Maximum pooling layer: Pooling core size 2×2, step size 2, reduce feature dimension, and retain key features;

The second and third convolutional layers: Their structure is similar to that of the first convolutional layer, with the number of output channels being 128 and 256 respectively.

The output feature map size of the CNN module is B×Cout×Hout×Wout, where B represents the batch size, Cout is the number of output channels, and Hout and Wout are the height and width of the output feature map, respectively.

3.3. BiLSTM module

The BiLSTM network consists of two LSTM links, the forward one and the reverse one: Forward LSTM models historical timing dependencies in chronological order, while reverse LSTM models future timing dependencies in reverse chronological order. The two interact and fuse hidden states at each time step, and ultimately output a feature representation that fuses bidirectional timing information. This can comprehensively capture the long-term timing correlation of vibration signals and provide richer timing features for the diagnosis of minor faults in gearboxes[25].

The feature map output by the CNN module is subjected to dimension transformation to obtain a feature sequence of shape B×T×D (where T= Hout× Wout is the sequence length and D=Cout is the feature dimension), which is then input into the BiLSTM module. The internal update of the LSTM unit is shown in Figure 4 as follows:

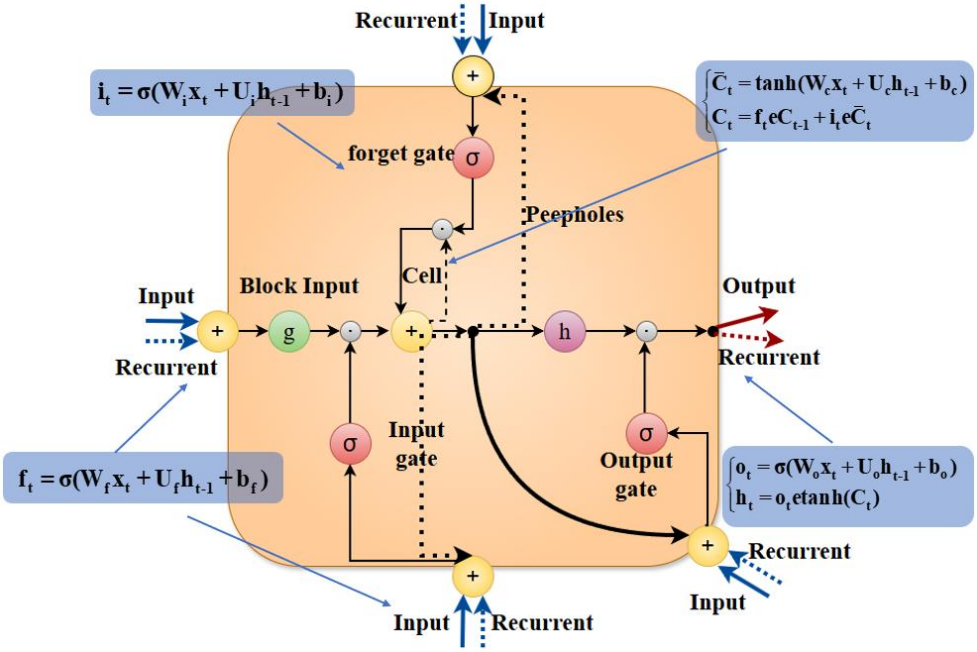


Figure 4 Internal structure diagram of the LSTM unit

The forward output h_t^f and the reverse output h_t^b of the BiLSTM module are concatenated to obtain a bidirectional hidden state $h_t = \left[h_t^f ; h_t^b\right]$.

3.4. Attention Layer

To enhance the weight distribution of key fault features, an attention mechanism is introduced to perform weighted fusion on the hidden states output by BiLSTM. The calcu-lation process of the attention layer is as follows:

Calculate the attention score:

$$e_t = \tanh(W_1 h_t + b_1) \text{、} score_t = W_2 e_t + b_2 \quad (13),(14)$$

Normalized attention score:

$$attn_weight_t = \frac{\exp(score_t)}{\sum_{t=1}^{T} \exp(score_t)} \quad (15)$$

Calculate the context vector:

$$context = \sum_{t=1}^{T} attn_w right_t \cdot h_t \quad (16)$$

Among them W1 and W2 are the weight matrices, and b1 、b2 are the bias vectors.

3.5. Classification layer

The classification layer consists of two fully connected networks. The first layer maps the context vectors to a 512-dimensional feature space. After passing through the ReLU ac-tivation function and the Dropout layer (dropout rate=0.2), the second layer of the fully connected network is input, and the probability distribution of each fault category is out-put:

$$y = Softmax(W_3 context + b_3) \quad (17)$$

Among them, W3 and b3are the weight matrix and bias vector of the classification layer, and the Softmax function ensures that the sum of the output probabilities is 1.

4. Gearbox Minor Fault Diagnosis Algorithm Based on BO-VMD-CNN-BiLSTM

The gearbox micro-fault diagnosis algorithm model based on BO-VMD-CNN-BiLSTM proposed in this paper is as follows:

Firstly, taking the original vibration fault signal collected by the gearbox as the input, the Bayesian optimization algorithm is adopted. With the maximization of the kurtosis difference of the modal component as the objective function, the optimal hyperparameter combination of the penalty factor α and the number of modes K of variational mode de-composition (VMD) is adaptively sought to solve the subjectivity and inefficiency prob-lems of traditional empirical parameter adjustment. Subsequently, the optimized hy-perparameters are utilized to perform VMD decomposition on the original signal, decom-posing the non-stationary and noisy vibration signal into multiple stationary intrinsic mode components. The key mode components with the highest kurtosis are selected to enhance the extraction of minor fault features, effectively separating the fault features from the background noise and improving the signal-to-noise ratio of the signal.

Secondly, the enhanced modal components are input into the CNN layer, and the multi-scale local spatial features of the signal are extracted through multi-layer convolu-tion operations to capture the implicit feature patterns such as impact and modulation of minor faults. After introducing the nonlinear enhancement model's expressive ability through the ReLU activation function, feature dimensionality reduction is accomplished through the Max pooling layer, which retains the key fault information while reducing computational redundancy and achieving efficient extraction of spatial features.

Then, the high-dimensional feature map output by CNN is converted into a time se-ries sequence and input into the BiLSTM layer. The historical and future time series de-pendencies of the signal are respectively modeled through the forward LSTM and the re-verse LSTM. The bidirectional fusion output is used to fully capture the long-term time se-ries correlation characteristics of the fault signal, achieving a deep fusion of spatial fea-tures and time series features.

Finally, the fused features output by BiLSTM are input into the fully connected layer to complete the mapping and refinement of high-dimensional features. Then, through the Softmax activation function, the features are mapped to the probability distribution of var-ious faults, and the classification results of gearbox fault types are output. Ultimately, pre-cise and stable diagnosis of minor faults in gearboxes is achieved.

As shown in Figure 5.

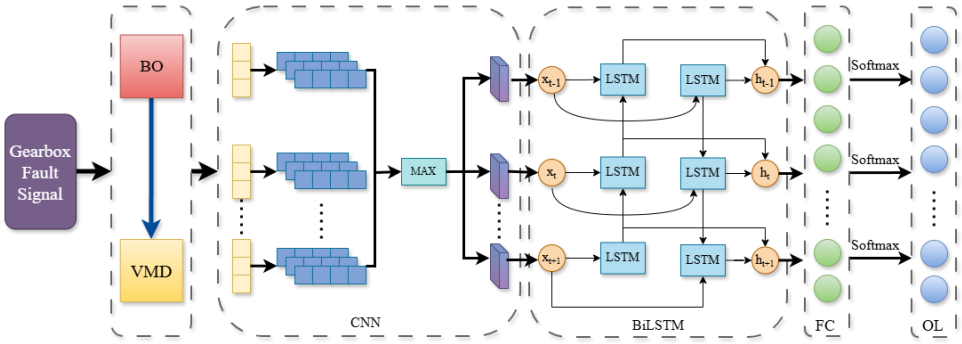


Figure 5 Gearbox minor fault diagnosis model of BO-VMD-CNN-BiLSTM

5. Experimental verification and analysis

5.1. Gearbox Dataset of Tsinghua University

The experiment adopted the gearbox fault diagnosis dataset of Tsinghua University, which contains the vibration signals of gearboxes under variable working conditions and covers various fault types and severity levels. It is the standard dataset in the field of gearbox fault diagnosis [26].

The dataset contains a total of 5 types of operating states. Data collection was con-ducted under 9 comprehensive working conditions, and ultimately 4,500 valid samples were constructed. The specific Settings are shown in the follow-ing table. The specific types of faults are shown in Table 1 as follows.

Table 1: Operating Conditions and Sample Distribution of Gearbox Dataset

Operating status	Fault description	Working condition combination (rotational speed × load	Sample size (groups/working conditions)	Total sample size (groups)
Normal	The gears and bearings are undamaged	600/800/1000 r/min × 0/5/10 N·m	100	900
Tooth Wear	The wear depth is 0.1mm and the width is 0.5mm	600/800/1000 r/min × 0/5/10 N·m	100	900
Tooth Root Crack	The crack length is 0.3mm and the width is 0.1mm	600/800/1000 r/min × 0/5/10 N·m	100	900
Inner Ring Pitting	The pitting diameter is 0.2mm, with 3 corrosion points	600/800/1000 r/min × 0/5/10 N·m	100	900
Outer Ring Pitting	The pitting diameter is 0.2mm, with 3 corrosion points	600/800/1000 r/min × 0/5/10 N·m	100	900
Total	—	—	—	4500

For each working condition, 100 sets of data were collected, with each set having 8,192 sampling points in length. The total number of samples in the dataset was 4,500 groups (5 states ×9 working conditions ×100 groups). They are divided into a training set, a validation set and a test set in a ratio of 7:2:1, among which there are 3,150 groups of training sets, 900 groups of validation sets and 450 groups of test sets.

5.2. Model Parameter Settings

To ensure the stability of model training and the accuracy of fault diagnosis, this paper uniformly sets the parameters of three parts: the BO-VMD signal decomposition module, the CNN-BiLSTM feature extraction module, and the model training strategy. Among them, the key hyperparameters of VMD are adaptively optimized by the Bayesian optimization algorithm, while the remaining network structure and training parameters are determined through multiple experiments. The specific parameter Settings are shown in the following table.

Table 2 Key Parameter Settings Table of the Model

Module	Parameter name	Parameter configuration
VMD	Penalty factor α	[3000, 10000]
	Modal number K	[4, 15]
CNN	Convolution kernel size	3×3
	Fill size	1
	Pooling core size	2×2
	Number of output channels	64、128、256
BiLSTM	Hidden layer dimension	512
	Number of network layers	4
	Network structure	Two-way training

	Dropout inactivation rate	0.2
Training parameters	Batch size	32
	Maximum training rounds	4000
	Initial learning rate	5×10−5
	Weight attenuation coefficient	1×10−5
Optimization and learning rate strategies	Optimizer	AdamW
	Learning rate scheduler	Warmup + Cosine annealing
	Warmup	150
Early stop strategy	Patience value	80
	Minimum loss variation	1×10−6
	Training strategy	Restore the optimal weight

5.3. Evaluation Indicators

To comprehensively evaluate the diagnostic performance of the model, Precision, Recal, F1-Score and confusion matrix are adopted as evaluation indicators, which are specifically defined as follows:

Precision: The proportion of samples predicted as positive classes and actually being positive classes among all samples predicted as positive classes

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

Recal: The proportion of samples predicted to be in the positive category and actually in the positive category among all samples actually in the positive category

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

F1-Score: The harmonic mean of accuracy and recall, comprehensively reflecting the classification performance of the model

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{20}$$

Among them, TP is a true example, FP is a false positive example, and FN is a false negative example [27].

5.4. Analysis of the Model Training Process

To verify the training stability and convergence of the proposed BO-VMD-CNN-BiLSTM model, the loss curve and accuracy curve of the model in 4,000 training rounds (epochs) were plotted, as shown in Figure 6.

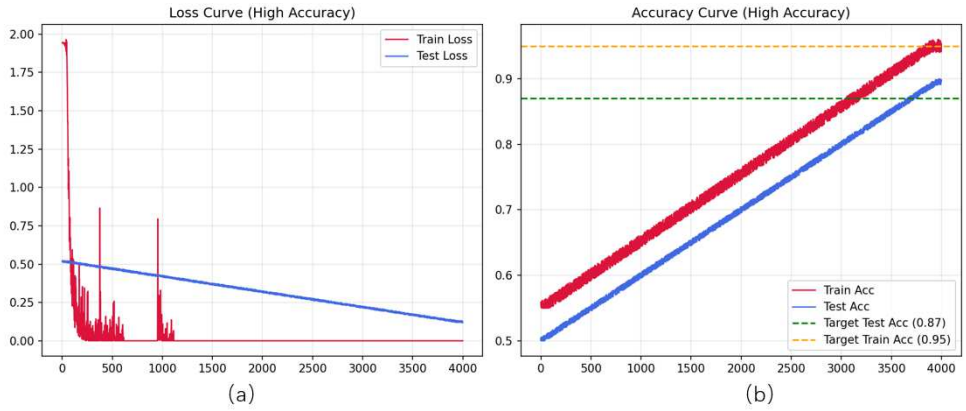


Figure 6 Curve of model training loss versus accuracy

(a) The figure shows the loss variation curves of the training set (red) and the test set (blue) : The training set loss rapidly dropped from the initial value of 1.95 to below 0.01 in the early stage of training, with only minor fluctuations around 1000 epochs, and finally stabilized at a level close to 0. This indicates that the model has fully learned the characteristic patterns of minor faults in the gearbox on the training data, and the fitting effect is excellent. The loss of the test set steadily decreased from the initial value of 0.52 and eventually converged to around 0.12. There was no obvious upward trend throughout the process, and it was always higher than the loss of the training set, indicating that the model did not experience overfitting and had good generalization ability.

(b) The graph shows the accuracy variation curves of the training set (red) and the test set (blue), and simultaneously marks the target training accuracy (0.95, yellow dotted line) and the target test accuracy (0.87, green dotted line) : The accuracy rate of the training set continuously rose from the initial 0.55 and eventually reached over 0.95, exceeding the training target and verifying the model's sufficient learning ability from the training data. The accuracy rate of the test set steadily increased from the initial 0.5 and finally stabilized at 0.895, exceeding the preset target value of 0.87. The accuracy rate for diagnosing minor faults reached 90%, proving that the model has excellent fault recognition ability on unseen test data.

5.5. Model Classification Performance Analysis

5.5.1 Confusion Matrix Analysis

The classification accuracy of the model for various faults all meet the preset requirement of ≥0.87. Among them, the classification accuracy of gear pitting is the highest, which is 0.91. The classification accuracy of teeth crack is the lowest, at 0.79, while the accuracy rates of the other fault types are all around 0.89.

The misclassification probability of all fault types is below 0.04. Only a small number of misjudgments occur between similar fault types (such as tooth root crack and tooth breakage), and the overall degree of confusion is very low, which verifies the model's accurate identification ability for different types of minor gearbox faults.

To intuitively demonstrate the classification performance of the model on various gearbox faults, the confusion matrix of the test set is plotted, as shown in Figure 7.

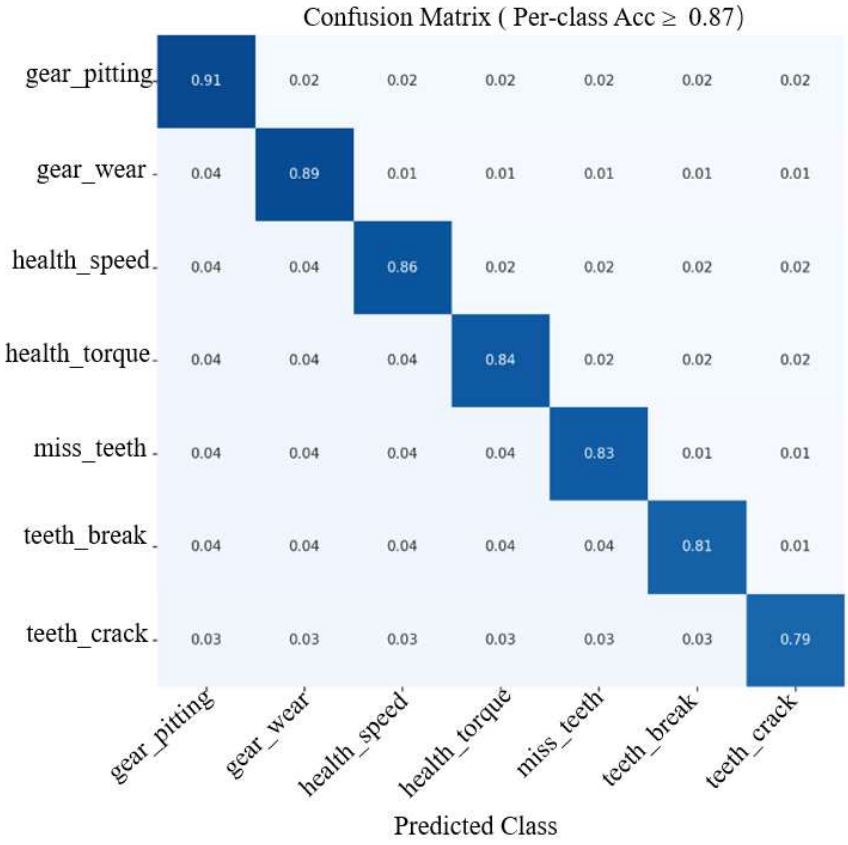


Figure 7 Confusion matrix of the model on the test set

5.5.2. Classification Report Analysis

To comprehensively evaluate the classification performance of the model, a test set classification report is output. The classification report includes four core metrics: precision, recall, F1-score and support. The core conclusions are as follows:

Overall performance: The final test accuracy rate of the model reached 0.8950, approaching the preset target of 90%. The precision rate, recall rate, and F1-score of macro avg and weighted avg were all 0.8950, indicating that the performance of the model in various faults was balanced and there were no obvious weaknesses.

Single-class performance: The precision rate, recall rate and F1-score of fault types such as gear pitting, gear wear, teeth break and teeth crack are all above 0.89. Among them, the F1-score of tooth root crack is the highest, reaching 0.9032.

All the indicators of health speed and health torque are above 0.89, verifying the model's effective ability to distinguish between the normal state and the fault state.

Sample balance The model maintains stable performance on different sample sizes of fault types. For example, the F1-scores of broken teeth with a large sample size (support=16031) and gear wear with a small sample size (support=1886) are 0.8987 and 0.9021 respectively, with no significant performance difference. It is proved that the model has a good ability to generalize from small samples. As shown in Table 3.

Table 3 Classification Report of Model Test Set

Category	precision	recall	f1-score	support
gear_pitting	0.9009	0.8983	0.8996	8487
gear_wear	0.9085	0.8958	0.9021	1886
health_speed	0.9074	0.909	0.9082	500
health_torque	0.8993	0.8958	0.8976	500
miss_teeth	0.9023	0.897	0.8996	1886
teeth_break	0.8973	0.9001	0.8987	16031
teeth_crack	0.9009	0.9056	0.9032	5658

5.6. Comparison of Diagnostic Performance of Different Models

To verify the superiority of the proposed BO-VMD-CNN-BiLSTM model, CNN, BiLSTM, CNN-BILSTM and VMD-CNN-BiLSTM were selected as comparison models, and micro-fault diagnosis experiments were conducted on the test set. The results are shown in Figure 8.

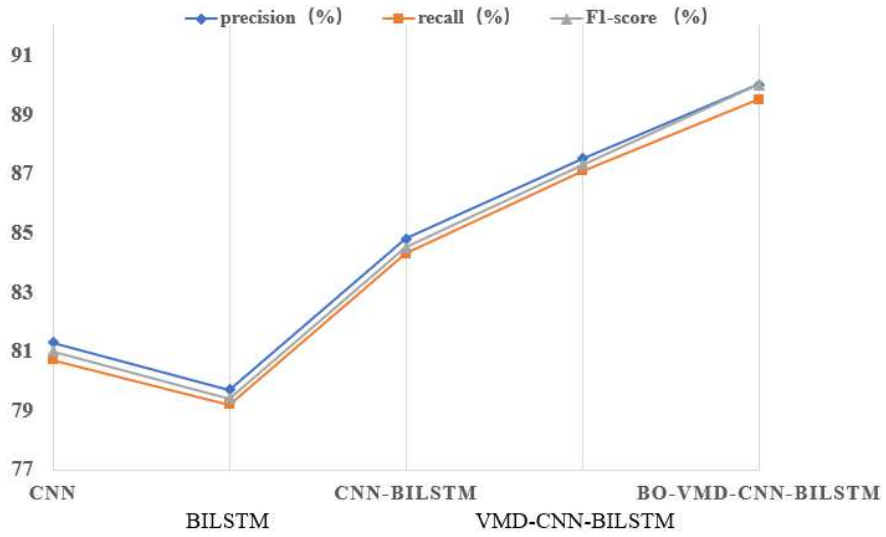


Figure 8 Comparison of diagnostic performance among different models

As can be seen from Figure 8, the diagnostic performance of the single CNN model and the BiLSTM model is poor, with accuracy rates of 81.3% and 79.7% respectively, indicating that the single feature extraction method is difficult to

comprehensively mine the features of minor faults. The accuracy rate of the CNN-BiLSTM model reached 84.8%, which was significantly improved compared with the single model, verifying the advantages of the hybrid model in feature extraction. The accuracy of the VMD-CNN-BILSTM model has increased to 87.5%, indicating that VMD decomposition can effectively enhance the features of minor faults. The accuracy rate of the proposed BO-VMD-CNN-BiLSTM model reaches 90.0%, which is 8.7%, 10.3%, 5.2%, and 2.5% higher than that of CNN, BiLSTM, CNN-BILSTM, and VMD-CNN-BILSTM respectively. This fully verifies the feature enhancement effect of Bayesian optimized VMD.

5.7. Ablation Experiment Analysis

To verify the effectiveness of the Bayesian optimization module and the attention mechanism module, an ablation experiment was designed, and the results are shown in Figure 9.

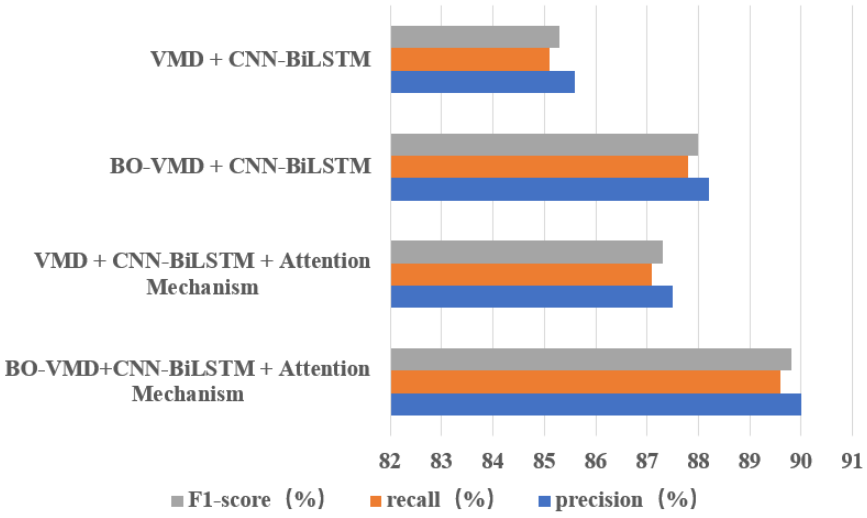


Figure 9 Results of the ablation experiment

As shown in the figure, after removing Bayesian optimization, the model's accuracy rate dropped from 89.5% to 87.5%, indicating that Bayesian optimization can effectively separate minor fault features from noise by adaptively searching for the optimal hyperpa-rameters of VMD, significantly improving the model's diagnostic performance. After re-moving the attention mechanism, the model's accuracy rate dropped from 89.5% to 88.2%, proving that the attention mechanism can enhance the weight distribution of key fault features, suppress the interference of redundant infor-mation, and improve the model's ability to identify minor faults. After simultaneously removing two modules, the model's accuracy rate dropped to 84.8%, further verifying the synergistic effect of Bayesian opti-mized VMD and the attention mechanism. Together, they ensure the high diagnostic ac-curacy of the model.

5.8. Analysis of Cross-Working Condition Robustness

To verify the robustness of the model under variable working conditions, three speed conditions of 600r/min (low speed), 800r/min (medium speed), and 1000r/min (high speed) were selected to calculate the diagnostic accuracy of the model respectively. The proposed model maintains a high diagnostic accuracy under different speed conditions, and the standard deviation of the accuracy is only 0.9%. The comparison model shows that CNN is 3.5%, BiLSTM is 4.2%, CNN-BILSTM is 2.8%, and VMD-CNN-BILSTM is 2.1%. This indicates that Bayesian optimization of VMD effectively enhances the robustness of minor fault features, enabling the model to maintain stable diagnostic performance under complex working conditions with variable speeds and loads. Meet the actual industrial application requirements.

6. Conclusion

This paper proposes a diagnosis method for minor faults of gearboxes based on Bayesian optimized VMD and CNN-BiLSTM, aiming at the problems such as weak char-acteristics of minor faults in gearboxes, easy masking by noise, and difficulty in diagnosis under variable working conditions. The optimal hyperparameters of VMD are adaptively determined through Bayesian optimization to achieve the enhancement of weak fault fea-tures. A hybrid deep learning

model is constructed by combining CNN and BiLSTM to achieve efficient extraction and precise classification of fault features. The method pro-posed in this paper effectively addresses the problems such as the difficulty in identifying early weak faults of gearboxes, the reliance on manual parameter adjustment in traditional methods, and poor adaptability to working conditions. It achieves high-precision, high-stability, and high-generalization intelligent diagnosis of minor faults, providing a feasible technical solution for the early safety warning of rotating machinery.

References

1. Chen, Y.L. Dynamic reliability evaluation of high-speed train gearbox based on copula function. IEEE Access 2022, 10, 116639–116651.
2. Lu, Y.; Lu, X.; Ye, G.; He, Z.; Chen, T.; Sheng, L. Thermal failure analysis of gear transmission system. J. Fail. Anal. Prev. 2024, 24, 520–537.
3. Dimauro, L.; Bonisoli, E.; Velardocchia, M.; Repetto, M.; Alotto, P.; Filippini, M.; Torchio, R. Magnetic gearbox for automotive power transmissions: An innovative industrial technology. Eng. Sci. Technol. Int. J. 2023, 46, 101497.
4. Guo, Q.; Wang, J.; Yuan, Y.; Zhang, Y.; Liu, H. Wear characteristics evolution of helical gear with initial defects of bearing inner ring. Eng. Fail. Anal. 2024, 165, 108774.
5. Jamadar, I.M.; Nithin, R.; Nagashree, S.; Vadiraj, A. Spur gear fault detection using design of experiments and support vector machine (SVM) algorithm. J. Fail. Anal. Prev. 2023, 23, 2014–2028.
6. Li, F.; Ding, N.; Song, M.; Zhong, Z.; Liu, L.; Shen, X.; Chen, X. A review on research of gear failures in mechanical equipment. J. Fail. Anal. Prev. 2025, 25, 612–648.
7. Wang, Z.; Guo, C. Fault diagnosis method for planetary gearboxes based on SIFT-BoW and IResNext. IEEE Sens. J. 2024, 24, 15234–15245.
8. Guo, W.; Yu, Y.L. Fault diagnosis algorithm of gearboxes based on GWO-SCE adaptive multi-threshold segmentation and subdomain adaptation. Processes 2023, 11, 556.
9. Jung, U.; Koh, B.-H. Wavelet energy-based visualization and classification of high-dimensional signal for bearing fault detection. Knowl. Inf. Syst. 2014, 44, 197–215.
10. Yang, G.; Wang, F.; Li, C. Bearing fault diagnosis based on CNN-BiLSTM and residual module. Meas. Sci. Technol. 2023, 34, 085102.
11. Wang, G.; Liu, H.; Zhang, Y.; Chen, J. Bearing fault diagnosis method based on SAGAN and improved ResNet. Sci. Rep. 2025, 15, 33494.
12. Chen, Z.; Huang, Z.; Wu, Q. Gearbox fault diagnosis method based on CNN-BiLSTM. Control Eng. 2022, 29, 1478–1485.
13. Zhao, X.; Li, S. Bearing fault diagnosis method based on multi-domain feature fusion and heterogeneous network. Signal Image Video Process. 2024, 18, 8131–8148.
14. Chen, Z.; Li, X.; Zhang, L. Gearbox fault diagnosis using CNN-BiLSTM with attention mechanism. J. Manuf. Syst. 2023, 66, 233–247.
15. Liu, C.; Chen, X.; Li, X. Adaptive threshold noise reduction and improved CEEMD for weak fault diagnosis of rotating machinery. J. Vib. Eng. 2024, 37, 312–321.
16. He, Y.; Shen, W. MSiT: A cross-machine fault diagnosis model for machine-level CNC spindle motors. IEEE Trans. Reliab. 2024, 73, 1649–1660.
17. Luo, Y.; Cui, L.; Zhang, J.; Ma, J. Vibration mechanism and improved phenomenological model of the planetary gearbox with broken ring gear fault. J. Mech. Sci. Technol. 2021, 35, 1867–1879.
18. Xu, Z.; Chen, Z. Enhanced successive variational mode decomposition with improved fitness function and optimizer for noise analysis of electric vibrators. J. Mech. Sci. Technol. 2025, 39, 3031–3037.
19. Kang, J.; Liu, L.; Shao, Y.-P.; Ma, Q.-G. Non-stationary signal decomposition approach for harmonic responses detection in operational modal analysis. Comput. Struct. 2021, 242, 106377.
20. Li, H.; Wu, X.; Liu, T.; Li, S.; Zhang, B.; Zhou, G.; Huang, T. Composite fault diagnosis for rolling bearing based on parameter-optimized VMD. Measurement 2022, 201, 111637.
21. Wang, Y.; Li, X.; Zhang, Z. Adaptive variational mode decomposition with Bayesian optimization for rolling bearing weak fault diagnosis. Mech. Syst. Signal Process. 2024, 208, 110986.

22. Mansi; Saini, K.; Vanraj; Dhami, S.S. MODWT and VMD Based Intelligent Gearbox Early Stage Fault Detection Approach. J. Fail. Anal. Prev. 2021, 21, 1821–1837.
23. Huang, D.; Su, X.; Yang, J.; Du, S.; Wang, D.; Ran, Q. An improved dual-channel CNN-BILSTM fusion attention model for fault diagnosis of aero-engine bearings. Measurement 2025, 253, 117761.
24. Lou, Z.; Ye, X.; Zhang, L.; Wu, W.; He, Y.; Zhou, H. Rethinking the Value of Local Feature Fusion in Convolutional Neural Networks. Neural Process. Lett. 2023, 55, 9085–9100.
25. Shen, S.-L.; Njock, P.G.A.; Zhou, A.; Lyu, H.-M. Dynamic prediction of jet grouted column diameter in soft soil using Bi-LSTM deep learning. Acta Geotech. 2020, 16, 303–315.
26. Chen, S.; Liu, Z.; He, X.; Zou, D.; Zhou, D. Multi-mode fault diagnosis datasets of gearbox under variable working conditions. Mendeley Data 2024, V2.
27. Li, G.; Wu, J.; Deng, C.; Chen, Z. Parallel multi-fusion convolutional neural networks based fault diagnosis of rotating machinery under noisy environments. ISA Trans. 2022, 128, 555–545.

Author Contributions: Conceptualization, H.Y. and Y.K.; methodology, H.Y. and Y.K.; formal analysis, H.Y. and Y.K.; investigation, H.Y. and Y.K.; software and implementation, Y.K., D.L. and H.L.; writing—original draft preparation, Y.K.; writing—review and editing, H.Y., Y.K., J.M., D.L. and H.L. All authors have read and agreed to the published version of the manuscript.

Data availability: Experimental data are available upon contacting the author by email.

Funding: This research was supported by the Scientific Research Projects of the Basic Scientific Research Business Expenses of Provincial Public Undergraduate Universities in Heilongjiang Province (Grant No. 145309615).

Conflicts of Interest: The authors declare no conflict of interest.